Computer Communications and Networks

Naeem Ramzan
Roelof van Zwol
Jong-Seok Lee
Kai Clüver
Xian-Sheng Hua *Editors*

Social Media

Computer Communications and Networks

For further volumes:
http://www.springer.com/series/4198

The **Computer Communications and Networks** series is a range of textbooks, monographs and handbooks. It sets out to provide students, researchers and non-specialists alike with a sure grounding in current knowledge, together with comprehensible access to the latest developments in computer communications and networking.

Emphasis is placed on clear and explanatory styles that support a tutorial approach, so that even the most complex of topics is presented in a lucid and intelligible manner.

Naeem Ramzan • Roelof van Zwol • Jong-Seok Lee
Kai Clüver • Xian-Sheng Hua
Editors

Social Media Retrieval

Springer

Editors
Naeem Ramzan
School of Electronic Engineering
and Computer Science
Queen Mary University of London
London
United Kingdom

Jong-Seok Lee
School of Integrated Technology
Yonsei University
Incheon
Korea, Republic of (South Korea)

Xian-Sheng Hua
Media Computing Group
Microsoft Research
Bellevue, Washington
USA

Series Editor
A.J. Sammes
Centre for Forensic Computing
Cranfield University
Shrivenham campus
Swindon, UK

Roelof van Zwol
Director of Product Innovation, Search
Netflix
Los Gatos, California
USA

Kai Clüver
Institut für Telekommunikationssysteme
Technische Universität Berlin
Berlin
Germany

ISSN 1617-7975
ISBN 978-1-4471-4554-7 ISBN 978-1-4471-4555-4 (eBook)
DOI 10.1007/978-1-4471-4555-4
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2012953932

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Multimedia content has become ubiquitous on the web, creating new challenges for indexing, access, search, and retrieval. At the same time, much of this content is made available on content-sharing websites such as YouTube or Flickr or shared on social networks like Facebook. In such environments, the content is usually accompanied with metadata, tags, ratings, comments, information about the uploader and their social network, etc. Analysis of these "social media" shows a great potential in improving the performance of traditional multimedia information analysis/retrieval approaches by bridging the semantic gap between the "objective" multimedia content analysis and "subjective" users' needs and impressions. The integration of these aspects, however, is non-trivial and has created a vibrant, interdisciplinary field of research.

This book presents in-depth knowledge that explicitly exploits the synergy between multimedia content analysis, personalisation, and next-generation networking and community aspects of social networks. We believe that this integration could result in robust, personalised multimedia services, providing users with an improved multimedia centric quality of experience (QoE) awareness. In response to the booming developments in social networks, this book offers a practical step-by-step walk through the various challenges, concepts, components, and technologies involved in the development of applications and services. Researchers and students interested in the field of social media retrieval will find this book a valuable resource, covering a broad overview of current state of the art, current research, and development trends in this area. Practical engineers from industry will find this book useful in envisioning and building innovative social media applications and services.

This book is divided in four major parts. The first part introduces the fundamentals of social media retrieval by presenting the most important area of research in this domain. The second part covers an essential area of multimedia tagging in social environment. Personalisation and privacy in social media domain is discussed in the third part. In the last part, applications related to social media is presented.

The aim of the Part I "Fundamentals of Social Media" is to give a comprehensive overview on the state of the art, the challenges, and the potential of social

media retrieval. This part opens with chapter "Social Video Retrieval: Research Methods in Controlling, Sharing, and Editing of Web Video" by K. Chorianopoulos, D.A. Shamma, and L. Kennedy. It presents a survey of the state of the art in social video retrieval areas. Three case studies are presented, and the findings through these case studies are discussed in detail with the perspective of few future research directions. Chapter "Social Media Recommendation" by Z. Wang, W. Zhu, P. Cui, L. Sun, and S. Yang presents the framework of social media recommendation, with a focus on two important types of recommendations: interest-oriented social media recommendation and influence-oriented social media recommendation. The following chapter "Multimedia Indexing, Search and Retrieval in Large Databases of Social Networks" by T. Semertzidis, D. Rafailidis, E. Tiakas, M.G. Strintzis, and P. Daras addresses multimedia indexing methods and a useful case study that demonstrates the challenges faced by the content-based multimedia retrieval research community. To address such challenges, the chapter focuses on state-of-the-art strategies that target the huge volume of heterogeneous data offered by social networks. Chapter "Survey on Social Community Detection" by M. Plantie and M. Crampes presents a review/survey of the many different methods in the literature of community detection based on semantics, output type, and the evaluation of communities. The last chapter of the Part I "Detecting Multimedia Contents of Social Events in Social Networks" by M. Rabbath and S. Boll summarises the state of the art of event detection and clustering approaches. It deals with several types of social media platforms and how do they deal with events. It also shows the features used in the multimedia content analysis to support event detection including metadata, visual, people, and structural-based features.

Part II of the book addresses converting social media into implicitly or explicitly tagging (textual description). It has long been a dream of a number of multimedia, computer vision, and machine-learning researchers that gain popularity exponentially due to the explosion of social media on the Internet, bringing us both challenges and opportunities. Online services based on users' geographical location are becoming very popular. Chapter "Georeferencing in Social Networks" by P. Kelm, V. Murdock, S. Schmiedeke, S. Schockaert, P. Serdyukov, and O. Van Laere provides an overview of issues in extracting and exploiting the geographical information from contents in social networks. By broadening the topic towards general content annotation, chapter "Predicting User Tags in Social Media Repositories Using Semantic Expansion and Visual Analysis" by T. Piatrik, Q. Zhang, X. Sevillano, and E. Izquierdo discusses how tags for multimedia contents can be automatically predicted based on the associated textual metadata and visual features and complementary information in external resources. Chapter "A Rule-Based Flickr Tag Recommendation System" by L. Cagliero, A. Fiori, and L. Grimaudo focuses on a rule-based tag recommendation approach that is robust to keyword data sparsity and able to capture different viewpoints of tags. In chapter "Sentic Computing for Social Media Analysis, Representation and Retrieval" by E. Cambria, M. Grassi, S. Poria, and A. Hussain, the concept of sentic computing, which is based on simultaneous usage of common sense knowledge and emotional information, is applied to bridge the gap between word-level natural language

data and concept-level opinions conveyed by these. Chapter "Highlight Detection in Movie Scenes Through Inter-users, Physiological Linkage" by C. Chenes, G. Chanel, M. Soleymani, and T. Pun shows how the problem of automatic video summarisation can be solved by aggregating users' physiological signals, based on the assumption that synchronised emotional excitation of users provides the information about the importance of the video segment corresponding to the time instance. Chapter "Toward Emotional Annotation of Multimedia Contents" by A. Yazdani, J.-S. Lee, and T. Ebrahimi also deals with the problem of multimedia annotation via emotion recognition using physiological signals, which is called implicit tagging. The potential of the electroencephalogram (EEG), peripheral physiological signals, and content features for extracting emotional aspects of contents is examined.

Part III covers different aspects of privacy and personalisation of social media. It is introduced by a chapter "Privacy in Recommender Systems" by A. Jeckmans, M. Beye, Z. Erkin, P. Hartel, R. Lagendijk, and Q. Tang. It discusses existing recommender systems, data that is used in recommender systems, risks to data privacy, privacy-protection techniques in the literature, and how they apply to recommender systems. Chapter "Geotag Propagation with User Trust Modeling" by I. Ivanov, P. Vajda, J.-S. Lee, P. Korshunov, and T. Ebrahimi presents an overview of methods for user trust modelling in automatic geotags propagation and then presents and evaluates their own approach for automatic geotagging. Chapter "Context-Aware Content Adaptation for Personalised Social Media Access" by H.K. Arachchi and S. Dogan elaborates issues related to content adaptation for personalised access to social media. The importance of using context-aware content adaptation is stressed, and several signal processing-based techniques for actual content adaptation are presented. Both technological and non-technological challenges for performing content adaptation in social media are also addressed.

Part IV of the book focuses on the different applications that exploit social media to enhance the overall acceptability of the system. Exchange of information in social media would not be possible without the enormous progress of communication techniques. First, efficient methods for flexible encoding and compression, especially of bitrate-intensive video data, are a prerequisite for acceptable distribution and retrieval speeds is explained. Further, intelligent network architectures ensure a minimum quality of experience (QoE) and help establish new applications within social networks. This part starts with the L. Anania from European Commission (EC) statement is elaborated how the advancement of social media is supported by EC. The following chapter "Video Technology for Storage and Distribution of Personalised Media" by G.V. Wallendael, J.D. Cock, D.V. Deursen, M. Mrak, and R.V. de Walle deals with scalable coding of video. The authors give an overview of current video compression techniques with the focus on strategies for scaling and adaptation of content delivery. The use of metadata for personalised adaptation of media is discussed, and standardisation efforts as well as test results are presented. Dealing with networking aspects, chapter "Social Aware TV Content Delivery Over Intelligent Networks" by F. Fraile, P. Arce, R. Belda, I. de Fez, J.C. Guerri, and A. Pajares discusses IP network architectures, quality criteria,

social aspects of TV services, and socially aware techniques for content delivery. Motivated by the increasing demand for video delivery, the authors present several approaches for ensuring QoE in video transmission by dynamic management of network resources. Video watching over a distributed social network requires synchronisation in order to ensure a shared experience for the participants. Chapter "Distributed Media Synchronisation for Shared Video Watching: Issues, Challenges and Examples" by Fernando Boronat, Rufael Mekuria, Mario Montagud, and Pablo Cesar presents an overview of the problem, results of measurements and subjective tests, a comprehensive description of possible synchronisation schemes, and an own proposal which is currently in the process of standardisation. Chapter "eGuided: Sharing Media in Academic and Social Networks Based on Peer-Assisted Learning e-Portfolios" by Paulo N.M. Sampaio, Rúben H. de Freitas Gouveia, and Pedro A.T. Gomes presents a software platform for peer-assisted learning. A student's "e-portfolio" may be used and/or edited by its owner, teachers, peers, etc., in order to assess the student's skills and weaknesses and derive suitable learning strategies. The chapter includes a discussion of different peer-assisted learning methods and of the authors' approach, together with a comparison of existing e-portfolio platforms. The last chapter of the book, "Exploiting Social Media for Music Information Retrieval" by M. Schedl, deals with an application of data mining, which deals with music information retrieval in social media. The chapter gives a comprehensive overview of approaches, describing similarity and popularity estimation and auto-tagging methods, including the results of recent work in the field.

We would like to thank all the authors and reviewers for their excellent original contributions and essential help in providing expert opinions and comments on the numerous chapter submitted to this book, respectively. It is our hope that the readers of this book will get useful knowledge, ideas, and insights for their own research.

Naeem Ramzan
Roelof van Zwol
Jong-Seok Lee
Kai Clüver
Xian-Sheng Hua

Contents

Part I
Fundamentals of Social Media

Social Video Retrieval: Research Methods in Controlling, Sharing, and Editing of Web Video

Konstantinos Chorianopoulos, David A. Shamma, and Lyndon Kennedy

Abstract Content-based video retrieval has been a very efficient technique with new video content, but it has not fully explored the increasingly dynamic interactions between users and content. We present a comprehensive survey on user-based techniques and instrumentation for social video retrieval researchers. Community-based approaches suggest there is much to learn about an unstructured video just by analyzing the dynamics of how it is being used. In particular, we explore three pillars of online user activity with video content: (1) Seeking patterns within a video is linked to interesting video segments, (2) sharing patterns between users indicates that there is a correlation between social activity and popularity of a video, and (3) editing of live events is automated through the synchronization of audio across multiple viewpoints of the same event. Moreover, we present three complementary research methods in social video retrieval: experimental replication of user activity data and signal analysis, data mining and prediction on natural user activity data, and hybrid techniques that combine robust content-based approaches with crowd sourcing of user-generated content. Finally, we suggest further research directions in the combination of richer user and content modeling, because it provides an attractive solution to the personalization, navigation, and social consumption of videos.

K. Chorianopoulos (✉)
Ionian University, T. Triantafyli, 15125, Marousi, Greece
e-mail: choko@ionio.gr

D.A. Shamma • L. Kennedy
Yahoo! Research, Santa Clara, CA, USA
e-mail: aymans@acm.org; lyndonk@yahoo-inc.com

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_1, © Springer-Verlag London 2013

1 Introduction

Every second millions of users enjoy video streaming on a diverse number of terminals (TV, desktop, smart phone, tablet) and create billions of interactions within video or between users. This amount of data might be converted into useful information for the benefit of all video users. In this chapter, we examine research methods for the main types of user interaction with video on the Web, such as controlling, sharing, and editing [3]. Indeed, Web-based video has become a popular medium for creating, sharing, and active interaction with video [4, 5, 20]. At the same time, Web-based video streaming has become available through alternative channels (e.g., TV, desktop, mobile, tablet). In the above diverse, but technologically converged scenarios of use, the common denominator is the increased interactivity and control that the user has on the video. For example, the users are able to seek forward and backward within a video, to post comments, to share with other users, and to post their own video recordings regardless of the transport channel (e.g., mobile, Web, broadcast, IPTV). In this work, we suggest that user-based video retrieval techniques are beneficial for all Web-based video systems.

In the next section, we present an outline of the most significant research findings in video retrieval. Moreover, we provide a summary of the research methods that have been employed by scientists in the exploration of video retrieval.

2 Related Work

Although online video availability is growing rather fast, there have been few research efforts to understand and leverage actual user behavior with video. Previous research has explored several techniques in order to improve users' navigation experience. One of the major goals in multimedia information retrieval is to provide abstracts (summaries) of videos. According to Money and Agius [21], there is a classification for video summarization techniques: (1) internal summarization techniques that analyze information sourced directly from the video stream and (2) external ones that analyze information not sourced directly from the video stream. Notably, Money and Agius suggest that the latter techniques hold the greatest potential for improving video summarization/abstraction, but there are rare examples of contextual and user-based works.

Abstraction techniques are a way for efficient and effective navigation in video clips [17]. For example, stationary images have proven an effective user interface in video editing [1] as well as in video browsing [12]. According to Truong and Venkatesh [32], those techniques are classified in (1) video skims, which provide moving images that stand for the important parts of the original video, and (2) key frames, which provide stationary pictures of key moments from the original the video. The evaluation of a key-frame extraction and video summarization systems has been considered a very difficult problem [19].

Table 1 Previous user-based research has established the significance of mapping user actions to video semantics, but there is no silver bullet because each approach has some drawback

	Advantages	Challenges
Ma et al. [19]	Assumes that viewers are interested in particular well-defined and easy-to-retrieve content features (e.g., faces)	Content-based and relies on a limited, preset vocabulary of what is interesting
Shaw and Davis [29]	User-contributed comments and tags	Most data lacks temporal indexing into the content
Shaw and Schmidt [30]	Community remix of popular video revels salient segments	Only a portion of users performs remix of video
Shamma et al. [27]	Micro-blogs are associated to many TV broadcasts	Deep timing information might lag against the video cue time
Kennedy and Naaman [14]	Audio fingerprinting on aggregated recordings of the same live event	Only a small portion of online content has been recorded and uploaded by multiple users
Carlier et al. [2]	Zoom denotes areas of interest within a video frame	Zoom is not a common feature
Yew et al. [35]	User comments accurately predict the category of the video	Only a portion of users post comments
Olsen and Moon [24]	Interest function	Explicit ratings are required for training the system
Chorianopoulos et al. [6]	Implicit and generic user interactions with video player (seek/scrub)	Hard to capture the needed information from public video Web sites
Peng et al. [25]	Eye tracking and face recognition	Requires an always on Web camera

In the following subsections, we present a comprehensive overview of the state-of-the-art social video retrieval (Table 1), in order to create a context for the study of more detailed case studies that follow immediately after.

2.1 Visual Feature Video Retrieval

Content-based video retrieval has been concerned with signal analysis of audio and video content. Moreover, content-based research has regarded the metadata that are being produced during the editing process of video content. In terms of research techniques, content-based researchers have defined a set of ground truths that are used as benchmarks during the evaluation of systems that focus on the fixed data and metadata of the video file. In this way, content-based systems have improved the quality of retrieving, adapting, and navigating in video content.

The main focus of content-based research has been the segmentation of video content by detecting key frames and important video segments. Content-based

Fig. 1 Video frames are an important part of user navigation within and between videos (*left*). The research issue is that content-based techniques produce too many video thumbnails, which might not be representative, because they are selected by the video uploader (*right*)

research has established the need for video thumbnails [10], video summaries [17], and the usefulness of automatic detection of key frames for user navigation [21,32]. There are several research works on content-based key-frame extraction from videos, because a collection of still images is easier to deliver and comprehend when compared to a long video stream. Girgensohn et al. [12] found that clustering of similar colors between video scenes is an effective way to filter through a large number of key frames. SmartSkip [11] is an interface that generates key frames by analyzing the histogram of images every 10 s of the video and looking at rapid overall changes in the color and brightness. Li et al. [16] developed an interface that generates shot boundaries using a detection algorithm that identifies transitions between shots.

The above research has found many practical applications in the industry of video retrieval. Nevertheless, in the case of Google Video, there are so many thumbnails that a separate scroll bar has been employed for navigating through them (Fig. 1, left). At the same time, search results and suggested links in popular video sites (e.g., YouTube) are represented with a thumbnail that the video authors have manually selected out of the three fixed ones (Fig. 1, right). Besides the threat of authors tricking the system, the author-based approach does not consider the variability of users' knowledge and preferences, as well as the comparative ranking to the rest of the video frames within a video.

The techniques that extract thumbnails from each shot are not always efficient for a quick browse of video content, because there might be too many shots in a video. On the other hand, content-based approaches provide robust technologies for quickly analyzing large numbers of new items.

2.2 Audio Feature Video Retrieval

A number of research efforts have addressed the domain of media from live music events. These efforts mostly looked at ways to present professionally produced or

"authoritative" video or audio content (e.g., a complete video capture provided by the event organizers). Naci and Hanjalic [22] provide a demo that utilizes audio analysis to help users find interesting moments from the concert. Detecting interesting moments automatically is approximated by detecting applause, instrument solos, and audio level of excitement; a browsing interface is provided that is supported by the extracted data. Snoek et al. [31], the authors, use the visual signal of produced content from live concerts to create concept detectors including "audience," "drummer," and "stage". Nevertheless, previous research on video recordings from concerts has not considered community-contributed content widely available on the Web.

A core audio identification technology known as audio fingerprinting [13, 33] is a rather robust and powerful technology that is already a reality in several commercial applications. In audio fingerprinting, audio recordings are characterized by local occurrences of particular structures. Given two recordings, one could rapidly identify if they were derived from the same original source material, despite a rather large amount of additive noise.

2.3 Text-Based Video Retrieval

Besides audio and visual features, researchers have leveraged existing techniques in text retrieval, in order to index and understand the contents of a video. Although text is not an inherent part of every video stream, there are several occasions that text complements a video [34]. For example, broadcast video usually includes closed captions (text with time code synchronized to the video), which have been mined to assign meaning to the respective video segments. Moreover, in some video segments, text might be part of the image. Then, optical character recognition might be employed in order to understand the respective text. Despite the robustness of the existing text-based retrieval techniques, we cannot assume that the majority of commercially available video streams are coupled or embed text. In the next subsection, we are also describing those text-based video retrieval techniques that are generated by the users in social media.

How a video is used, interacted with, and commented is often indicative of the nature of its content. Shamma et al. [27] have explored whether micro-blogs (e.g., Twitter) could structure a TV broadcast. Video sharing sessions leave behind digital traces in the form of server logs and metadata. The ability to share videos in real time while in an instant messaging (IM) session is an attempt to replicate the social experience of watching videos in a virtual environment. Although there are various methods that collect and manipulate user-based data, the majority of them are considered burdensome for the users, because they require an extra effort, such as writing a micro-blog or posting a comment. Nevertheless, the percentage of users leaving a comment are rather small when compared to the real number of viewers [20].

2.4 User-Based Video Retrieval

User-based techniques approach the problem of video retrieval differently to the established content-based ones. Rather than paying attention to the content of the video, its metadata, or its position in a network, they focus mainly on identifying particular video interaction patterns, such as video seeking and sharing between users.

The media is often experienced socially, and a large part of that experience involves frequent commentary, backchannel conversations, and starts/stops to review key moments in the media content. Social video interactions on Web sites are very suitable for applying community intelligence techniques [37]. Levy [15] outlined the motivation and the social benefits of collective intelligence, but he did not provide particular technical solutions to his vision. In the seminal user-based approach to Web video, Shaw and Davis [29] proposed that video representation might be better modeled after the actual use made by the users. In this way, they have employed analysis of the annotations as well as of the reuse of video segments in community remixes and mash-ups [30] to understand media semantics.

2.5 Research Issues and Methods in Social Video Retrieval

2.5.1 Controlling Video and Controlled Experiments

The concept of analyzing implicit user interaction in computing activities, in order to develop user models and to provide intelligent interactions, is not new. Liu et al. [18] have improved the personalization of news items by analyzing previous users, interactions with news items. In the context of multimedia, previous research has considered both content- and user-based methods for video retrieval. The most generic user interaction with social video is the seeking behavior within video. Notably, the video seeking behavior has been employed as a research granule in key-frame detection. The evaluation of a key-frame extraction and video summarization systems has been considered a very difficult problem, as long as user-based systems are concerned. Notably, Ma et al. [19] have argued that "Although the issues of key-frame extraction and video summary have been intensively addressed, there is no standard method to evaluate algorithm performance. The assessment of the quality of a video summary is a strong subjective task. It is very difficult to do any programmatic or simulated comparison to obtain accurate evaluations, because such methods are not consistent with human perception." In content-based research (e.g., TRECVID), researchers have defined a set of ground truths that are used as benchmarks during the evaluation of novel algorithms. Chorianopoulos et al. [6] propose that the evaluation of user-based key-frame extraction systems could be transformed into an objective task as long as there is a set of experimentally replicated ground truths about the content (e.g., questions about specific parts of the video).

2.5.2 Sharing Video and Data-Mining

In online multimedia sharing contexts, one-to-one chatting provides a rich context for social exchange and data collection. While still emerging, several systems support various real-time multimedia sharing interactions, like TuVista, Zync, and Google Hangouts. In effect, as people chat while sharing online videos, they leave traces of activity (clicks and chats) and inactivity (pauses), which can reveal more about the underlying multimedia, which is fueling the conversation. To discover the content categories of videos in one-to-one sharing systems such as Zync, Yew et al. [35] examined the types of noncontent data available. They collected an aggregate volume of chat activity as number of characters typed during a playback moment of the video. Beyond chat, play, pause, and scrubs were also logged. Most importantly, they looked at the length of the chat session while the embedded video player was open. This is to be distinguished from the length of the video. Using the collected data, they modeled each video as a feature vector that was informed from qualitative surveys and semi-structured interviews. This vector was then used to predict the video's content category, like news, sports, film, or TV.

2.5.3 Editing Video and Hybrid Techniques

In some cases, we can consider the video object itself to be a proxy for an interaction with an actual live event. In particular, many online video users commonly record videos of events that they are attending and later share those clips online in order to express presence to their friends and others. In a system by Kennedy and Naaman [14], this application was explored in the context of videos recorded at live concert events. This collection of videos from a concert event can tell us a lot about the relative importance of any particular moment in the event and give us some clues for understanding semantically why the given moment is important. To uncover these importance cues, one has to first discover which videos were actually recorded at the same time during a given concert event. This can be approached by utilizing an audio fingerprinting system, which can detect replicated audio tracks under extreme noise conditions with very high precision. The insight here is that the videos are not just videos, but rather an expression of interest by an individual in a particular point in an event. Aggregating across many different individuals expressing their interest in various points in the media, we can arrive at a general level of interest spread temporally across the event. Furthermore, the words that many different individuals use to describe each point in the event can be aggregated to give us clues about why, semantically, the point in the event is of interest.

In the next sections, we examine in more detail three indicative case studies in social video retrieval. The selected case studies stand for the diversity of methodologies and research instrumentation found in the area of social video retrieval. Therefore, we highlight the complementary research methods in each one of the three case studies, and we encourage the reader to elaborate into the detailed results by visiting the respective publication, which is indicated at the end of each case study.

3 Case Study 1: SocialSkip

SocialSkip [6] is an open-source Web-based system that collects and visualizes activity data from simple user interactions such as play/pause and seek/scrub. SocialSkip (Fig. 2) employs few buttons, in order to be simpler to associate user actions with video semantics. We have simplified the standard random-seek bar to the GoForward and GoBackward buttons. The first one goes backward 30 s, and its main purpose is to replay the last viewed seconds of the video, while the GoForward button jumps forward 30 s, and its main purpose is to skip insignificant video segments. The 30 s step is a popular seek window in previous research and commercial work due to the fact that it is the average duration of commercials. Furthermore, we have observed replay functions and buttons in mobile devices such as Apple's iPhone and Safari QuickTime video players, which has the default time of 30 s as a replay.

We did not use a random-seek timeline because it would be difficult to analyze users' interactions. Li et al. [16] observed that when seek thumb is used heavily, users had to make many attempts to find the desirable section of the video and thus caused significant delays. Drucker et al. [11] and Li et al. [16] tested different levels of speed for the functions of forward and rewind, too. User could make the choice of speed and locate more quickly the segment he wanted. For example, there have been commercial systems such as ReplayTV and TiVo that provide the ability to replay segments or to jump forward in different speeds. Next to the player's button, the current time of the video is shown followed by the total time of the video in seconds. Although we did not have a seek bar, we suggest that the data collected from the fixed skip could simulate the use of random seek, because any random

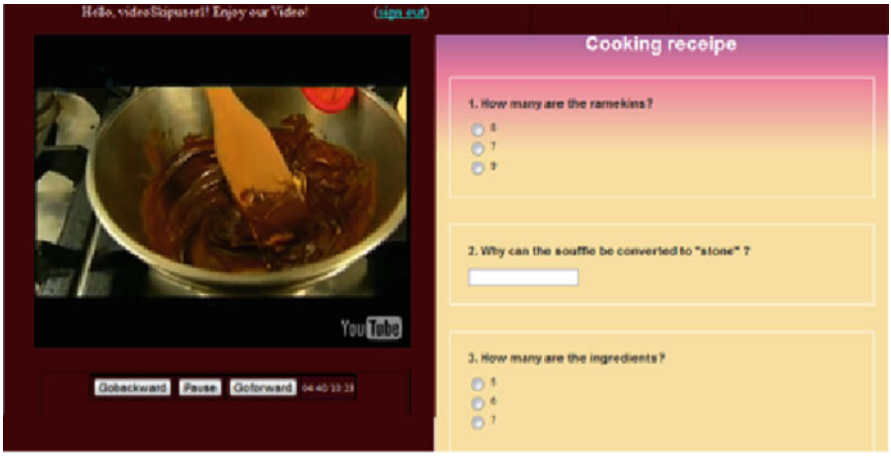


Fig. 2 SocialSkip player is focused on skipping buttons and questionnaire functionality

seek activity can be modeled as a factor of fixed skipping actions (e.g., a random seek of 180 s is equal to 6 skips of 30 s).

In this case study, we selected three videos (lecture, how-to, documentary) that are as much visually unstructured as possible, because content-based approaches have already been successful with those videos that have visually structured scene changes.

In order to experimentally replicate user activity, we added an electronic questionnaire that corresponds to a few segments of the video. According to Yu et al. [36], there are segments of a video clip that are commonly interesting to most users, and users might browse the respective parts of the video clip in searching for answers to some interesting questions. In other words, it is expected that in a future field study, when enough user data is available, user behavior will exhibit similar patterns even if they are not explicitly asked to answer questions. The experiment took place in a lab with Internet connection, general-purpose computers, and headphones. Twenty-five users spent approximately 10 min to watch a video with all buttons muted, so they could not skip or pause. Next, there was a time restriction of 5 min, in order to motivate the users to actively browse through the video and answer the questions that corresponded to a few key frames. We informed the users that the purpose of the study was to measure their performance in finding the answers to the questions within time constraints.

In order to understand video pragmatics, we visualized the user activity data with a simple user heuristic. Firstly, we considered that every video is represented with an array with a length that equals the duration of the video in seconds. Next, we modified the value of each cell, depending on the type of interaction. For each play, pause, and GoBackward, we increased the value. We decreased the value for each GoForward. In this way, we have created the following activity graph (lecture video) that assist, the understanding of video content, based on the pragmatics (user video browsing actions) rather than the content itself (Fig. 3). The main benefit of this technique is that user interactions within a video have been transformed into a time-based signal, which might be further analyzed with techniques from signal processing.

In comparison to previous research, the proposed user activity heuristic is more malleable, because researchers can make various combinations and give different meaning to them. Yu et al.'s [36] experimental process used some questions to help mimic user interests and focus user behavior. Their algorithm should work with any video as long as it contains some commonly attractive content. SocialSkip has been developed with the same assumption. On the other hand, they have implemented a system with a custom video browsing applications. Peng et al. [25] have examined the physiological behavior (eye and head movement) of video users, in order to identify interesting key frames, but this approach is not practical because it assumes that a video camera should be available and turned on in the home environment. In contrast, the majority of users browse Web video in more traditional ways that require no extra interactions or extra equipment, besides play, pause, and seek, which are the main controls of SocialSkip.

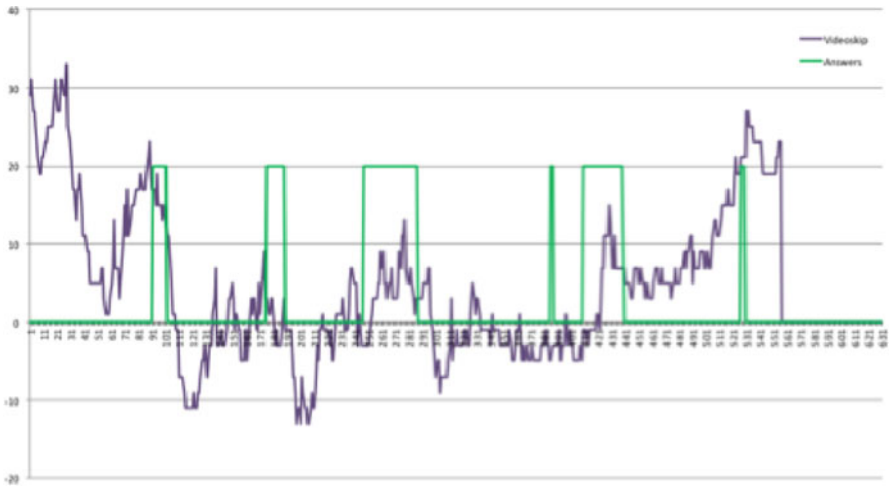


Fig. 3 The user activity graph provides a comprehensive visualization of cumulative user interactions and direct comparison to the experimentally defined ground truth

4 Case Study 2: Viral Actions

In the case of Yahoo!'s Zync [28], two people in an IM conversation can watch an embedded video together in real time; videos can be from Yahoo!, Flickr, or YouTube. The video stays in sync across two people, and both individuals share control. In effect, the video becomes a reified synchronous context for the conversation. The chat and play, pause, and scrub behavior become one trace around the media object.

We argue that the implicit social sharing activity that occurs while sharing a video in a real-time IM conversation would result in more accurate predictions of a video's potential viewership. Implicit social sharing activity here refers to the number of times a video was paused, rewound, or fast forwarded as well as the duration of the IM session while sharing a video. We believe that implicit social sharing activity is indicative of deeper and more connected sharing constructs and hence better fidelity data to predict how much viewership a particular video is likely to attract. How a video is interacted with and shared between users is often indicative of how popular it is likely to be in the future. For instance, videos that have great appeal and potential to be popular will mostly likely be interacted with more and generate more conversation than others. Taken in aggregate across all users, patterns and "signatures" [8] of interactions found in the implicit social sharing data can point to how popular and even viral a video is likely to be.

Viral videos are those that have gained outsized prominence and viewership as a result of an epidemic-like social transmission. In this case, we argue that the usage data that surrounds such viral videos can be used to predict the popularity of the

video. Here we capitalize on the "wisdom of the masses" by identifying patterns in the metadata to make predictions about the future popularity of that content [26].

New social media systems allow users to synchronously interact with each other and share videos simultaneously. These real-time interactions leave behind large amounts of contextual usage data that, we believe, are reflective of the deeper and more connected social interaction that accompanies synchronous content sharing. In this chapter, we present a method of utilizing usage data from synchronously sharing videos to make predictions about the popularity of a particular video. In particular, we use play/pause behavior and chat volume pulled from a real-time video sharing environment, Zync (a plug-in for the Yahoo! instant messaging (IM) client that allows participants to view and interact with a video simultaneously during a chat session). We argue that the usage data from synchronous video sharing tools provides robust data on which to detect how users are consuming and experiencing a video. By extension, we can predict a video's popularity based on how it has been shared in a handful of sessions. To do this, we trained a naïve Bayes classifier, informed by synchronous sharing features, to predict whether a video is able to garner ten million views on its hosting site. Our goal is to eventually predict a video's viral potential based on how it is being shared.

The ability to socially share videos online has enabled select videos to gain a viewership of thousands in a very short period of time. Often, but not always, these videos take on a viral nature and gain tens of millions of views, while other videos only receive a fraction of the attention and viewing. These popular viral videos also benefit from rich-get-richer dynamic where the more popular they become, the more views they are likely to attract. Viral videos attract not only a disproportionate amount of attention; they also consume greater amounts of resource and bandwidth as well. Thus, it would be helpful to be able to predict and identify which videos are most likely to go viral for recommendation, monetization, as well as systems performance.

We acquired a 24-h sample of the Zync event log for Christmas day 2009. Zync allows two people to watch a video together in an instant message session; both participants share playback and control of the video, and the video stays in sync across both participants IM windows; see Fig. 4. The dataset provides a list of watched videos from YouTube as well as synchronous activity from the shared control of the video. These features are anonymous user id hashes, session start/stop events, the session duration, the number of play commands, the number of pause commands, the number of scrubs (fast forward or rewinds), and the number of chat lines typed as a character and word count. For the chat lines, the dataset contained no actual text content, only the aggregate count of characters, words, and lines. The only textual content that is collected is video URLs and emoticons. Each activity collected is a row in the dataset and is associated with the time of the event and the playback time on the video.

The final test sample contained 1,580 videos with valid YouTube metadata and valid session data. The data collected from YouTube consisted of a video identifier, the video's title, its published date, its description, the genre category, the uploader

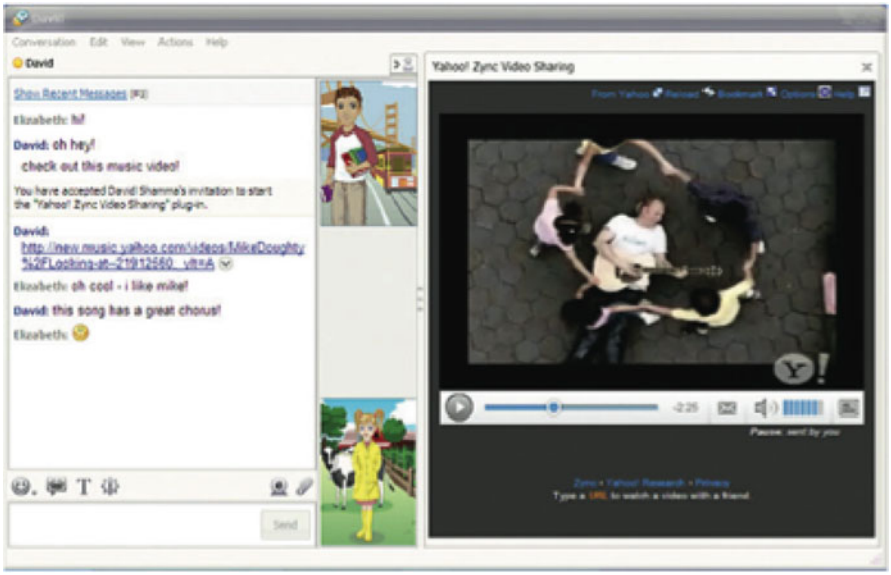


Fig. 4 The Zync plug-in allows two instant messenger users to share a video in sync while they chat. Playback begins automatically; the users share control over the video

used, the tags the video was labeled with, the video's duration, and the 5-star rating score it attained. Of these data, we only use the video's YouTube view count. For this case, the view count will be the predictive variable for the classifier and allows us to investigate if there is a match between YouTube view count and the synchronous social session actions.

As mentioned earlier, each single event from every session is a row in the dataset. This data needs to be aggregated into a feature vector for training a classifier. To do this, every session was divided into segments (sub-sessions) where a single video was viewed. This was necessary as many sessions contained multiple videos. The sub-sessions were then grouped by their representative video, mixing all the sessions that watched the same video. Finally, each event type and the overall sub-session durations were averaged into a single feature vector. Lastly, we assign a label indicating if the YouTube view count is over ten million views; see Table 2 for some sample feature sets.

In the model and results, we have addressed our research question: We can predict the view count of a video based on how it is viewed in a rich, shared, synchronous environment. In total, 100 of the 1,580 had over ten million views. The naïve Bayes classifier correctly identified 81 of these popular videos. There are far more videos that have less than ten million views and thus higher prediction accuracy. It is important to note that our classifier produces a larger increase over a fair random prediction. We believe the session's duration to be the dominant

Table 2 Random and naïve Bayes prediction accuracies. Random guess is calculated by using the distribution bias (6.3% of guessing yes). The F1 score illustrates the overall performance accounting for both precision and recall. The naïve Bayes predictions use cross-folded verification

Method	Training sample	Accuracy (%)	F_1 score
Guessing	All yes	6.3	0.119
	Random	88.3	0.041
	All no	93.7	*NaN*[a]
Naïve Bayes	25%	89.2	0.345
	50%	95.5	0.594
	60%	95.6	0.659
	70%	95.8	0.778
	80%	96.6	0.786

[a] Divide by zero

feature in the predictive model as the correlation between Zync session duration and YouTube view count had the highest correlation ($p < 0.12$). While not quite significant, the average session duration in the feature vector is completely independent of the view count. Furthermore, there is no significant or near significant correlation between the session duration and the video's playback time. Similarly, no significant correlations were observed within the other metadata from YouTube (ratings and playback time) and the YouTube view count.

5 Case Study 3: Less Talk, More Rock

In this case study, we highlight the methods that we employed in a study of user-generated music video recordings [14]. The availability of video capture devices and the high reach and impact of social video sharing sites like YouTube make video content from live shows relatively easy to share and find [9]. Users of YouTube share millions of videos capturing live musical performances, from classical pianist Ivo Pogorelich to metal rockers Iron Maiden. Potentially, such an abundance of content could enable comprehensive and deeper multimedia coverage of captured events. However, there are new challenges that impede this new potential: The sample scenario above, for example, illustrates issues of relevance, findability, and redundancy of content.

The lack of detailed metadata associated with video content presents several interesting challenges. First, with no accurate, semantic event-based metadata, it is not trivial to automatically identify a set of video clips taken at a given event with high recall and precision. Second, with no dependable time-based metadata associated with the clips, aligning and synchronizing the video clips from the same event cannot be done using simple timestamps.

In this case study, we report on an approach for solving the synchronization problem and how we leverage the synchronization data to extract additional metadata. The metadata would help us organize and present video clips from live music shows. We start by assuming the existence of a curated set of clips, having already identified the video clips from each event.

Fig. 5 A sample interface for synchronized playback of concert video clips

We use audio fingerprinting [13, 33] to synchronize the content. In other words, we use the clips' audio tracks to detect when the same moment is captured in two different videos, identify the overlap, and specify the time offset between any pair of overlapping clips. The synchronization of clips allows us to create a novel experience for watching the content from the event, improving the user experience and reducing the redundancy of watching multiple clips of the same moment. Figure 5 presents one possible viewing interface.

Once synchronized, we use both the relative time information and links between overlapping clips to generate important metadata about the clips and the event. First, we show how we identify the level of interest [23] and significant moments in the show as captured by the users. Second, we mine the tags of videos associated with a single point in time to extract semantically meaningful descriptive terms for the key moments in the show; these terms can be used to represent or explain the aggregated content. Third, we use the link structure created by the audio fingerprinting when a clip matches another to find the highest-quality audio recording of any time segment, given multiple overlapping recordings.

The clip overlap structure, created by the community activity, can help identify moments in an event that are likely interesting to consumers of content [23]. In particular, we hypothesize that the segments of concerts that are recorded by more people might be of greater appeal to content consumers. Identifying these segments can be helpful for search, summarization, key-frame selection [7], or simple exploration of the event media. Videos of the most important segments or other aspects of the concert could be highlighted, while filtering lower-scoring clips that are either unrelated or, presumably, less interesting.

Our hypothesis is that larger clusters of matches between clips typically correspond to segments of the concert that are subjectively most "interesting." In the case of live music, these clusters could reflect significant moments in the show where a hit song is being played or something particularly interesting is happening on stage.

We use the synchronization data to select the highest-quality audio for each overlapping segment. The synchronization between video clips can be used for

playback, remixing, or editing content. Inevitably, given the nature of user-generated recordings, the video and audio quality and content can be highly variant between clips as well as from minute to minute within clips. Interestingly, low-quality audio tracks cause the audio fingerprinting method to fail in systematic ways that can be leveraged to point us toward higher-quality recordings.

We aggregate the textual information associated with the video clips based on the cluster structure to extract descriptive themes for each cluster. On many social media Web sites, users often provide lightweight annotations for the media in the form of titles, descriptions, or tags. Intuitively, if the overlapping videos within our discovered clusters are related, we expect the users to choose similar terms to annotate their videos such as the name of the song being captured or a description of the actions on stage. We can identify terms that are frequently used as labels within a given cluster, but used relatively rarely outside the cluster. These terms are likely to be useful labels/descriptions for the cluster and can also be used as suggested metadata for unannotated clips in that cluster.

We have applied our system to a large set of real user-contributed videos from three concerts crawled from the popular video sharing Web site, YouTube. Each concert collection contains several hundred video clips, providing for a total of just over 600 clips and more than 23 h of video footage. The three concerts that we have investigated are Arcade Fire in Berkeley, CA; Daft Punk in Berkeley, CA; and Iron Maiden in Bangalore, India. All three concerts occurred during the spring or summer of 2007.

We find that the proposed method is able to identify clusters of videos taken at the same point in time, with a near-perfect precision up to a recall in the range of 20–30 %, which is sufficient for many discovery and browsing applications. We compare the number of people recording each song in a concert against the number of plays for the song on social music site last.fm and find a significant correlation between the number of plays and the size of the cluster ($r^2 \sim 0.44$, $p < 0.001$, $N = 41$), suggesting that the number of people recording is a reasonable estimate of the level of interest. We ask human subjects to score the relative audio quality of various segments and find that our system for scoring audio quality significantly correlates with human assessments ($r^2 \sim 0.26$, $p < 0.001$, $N = 50$). Finally, we find that repeated text terms in the videos associated with a moment in the concert often correspond to words from the names of the songs or to actions taking place on stage.

Our primary focus in this work is an in-depth exploration of the different methods, rather than building and evaluating a browsing system. We shift the focus of our system from developing matching algorithms and focus on mining the structure of the discovered overlaps and audio reuse to create compelling new ways of aggregating and organizing community-contributed Web data. The ideas above could be used in the design and implementation of a system for sharing live concert videos and content. We would also imagine such an application to elicit more accurate or structured metadata and contributions from users, contributions that might exceed and extend the social media tools available on YouTube.

6 Directions for Further Research

In this section, we make suggestion for further research, which has been organized according to the research instrumentation and method employed in the above case studies.

First, video key frames provide an important navigation mechanism and a summary of the video, either with thumbnails or with video skims. There are significant open research issues with video skims: (1) the number and relative importance of segments that are needed to describe a video and (2) the duration of video skims. The number of segments depends on several parameters, such as the type and length of the video. Therefore, it is unlikely that there are a fixed number of segments (or a fixed video skim duration) that describe a particular category of videos (e.g., lectures). If the required number of segments is different for each video, then, besides the segment extraction technique, we need a ranking to select the most important of them. Moreover, the duration of each video skim should not be fixed but should depend on the actual duration of user interest for a particular video segment. The above research issues might be addressed by means of signal processing techniques on the user activity signal.

Secondly, we have demonstrated the possibility to a classifier, based on social synchronous sharing patterns, to predict if a video has a high view count. Our goal in this research is to predict if a video will go viral based on how it is shared within a conversation. The successful predictions in our classifier are based on most videos (85 %) viewed once in only one session. The next step in this work is to collect various data samples over time and investigate how a video's implicit sharing behaviors change as it becomes viral. In effect, this is somewhat of a fishing exercise over time; we need to collect data on videos as they turn viral to train a classifier on how to predict them. We expect the temporal changes between the feature vectors (the deltas and rate of change across our video feature vectors) to enable accurate viral predictions for recommendations. Additionally, when socially filtered, unique viral patterns found in some social groups and networks could bring socially targeted recommendations and content promotion.

Finally, further research in dynamic editing of live video events should consider the fusion of context provided by social media sites and content analysis. In particular, the combination of content-based (e.g., audio or video) matching with metadata (e.g., tags, comments) from social media might create a better representation of the content. For example, audio fingerprinting features can be extracted only for clips identified as events using the social media context, and clips are pairwise compared only within a single show, significantly reducing the required scale as well as potentially improving precision. This approach for multimedia analysis leveraging social media contribution promises to change the way we consume and share media online.

It is important to underscore that there is a use for traditional content analysis to discover the social interaction. While we suggest social interaction, analysis can supersede many content analysis techniques, and content techniques can identify and connect disassociated social actions, providing a reified context.

7 Conclusion

As long as the community of users watching videos on social video systems is growing, more and more interactions are going to be gathered, and therefore, we are going to have a better understanding of a video according to evolving user interests. We also expect that the combination of richer user profiles and content metadata provides opportunities for personalization. Overall, our findings support the concept that we can learn a lot about an unstructured video just by analyzing how it is being used, instead of looking at the content item itself.

In contrast to content-based video retrieval, we have employed few videos in the research methods of the case studies. Previous work on video retrieval has emphasized the large number of videos, because the respective algorithms treated the content of those videos. In this user-based work, we are not concerned with the content of the videos, but with the user activity on videos. Nevertheless, it is worthwhile to explore the effect of more videos and interaction types. Therefore, the small number of videos used in the case studies is not an important limitation, but further research has to elaborate on different genres of video (e.g., news, sports, comedy, music, lecture) and on the number of user interactions that are necessary to obtain meaningful user activity patterns.

The methodological approaches of the three case studies provide a balance between two very different research philosophies: the employment of big natural data versus the design of controlled user experiments. We suggest that data mining on a large-scale Web-video database is the most effective approach, because the data and the techniques have high external validity. Nevertheless, we found that the experimental approach is very flexible during the development phase of a new system. Moreover, the iterative and experimental approach is very suitable for user-based information retrieval, because it is feasible to associate user behavior to the respective data logs.

Although we suggest the employment of user-based video retrieval techniques, we have also considered the benefits of content-based ones. Content-based techniques, such as pattern recognition algorithms that focus on the contents of a video (e.g., detection of changes in shots and scenes), are static, because they produce the same result all the time, but they are also very efficient in the analysis of new videos that do not have any interactions, such as pause, rewind, or sharing with other users. In contrast, the community (or crowd-sourced) intelligence of implicit user activity with Web video is dynamic (e.g., scrubs, comments, remixes), because it continuously adapts to evolving users' preferences, but it is also more difficult to analyze and evaluate. In the end, we expect that a balanced mix of hybrid algorithms (content-based and user-based) might provide an optimal solution for editing, sharing, and navigating through video content on social media Web sites and applications.

References

1. Baecker, R., Rosenthal, A.J., Friedlander, N., Smith, E., Cohen, A.: A multimedia system for authoring motion pictures. In: Proceedings of the Fourth ACM International Conference on Multimedia, MULTIMEDIA '96, pp. 31–42. ACM, New York (1996). doi:10.1145/244130.244142. http://doi.acm.org/10.1145/244130.244142
2. Carlier, A., Charvillat, V., Ooi, W.T., Grigoras, R., Morin, G.: Crowdsourced automatic zoom and scroll for video retargeting. In: Proceedings of the International Conference on Multimedia, MM '10, pp. 201–210. ACM, New York (2010). doi:10.1145/1873951.1873962. http://doi.acm.org/10.1145/1873951.1873962
3. Cesar, P., Chorianopoulos, K.: The evolution of tv systems, content, and users toward interactivity. Found. Trends Hum.-Comput. Interact. **2**(4), 373–95 (2009). doi:10.1561/1100000008. http://dx.doi.org/10.1561/1100000008
4. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC '07, pp. 1–14. ACM, New York (2007). doi:http://doi.acm.org/10.1145/1298306.1298309. http://doi.acm.org/10.1145/1298306.1298309
5. Cheng, X., Dale, C., Liu, J.: Statistics and social network of youtube videos. In: Quality of Service, 2008. IWQoS 2008, 16th International Workshop on, pp. 229–238 (2008). doi:10.1109/IWQOS.2008.32
6. Chorianopoulos, K., Leftheriotis, I., Gkonela, C.: Socialskip: pragmatic understanding within web video. In: Proceddings of the 9th International Interactive Conference on Interactive Television, EuroITV '11, pp. 25–28. ACM, New York (2011). doi:http://doi.acm.org/10.1145/2000119.2000124. http://doi.acm.org/10.1145/2000119.2000124
7. Christel, M.G., Hauptmann, A.G., Wactlar, H.D., Ng, T.D.: Collages as dynamic summaries for news video. In: Proceedings of the Tenth ACM International Conference on Multimedia, MULTIMEDIA '02, pp. 561–569. ACM, New York (2002). doi:10.1145/641007.641120. http://doi.acm.org/10.1145/641007.641120
8. Crane, R., Sornette, D.: Viral, quality, and junk videos on youtube: separating content from noise in an information-rich environment. In: Proceedings of AAAI Symposium on Social Information Processing, Menlo Park (2008)
9. Cunningham, S.J., Nichols, D.M.: How people find videos. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '08, pp. 201–210. ACM, New York (2008). doi:10.1145/1378889.1378924. http://doi.acm.org/10.1145/1378889.1378924
10. Davis, M.: Human-computer interaction. In: Baecker, R.M., Grudin, J., Buxton, W.A.S., Greenberg, S. (eds.) Readings in Human-Computer Interaction: Toward the Year 2000, Chap. Media Streams: An Iconic Visual Language for Video Representation, pp. 854–866. Morgan Kaufmann, San Francisco (1995). http://dl.acm.org/citation.cfm?id=212925.213009
11. Drucker, S.M., Glatzer, A., De Mar, S., Wong, C.: Smartskip: consumer level browsing and skipping of digital video content. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves, CHI '02, pp. 219–226. ACM, New York (2002). doi:10.1145/503376.503416. http://doi.acm.org/10.1145/503376.503416
12. Girgensohn, A., Boreczky, J., Wilcox, L.: Keyframe-based user interfaces for digital video. Computer **34**(9), 61–67 (2001). doi:10.1109/2.947093. http://dx.doi.org/10.1109/2.947093
13. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system with an efficient search strategy. J. New Music Res. **32**(2), 211–221 (2003). doi:10.1076/jnmr.32.2.211.16746
14. Kennedy, L., Naaman, M.: Less talk, more rock: automated organization of community-contributed collections of concert videos. In: Proceedings of the 18th International Conference on World Wide Web, WWW '09, pp. 311–320. ACM, New York (2009). doi:http://doi.acm.org/10.1145/1526709.1526752. http://doi.acm.org/10.1145/1526709.1526752

15. Levy, P.: Collective Intelligence: Mankind's Emerging World in Cyberspace. Perseus Books, Cambridge (1997)
16. Li, F.C., Gupta, A., Sanocki, E., He, L.W., Rui, Y.: Browsing digital video. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '00, pp. 169–176. ACM, New York (2000). doi:10.1145/332040.332425. http://doi.acm.org/10.1145/332040.332425
17. Lienhart, R., Pfeiffer, S., Effelsberg, W.: Video abstracting. Commun. ACM **40**(12), 54–62 (1997). doi:10.1145/265563.265572. http://doi.acm.org/10.1145/265563.265572
18. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI '10, pp. 31–40. ACM, New York (2010). doi:10.1145/1719970.1719976. http://doi.acm.org/10.1145/1719970.1719976
19. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the Tenth ACM International Conference on Multimedia, MULTIMEDIA '02, pp. 533–542. ACM, New York (2002). doi:10.1145/641007.641116. http://doi.acm.org/10.1145/641007.641116
20. Mitra, S., Agrawal, M., Yadav, A., Carlsson, N., Eager, D., Mahanti, A.: Characterizing web-based video sharing workloads. ACM Trans. Web **5**(2), 8:1–8:27 (2011). doi:10.1145/1961659.1961662. http://doi.acm.org/10.1145/1961659.1961662
21. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. J. Vis. Commun. Image Represent. **19**(2), 121–143 (2008). doi:10.1016/j.jvcir.2007.04.002. http://dx.doi.org/10.1016/j.jvcir.2007.04.002
22. Naci, S.U., Hanjalic, A.: Intelligent browsing of concert videos. In: Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07, pp. 150–151. ACM, New York (2007). doi:10.1145/1291233.1291264. http://doi.acm.org/10.1145/1291233.1291264
23. Nair, R., Reid, N., Davis, M.: Photo loi: browsing multi-user photo collections. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05, pp. 223–224. ACM, New York (2005). doi:10.1145/1101149.1101187. http://doi.acm.org/10.1145/1101149.1101187
24. Olsen, D.R., Moon, B.: Video summarization based on user interaction. In: Proceedings of the 9th International Interactive Conference on Interactive Television, EuroITV '11, pp. 115–122. ACM, New York (2011). doi:10.1145/2000119.2000142. http://doi.acm.org/10.1145/2000119.2000142
25. Peng, W.T., Chu, W.T., Chang, C.H., Chou, C.N., Huang, W.J., Chang, W.Y., Hung, Y.P.: Editing by viewing: automatic home video summarization by viewing behavior analysis. Multimed. IEEE Trans. **13**(3), 539–550 (2011). doi:10.1109/TMM.2011.2131638
26. Segaran, T.: Programming Collective Intelligence: Building Smart Web 2.0 Applications, 1st edn. O'Reilly, Beijing/Sebastapol (2007). http://www.amazon.com/gp/product/0596529325
27. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: understanding community annotation of uncollected sources. In: WSM '09: Proceedings of the International Workshop on Workshop on Social Media. ACM, Beijing (2009)
28. Shamma, D.A., Kennedy, L., Churchill, E.F.: Viral actions: predicting video view counts using synchronous sharing behaviors. In: ICWSM 11: Proceedings of the International Conference on Weblogs and Social Media Data. AAAI, Barcelona (2011)
29. Shaw, R., Davis, M.: Toward emergent representations for video. In: MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 431–434. ACM, New York (2005). doi:http://doi.acm.org/10.1145/1101149.1101244
30. Shaw, R., Schmitz, P.: Community annotation and remix: a research platform and pilot deployment. In: HCM '06: Proceedings of the 1st ACM International Workshop on Human-Centered Multimedia, pp. 89–98. ACM, New York (2006). doi:http://doi.acm.org/10.1145/1178745.1178761
31. Snoek, C., Worring, M., Smeulders, A., Freiburg, B.: The role of visual content and style for concert video indexing. In: Multimedia and Expo, 2007 IEEE International Conference on, pp. 252–255. IEEE, Washington, DC (2007)

32. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimed. Comput. Commun. Appl. **3**(1) (2007). doi:10.1145/1198302.1198305. http://doi.acm.org/10.1145/1198302.1198305
33. Wang, A.: The shazam music recognition service. Commun. ACM **49**(8), 44–48 (2006). doi:10.1145/1145287.1145312, http://doi.acm.org/10.1145/1145287.1145312
34. Yan, R., Hauptmann, A.G.: A review of text and image retrieval approaches for broadcast news video. Inf. Retr. **10**, 445–484 (2007). doi:10.1007/s10791-007-9031-y
35. Yew, J., Shamma, D.A., Churchill, E.F.: Knowing funny: genre perception and categorization in social video sharing. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, CHI '11, pp. 297–306. ACM, New York (2011). doi:http://doi.acm.org/10.1145/1978942.1978984. http://doi.acm.org/10.1145/1978942.1978984
36. Yu, B., Ma, W.Y., Nahrstedt, K., Zhang, H.J.: Video summarization based on user log enhanced link analysis. In: Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA '03, pp. 382–391. ACM, New York (2003). doi:10.1145/957013.957095. http://doi.acm.org/10.1145/957013.957095
37. Zhang, D., Guo, B., Yu, Z.: The emergence of social and community intelligence. Computer **44**(7), 21–28 (2011). doi:10.1109/MC.2011.65

Social Media Recommendation

Zhi Wang, Wenwu Zhu, Peng Cui, Lifeng Sun, and Shiqiang Yang

Abstract Social media recommendation is foreseen to be one of the most important services to recommend personalized contents to users in online social network. It imposes great challenge due to the dynamical behavior of users and the large-scale volumes of contents generated by the users. In this chapter, we first present the principal concept of social media recommendation. Then we present the framework of social media recommendation, with a focus on two important types of recommendations: interest-oriented social media recommendation and influence-oriented social media recommendation. For each case, we present the design of the recommendation that takes both social property and content property into account, such as user relations, content similarities, and propagation patterns. Furthermore, we present theoretical results and observations on the social media recommendation approaches.

1 Social Media Recommendation in Online Social Network

Online social network is attracting more and more people in today's Internet, where users can share and consume all kinds of multimedia contents. Social media sharing is based on online social network, where users are able to reach various contents shared by others. With the exponential growth in social media contents, such as images and videos generated by users, it is of great importance to study how to provide personalized contents in the social media service. Recommendation is foreseen to be one of the most important services that can provide such personalized multimedia contents to users. However, social media recommendation is different

Z. Wang (✉) • W. Zhu • P. Cui • L. Sun • S. Yang
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
e-mail: wangzhi04@mails.tsinghua.edu.cn; wwzhu@tsinghua.edu.cn; cuip@tsinghua.edu.cn; sunlf@tsinghua.edu.cn; yangshq@tsinghua.edu.cn

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_2, © Springer-Verlag London 2013

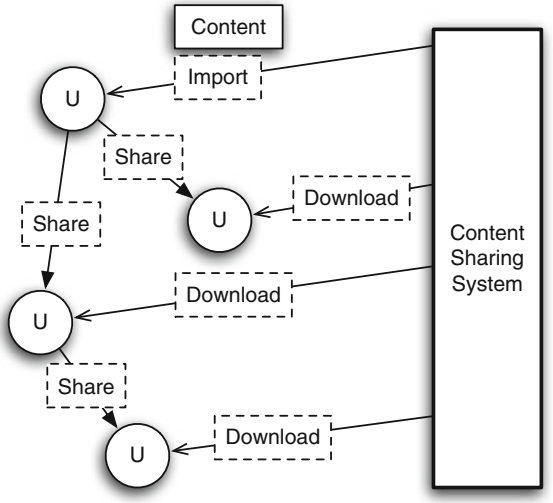


Fig. 1 Contents shared in online social network

from traditional content recommendation in that social media recommendation needs to take not only the content information but also users' social relationship and behavior into account.

1.1 Why Do People Need Recommendation in Social Media?

In online social network, many contents are generally "imported" by users from other systems, i.e., the *content sharing system* where contents are hosted to serve users. Online social network service and online content sharing service are two of the most popular applications in today's Internet. For example, online social network services like Facebook[1] and Twitter[2] have attracted hundreds of millions of users all over the world, and online content sharing system like YouTube[3] and Flickr[4] are also providing contents to billions of viewers per day.

Recent years have witnessed a rapid convergence of online social network and online content sharing network. An example of the convergence is illustrated in Fig. 1. We observe that contents are first generated by users and uploaded to the online content sharing system, e.g., more than 60 h worth of videos is uploaded every minute to YouTube.[5] Then the contents which are originally hosted by the

[1] 2012: http://www.facebook.com

[2] 2012: http://www.twitter.com

[3] 2012: http://www.youtube.com

[4] 2012: http://www.flickr.com

[5] http://www.telegraph.co.uk/technology/news/9033765/YouTube-uploads-hit-60-hours-per-minute.html

content sharing system are imported by users into the online social network and shared between users. Due to the dynamical behavior of users in the online social network and the massive number of contents generated by users, it imposes great challenge for traditional recommendations to provide personalized contents to users.

To effectively perform recommendation in the online social network, in this chapter, we present the conceptual design of social media recommendation, which takes both users and contents into account. We illustrate the advantages of social media recommendation by two types of social media recommendations in the online social network, namely, *interest-oriented social media recommendation* and *influence-oriented social media recommendation*, as follows:

- *Interest-oriented social media recommendation* answers the question "how to find the contents that can interest users in the online social network?" In this recommendation, the relevance between a user and a content is to be evaluated, so that contents that can mostly interest a user are recommended. The recommendation is based on the user relations and actions and the content similarity.
- *Influence-oriented social media recommendation* answers the question "what contents should one share to maximize one's influence?" which can be extended into two information retrieval scenarios: (1) ranking of users, given a content item, who should share it so that its diffusion range can be maximized in a social network; and (2) ranking of social media contents, given a user, what should one share to maximize one's influence among one's friends.

1.2 Traditional Approaches for Social Media Recommendation

Traditional recommendation relies on the content similarities and the collaborative references from users. Collaborative filtering-based [6, 28] and content-based [24,26] approaches have been widely used in the existing recommendation systems, where users' rating scorings can be used to predict others' interests. Since both approaches have their shortcomings when being used individually, some designs are proposed to combine their advantages. Melville et al. [23] have proposed to use the content-based predictor to enhance the existing user data so as to provide better personal suggestions through collaborative filtering. Debnath et al. [11] have proposed to improve the content-based recommendations by some weights which determine the content attributes' importance to users. The weight values are estimated from a set of linear regression equations obtained from a social network graph which captures human judgment about similarity of items.

In online social network, users' rating scores on contents are usually modeled as a user-content matrix with each entry indicating the score a user rates a content item. The user-content matrix is highly sparse, as there are a large number of missing entries in the matrix. The recommendation system is to estimate the values of the missing entries. Matrix factorization is proposed to perform the recommendation

in such model, where users and contents can be represented by vectors that are indicated in the factor matrices [18]. To finally suggest interesting contents to users, the product of the vector representing a user and the vector representing a content can be used to evaluate the relevance between them. The performance of such matrix-based model has been verified by the Netflix prize.[6]

In the context of recommendation for users' interests, Davidson et al. [10] have studied the challenges in the recommendation for such user-generated contents by taking YouTube as an example. Videos on YouTube are mostly short form, and user interactions are thus relatively short and noisy, making the traditional recommendations less effective. Baluja et al. [4] have proposed to use random walk through a co-view graph concluded from the viewing links to give video suggestions to users. Zhou et al. [40] have investigated the impact of such recommendation systems on the diversity of videos on YouTube. They observe that recommendation is the main source of views of videos on YouTube, and the number of views and the rank of recommendation are highly correlated. Walter et al. [32] have proposed a model to use the users' social connections to reach contents and filter the contents by their trust relationship. Golbeck et al. [13] have considered the social network as a recommendation network since the cascade of information is phenomenal in online social network. DuBois et al. [12] have proposed to improve the collaborative filtering recommendations by using the trust information as the weights between users.

In the context of influence maximization, there are studies on structure-level analysis [25, 29] and topic-level analysis [14, 30]. The goal of influence-oriented recommendation is to predict users' social influence for unobserved data [9] or to analyze the influence patterns from observed data [2, 3]. The existing social influence analysis research can be summarized into a diagram: *Who (A) influences whom (B) given what (C)*. A is often regarded as a single user (or node). For B, previous works can be categorized as *macroscale*, where B is the whole network [2, 37]; *microscale*, where B is a single user [8, 30]; and *mesoscale*, where B is the community or A's friends (neighborhoods) [1, 22]. From the side of C, contents can be analyzed at structure level [25, 29] and at topic level [14, 30].

1.3 What Information Can Be Utilized in Social Media Recommendation?

Traditional approaches have not taken users' social relations and their social actions into account, which is less effective in social media recommendation. In online social network, *users* and *contents* are the two core dimensions. On one hand, users can import and re-share contents in the online social network; their behavior not

[6] http://www.netflixprize.com

only indicates their interests but also reflects how contents can influence the users. On the other hand, the rapid convergence of online social network service and online content sharing service makes it possible to perform social media recommendation using information from the contents.

Our social media recommendation is based on the users and the contents generated by the users. To perform the interest-oriented recommendation and influence-oriented recommendation, we refer to the social relations, the user actions, and the content similarities:

- **Social connections**. The unique information that is available in the online social network is the set of social connections between users. On one hand, social connections reflect the social relations between people in the real world; on the other hand, social connections reflect users' interests and determine how contents can propagate among users. In different types of social network services, social connections are different, e.g., the "friending" connections in Facebook-like systems indicate users' general social interests, the "following" connections in a Twitter-like systems reflect users' information preference, and the "professional" connections in LinkedIn-like[7] systems usually show people's career interests.
- **Social actions**. The online social network has recorded valuable information on how contents are utilized by users, including which contents are *imported* by users, how these contents are *re-shared* by users, how users view these contents, and how users *comment* on these contents. On one hand, such actions make the contents propagate through the social connections and reach other users in the online social network; on the other hand, they reflect users' interests in these contents and determine the influences of users and contents, e.g., when a content item is being shared by a user, the user probably likes the content (interested) and wants his/her friends to see the content (influenced).
- **Social media analysis**. Based on the content analysis, we are able to investigate the similarities between content items. In the content analysis, multiple similarity analysis approaches can be utilized. For example, in Twitter-like systems where texts are the main form of media contents, the similarity can be evaluated using the following intuitive approaches: (1) Contents are more similar to each other if they are published in the same category, which can be inferred from the text tags in the content items, and (2) contents are similar to each other if they are published at the same location, which can be retrieved from the geographic tags. Such content similarity is not limited to text-based contents, e.g., images and videos can be analyzed using their visual features. In social media recommendation, on one hand, when performing interest-oriented recommendation, it is possible that a user likes contents that are similar to the ones he/she likes before; on the other hand, for the influence-oriented recommendation, similar contents are supposed to have similar influence properties.

[7] 2012: http://www.linkedin.com/

1.4 Basic Formulation in the Social Media Recommendation

We formulate the basic models for the social media recommendation in the online social network.

1.4.1 Interest-Oriented Recommendation

In the interest-oriented recommendation, we formulate the social connections, social actions, and content analysis used in the social media recommendation as follows:

- The *social graph* is used to represent the social connections [37], e.g., how users follow each other in an online microblogging system. In the social graph, users are represented by the nodes and the social connections are represented by the edges. We use a user-user matrix $\mathbf{A}$ to denote the social graph. In an online microblogging system, $\mathbf{A}$ can be defined as follows:

$$\mathbf{A}_{ij} = \begin{cases} 1, & i \text{ follows } j, \text{ or } i = j \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

- The *user-content graph* is used to represent how users import and share contents, which are motivated by traditional matrix completion-based recommendation approaches [7]. Users and contents are represented by the nodes, and the importing/sharing actions are represented by the edges. We use a user-content matrix $\mathbf{B}$ to represent the user-content graph, which records the importing/sharing history of users as follows:

$$\mathbf{B}_{ij} = \begin{cases} 1, & \text{user } i \text{ has imported/shared content } j \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

- The *content graph* is used to represent the similarities between the contents according to content analysis [19]. Contents are represented by the nodes, and the similarities between each other are represented by the weights of the edges. We use a content-content matrix $\mathbf{C}$ to represent the content graph, which records the "similarities" between the contents as follows:

$$\mathbf{C}_{ij} \begin{cases} = 1, & i = j \\ \in [0, 1), & \text{otherwise} \end{cases}, \tag{3}$$

  where larger $\mathbf{C}_{ij}$ indicates that content i is more similar to content j.

The three matrices are then used to describe users' interests in media contents: (1) Based on the social graph $\mathbf{A}$ and user action $\mathbf{B}$, we can infer which content items a user may like according to people one follows (friends) by using $\mathbf{A} \times \mathbf{B}$;

(2) based on the user action $\mathbf{B}$ and content similarity $\mathbf{C}$, we can also infer users' interests, since a user may like content items that are similar to the ones he/she has already imported or shared before, by using $\mathbf{B} \times \mathbf{C}$. The operation "$\times$" varies for different applications and different sparsities of the matrices. Using (1) and (2), user actions can be updated to a new user-content matrix $\mathbf{B}'$, where some missing entries can be estimated. Based on $\mathbf{B}'$, we are able to factorize user actions, i.e., the relevance between a user and a content can be calculated to perform the interest-oriented recommendation.

1.4.2 Influence-Oriented Recommendation

To formally define the recommendation for influence maximization problem, suppose we have M users with the ith user denoted as u_i and N content items with the jth item denoted as p_j. Let $\mathcal{N}(u_i)$ denote the collection of u_i's friends (in Facebook-like systems) or followers (in Twitter-like systems). There are two key factors involved in the influence-oriented recommendation:

- **Item-level social influence**: A straightforward way to define the strength of u_i's influence on $\mathcal{N}(u_i)$ given the content item p_j, denoted as f_{ij}, is the number of u_i's friends/followers who have viewed content p_j.
- **Social influence prediction**: There are $M \times N$ potential social influences in total. However, in practice, only a tiny fraction of them can be observed. The social influence prediction is to predict the unobserved social influences $\hat{f}_{ij}$ based on the observed f_{ij}'s and those predictive factors.

With the above terminologies, let us formally define the task of item-level social influence prediction. We denote the user-content influence matrix as $\tilde{\mathbf{X}} \in \mathbb{R}^{M \times N}$, with its (i, j)th entry

$$\tilde{X}_{ij} = \begin{cases} f_{ij} & \text{if } u_i \text{ shared } p_j \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

If we use g_i to denote the number of u_i's friends (i.e., $g_i = |\mathcal{N}(u_i)|$, where $|\cdot|$ is the cardinality of a collection), then $f_{ij} \leq g_i$. Also, it should be noted that in the matrix $\tilde{\mathbf{X}}$, there are two cases where an entry $\tilde{X}_{ij} = 0$. First, user u_i has not shared the content item p_j, and second, user u_i has shared content item p_j, but none of u_i's friends have viewed it.

Since different users have different numbers of friends/followers, the strength of social influence (if measured by f_{ij}) for each user-content pair should be evaluated in different scales. To alleviate its effect on the final performance, we use the following *percentile* influence matrix:

$$X_{ij} = \begin{cases} \frac{f_{ij}}{g_i} & \text{if } u_i \text{ shared } p_j \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

so that X_{ij}'s are normalized into the range of $[0, 1]$.

The user-content influence matrix $\tilde{X}$ can be reconstructed by

$$\tilde{\mathbf{X}} = Diag(\mathbf{g}) \cdot \mathbf{X}, \tag{6}$$

where $\mathbf{g} = [g_1, g_2, \cdots, g_N]^\top \in \mathbb{R}^N$ and $Diag(\mathbf{g})$ is the diagonal matrix with $\mathbf{g}$ on the diagonal line.

In this way, the recommendation of influential content items is converted to the problem of predicting the unobserved entries in $\mathbf{X}$.

Next, we will discuss the interest-oriented and influence-oriented social media recommendations, respectively.

2 Interest-Oriented Social Media Recommendation

With the large number of user-generated contents in online social network, it is crucially important for social media providers to recommend users the ones that can interest them. In this section, we first show how users' interests are represented by their social behavior in online social network. Then, we present a joint social-content recommendation framework [34].

2.1 Representation of Users' Interests in Online Social Network

Explicit scores can be provided by users to indicate their interests in multimedia contents, which are commonly used in traditional recommendation frameworks. However, such rating mechanism is not suitable in the social media recommendation as follows. (1) Ratings are only a small fraction of users' social actions in online social network—many users do not rate a content even after having imported or shared it. (2) The scores cannot always accurately reflect users' interests, e.g., users can provide different scores to the same content at different times. To address the first problem, in the social media recommendation, we use the dominant social actions to study users' interests instead of the rating scores. To address the second problem, we separate the actions so that the recommendation is performed for each action, where the binary actions can be referred to in the recommendation.

We present social media recommendation for two important social actions: importing and sharing. Importing is used by users to generate new contents in the social network, while sharing is used by users to distribute contents that are already imported by others. Correspondingly, contents are recommended for users to import or share as follows:

- *Importing recommendation* which recommends users the contents to import to their profiles in online social network. Since in popular online social network

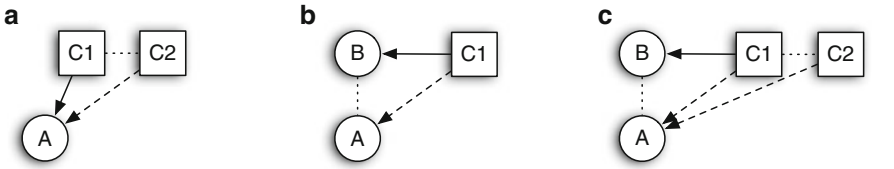


Fig. 2 Content-based, social-based, and joint social-content recommendations

systems such as Facebook and Twitter, contents (e.g., videos) are not hosted by the systems directly, instead, they are imported from other content sharing systems, the importing recommendation helps users in online social network to discover contents that they want to import to online social network, among all the contents from the external content sharing systems.

- *Sharing recommendation* which recommends users the contents to share in online social network. After users have imported contents to online social network, such contents will be distributed through the social connections. In online social network, users who obtain the contents shared by their friends or people they follow can further share the contents to their (other) friends or users following them, making the contents propagate in a cascade way [17]. The sharing recommendation helps a user discover the contents that he/she wants to share, among all the contents shared from his/her friends.

As important social actions in online social network [35], importing and sharing are implicit in representing users' interests, e.g., if a user imports a content, it only indicates that the user is interested in the content but cannot evaluate how much the user likes the content. Since rating scores are not available to connect the users and the contents, in the interest-oriented social media recommendation, social actions are divided into different groups so that the recommendation can be performed separately. The rationale is that the actions in the same group can be used to evaluate each other.

2.2 Recommendation Based on Users and Contents

Traditional interest-oriented recommendation for social media includes both *content-based approach* [26] and *social-based approach* [39]. In the content-based recommendation, content filtering and collaborative filtering [28] have been widely used. They make use of either the similarities based on content analysis or the similarities based on the historical users' ratings. Such recommendation can provide a user with contents that are similar to the ones he/she has viewed before, as illustrated in Fig. 2a. On the other hand, in the social-based recommendation, social network is used to filter the information distributed through the social connections,

so that content items that one likes can be recommended to their friends [32]. Such recommendation is able to provide users with the contents that have previously interested their friends or people they follow, as illustrated in Fig. 2b. In existing social media recommendations, users' social connections and contents' similarities are used separately.

Figure 2c illustrates the new conceptual design of a *joint social-content recommendation*, where users and the contents are used jointly to perform the recommendation. The design has the following advantages:

- Using social graph and social actions in the recommendation makes it possible to take the propagation patterns into consideration [41]. Propagation determines how contents reach different people in online social network, which can be very important to the recommendation, e.g., if a content item repeatedly appear to a user since it is repeatedly shared by his/her friends, he/she may get interested in that content item as well [33]. Given the social relations and social actions, such propagation of content items can be simulated and the strength of the propagation can be evaluated for the recommendation.
- Since importing and sharing contents in online social network are implicit, i.e., users do not rate the contents they have imported or shared, which are required in many existing recommendation approaches, we need to further refer to other information from the social network and the content sharing network, e.g., comments of users in online social network can be used to further analyze users' emotions [31].
- Cold start is even more challenging in social media recommendation for two reasons as follows. (a) Users who have just joined the system have hardly imported or shared any content in the system. It is difficult to recommend any content for them, since existing recommendation systems rely on users' historical preferences. (b) A large number of user-generated contents have no or fewer viewers. It is difficult to decide which users these contents should be recommended to. In the joint social-content recommendation, (a) and (b) can be solved by updating the user-content matrix, based on the original matrices **A**, **B**, and **C**, where both users' social connections and the content similarities can be used to assist the updating of the missing entries in **B**.

In the joint social-content recommendation, users' social connections, user actions, and content similarities are utilized to derive an updated *user-content graph* to perform joint recommendation. More specifically, first, the joint social-content recommendation utilizes a user-content matrix completion to predict which contents are to be imported/shared by which users. In the completion, entries for cold users and cold contents are updated so as to perform recommendation for such users and contents. Second, based on the updated user-content matrix and the historical importing and sharing records of users, a joint *user-content space* can be built to measure the relevance between a user and a content for the recommendation. Next, we discuss the details of the joint social-content recommendation.

2.3 The Joint Social and Content Recommendation

The joint social-content recommendation is proposed to recommend interesting contents to users by using their social relations, social actions, and the content similarities. The joint social-content recommendation framework is facing the following challenging problems: (1) How to make use of the propagation patterns of contents to update the user-content matrix? (2) How to make use of the implicit actions of users' importing and sharing contents to perform the recommendation? (3) How to improve the recommendation for newly joint users who have little historical importing/sharing information and newly published contents which are viewed by fewer users? The answer to these questions is the design of the joint social-content recommendation framework.

First, a *completion* scheme based on the social propagation and the content similarity is used to update the user-content matrix. In the user-content matrix completion, the missing entries corresponding to some cold users/contents can be estimated. On one hand, the more a user's friends have imported or shared a content, the more possible it is for the user to import or share the content as well. On the other hand, if a user has imported or shared a content, the user will probably import or share contents similar to that content as well. The matrix update procedure can be executed based on new entries that are updated, simulating the propagation of contents in online social network. The user-content matrix completion solves the aforementioned problems (1) and (3).

Second, to connect the users and the contents by their implicit social actions, we use the importing and sharing records as input to construct a dynamical user-content space to measure the relevance between a user and a content. In the user-content space, a user can be represented by a vector, each entry in which is calculated according to the contents the user has imported/shared. Similarly, a content can be represented by a vector according to the user actions as well. Thus, the construction of the user-content space only relies on the binary entries in $\mathbf{B}'$. Each import or share will be regarded as a small increase of the user's interest in a particular dimension in the user-content space. The user-content space construction solves the aforementioned problem (2).

2.3.1 Completion of the User-Content Matrix

The user-content matrix which indicates users' social actions on the contents can be very sparse [15]. It is difficult for the traditional recommendation algorithms to recommend contents to users who have imported/shared no or fewer contents, which are based on users' historical preferences. In a joint social-content recommendation, the user-content matrix ($\mathbf{B}$) can be updated by the collaboration of social graph ($\mathbf{A}$), social actions (the original $\mathbf{B}$), and content analysis ($\mathbf{C}$). The completion makes use of the original three matrices simultaneously. On one hand, the social

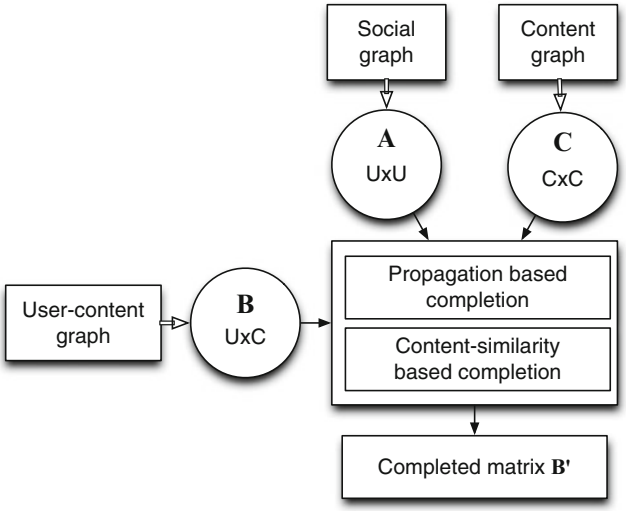


Fig. 3 User-content matrix completion based on propagation and content similarity

propagation model can be used to connect the contents to the users, i.e., contents imported/shared by a user's friends are likely to be imported/shared by the user as well. On the other hand, contents can be connected to users according to the content similarity analysis. The reason is that a user is likely to enjoy contents that are similar to the ones he/she has imported or shared before. Figure 3 illustrates the user-content completion framework. When updating the user-content matrix **B**, besides the original user-content matrix **B**, the user-user matrix **A** and the content-content matrix **C** are all used. In particular, **A** and **B** will be used for the social propagation-based completion, and **C** and **B** will be used for the content similarity-based completion.

2.3.2 Construction of the User-Content Space

The recommendation is based on constructing a joint user-content space to measure the relevances between users and content items in the space. Figure 4 illustrates how the relevance between a user and a content item is measured. The joint user-content space is constructed by combining a user space and a content space. The user space depends on the user-user matrix, and the content space depends on the content-content matrix. A user or a content item can be mapped to two *description vectors* [38] in both spaces. Since both the user space and the content space are constructed using a similar procedure, we take the user space as an example to illustrate how the space is constructed. First, several representative users are selected, which are the top-followed users in the online social network. Since these users can be followed by many other users, it is effective to use them to represent other users' interests.

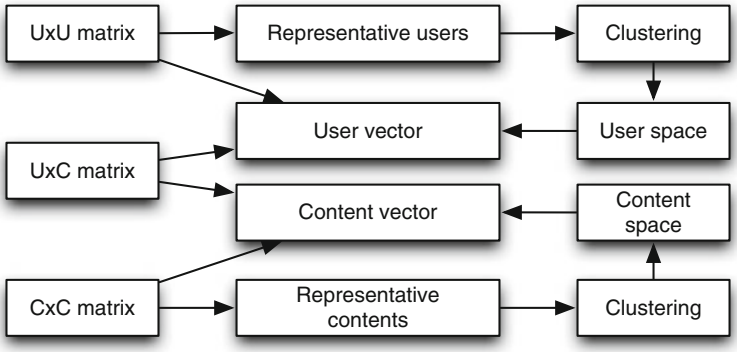


Fig. 4 Construction of the joint user-content space based on user actions

Second, these representative users are clustered into several groups, according to a similarity defined between any two representative users (two representative users are more similar to each other if they are followed by more common fans). Third, based on the clustered groups and the user-user matrix, we can define a user's description vector in the user space by counting the number of representative users he/she follows in each group, and based on the user-content matrix, we can define a content's description vector in the user space by combining the description vectors of users who have imported/shared the content.

The *relevance* between a user and a content is then measured by combining their dot products in both spaces, where a recommendation parameter is utilized to control their impacts for different social actions, i.e., importing and sharing. Finally, contents can be recommended according to their relevances to the user.

3 Influence-Oriented Social Media Recommendation

With the rapid proliferation of social applications, more and more user profiles, interactions, and collective intelligence (such as social tags and comments) are available online, which opens a new perspective for recommendation applications where more focus should be put on user collaborative information. At the same time, new recommendation scenarios, such as friend recommendation [16, 21], have also emerged. These scenarios propose a new challenge to traditional recommendation: how to effectively handle it in the social media scenario?

One key concept related to this challenge is *recommendation for influence maximization*, which has been becoming a prevalent and complex force governing the dynamics of social relationship or social network [37]. It is also a key dimension for modern recommendation in multiple aspects. To mention a few, (1) each user acts as an information source in social network, and the influence of a user is meaningful for the authority of the generated information; (2) in the

influence-oriented recommendation, the social influence is the key indicator for content recommendation; (3) as different contents vary with the power to affect users to change their actions, they can be recommended by influence ranking for social purpose. Therefore, there is a clear need for techniques to analyze social influence and, more importantly, in recommendation field.

In the influence-oriented social media recommendation, we need to answer the question "who should share what?" which can be extended into two scenarios as follows. (1) Users ranking: given an item, who should share it so that its diffusion range can be maximized in a social network. (2) Content ranking: given a user, what should one share to maximize one's influence among one's neighbors.

Item-level social influence is not a general measure on users but on the interactions of users and contents. That is, we need to discriminate a user's social influences with respect to different contents. Different from most of the existing research works focusing on users' overall social influence analysis [25, 29] and topical social influence mining [14, 30], the social influence in this section is a finely grained measure of influence.

3.1 Motivating Recommendation Scenario

In social media, users and contents are two core dimensions, and users' sharing of contents (such as blog, news, and album) is the basic behavior. Actually, the spreading out of contents is because of the user sharing in social network. The owner of the contents, e.g., the advertisers, hopes to maximize the diffusion range of the contents [5]. This goal makes them desire to target the influencers, who are able to let many friends to click the information they share or even share further to extend the sharing cascades. Psychologically, users share contents with their friends mainly because they want to build their reputations and help others, in which *to influence others* is the important motivation for sharing [35].

According to the definition of social influence on Wikipedia,[8] social influence occurs when "an individual's thoughts, feelings or actions are affected by other users." In the context of online social network systems like Facebook and Twitter, when a user shares a content item, a portion of his/her friends (or neighbors) will click, comment, or even re-share the content, which are three levels of influence [36]. The resulted predictive model can be used in two angles. On one side, given a content item, we can find out the influencers for the diffusion. On the other, given a user, we can recommend a list of content items to share, which can improve the interactions between the user and his/her friends.

In predicting the item-level social influence, we mainly face the following challenges:

[8] 2012: http://en.wikipedia.org/wiki/Social_influence

- **User-content specific**. Item-level social influence is not a general measure on users but on the interactions of users and contents. That is, we need to discriminate a user's social influences with respect to different content items. Different from most of the existing research works focusing on users' overall social influence analysis and topical social influence mining, the social influence in this chapter is a finely grained measure of influence.
- **Sparsity**. The interactions between users and content items are extremely sparse compared with the total number of user-content pairs. According to our statistics of 34 K users in Renren,[9] which is a Facebook style social network site in China, each user only shares six web contents in average during a month, compared with a total of 43 K content items, and each item is only shared by four users, compared with a total of 34 K users. Thus, it is clear that we need subtle and effective prior knowledge for user and content grouping to alleviate the sparsity problem.
- **Complex factors**. There are a volume of factors that affect how many friends will click a shared content item and provide potential clues for user and content grouping, for example, the total number of friends, the tie strength between the user and his/her friends and the semantics of content items, which are often in different scales. How to select the effective factors and integrate these complex factors in one predictive model is also one of the focus in the influence-oriented recommendation.

We formulate the recommendation for influence maximization problem as the estimation of a user-content matrix, in which each element (i, j) represents the number of clicks by friends of user i on his/her jth shared content item. A *hybrid factor nonnegative matrix factorization (HF-NMF)* algorithm is designed for item-level social influence modeling [9]. The algorithm tries to find out the common hidden vector space for both the users and the contents, where their multiplication can well approximate the observed training interaction matrix. Meanwhile, in order to deal with the sparsity problem, we construct the priors on users and contents by incorporating the user-user similarity matrix and content topic distribution matrix. Also, in order to alleviate the over-fitting problem, we introduce the $L2$-norm as regulations for the hidden vector space to improve the generalization ability. We apply *projected gradient* [20] to solve the HF-NMF problem and carry out intensive experiments to demonstrate the effectiveness of the proposed method. To summarize, the algorithm for recommendation for influence maximization includes:

- The formulation of the item-level recommendation for influence maximization problem with HF-NMF and an efficient projected gradient method to solve it.
- The predicted item-level social influence from HF-NMF can support the applications, such as influencer ranking and contents recommendation by user-content matrix ranking in two directions.

[9] http://www.renren.com

- The strength of social influence in this method is well interpreted, which makes it easy to understand and extendable to higher-order social influence, e.g., the influence on all the friends and the friends of friends.

3.2 Recommendation for Influence Maximization

Let us demonstrate the necessity of the item-level social influence and validate the rationality of predictive factors by preliminary statistical analysis.

The dataset is acquired from the real social network Renren. Till now, the web site already owns more than 150 million active users. In this web site, a user can generate a content or share a web page as a content, and the user's friends will be informed through the news-feed mechanism. Then some of the friends will click, comment, or share the content. In this section, we only consider the click action as the manifestation of influence, and the number of clicks corresponds to the strength of influence. As the number of friends is different for each user, the upper bound of users' social influence strengths is also different. In order to make the strength of influence be measured in a unified scale for different users for the sake of observational and modeling study, we use the proportion of friends (of the user who publish a content) who click the shared content as the measure.

In the influence-oriented recommendation, we first assume that the influence should be specific on each user-content pairs. In order to validate the hypothesis, we randomly select three active users. Given a user, we calculate the proportion of his/her influenced friends (who clicked the shared content) for each of the shared contents. Then, we randomly select three popular contents. Given a content, we calculate the proportion of influenced friends when the content is shared by different users. In our observations, the social influence notably varies with different users and contents, which implies that (1) different users have different influence power to their friends, (2) different contents have different influence power (more intuitively, attraction) to users who are interested in, and (3) users' influences manifest differently for different contents. Therefore, only item-level social influence can reveal the users' real influence on friends, and the strength of influence should definitely be user-content specific.

3.3 Predictive Factors for Influence-Oriented Social Media Recommendation

The factors that affect the strength of social influence include the following three aspects:

- **User-specific factors**. Although users' social influence varies with the shared contents, the average of the social influences over contents determines the overall

social influence of a user. We regard the factors that affect users' overall social influence (excluding the contents) as the user-specific factors.
- **Content-specific factors**. Similar as user-specific factors, we regard the factors affecting contents' overall social influence (excluding the users) as the content-specific factors.
- **User-content-specific factors**. As mentioned above, the social influence is user-content specific. The social influence of a user given a content cannot be well estimated only by the user and content-specific factors. The factors indicating the interactions between users and contents are also important for social influence-oriented recommendation.

One issue that is worthy of emphasizing here is that both the users and contents are essential for the predictive modeling. On one hand, the user-content interactions are very sparse. We need to find effective factors to "group" those users and contents to alleviate the sparsity problem. On the other hand, the user and content-specific factors also provide some effective prior knowledge to complement the inference from pure user-content interactions.

In order to find out the effective predictive factors for the influence-oriented recommendation, we first prepare a factor pool, which includes the available potential predictive factors including user profiles, number of users' friends, visiting frequency between users, and contents' topic distributions. Given each factor, we measure the correlation between the strength of social influence and the factor value. Finally, we select two user-oriented factors: the percentage of active friends, the average social tie strength (the interaction frequency) between a user and his/her friends, and one content-specific factor—the topic distribution of a content's content.

3.4 Influence Maximization: A Matrix Factorization Approach

Based on the symbols and basic model we have illustrated in Sect. 1, the influence maximization can be formulated as a matrix factorization problem. Let $\mathbf{U} \in \mathbb{R}^{M \times k}$ be the latent user feature matrix and $\mathbf{V} \in \mathbb{R}^{N \times k}$ be the latent content feature matrix, where k is the number of latent features. Then given the observed user-content-specific social influence matrix $\mathbf{X}$, the objective is to find the optimal latent user matrix $\mathbf{U}$ and latent content matrix $\mathbf{V}$ by minimizing the following objective:

$$\mathscr{J}_1 = \left\| \mathbf{X} - \mathbf{U}\mathbf{V}^\top \right\|_F^2, \tag{7}$$

where $||.||_F$ denotes the matrix Frobenius norm.

The objective function $\mathscr{J}_1$ can be regarded as the quality of approximating $\mathbf{X}$ by the inner product of $\mathbf{U}$ and $\mathbf{V}$. However, in real cases, most of the elements in $\mathbf{X}$ are zero because of the sparse interactions between users and contents. Thus, in order to focus more on the valid elements, we propose to only measure the approximation

loss on observed elements on $\mathbf{X}$. To formulate this, we introduce the *sharing matrix* $\mathbf{Y} \in \mathbb{R}^{M \times N}$ with its (i, j)th entry defined as

$$Y_{ij} = \begin{cases} 1 & \text{if } u_i \text{ shared } p_j \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

The objective function is converted to

$$\mathscr{J}_2 = \left\| \mathbf{Y} \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^\top) \right\|_F^2, \tag{9}$$

where $\odot$ is the Hadamard product.

As mentioned above, the severe sparsity of $\mathbf{X}$ makes it very challenging to directly learn the latent spaces for users and contents from only observed user-content interaction entries. We need to make full use of the user-specific and content-specific factors to compress the degrees of freedom, so that the correlation within users and contents can be exploited to alleviate the sparsity problem. The solution to the optimization is similar to the *maximum margin matrix factorization* approach in [27] and the *joint matrix factorization* approach in [18].

4 Conclusions

In this chapter, we introduce the principal concept and the framework of social media recommendation. We first present why recommendation is demanded in social media and what unique information is needed for effective social media recommendation from both users and contents. We then provide the theoretical formulation and key insights on the social media recommendation. In particular, we have presented two types of recommendations in online social network: interest-oriented social media recommendation and influence-oriented social media recommendation. To deepen the understanding of each social media recommendation, we further provide the following results:

- In the interest-oriented social media recommendation, we present the joint social-content recommendation. We have shown that user relations, user actions, and content similarities can be simultaneously utilized in the recommendation, so that two problems can be addressed as follows. (1) The completion of the user-content matrix based on the social propagation and the content similarity makes it possible to recommend newly published contents to newly joint users. (2) The construction of the joint user-content space enables recommendation for implicit social actions including importing and sharing.
- In the influence-oriented social media recommendation, we have presented the effective user-specific and content-specific predictive factors and used a matrix factorization method to incorporate these predictive factors for user-content-specific recommendation for social influence maximization at content-item level.

We see such an exploration will shed light on future applications in social media recommendation.

References

1. Agarwal, N., Liu, H., Tang, L., Yu, P.: Identifying the influential bloggers in a community. In: Proceedings of the ACM International Conference on Web Search and Web Data Mining, Palo Alto (2008)
2. Anagnostopoulos, A., Kumar, R., Mahdian, M.: Influence and correlation in social networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas (2008)
3. Bakshy, E., Karrer, B., Adamic, L.: Social influence and the diffusion of user-created content. In: Proceedings of the ACM Conference on Electronic Commerce, Stanford (2009)
4. Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., Aly, M.: Video suggestion and discovery for Youtube: taking random walks through the view graph. In: Proceedings of the ACM WWW, Beijing (2008)
5. Bao, H., Chang, E.: Adheat: an influence-based diffusion model for propagating hints to match ads. In: Proceedings of the ACM WWW, Raleigh (2010)
6. Breese, J., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco (1998)
7. Cai, J.-F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
8. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas (2008)
9. Cui, P., Wang, F., Liu, S., Ou, M., Yang, S., Sun, L.: Who should share what? Item-level social influence prediction for users and posts ranking. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information. ACM, New York (2011)
10. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al.: The Youtube video recommendation system. In: Proceedings of the ACM Conference on Recommender Systems, Barcelona (2010)
11. Debnath, S., Ganguly, N., Mitra, P.: Feature weighting in content based recommendation system using social network analysis. In: Proceedings of the ACM WWW, Beijing (2008)
12. Golbeck, J., Kleint, J., Srinivasan, A.: Improving recommendation accuracy by clustering social networks with trust. ACM RecSys'09 Workshop on Recommender Systems & the Social Web, New York (2009)
13. Golbeck, J., Hendler, J.: FilmTrust: movie recommendations using trust in web-based social networks. In: Proceedings of the IEEE Consumer Communications and Networking Conference, Las Vegas (2006)
14. Goyal, A., Bonchi, F., Lakshmanan, L.: Learning influence probabilities in social networks. In: Proceedings of the ACM International Conference on Web Search and Data Mining, New York (2010)
15. Johnson, C.: Matrix completion problems: a survey. In: Matrix Theory and Applications, vol. 40, pp. 171–198. American Mathematical Society, Providence (1990)
16. Kalashnikov, D., Chen, Z., Mehrotra, S., Nuray-Turan, R.: Web people search via connection analysis. IEEE Trans. Knowl. Data Eng. **20**(11), 1550–1565 (2008)
17. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington (2003)

18. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *Computer* **42**(8), 30–37 (2009)
19. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage, Thousand Oaks (2004)
20. Lin, C.: Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**(10), 2756–2779 (2007)
21. Macdonald, C., Ounis, I.: Voting for candidates: adapting data fusion techniques for an expert search task. In: Proceedings of the ACM International Conference on Information and Knowledge Management. ACM, New York (2006)
22. McMillan, D., Chavis, D.: Sense of community: a definition and theory. J. Community Psychol. **14**(1), 6–23 (1986)
23. Melville, P., Mooney, R., Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. In: Proceedings of the National Conference on Artificial Intelligence. AAAI, Menlo Park (2002)
24. Mooney, R., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the ACM Conference on Digital Libraries, San Antonio (2000)
25. Newman, M.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
26. Pazzani, M., Billsus, D.: Content-based recommendation systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, pp. 325–341. Springer, Berlin/New York (2007)
27. Rennie, J., Srebro, N.: Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of the ACM International Conference on Machine Learning. ACM, New York (2005)
28. Schafer, J., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, pp. 291–324. Springer, Berlin/New York (2007)
29. Strogatz, S.: Exploring complex networks. *Nature* **410**(6825), 268–276 (2001)
30. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris (2009)
31. Thelwall, M., Wilkinson, D., Uppal, S.: Data mining emotion in social network communication: gender differences in myspace. J. Am. Soc. Inf. Sci. Technol. **61**(1), 190–199 (2010)
32. Walter, F., Battiston, S., Schweitzer, F.: A model of a trust-based recommendation system on a social network. Auton. Agents Multi-Agent Syst. **16**(1), 57–74 (2008)
33. Wang, Z., Sun, L., Yang, S., Zhu, W.: Prefetching strategy in peer-assisted social video streaming. In: Proceedings of the ACM Multimedia, Scottsdale (2011)
34. Wang, Z., Sun, L., Zhu, W., Yang, S., Li, H., Wu, D.: Joint social and content recommendation for user generated videos in online social network. Technical report
35. Wasko, M., Faraj, S.: Why should i share? Examining social capital and knowledge contribution in electronic networks of practice. Mis Q. **29**, 35–57 (2005)
36. Weng, J., Lim, E., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the ACM International Conference on Web Search and Data Mining, New York (2010)
37. Wolfe, A.: Social network analysis: methods and applications. Am. Ethnol. **24**(1), 219–220 (1997)
38. Xu, R., Wunsch, D., et al.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
39. Yang, W., Dia, J., Cheng, H., Lin, H.: Mining social networks for targeted advertising. In: Proceedings of the IEEE Annual Hawaii International Conference on System Sciences. IEEE Computer Society Press, Los Alamitos (2006)
40. Zhou, R., Khemmarat, S., Gao, L.: The impact of YouTube recommendation system on video views. In: Proceedings of the ACM IMC, Melbourne (2010)
41. Ziegler, C., Lausen, G.: Propagation models for trust and distrust in social networks. Inf. Syst. Front. **7**(4), 337–358 (2005)

Multimedia Indexing, Search, and Retrieval in Large Databases of Social Networks

Theodoros Semertzidis, Dimitrios Rafailidis, Eleftherios Tiakas, Michael G. Strintzis, and Petros Daras

Abstract Social networks are changing the way multimedia content is shared on the Web, by allowing users to upload their photos, videos, and audio content, produced by any means of digital recorders such as mobile/smartphones and Web/digital cameras. This plethora of content created the need for finding the desired media in the social media universe. Moreover, the diversity of the available content inspired users to demand and formulate more complicated queries. In the social media era, multimedia content search is promoted to a fundamental feature toward efficient search inside social multimedia streams, content classification, and context and event-based indexing. In this chapter, an overview of multimedia indexing and searching algorithms, following the data growth curve, is presented in detail. This chapter is thematically structured in two parts. In the first part, pure multimedia content retrieval issues are presented, while in the second part, the social aspects and new, interesting views on multimedia retrieval in the large social media databases are discussed.

T. Semertzidis (✉) • M.G. Strintzis
Information Processing Laboratory, Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki, Greece

Information Technologies Institute, Centre For Research and Technology Hellas, Thessaloniki, Greece
e-mail: theosem@iti.gr; strintzi@eng.auth.gr

D. Rafailidis • E. Tiakas • P. Daras
Information Technologies Institute, Centre For Research and Technology Hellas, Thessaloniki, Greece
e-mail: drafail@iti.gr; tiakas@csd.auth.gr; daras@iti.gr

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_3,

1 Introduction

Social networking sites enabled multimedia content sharing in large volumes, by allowing users to upload their photos, videos, and audio data, produced by any means of digital recorders. Moreover, the huge volumes of information in the social media inspired users to formulate new types of queries that pose complex questions to these heterogeneous databases.

An attempt to index and retrieve multimedia content shared through the social media, using techniques of the content-based multimedia retrieval community, shows clearly the inability to do so. The parameters and constraints posed from the social media aspect reformed the multimedia indexing and retrieval processes in a new problem seeking for new solutions. The major problems of social multimedia indexing and retrieval exist due to (a) the enormous volumes of information that push existing techniques to their edges and (b) the well-known semantic gap [69] between the low-level multimedia descriptors and the higher-level concepts that exist in each multimedia content. These facts do not imply that previous knowledge and tools are totally useless, rather that they should be used in a different way.

This chapter aims to structure the concept of multimedia indexing, search, and retrieval, based on the data growth curve from the small, locally stored, multimedia collections to the huge, heterogeneous, and context-rich social multimedia collections, and to present interesting works along the way.

Multimedia indexing, search, and retrieval is a multistep process that deeply depends on the content type and its characteristics. The typical content-based multimedia indexing (CBMI) and retrieval methods apply the well-studied query-by-example (QBE) paradigm, where a multimedia (MM) object is used as a query to retrieve similar multimedia objects (See Fig. 1). In this chapter, we mainly focus on image indexing, search, and retrieval, yet the concepts, workflows, and conclusions are valid also for other multimedia objects such as video, audio, and 3D. A common

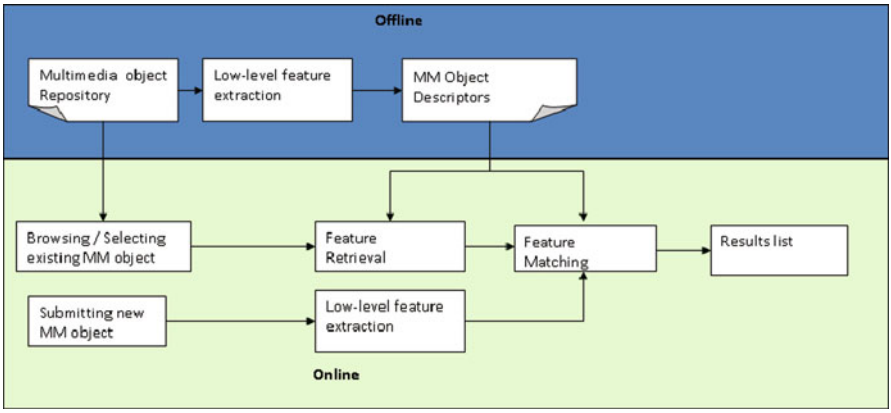


Fig. 1 A typical content-based multimedia retrieval process

initial step of multimedia indexing is the extraction of characteristic features from the content in order to describe it in a more compact and discriminative manner. A plethora of works has been published on the field, and many robust and well-evaluated multimedia content descriptors exist that encode dominant features such as color information, texture and edge, spectral characteristics, and motion (e.g., SIFT, self-similarity, CEDD [15, 44, 64, 74]) targeted to specific application areas [22]. The next step is to define a distance function (such as L1, L2, Mahalanobis) between these descriptors in order to compare the similarity between multimedia objects. As it is obvious, when volumes of data increase dramatically (as in the case of social media data), this part of the process becomes extremely time-consuming to be performed in real time. Indexing structures (discussed in Sect. 2) are data structures aiming to reduce comparisons and consequently reduce the search time. However, content-based multimedia retrieval restricts user queries to the QBE approach which is not sufficient for the environment of the large social multimedia databases for various reasons. The responsiveness of the services in a timely manner, the inability to map with the *subjective* user semantics to enhance the quality of the retrieved results, and the new, complex types of user queries are some of them. In the context of large social multimedia databases, the available social metadata are used to give answers to these challenging problems.

Sections 2 and 3 present, in a compact way, indexing structures aiming to address the problem of searching inside large collections of multimedia objects. Section 2 discusses multidimensional indexing structures, by classifying them either in exact or approximate approaches. Next, in Sect. 3, a use case of one million images from the Flickr collection is examined to evaluate indexing in both exact and approximate approaches. In these two sections, the aim of the algorithms is to reduce search time and storage/memory overhead while keeping accuracy as close as possible to the baseline which is the exhaustive search. Exhaustive search is the process of comparing the query descriptor vector with the descriptor vectors of all multimedia objects of a database, i.e., one-to-one comparison without using any indexing structure. Although this gain is very important for the volumes of real-world databases such as the social multimedia databases, the other major issue of bridging the semantic gap between the low-level multimedia descriptors and the concepts included in the multimedia objects remains unsolved. Toward these objectives, a new area of multimedia computing that clearly takes into consideration the social aspects of the social media data has recently emerged [48, 71]. Section 4 examines exactly this part of social multimedia indexing and retrieval and categorizes the methods presented in three subsections: (a) context-based multimedia retrieval, where the contextual information stored in social media is used to semantically enhance the retrieved results and thus improve the overall accuracy; (b) event-based indexing, as another way of indexing multimedia content in a more user-oriented conceptualization; and (c) time-related multimedia indexing and search for evolving social multimedia collections. Finally, Sect. 5 discusses the conclusions drawn from this work and presents future challenges toward efficient social multimedia indexing, search, and retrieval in the ever-growing social multimedia collections.

2 Content-Based Multimedia Indexing

The availability of multimedia in the large databases of social networks emerged the imperative need to address the challenge of content-based searching, where users pose multimedia objects as queries, in order to find relevant content (see Fig. 1). However, exhaustive searching is infeasible for the large-scale applications of social networks due to the extensive time consumption it requires. Thus, the large databases of social networks should be supported by indexing schemes which are able to provide (a) low space requirements for storing the multimedia content within the indexing scheme, (b) efficient search time, and (c) high retrieval accuracy. Nevertheless, multimedia objects like compressed images, and video and audio streams are usually described by sequences of descriptor vectors with over than a thousand dimensions. In this high-dimensional space, the performance of existing indexing schemes deteriorates significantly, since content-based similarity search in high dimensions ($\geq$1,000) is challenging, due to the well-known problem of dimensionality curse [7, 16]. In order to address the aforementioned challenges, the existing indexing schemes are divided into two main categories: those that follow (a) exact and (b) approximate similarity search strategies.

The family of exact similarity search, despite achieving identical retrieval accuracy to exhaustive search, fails to support the high dimensionality. Meanwhile, storage space and search time are dramatically increased. The family of indexing schemes that follow the approximate search strategy, despite reducing the space and search time requirements, fails to preserve the retrieval accuracy of the exhaustive search.

2.1 Exact Similarity Search

In the family of exact similarity search indexing schemes, a state-of-the-art approach is the M-tree and its variants [20, 21, 47]. The most efficient way to construct the M-tree is using the bulk-load method [21]. The M-tree structure manages the query processing according to the distances between multimedia objects, which are stored as nodes. Additionally, the M-tree has been further extended, in order to support both exact and approximate strategies, while preserving the same indexing structure of the M-tree file. Furthermore, M-trees have been introduced to perform indexing and searching, not only in content-based multimedia applications, but also in other similarity search applications due to their dynamic ability to support insertions and deletions efficiently [17]. Recently, M-trees were evaluated in distributed environments, in order to support Web-scale applications [5]. The idea was to build small M-trees in each node of the distributed environment and perform the similarity search strategy to all relative nodes.

Moreover, a plethora of alternative exact similarity search indexing schemes have been proposed in the literature. The most important of them are presented in

Table 1 Exact similarity search indexing schemes

Method name	Abbreviation	Reference
R-tree	R-tree	[32]
KD-tree	KD-tree	[8]
Quad-tree	Quad-tree	[61]
Burkhard-Keller tree	BKT	[13]
Fixed queries tree	FQT	[2]
Fixed height FQT	FHQT	[3]
Fixed queries array	FQA	[16]
Vantage point tree	VPT	[18]
Multi-vantage point tree	MVPT	[11]
Vantage point forest	VPF	[81]
Bisector tree	BST	[50]
Generalized hyperplane tree	GHT	[12]
Geometric NN access tree	GNAT	[14]
Voronoi tree	VT	[24]
Pivoting M-tree	PMT	[67]
Nearest-neighbor graphs	NNG	[68]

Table 1. However, in these schemes there are several constraints. In particular, the family of BKT, FQT, FHQT, and FQA is constrained to metric spaces derived by a distance measure that necessary returns discrete values. In case that a continuous distance function is applied or a large amount of different discrete values are returned, it is infeasible to exploit these indexing schemes, as explained in [17]. The rest of the aforementioned methods support continuous distances, applicable to general metric spaces. However, the dynamic capabilities of insertions and deletions and the input/output (I/O) cost must also be considered. The VPT, MVPT, and VPF methods support latter insertions insufficiently [17]. Additional and more complicated problems appear in methods like GHT, BST, VT, GNAT, VPT, MVPT, VPF, PMT, and NNG for deletion operations [17].

Therefore, the M-tree is the unique exact indexing scheme, which supports dynamic operations efficiently [17]. The M-tree and its variants have been designed specifically for secondary memory operations and can be balanced to maintain the I/O cost low, compared with the aforementioned exact similarity search indexing schemes. However, due to the problem of dimensionality curse [7, 16], M-trees are transformed to high-level trees with many internal nodes, resulting in enormous increase of the I/O cost, unsuitable for performing efficient content-based search and retrieval in databases with billions of records such as the ones in social networking sites.

2.2 Approximate Similarity Search

In the family of approximate methods, a state-of-the-art approach is the locality-sensitive hashing (LSH) [31] and its variants, which are used for indexing of

high-dimensional data for multimedia search and retrieval. The basic idea of LSH is (a) to encode the distances between the multimedia objects into the form of compressed sequences of bits, while using hash functions, and (b) to store the encoding distances into tables, in order to ensure that the probability of collision is much higher for multimedia objects that are close to each other than those that are far apart. Therefore, the LSH-based indexing schemes vary according to the respective hashing function, trying to reduce the search time, while maximizing the retrieval accuracy, by minimizing the approximation error. Thus, the LSH variants are categorized in data-dependent [19,23,35,38,55,58,63,77] and data-independent ones [36, 37, 39, 51, 52, 60, 78], where efficiency improvements of data-dependent methods over independent ones have been proved in several studies [36, 60, 78].

Alternative approximate techniques have also been introduced in the literature, such as the spatial approximation tree (SAT) [49], the approximating eliminating search algorithm (AESA) [75], and the linear approximating eliminating search algorithm (LAESA) [46]. The main advantage of these approximate methods is the significant reduction of search time, while their main disadvantage is the low retrieval accuracy. Moreover, such approximate methods require a time-consuming preprocessing step of the multimedia content, in order to increase the retrieval accuracy [45]. Additionally, these methods are not able to efficiently support dynamic changes of the insertion and deletion operators, since for each change a full preprocessing step is required [17]. For example, SAT does not support insertions, and consequently, the whole indexing structure has to be built from scratch [17]. Despite the fact that AESA and LAESA support dynamic operations, during their search approach, all disk pages of the indexing structure have to be read, which results in limited I/O performance, and consequently, the search time is highly increased [17]. Therefore, research has been focused on LSH-based approaches, which are able to support both dynamic operations and efficient search time. However, their achieved retrieval accuracy is rather low.

3 Use Case: Flickr's One Million Images

In this section, we present a case study of Flickr's one million image dataset[1] [26, 27]. The photo sharing Website Flickr has over six billion images and is a representative example of the large-scale problem in multimedia indexing, search, and retrieval in the large databases of social networks.

In order to build the evaluation dataset, a recent variant of SIFT descriptors [74] was used. Consequently, several collections of descriptor vectors were constructed, by varying the number of dimensions from 64 to 1,024, where the resulted datasets are denoted by SIFT-64dim, SIFT-128dim, and SIFT-256dim, SIFT-512dim, SIFT-1024dim, respectively. All collections were indexed by the exact approach of M-tree

[1] For further details visit Image CLEF 2011, the "Visual Concept Detection and Annotation" task.

Table 2 Disk space requirements in GB

SIFT dataset	Exhaustive	M-tree	LSH-1L	LSH-2L
SIFT-64dim	0.238418579	2.088517205	0.556949615	0.871504784
SIFT-128dim	0.476837158	4.239689925	0.795368195	1.109923363
SIFT-256dim	0.953674316	8.097807758	1.272205353	1.586760521
SIFT-512dim	1.907348633	15.95268128	2.225879669	2.540434837
SIFT-1024dim	3.814697266	30.94820169	4.133228302	4.44778347

and the approximate approach of locality-sensitive hashing, since both methods are superior over other indexing schemes in exact and approximate similarity search approach, respectively.

However, in the case of M-tree, it is infeasible to preprocess the 1M SIFT-1024dim dataset for any parameter combination, like node size, utilization, and split strategy [21]. Therefore, we measured the corresponding results similar to the case of 100 K multiplied by 10, assuming that 10 M-trees are built, by splitting the SIFT-1024dim dataset into equal size datasets, of 100 K. For the remaining datasets we report directly the performance of M-tree, by identifying the optimal parameter selection.

Moreover, in the case of LSH, we varied the number of hash tables L, to achieve the maximum retrieval accuracy, while preserving the search time below the respective search time of exhaustive search. Therefore, we concluded that in the case of SIFT descriptors, the maximum number of L hash tables equals 2, since further increase results in exceeding the search time of exhaustive search. Additionally, in order to encode the multimedia distances, hash keys of 1,024 bits were used, resulting in the maximum retrieval accuracy of LSH.

In order to demonstrate and identify the aforementioned challenges, as presented in Sect. 2, three respective experiments were conducted, concerning (a) space requirements, (b) search time, and (c) retrieval accuracy.

Firstly, in Table 2, we present the construction requirements for (a) the exact method of the M-tree family and (b) the approximate method of the LSH family. The column "exhaustive" denotes the case of performing exhaustive search, and consequently, an indexing scheme is not required. Therefore, the required disk space is equal to the size of each dataset. Based on the experimental results shown in Table 2, we conclude that (a) the M-tree indexing scheme requires a significant amount of space, ten times greater than the corresponding dataset space, since high-level trees are constructed, consisting of many internal nodes, and (b) LSH requires an important amount of additional space for storing the constructed hash tables, linked to buckets, in which the IDs of the corresponding multimedia are stored.

Next, we evaluate M-tree and LSH against the exhaustive search approach, in terms of search CPU time and disk accesses, by varying the size of the SIFT-1024dim dataset. For measuring the search time, CPU and I/O time are separately reported, following either a memory-based or a disk-based approach. Since, the I/O time is completely dependent on (a) the running operating system, (b) memory/disk cache, and (c) hard disk specifications, the corresponding disk accesses (DA) are

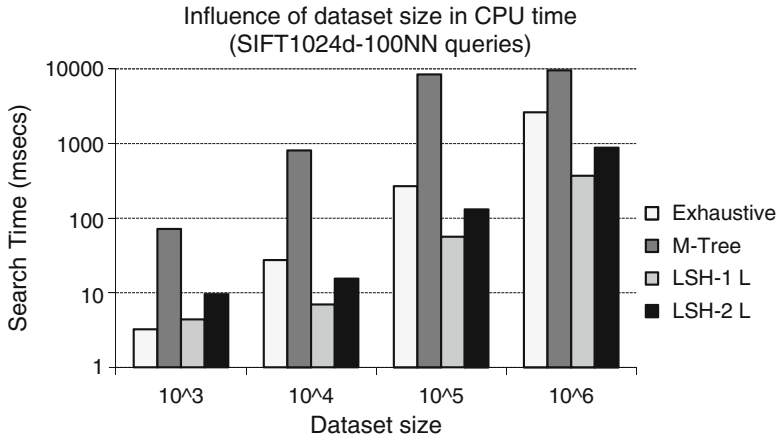


Fig. 2 Search CPU time versus dataset size for 100-NN queries in SIFT-1024dim dataset

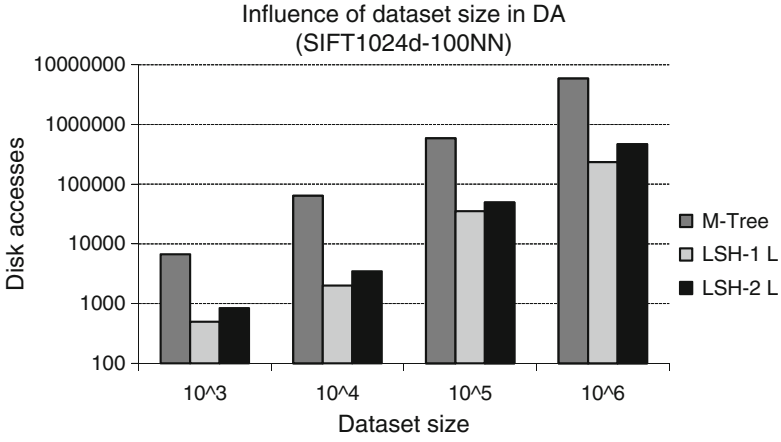


Fig. 3 Search disk accesses versus dataset size for 100-NN queries in SIFT-1024dim dataset

measured, assuming that a disk page is equal to 4 KBytes. Moreover, to produce the different datasets, 10^3, 10^4, and 10^5 images were randomly selected from the initial 10^6 SIFT-1024dim dataset. In Figs. 2 and 3, the respective results are presented (note that the exhaustive search method is omitted from Fig. 3, since the disk-level implementation is not feasible without using an external indexing scheme). We can make the following observations: (a) in each method, search CPU time and disk accesses are increased with respect to the dataset size; (b) the M-tree requires significantly high search times and disk accesses, even higher than the respective times of the exhaustive approach, verifying the dimensionality curse problem faced by the family of exact similarity search methods; and (c) LSH has high performance, since the respective search time is highly reduced.

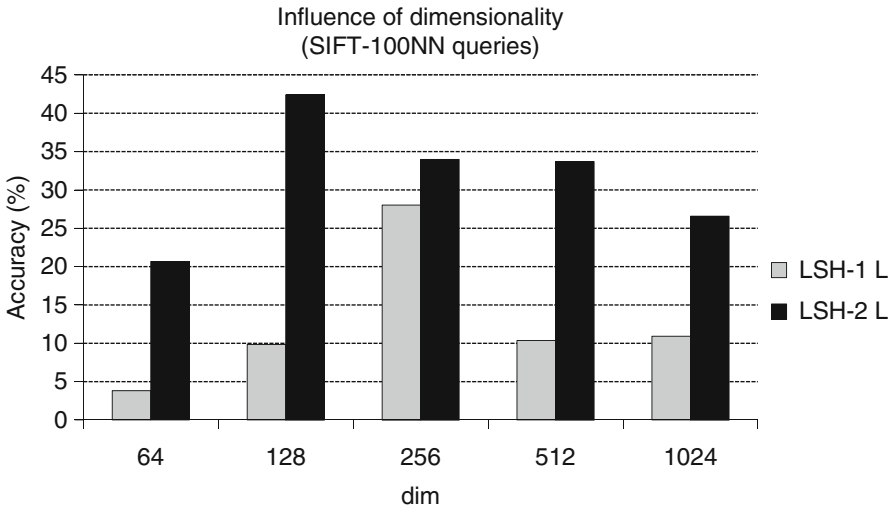


Fig. 4 LSH's retrieval accuracy versus dimensionality for 100-NN queries in SIFT datasets

Finally, in Fig. 4, we evaluate the retrieval accuracy of LSH, by performing 1,000 top-100 queries (denoted by 100-NN queries) and varying the dimensionality of the SIFT datasets. The retrieval accuracy is measured according to the ratio of the top-k results retrieved by the exhaustive search over the top-k results retrieved by the proposed indexing method. Based on the experimental results in Fig. 4, an important observation is that LSH, despite using the maximum number of hash tables ($L = 2$), achieves retrieval accuracy below 40% in all datasets. The low accuracy of LSH is affected by the poor encoding of the multimedia distances. Additionally, by increasing the number of dimensions, LSH's accuracy is reduced. Note that the comparison with M-tree is omitted, since M-tree belongs to the family of exact similarity search and its retrieval accuracy is always equal to 100%.

Summarizing our conclusions, in the case study of Flickr's one million images, the family of exact similarity search, despite achieving identical retrieval accuracy to exhaustive search, fails to support the high dimensionality and as a consequence, storage space and search time are dramatically increased. Moreover, the approximate search strategy of LSH achieves to reduce the search time requirements. Nevertheless, there is no analogous progress in terms of retrieval accuracy, since LSH fail to preserve the retrieval accuracy of the exhaustive search.

Clearly, this case study revealed the fragility of multimedia retrieval in large volumes of data at the scale of millions of records. However, the amounts of multimedia objects that should be indexed and retrieved in the large social multimedia databases such as Flickr and Youtube are in the scale of dozens of billions making it extremely challenging. Moreover, the retrieved results in the presented multimedia retrieval tasks did not preserve any of the "subjective" user semantics that were made available through the sharing process in social networking sites.

4 The Social Media Era

When social networking sites enabled users to share and publish their content online, multimedia search and retrieval became one of the most important and desired features on the Web. However, the volumes of data shared introduce new challenges in multimedia retrieval. Some recent YouTube statistics [82] show clearly that the existing multimedia indexing and retrieval methods are not adequate for these volumes of information. According to YouTube, 48 h of video are uploaded every minute. This is approximately 8 years of content every day or the equivalent of 240,000 full-length films every week. The video uploaded on YouTube per month exceeds the content created by the three major US networks in the last 60 years. In August 2011, Flickr reported that it reached more than six billion uploaded images and the number continuous to grow steadily [28, 29].

Moreover, the amount and diversity of metadata collected and shared through these enormous social media collections pose more parameters to consider toward efficient indexing and retrieval. In a first evaluation, this extra information increases the complexity of the retrieval tasks dramatically. However, the differentiation of the metadata sources (user tags, sensors' information, social graph relations, etc.) constructs a rich environment that helps to narrow down these sets to manageable clusters of information.

All these numbers and facts reveal that multimedia usage and applications changed drastically on the social media era, thus revolutionary multimedia indexing, search and retrieval approaches are needed. What we should clearly state here is that not only did the multimedia collection change shape and size but also the users' needs and goals evolved through the available Web applications.

In the following subsections, recent works are presented roughly classified based on the usage of social metadata and targeted applications. Section 4.1 presents some recent works which aim at involving contextual information to semantically enhance the retrieved results. Section 4.2 discusses works on the relevantly new approach of recognizing and indexing events recorded in social multimedia content. Finally, Sect. 4.3 discusses the works conducted on time-related multimedia retrieval from social streams and ephemeral collections.

4.1 Context-Based Multimedia Indexing and Retrieval

In an environment as wide and heterogeneous as the World Wide Web is, contextual information is known to be inherently noisy, subjective, and ambiguous. However, the information carried out through the context of the various Web applications may be overly helpful after applying some filtering and post-processing.

One of the most used and well-studied contextual data is the user tags. Users tend to tag content in a very personalized way based on their interests, culture, education, etc., thus the relevance of each tag to the actual content is clearly subjective. In order

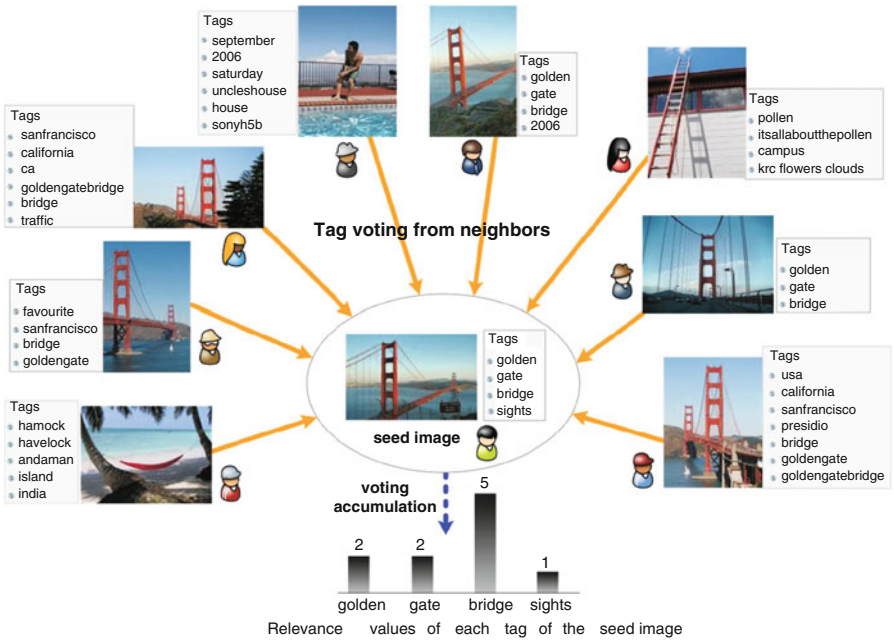


Fig. 5 Learning tag relevance by neighbor voting. The relevance is computed as the accumulated neighboring votes from visually similar image of the seed image [41]

to build a system able to exploit user tags, so as to enhance multimedia retrieval, tags should be found that are relevant to the majority of the users of the system in an objective way [41]. Toward this goal, several methods [4, 40] proposed learning a mapping of low-level visual features to semantic concepts.

Li et al. [41] proposed a technique for learning tag relevance by neighbor voting. The authors rely on the intuition that a relevant tag may be inferred based on the tags of the visual neighbors of that image. The major difference from the related works of [72, 76] is that only common tags between visually similar images are propagated. With this approach, no new tags are introduced to the image, and thus, the technique protects from incorrectly assigning irrelevant tags.

The algorithm reads as follows: Firstly, top-K visually similar neighbors of the image are found, using common visual features and k-means clustering to divide the dataset into small blocks of clusters. Then, for each neighboring image, only the common tags vote to the examined image tags, i.e., each tag of the examined image accumulates votes from the common tags of the neighbor images (see Fig. 5). Since the approach started with the intuition that common tags from *different* users impose a strong relevance of the tags on the visual content, images in the K-nn set that come from the same user as the examined image are ignored.

For evaluating their method, the authors compiled a database of one million images with tags from Flickr and separated a ground truth set for evaluation. Their

experiments were evaluated in a tag-based social image retrieval framework where the well-known Okapi BM25 ranking function for text retrieval was used [70]. The authors compiled three different experimental setups, aiming to identify different aspects of tag-based multimedia retrieval. In the first experiment, a single word was selected as query and various numbers of neighbors were also selected. In the second experiment, the initial queries were expanded with synonyms using WordNet [30] and an online dictionary (http://dico.isc.cnrs.fr/dico/en/search). Finally in the third experiment, the impact of database size was examined by dividing the database in 100 K parts and increasing the database by incrementally adding the parts to reach the whole 1 M images. Both experiments showed clearly the advantage of learning the relevant tags from the visually similar neighbor images, since there was a significant improvement in retrieval accuracy. However, the most interesting result, in terms of scalability of the method, came from the third experiment showing that search performance (in terms of mean average precision) increases as the database size does.

4.1.1 Latent Semantic Spaces

A popular approach used in multimedia indexing and retrieval, cast as clustering and classification, is the extraction of the latent semantics of the explored data to reveal hidden relationships, concepts, and possible structures that a human mind would easier understand [65, 80]. Revealing the hidden semantics in data is a well-studied research field with some interesting statistical tools available [9, 33, 34]. However, the formation of a problem to fit such a tool and the decisions on the design and the social media data to be used are a very interesting and challenging task.

Bosch et al. [10] performed scene classification using probabilistic latent semantic analysis (pLSA) [34] on visual vocabularies extracted directly from the images. The visual vocabularies were extracted by quantizing content descriptors using k-means and a bag-of-words model. The results of the classification showed promising performance in categorizing images; however, the algorithm does not take advantage of the available knowledge in social media, and thus, it is not capable to accurately reveal the semantic concepts. On the contrary, Sizov [66] proposed the GeoFolk model for classification of social media documents using only the contextual information of tags, unstructured text, etc. and spatial knowledge (geotags and geo-coordinates). Moreover, in order to reduce problems such as ambiguity of tags or geolocation information and the sparseness of the available metadata, a model that use both meta-information was developed. Yang et al. [80] proposed heterogeneous transfer learning for image classification using both contextual and content information. Their approach was to extract visual words from the images and additional tags from the social Web in order to build their annotation-based pLSA implementation, which showed significant improvement compared to k-means clustering and plain pLSA. The authors introduced aPLSA as a combination of two separate pLSA models, one for image to visual features co-occurrence matrix and one for text feature to visual feature co-occurrence matrix, with the same latent variables.

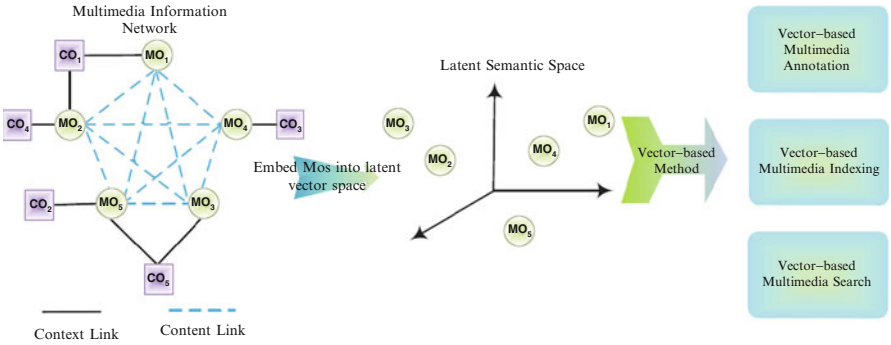


Fig. 6 Learning latent semantic space from a content and context links [56]

A very interesting study on the combination of content and context information to learn a latent semantic space for use in social multimedia retrieval environments is the work of Guo-Jun Qi et al. [56]. Most of the works in the field exploit social metadata (tags, geolocation, etc.) or content features to learn latent spaces; however, they have not addressed problems directly inherited from the use of content or context data sources. Moreover, the sparsity of the annotated objects (contextual information) is one of the major problems that machine learning algorithms suffer from. In their work, Qi et al. [56] aim to address the metadata sparsity problem. They present the multimedia resources in the form of multimedia information networks with two types of objects—multimedia and context objects linked together (see Fig. 6). While the content similarity links are important for retrieval, the context links are the ones that bring quality to the retrieved results. The proposed algorithm learns a latent space where content and context information is encoded and mapped to it. Then, the multidimensional vectors describing the multimedia objects can be indexed, classified, and retrieved with common vector-based methods as the ones described in Sect. 2. The authors make the assumption that similar multimedia objects should be closer to each other in the latent semantic space. This assumption acts as a regularization factor to avoid overfitting problems derived from the sparse context links. To evaluate the algorithm, a database of Flickr images along with their tags was used. A comparison of the different multimedia retrieval schemes was conducted to show the promising results of the approach. The tested retrieval schemes were (a) content-based multimedia retrieval (CMR), (b) context-based multimedia retrieval (CxMR), and (c) both content and context multimedia retrieval (C2MR). Eighty one concepts were manually defined in the database, and experiments were formulated as multimedia annotation problems. As a performance metric, average precision (AP) was selected to measure the retrieval accuracy of each concept. A supervised and an unsupervised method of the model were presented (S-C2MR, U-C2MR), and both had clear improvements against CMR and CxMR. The U-C2MR method improved CMR results by 246.8 and 37.6% CxMR, while S-C2MR improved CMR by 264.2% and CxMR 44.5%.

4.2 Event-Based Multimedia Indexing and Retrieval

Event detection from Web data has attracted a lot of research attention recently [53, 59, 62] due to the immense amount of available information and desire of users to extract/exploit structured information. Moreover, a relatively new approach in detecting, identifying, and indexing social events [6, 43, 73, 79] is through the usage of social metadata along with the shared multimedia content. Toward bridging the semantic gap between human perception and plain multimedia analysis, social multimedia researchers developed methods to detect and link events to multimedia objects in order to support a more human-centered retrieval process and new query types. Since humans tend to structure their knowledge and memory based on specific events and experiences, event identification and indexing should become a realistic way to retrieve multimedia content that we perceive as relevant.

Users aiming to decide whether they will attend a concert/show are interested to feel the atmosphere of previous events based on images and videos available in social networks. In [43], Liu et al. use content and metadata information from Flickr, Last.fm, Eventful, and Upcoming to identify events (concerts, shows, etc.) in nine preselected venues, in order to present characteristic content to interested users. Their work is a two-step process. In the first step, they measure photo sharing activity in known venues, to detect an event occurrence. As a prior knowledge for this measure, a bounding box of a venue geolocation is used. To extract this information, the authors use GPS information from Last.fm and Flickr metadata so as to discard any other information that was recorded outside this area.

This approach achieved to reduce the initial collection of images about the venues to only 4,604 geo-tagged images from the huge collection of Flickr images. Next, they collected more relevant photos, by querying Flickr with each venue name to finally build a photo collection of approximately 9,000 images in total. By tracking the number of shared images per day, number of owners, and the product of the previous two, the event days were identified with appropriate thresholding. The next step of the process was to use seeds images (representative images) that were explicitly linked to an event (in the events database), to retrieve visually similar images from Flickr. In order to calculate visual similarity, low-level visual descriptors such as color moments, gabor texture, and edge histograms were used. Note here that the search space was largely reduced due to the first step of keeping only images that were inside the venue bounding box or had venue name as their tag.

In [6], Becker et al. discussed also the problem of event detection and identification by learning similarity metrics and cast it as a clustering problem. The features used for clustering the social media objects to events' clusters were context features such as tags, descriptions, time/date, and location. Moreover, they used an *all-text* feature where all the textual representation of document features are included (title, description, tags, time/date, location). The textual features were used as $tf \cdot idf$ weight vectors, while typical text processing (stop-word elimination, stemming) steps were applied when needed. In order to support a scalable clustering approach, they proposed a single-pass incremental clustering and tested it into two

different scoring cases. In the first case, each examined document was compared to every other in the cluster so as to generate an overall score, while in the second approach, only a document to cluster centroid comparison was performed. For the selection of similarity metrics, ensemble-based similarity and classification-based similarity were also examined for the final results. The experimental results showed that *all-text* individual clusterer outperformed the other clusterers, while the similarity-based combination outperformed the individual clusterers (include *all-text*). Overall, the classification method showed significant improvement over the typical text-based similarity approaches.

The work of Papadopoulos et al. [52] approached event detection in social multimedia as a graph-based image clustering problem. Their approach combined visual similarity along with tag-based similarity to build two image similarity graphs which are then merged to a unified hybrid image similarity graph. Then, a community detection algorithm was applied to detect clusters of similar images in the hybrid image similarity graph. The authors examined two different cases of image clusters, which are commonly found in social multimedia sharing sites such as Flickr. These are "landmarks" and "events." For classifying the image clusters as events or landmarks, four features were used. The first two, introduced in Quack et al. [57], are the duration of the cluster in terms of creation time-stamps of the included images and the ratio of owners over the number of the images in the cluster. According to the authors, these two features were not adequate to discriminate efficiently "landmark" clusters from "event" clusters. Additionally, they also created two tag vectors corresponding to each class and then removed the common tags to build class-specific tag vectors. Finally, the other two features were the counts of tags of the images clusters that belong to the one set of the other. With the classification step, the event detection task was finished. However, in the case of landmark detection, another step was also required. Although the method aimed to cluster the image collections in meaningful groups, the authors observed that some of the landmark clusters referred to the same object. Toward facing this inefficiency, a merging step was applied. By using geolocation information, a new spatial proximity graph was built, and the community detection algorithm was used to form new clusters of images. The experimental results compared with k-means clustering of both the visual and the tag features appear to have better geo-spacial focus (which is significant to events and landmarks) and overall higher precision in the subjective evaluation.

4.3 Time-Related Multimedia Indexing and Search for Evolving Social Multimedia Collections

The speed of multimedia content sharing in the social Web, bloats the social media databases with enormous volumes of data in very short time windows. Thus, a major difference between the social Web databases and the common multimedia databases is the fact that they are constantly increasing, populated

with fresh content. This feature inspires users to ask for complicated queries that include time-/date-related information. Moreover, these ever-evolving databases store millions of records every day, with ephemeral interest to the users and useless if not consumed in a certain period of time. Thus, queries that filter the retrieved content in specified time windows are also needed. However, the time-evolved social multimedia databases require also new approaches of organizing the content in order to enable for efficient search and retrieval.

The study of Lin et al. [42] addressed the social multimedia retrieval problem from exactly this viewpoint. The authors consider Flickr's photo groups as mixtures of themes with similarities in content and context. Their goals were (a) to better organize the content inside each photo group, since the exponential growth of the content makes exploration and searching inside a group a challenging task, and (b) to reveal the changing interests and trends inside the photo group and reveal the photo genres that a group contains. In order to exploit both content and context information, the authors extracted content features from images, tags, owner information, and post time to build four matrices: a photo-features matrix, a photo-user matrix, a photo-tag matrix, and a photo-time matrix. Then, a nonnegative joint matrix factorization procedure was applied on these matrices to extract "themes" of photos that change over time. As they clearly stated, their motivation was the development of a method to answer difficult, for pure multimedia retrieval systems, questions. Such questions were as follows: *Are there typical patterns in the photo stream? How these patterns evolve over time? How can we extract the patterns and which users and photos follow them?* As it is clear, such questions are becoming typical and extremely useful in global-scale databases, and as such, they add value to the existing social multimedia sharing services.

Another interesting study on multimedia retrieval from the social Web is the work of De Silva et al. [25], which proposed interactive spatiotemporal query formulation for quick multimedia retrieval from large multimedia databases. De Silva et al. stated that the presentation of an one-dimensional results list is inadequate for such rich multimedia databases since the query was tested against multiple dimensions (content, tags, time, space, etc.). The proposed interface enables the iterative searching and browsing of content with query refining in any of the available modalities (content, social relations, time, etc.).

Since temporal information is widely available in the shared multimedia content through social media, new approaches emerge exploiting the temporal information to extract usage and sharing patterns or visualize the content and extract valuable information that may be used to cluster multimedia content such as the works of [1, 54].

5 Conclusions and Future Challenges

Multimedia indexing and retrieval is a challenging task on its own, and thus different solutions have been proposed, trying to address different angles of the problem. Further, social multimedia indexing and retrieval in the large databases

of the social networks advanced the challenges to form a new problem that needs special handling. Multimedia indexing, search and retrieval in the large databases of social networks clearly stated its "uniqueness" mainly in two dominant axes: a) as multimedia analysis task with heterogeneous, noisy and ambiguous modalities such as text, images, videos, audio along with tags, free-text, geotags, geo-coordinates, time information, social relations and communities; and b) as a web-scale information retrieval task with all the scalability and performance issues carried along.

The majority of the works presented in this chapter were evaluated using Flickr images. The Flickr image sharing service has a characteristic, socially "sound," design that enables the evolution of the database in terms of time, themes, groups, trending tag annotations, and of course users that form communities and groups, follow other users' works, and give ratings.

In this chapter, social multimedia indexing, search, and retrieval techniques and algorithms were presented, aiming to shed light to different viewpoints of the problem. Section 2 discussed shortly the state-of-the-art multidimensional indexing structures by classifying them in the exact and approximate approaches. Then, in Sect. 3, social multimedia content was used to present a case study of indexing one million images from Flickr photo sharing site. In this case study, indexing was performed based only on the content of the multimedia objects. This approach showed that in the large social multimedia databases of billions of records, these approaches are inefficient in terms of response time and memory/storage needs and/or accuracy of the retrieved results. The major issue though is that such indexing methods do not consider the available social metadata that enclose "subjective" user semantics. However, these semantics are crucial for improving both the quality of the retrieved results and the performance in qualitative aspects. Section 4 presented exactly these aspects of social multimedia retrieval. Moreover, in the context of social media, multimedia retrieval was explored as another means of a higher-level query formulation to answer complex questions.

Since the social multimedia databases will continue to grow exponentially to unmanageable volumes, revolutionary approaches in content search and retrieval are necessary. Future challenges to this field include searching in multimedia streams, classification of ephemeral data for subscription purposes, and trending algorithms for identifying popular multimedia content. Moreover, algorithms, techniques, and search schemes that enable users to improvise in querying the enormous social multimedia collections are also sought.

Acknowledgements This work was partially supported by the EC FP7-funded project CUBRIK, ICT-287704 (www.cubrikproject.eu).

References

1. Andrienko, G., Andrienko, N., Bak, P., Kisilevich, S., Keim, D.: Analysis of community-contributed space-and time-referenced data (example of Panoramio photos). In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09), pp. 540–541. ACM, New York (2009)

2. Baeza-Yates, R., Cunto, W., Manber, U., Wu, S.: Proximity matching using fixed-queries trees. In: Proceedings of the 5th Combinatorial Pattern Matching (CPM), Asilomar. LNCS, vol. 807, pp. 198–212. (1994)
3. Baeza-Yates, R., Navarro, G.: Fast approximate string matching in a dictionary. In: Proceedings of the 5th South American Symposium on String Processing and Information Retrieval (SPIRE), pp. 14–22. IEEE CS Press, Los Alamitos (1998)
4. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. JMLR **3**, 1107–1135 (2003)
5. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubsky, J., Zezula, P.: Building a web-scale image similarity search system. J. Multimed. Tools Appl. **47**(3), 599–629 (2010)
6. Becker, H., Naaman, M., Gravano, L.: Learning similarity metrics for event identification in social media. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10), pp. 291–300. ACM, New York (2010). doi:10.1145/1718487.1718524, http://doi.acm.org/10.1145/1718487.1718524
7. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)
8. Bentley, J.L.: Multidimensional binary search trees in database applications. IEEE Trans. Soft. Eng. **5**(4), 333–340 (1979)
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
10. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. Computer vision ECCV 2006. Lecture Notes in Computer Science, vol. 3954, pp. 517–530. Springer, Berlin/Heidelberg (2006)
11. Bozkaya, T., Ozsoyoglu, M.: Distance-based indexing for high-dimensional metric spaces. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Tucson, pp. 357–368 (1997)
12. Bugnion, E., Fhei, S., Roos, T., Widmayer, P., Widmer, F.: A spatial index for approximate multiple string matching. In: Proceedings of the 1st South American Workshop on String Processing (WSP), Belo Horizonte, pp. 43–53 (1993)
13. Burkhard, W.A., Keller, R.M.: Some approaches to best-match file searching. Commun. ACM **16**(4), 230–236 (1973)
14. Brin, S.: Near neighbor search in large metric spaces. In: Proceedings of the 21st International Conference on Very Large Data Bases (VLDB), Zurich, pp. 574–584 (1995)
15. Chatzichristofis, S.A., Boutalis, Y.S.: CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) Proceedings of the 6th International Conference on Computer Vision Systems (ICVS'08), pp. 312–322. Springer, Berlin/Heidelberg (2008)
16. Chavez, E., Marroquin, J., Navarro, G.: Overcoming the curse of dimensionality. In: Proceedings of the European Workshop on Content-Based Multimedia Indexing (CBMI), Toulouse, pp. 57–64 (1999)
17. Chavez, E., Navarro, G., Baeza-Yates, R., Marroquin, J.L.: Searching in metric spaces. ACM Comput. Surv. **33**(3), 273–321 (2001)
18. Chiueh, T.: Content-based image indexing. In: Proceedings of the 20th Conference on Very Large Databases (VLDB), Santiago, pp. 582–593 (1994)
19. Chum, O., Philbin, J., Zisserman, A.: Near duplicate image detection: min-hash and tf-idf weighting. In: Proceedings of the British Machine Vision Conference, Leeds (2008)
20. Ciaccia, P., Patella, M., Zezula, P.: M-tree: an efficient access method for similarity search in metric spaces. In: Proceedings of the 23rd Conference on Very Large Databases (VLDB), Athens, pp. 426–435 (1997)
21. Ciaccia, P., Patella, M., Zezula, P.: Bulk loading the M-tree. In: Proceedings of the 9th Australasian Database Conference (ADC), Perth, pp. 15–26 (1998)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE Computer Society, Los Alamitos (2005)

23. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Symposium on Computational Geometry, Brooklyn, pp. 253–262 (2004)
24. Dehne, F., Nolteimer, H.: Voronoi trees and clustering problems. Inf. Syst. **12**(2), 171–175 (1987)
25. de Silva, G.C., Aizawa, K., Arase, Y., Xing X.: Interactive social, spatial and temporal querying for multimedia retrieval. In: Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on, Madrid, 13–15 June 2011, pp. 7–12 (2011)
26. Flickr's 1 million image dataset, visual concept detection and annotation, ImageCLEF 2011, [online]. Available http://www.imageclef.org/2011/Photo
27. Flickr's website, [Online]. Available: http://www.flickr.com/
28. Flickr website. http://www.flickr.com/.Cited20Feb2012
29. Flickr record in Wikipedia. http://en.wikipedia.org/wiki/Flickr#cite_note-3Cited20Feb2012
30. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT, Cambridge (1998)
31. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), Edinburgh, pp. 518–529 (1999)
32. Guttman, A.: R-trees: a dynamic index structure for spatial searching. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Boston, pp. 47–57 (1984)
33. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, pp. 50–57, (1999)
34. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **41**, 177–196 (2001)
35. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Proceedings of the European Conference on Computer Vision. LNCS, vol. I, pp. 304–317. Springer, Berlin (2008)
36. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 117–128 (2011)
37. Joly, A., Buisson, A.O.: Random maximum margin hashing. In: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, pp. 873–880 (2011)
38. Joly, A., Frelicot, C., Buisson, O.: Feature statistical retrieval applied to content-based copy identification. In: Proceedings of the International Conference on Image Processing, Singapore, pp. 681–684 (2004)
39. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing for scalable image search. In: Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV), Kyoto, pp. 2130–2137 (2009)
40. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. TPAMI **30**(6), 985–1002 (2008)
41. Li, X., Snoek, C.G.M., Worring, M.: Learning tag relevance by neighbor voting for social image retrieval. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR '08), pp. 180–187. ACM, New York (2008)
42. Lin, Y.-R., Sundaram, H., De Choudhury, M., Kelliher, A.: Temporal patterns in social media streams: theme discovery and evolution using joint analysis of content and context. In: IEEE International Conference on Multimedia and Expo, ICME 2009, June 28 2009–July 3 2009, New York, pp. 1456–1459 (2009)
43. Liu, X., Troncy, R., Huet, B.: Using social media to identify events. In: Proceedings of the 3rd ACM SIGMM International Workshop on Social media (WSM '11), pp. 3–8. ACM, New York (2011). doi:10.1145/2072609.2072613, http://doi.acm.org/10.1145/2072609.2072613
44. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)

45. Lv, Q., Josephson, W., Wang, Z., Charikar, M., Li, K., Multi-probe LSH: efficient indexing for high-dimensional similarity search. In: Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB), Vienna, pp. 950–961 (2007)
46. Mico, L., Oncina, J., Vidal, E.: A new version of the nearest-neighbor approximating and eliminating search (AESA) with linear preprocessing-time and memory requirements. Pattern Recognit. Lett. **15**, 9–17 (1994)
47. M-tree web site, the M-tree project, [Online]. Available: http://www-db.deis.unibo.it/Mtree (2008)
48. Naaman, M.: Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. Multimed. Tools Appl. **56**(1), 9–34 (2012). Springer, Computer Science
49. Navarro, G.: Searching in metric spaces by spatial approximation. VLDB J. **11**(1), 28–46 (2002)
50. Nolteimer, H., Verbarg, K., Zirkelbach, C.: Monotonous bisector trees: a tool for efficient partitioning of complex schemes of geometric objects. In: Monien, B., Ottmann, T. (eds.) Data Structures and Efficient Algorithms. LNCS, vol. 594, pp. 186–203. Springer, Berlin/New York (1992)
51. Pauleve, L., J'egou, H., Amsaleg, L.: Locality sensitive hashing: a comparison of hash function types and querying mechanisms. Pattern Recognit. Lett. **31**(11), 1348–1358 (2010)
52. Papadopoulos, S., Zigkolis, C., Kompatsiaris, Y., Vakali, A.: Cluster-based landmark and event detection for tagged photo collections. IEEE Multimed. **18**(1), 52–63 (2011)
53. Popescu, A.-M., Pennacchiotti, M.: Detecting controversial events from twitter. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10), pp. 1873–1876. ACM, New York (2010). doi:10.1145/1871437.1871751, http://doi.acm.org/10.1145/1871437.1871751
54. Popescu, A., Grefenstette, G., Moellic, P.-A.: Mining tourist information from user-supplied collections. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09), pp. 1713–1716. ACM, New York (2009)
55. Poullot, S., Buisson, O., Crucianu, M.: Z-grid-based probabilistic retrieval for scaling up content-based copy detection. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR), Amsterdam, pp. 348–355 (2007)
56. Qi, G.-J., Aggarwal, C., Tian, Q., Ji, H., Huang, T.: Exploring context and content links in social media: a latent space method. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 850–862 (2012)
57. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval (CIVR '08), pp. 47–56. ACM, New York (2008). doi:10.1145/1386352.1386363, http://doi.acm.org/10.1145/1386352.1386363
58. Raginsky, M., Lazebnik, S.: Locality-sensitive binary codes from shift-invariant kernels. In: Proceedings of the ACM NIPS, Vancouver, pp. 1509–1517 (2009)
59. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web (WWW '10), pp. 851–860. ACM, New York (2010). doi:10.1145/1772690.1772777, http://doi.acm.org/10.1145/1772690.1772777
60. Salakhutdinov, R., Mnih, A., Hinton, G.: Restricted boltzmann machines for collaborative filtering. In: Proceedings of the 24th ACM International Conference on Machine learning, Oregon, pp. 791–798 (2007)
61. Samet, H.: The quadtree and related hierarchical data structures. ACM Comput. Surv. (CSUR) **16**(2), 187–260 (1984)
62. Sayyadi, H., Hurst, M., Maykov, A.: Event detection and tracking in social streams. In: Proceedings of the International AAAI Conference on Weblogs and Social Media. AAAI Press, Menlo Park (2009)
63. Shakhnarovich, G., Darrell, T., Indyk, P.: Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press, Cambridge, USA (2006)

64. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR '07, 17–22 June 2007, pp. 1–8. IEEE, Piscataway (2007)
65. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: Proceedings of the International Conference on Computer Vision, Beijing (2005)
66. Sizov, S.: GeoFolk: latent spatial semantics in web 2.0 social media. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10), pp. 281–290. ACM, New York (2010). doi:10.1145/1718487.1718522, http://doi.acm.org/10.1145/1718487.1718522
67. Skopal, T.: Pivoting M-tree: a metric access method for efficient similarity search. In: Proceedings of the Annual International Workshop on Databases, Texts, Specifications and Objects (DATESO), Desna, pp. 27–37 (2004)
68. Skopal, T., Hoksza, D.: Improving the performance of M-tree family by nearest-neighbor graphs. In: Proceedings of the 11th East European Conference on Advances in Databases and information Systems (ADBIS), Varna, pp. 172–188 (2007)
69. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000). doi:10.1109/34.895972, http://dx.doi.org/10.1109/34.895972
70. Sparck Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manag. **36**(6), 779–808 (2000)
71. Tian, Y., Srivastava, J., Huang, T., Contractor, N.: Social multimedia computing. Computer **43**(8), 27–36 (2010)
72. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1958–1970 (2008)
73. Troncy, R., Malocha, B., Fialho. A.T.S.: Linking events with media. In: Paschke, A., Henze, N., Pellegrini, T. (eds.) Proceedings of the 6th International Conference on Semantic Systems (I-SEMANTICS '10), p. 4. ACM, New York (2010). doi:10.1145/1839707.1839759, http://doi.acm.org/10.1145/1839707.1839759
74. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: Real-time visual concept classification. IEEE Trans. Multimed. **12**(7), 665–681 (2010)
75. Vidal, E.: An algorithm for finding nearest neighbors in (approximately) constant average time. Pattern Recognit. Lett. **4**, 145–157 (1986)
76. Wang, X.-J. Zhang, L., Li, X., Ma, W.-Y.: Annotating images by mining image search results. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1919–1932 (2008)
77. Weber, R., Schek, H.J., Blott, S.: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: Proceedings of the 24th International Conference on Very Large Data Bases (VLDB), New York, pp. 194–205 (1998)
78. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, pp. 1753–1760 (2008)
79. Westermann, U., Jain, R.: Toward a common event model for multimedia applications. IEEE Multimed. **14**(1), 19–29 (2007)
80. Yang, Q., Chen, Y., Xue, G.-R., Dai, W., Yu, Y.: Heterogeneous transfer learning for image clustering via the social web. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 – Volume 1 (ACL'09), vol. 1, pp. 1–9. Association for Computational Linguistics, Stroudsburg (2009)
81. Yianilos, P.N.: Excluded middle vantage point forests for nearest neighbor search. NEC Research Institute, Princeton University, technical report (1998)
82. YouTube statistics. http://www.youtube.com/t/press_statistics.Cited20Feb2012

Survey on Social Community Detection

Michel Plantié and Michel Crampes

Abstract Community detection is a growing field of interest in the area of social network applications. Many community detection methods and surveys have been introduced in recent years, with each such method being classified according to its algorithm type. This chapter presents an original survey on this topic, featuring a new approach based on both semantics and type of output. Semantics opens up new perspectives and allows interpreting high-order social relations. A special focus is also given to community evaluation since this step becomes important in social data mining.

1 Introduction

As social networks gain prominence, the first obvious question that comes to a researcher's mind in observing these networks is how to extract meaningful knowledge from these data. In seeking a response, the network structure proves to be of utmost importance. Identifying high-order structures within networks yields insights into their functional organization, which in turn contributes more knowledge while offering many possible actions, including marketing plans, recommendations, and user interface adaptations. Community detection may become a more complicated task given that social networks can be structured on many different levels, yet communities reduce the complexity of a network's original graph in a substantial way, thus revealing its macrostructure and uncovering more semantic knowledge. A growing number of community detection methods have recently been published. The goal here is to assess the state of the art in this area, by focusing on the qualities and shortcomings of each method. A number of partial surveys have been conducted

M. Plantié (✉) • M. Crampes
Laboratoire de Genie Informatique et d'Ingenierie de Production (LGI2P), EMA - Ecole des Mines, Site EERIE, Parc Scientifique Georges Besse, F-30035 Nîmes cedex 1, France
e-mail: michel.plantie@mines-ales.fr; michel.crampes@mines-ales.fr

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_4, © Springer-Verlag London 2013

over the past few years; though this body of work has exposed different approaches in the field, such efforts are often limited to specific network structures. This chapter is intended to present three analytical approaches to community detection that encompass most of the main methods and techniques. The first approach, which is also the most widespread, considers the social network as a graph and then analyzes its structure with graph properties and algorithms built around the graph structure. The second approach associates the social network with a hypergraph and analyzes its structure through hypergraph properties and algorithms based on hypergraph structures, as exemplified in [53]. The third and final approach uses the properties of concept lattices in order to analyze the social network structure in association with hypergraph properties and algorithms based on Galois lattices and hypergraph structures, for example, [54,60]. As opposed to graphs, both hypergraphs and Galois lattices have been poorly analyzed in surveys on community detection strategies. These structures offer very efficient tools for managing communities, and this discussion will demonstrate how researchers have applied them. This chapter will be organized as follows. To ensure a good understanding of all elements being addressed in this survey, Sect. 2 will give all necessary definitions, and Sect. 3 will then deliver a state of the art from previous surveys on the community detection topic. The next section will classify each community detection method according to a graph type of classification, and lastly Sect. 5 will lend insight into all possible evaluations of community detection algorithms, as this area of investigation has only been sparsely studied in previous surveys.

2 Definitions: Social Network Community Detection and Other Definitions

2.1 Unipartite and Bipartite Graphs

A graph is a representation of a set of objects called vertices, some of which are connected by links. Object connections are depicted by links, also called edges. Such a mathematical structure may be referred to as a unipartite graph. A good example of this type of graph is the well-known Zachary's Karate Club [77] (shown Fig. 1).

A special case of this graph is known as the bipartite graph, that is, whose vertices can be divided into two disjoint sets, A and B, such that the edges only connect one vertex in A to one in B, in considering that A and B are independent sets. Vertices of A are not connected to any other vertices within A, and the same applies for B. For example, let A be a set of individuals and B a set of photos showing these same individuals. Bipartite graphs may take the form of graphs, hypergraphs, or Galois lattices.

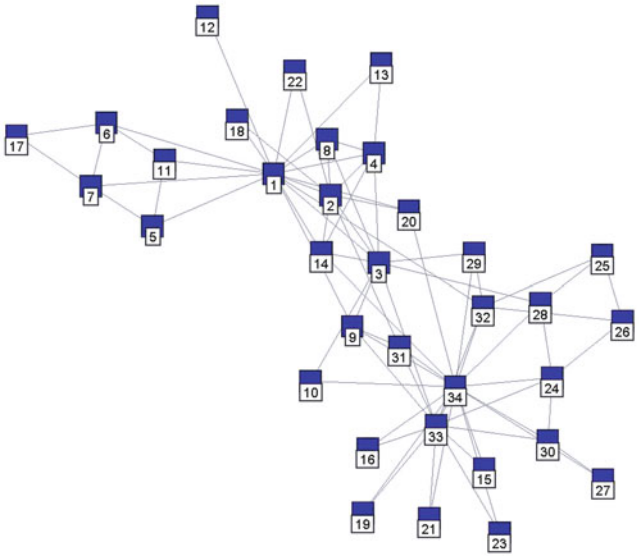


Fig. 1 Depiction of the Zachary's Karate Club graph example (with a different display for better visibility)

2.2 *Hypergraph*

A hypergraph [4] H is a pair (V, E) where V = v1, v2, …, vn is a nonempty (usually limited) set and E = E1, E2, …, Em is a family of not empty subsets of V. The elements of V are the vertices of H. The elements of E are the edges (also called hyperedges) of H. A set of social communities can be viewed as a hypergraph whose vertices are the individuals and whose hyperedges are the communities. Most researchers in the field of community detection seek to partition individuals into communities, that is, nonintersecting hyperedges. Some authors have attempted to find overlapping communities, that is, connected hypergraphs. A bipartite graph can be displayed as a hypergraph with individuals at the vertices and properties at the hyperedges. Alternatively, the properties can be considered as vertices and the individuals as hyperedges. An example of a simple hypergraph is shown in Fig. 2.

2.3 *Galois Lattice*

Freeman [19] was the first to use Galois lattices in order to represent network data. The underlying assumption is that individuals sharing the same subset of properties define a community. The approach adopted consists of the following: objects, attributes, and the set of relations between objects and attributes form a

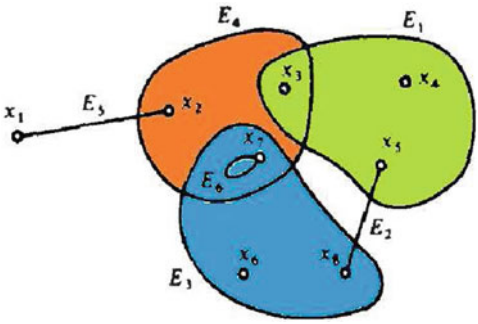


Fig. 2 Example of a hypergraph (a colored version of example in [4] p. 2)

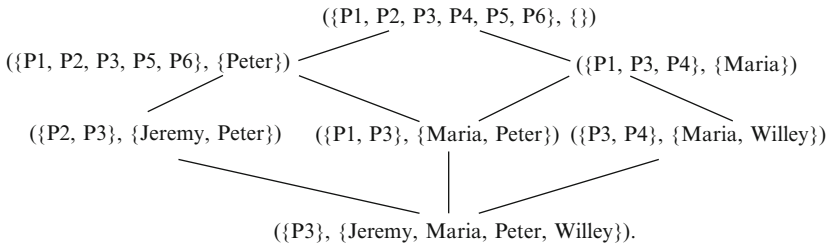


Fig. 3 Example of a Galois lattice (with photos Pi and individuals)

"context," in accordance with formal concept analysis [20]. This set of relations can then be represented by a binary bi-adjacency matrix, whereby objects o are the columns, attributes a are the rows, and a "1" is placed at the cell corresponding to a pair (o_i, a_j) if o_i possesses a_j. A maximum subset of objects that contain a subset of attributes is defined as a "concept," that is, a group of objects for which the addition or removal of an attribute changes its constitution. All objects of a concept then form the "extent," and all attributes of a concept give rise to the "intent." A partial order is applied to concepts and serves to establish a hierarchy. According to the definition of Galois hierarchies, an object can appear in all the concepts where it can share the same set of attributes with other objects belonging to other concepts. Figure 3 illustrates a simple example of a Galois lattice in which several individuals are sharing several photos.

2.4 The Concept of Modularity

Modularity has been introduced to measure the quality of community algorithms. Newman [46] proceeded with the initial introduction, in providing the following formula:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{1}$$

where

- e_{ij}: number of edges having one end in group i and the other end in group j
- $a_i = \sum_j e_{ij}$: number of edges having one end in group i

This quantity Q measures the fraction of edges in the network that connect vertices of the same type (i.e., "intra-community" edges) minus the expected value of the same quantity in a network with the same community divisions yet with random connections between vertices.

Modularity measures the capacity of a given graph partition to yield the densest groups. This formula has mainly been used by researchers in order to measure the ability of a community detection algorithm to obtain a satisfactory partition of a given graph. Moreover, the formula may be adapted to weighted graphs (i.e., graphs whose edges display different weights or lengths), like in [5].

3 State of the Art Assessment: Existing Surveys

This section will address the body of existing surveys on community detection. Most surveys primarily focus on the graph structure aspect of communities. Seven of the main existing surveys have in fact been recently conducted. Since the concept of community may differ, this existing body of surveys defines communities before classifying the methods employed according to various classification systems. This study will start by determining which definitions are provided for these different community detection methods before presenting a state of the art on existing surveys.

3.1 Community Definitions

Defining a community is quite a challenging task. Definitions vary from author to author and from algorithm to algorithm. The most commonly used definition is that of Yang [76]: "a community as a group of network nodes, within which the links connecting nodes are dense but between which they are sparse." This definition is applicable for graphs and could be extended to bipartite graphs.

Fortunato [18] identifies three levels to define a community: local definitions, global definitions, and definition based on vertex similarity. In the local definition group, a definition of communities consists of "parts of the graph with few ties to the rest of the system." In this partition, communities are studied from their inner structure independently of the remaining part of the graph. In the global definition group, a global criterion associated with the graph is used to compute communities. This global criterion is dependent on the algorithm implemented to locate communities. Either a clustering criterion or a distance-based criterion may be introduced; more often, the criterion commonly used shows that the graph

contains a community structure different from that of a random graph. In the vertex similarity-based community definition group, communities are considered as groups of vertices similar to one another.

Fortunato [18] further defines communities, in also calling them clusters or modules, as "groups of vertices that probably share common properties and/or play similar roles within the graph." His assigned definition depends on the algorithm employed, resulting in the identification of at least eight different definitions:

- Clique: subgroups whose members are all "friends" (i.e., connected with an edge) to each other [37]
- n-clique with two variants: maximal subgraphs such that the distance of each pair from its vertices is not greater than n [36]
- k-plex: maximal subgraph in which each vertex is adjacent to all other vertices of the subgraph except at most k of them [65]
- LS-set (weak community): subgraph such that the internal degree is greater than the external degrees [35]
- Lambda set: subgraph where each pair of vertices has a greater edge connectivity than any pair formed by one vertex of the subgraph and one outside the subgraph [6]
- Communities based on either a fitness measure or a quality measure
- Communities determined by means of modularity-based algorithms
- Clusters: communities derived using well-known clustering methods [64]

Porter [56] recalls the origins of community study in the fields of sociology and anthropology. He defines communities as "cohesive groups of nodes that are connected more densely to each other than to the nodes in other communities." The difference in methods highlighted in his survey relies on a definition of the expression "more densely," which is identified with five types of algorithms, namely, clustering techniques [64], quality function algorithms [30], centrality-based community detection algorithms [46], and other similar ones, clique percolation algorithms [15] and lastly modularity optimization algorithms [44].

Gulbahce and Lehmann [26] define a community as "a densely connected subset of nodes that is only sparsely linked to the remaining network."

Papadopoulos [53] ultimately defines communities as "groups of vertices that are more densely connected to each other than to the rest of the network."

It can be seen that all definitions are quite similar yet may still differ in their associated formal mathematical definition. Communities may also be considered from a different perspective. The initial approach partitions the underlying graph, that is, by dividing the existing graph or network structure into distinct communities using the optimal algorithms. The second approach then detects overlapping communities and seeks the best community arrangement.

In [61], Roth defines a new type of community, "epistemic communities," which are "knowledge communities or groups of agents sharing common knowledge concerns," for instance, a group of researchers investigating a single precise topic. This new type of community concept requires new kinds of structures to proceed with their description. Roth has opted to use Galois lattices.

3.2 State of the Art in Community Detection Surveys

Most surveys classify research papers and methods according to the type of community detection algorithm.

The first of these seven main surveys by Fortunato [18] is exhaustive with respect to many community detection methods and has been based on a graphic representation. This survey provides an effective overview of the field and describes the methodological foundations of community detection, adopting a statistical physics perspective and specifically focusing on techniques designed by statistical physicists. His discussion also includes critical issues like the significance of clustering, the procedure by which methods should be tested and compared to one another, and applications to real networks. Methods are classified into eight families, that is:

- Traditional methods based on clustering like k-means [39] and other applications [30]
- Divisive algorithms mainly based on hierarchical clustering [51]
- Modularity-based algorithms [7, 11, 46] and other similar algorithms
- Spectral algorithms [57]
- Dynamic algorithms [27, 73] and other similar algorithms
- Statistical inference-based methods [2]
- Multi-resolution methods [55]
- Methods to find overlapping communities [15, 52] and other miscellaneous methods

The second survey, conducted by Porter [56], only includes graph partitioning approaches and offers insight into graphical techniques through citing the first survey. An extensive set of techniques is highlighted, as are some of the most important unresolved issues remaining. Application examples are also given on some of the largest social networks in addition to grouping community detection into five main techniques:

- Centrality-based techniques built around the Newman algorithm [46]
- Local methods around the k-clique percolation method [52]
- Modularity optimization methods around the Newman algorithm [44]
- Spectral partitioning methods around Simon's algorithm [57]
- Physics-based methods inspired by Potts law [73]

The third survey by Yang [76] is quite exhaustive relative to all techniques relying on graphical representation and produces a good overview of the field through classifying all techniques in a tree structure, according to three categories:

- Optimization-based algorithms [25, 30, 44]
- Heuristic algorithms [21, 69]
- Similarity-based algorithms and hybrid methods [55]

The fourth one from Gulbahce and Lehmann [26] is a partial survey analyzing hierarchical-type community detection methods and provides a number of leads for future community detection approaches.

The fifth survey by Pons [55] incorporates several community detection methods and classifies them into five different families:

- Classical approaches including classical graph partitioning, for example, spectral bisection from [57], Kernighan and Lin [30], clustering [28], and hierarchical clustering like [72].
- Separative approaches attempting to split a graph into several communities by deleting the edges connecting distinct communities. In this group, Pons places the well-known Girvan-Newman algorithm [46] and other variant approaches.
- Agglomerative approaches quite similar to their hierarchical counterparts and include a method based on optimized modularity by Newman [44] and other algorithms.
- Random walk-type algorithms [14] and others based on the mean time required to reach a vertex [79].
- A broad group of miscellaneous approaches.

The Pons survey has only examined graph partitioning approaches.

The sixth survey by Papadopoulos [53] classifies community detection techniques in five methodological categories:

- Cohesive subgraph discovery [52]
- Vertex clustering [55]
- Community quality optimization [10, 46]
- Divisive [46]
- Model-based methods [23]

The seventh and last survey was undertaken by Danon [13] and primarily focuses on the performance of each type of algorithm.

Among these seven surveys, only one offers leads on overlapping communities, while none make use of hypergraph structures. Another domain overlooked by these surveys is Galois lattice structures. As will be described below, Galois lattices are more complex structures than regular graphs, yet they provide more semantics to network structures.

4 Social Network Community Detection Methods

4.1 Approach Classification

All community detection methods will be classified in a grid divided into six categories based on the types of input data and output data; it will then be shown how each author can be placed in this grid.

Table 1 Methods according to representations

$\frac{Input \Rightarrow}{Output \Downarrow}$	**1:** graph	**2:** bipartite graph/hypergraph
A: partition	Input: unipartite graph (see Sect. 2.1) Output: partition	Input: bipartite graph (see Sect. 2.1) Output: partition
B: hypergraph (see Sect. 2.2)	Input: unipartite graph (see Sect. 2.1) Output: overlapping communities represented by a hyper-graph	Input data: bipartite graph (see Sect. 2.1) Output: overlapping communities represented by a hypergraph (see Sect. 2.2)
C: galois hierarchy (see Sect. 2.3)	No method eligible except [12] With partial results	Input: bipartite graph (see Sect. 2.1) Output: Galois lattice of communities (see Sect. 2.3)

In order to detect communities, the initial input data were considered in network form (whether a social network, biological network, etc.) and represented by different mathematical structures relative to graph structures, which may be of three distinct types:

- Unipartite graph: This is a normal graph whose vertices are individuals and whose edges are links connecting the individuals (these links may be of various types: friend, family, club, sport, university, etc.).
- A bipartite graph: This type of graph may be generated whenever individuals share tags (i.e., terms assigned by users), web pages or links. The first set contains individuals, and the second is a set of tags, web links or documents, etc.
- A multipartite graph: Very similar to bipartite graphs, this graph however is composed of several disjoint sets. In this survey, this type of graph has not been directly taken into account since it may be reduced to a bipartite graph, as shown in [48] and elsewhere.

The output data of a community detection method consists mainly of a set of node groups representing communities. The following merit consideration:

- Graph partition, where each node is associated with just one group of nodes and where no overlap exists between groups. Partitions are the primary result of most community detection algorithms.
- Hypergraph with overlapping communities.
- Concept graphs or Galois lattices where nodes share several common properties.

The following Table 1 explains which type of input and output data each method is capable of accepting.

Most existing surveys included in our state of the art evaluation describe community detection methods that may be classified within the "A1" cell of Table 1. Table 2 below summarizes the major methods according to the above classification.

Table 2 Papers classified according to their community detection methods

$\frac{Input \Rightarrow}{Output \Downarrow}$	**1:** graph	**2:** bipartite graph/hypergraph
A: partition	[8,45,70,71,75,78] **S[18,26,53,56,76]:** *[3, 9, 11, 21, 37, 46, 55, 64]* *[5, 6, 33, 44, 45, 47, 50, 74]*	[41,58,61,66,67]
B: hypergraph (overlapping communities)	**S[18,56]:** *[1, 16, 33, 49, 52]*. [17], **S[53]**: *[15, 23, 52]*	[34,40,48,59] **S[53]**: *[34]*
C: galois hierarchy	[12]: partial results	[19,29,62,68]

Table legend: Letter **S** preceding a publication number indicates that this paper is a survey. The expression **S**[2]:[1] indicates that the survey referenced in paper number 2 cites and comments on the community detection method described by paper number 1, and moreover paper number 1 can be classified according to the table cell where it has been placed.

4.2 From Graphs to Partitions (Cell A1)

Most community detection algorithms lie in this class of methods. The input data is a normal graph (i.e., a set of vertices representing individuals, who are connected by edges), and the output is a list of node groups representing the communities on the initial graph. Each individual belongs to one and only one community. All surveys mentioned in Sect. 3 describe these algorithms in full detail; the following algorithm classification has been adopted: top-down (separate) methods in S [18, 55]; bottom-up (agglomerative) and/or clustering methods in S [18, 53, 55, 56, 76]; optimization-based algorithms [76]: and heuristic algorithms [76]. The three most popular algorithms will be discussed hereafter, that is, the Girvan-Newman algorithm [21] based on intermediate centrality, the Newman algorithm [44, 46] based on modularity, and the Louvain algorithm [5].

4.2.1 Girvan and Newman Algorithm

According to survey [18], this algorithm belongs to the category of divisive algorithms. Its underlying principle calls for removing the edges that connect different communities. In the algorithm described in [46], several measures of edge centrality are computed, in particular the so-called intermediate centrality, whereby edges are selected by estimating the level of edge importance based on these measures. As an illustration, intermediate centrality is defined as the number of shortest paths using the edge under analysis. The steps involved are as follows:

1. Compute centrality for all edges.
2. Remove edges with the greatest centrality (when ties exist with other edges, one edge is to be chosen at random).
3. Recalculate centralities on the remaining graph.
4. Iterate beginning at step 2.

This work has exerted great influence on research, and, consequently, edge centrality has been a key field of study for many scientists, resulting in the proposal of several measures [69].

4.2.2 Modularity-Based Algorithm

This algorithm introduced by Girvan and Newman [46] and then improved in [11] is based on modularity (see Sect. 2.4). The "glutton"-type algorithm maximizes modularity by merging communities at each step in order to get the greatest value increase. Only those communities sharing one or more edges are allowed to merge at each step. This method is performed in linear time; however, the community quality is less than that of other more costly methods.

4.2.3 Louvain Algorithm

The main benefit of the Louvain algorithm [5] lies in its capacity to operate very quickly on extremely large weighted graphs. This property however does not guarantee an optimal graph partition; an adapted modularity formula, derived from the initial formula presented in Sect. 2.4, is used for weighted graphs. Initially, all vertices are placed in different communities. At first, all vertices are taken into consideration. For each node i, the algorithm computes the gain in weighted modularity when placing i in the community of its neighbor node j and then chooses the community offering maximal gain. At the end of this first loop, the algorithm yields the first partitioning scheme before repeating the same step while already considering formed communities as new nodes. The algorithm stops once additional increases in modularity are no longer possible. This method has been used to process very large social networks extracted from phone companies, for example, with over 2.6 million customers (see [5] for more details). Its processing time is very short.

4.3 From Graphs to Overlapping Communities, Hypergraphs (B1)

This class of methods remains atypical, yet a number of authors have attempted its implementation. The input data is a normal graph, while the output is a list of node groups representing the communities of the initial graph; these communities may indeed overlap. The result could be represented as a hypergraph. The survey [56] mentions several methods without providing any description along with two surveys [18, 53] that describe methods found in this class. The most famous of such methods is "clique percolation"; this algorithm devised by Palla et al. [15] speculates that the internal edges of a community are likely to form cliques due to

their high density. On the other hand, it is unlikely that intercommunity edges form cliques. Palla et al. use the term k-clique in reference to a complete graph with k vertices. Two k-cliques are adjacent if they share k-1 vertices. The union of adjacent k-cliques constitutes a k-clique chain. Two k-cliques are connected if they form part of a k-clique chain. Moreover, a k-clique community is the largest connected subgraph obtained by uniting a k-clique with all k-cliques connected to it. Several authors have proposed improvements to this method given that the computing time may increase exponentially with the number of nodes or edges in the graph. This method has been determined to provide good results.

Other authors have found alternative ways to extract overlapping communities, such as [17, 22, 23]. As an example, [23] enhanced the Girvan-Newman algorithm (see above) in its ability to detect overlapping communities.

4.4 From Bipartite Graphs to Partitions (A2)

In this class, the inputs are bipartite graphs, representing, for example, individuals sharing common properties (photos, tags, etc.). The output contains a list of communities from the initial graph. Each node belongs to just one community. No surveys directly describe this particular case; however, [67], which is based on work performed by Murata [42], adapted a new modularity measure for bipartite graphs in order to build separate communities. Another effort in [58] uses cluster-type local density to extract communities from bipartite graphs modeled as hypergraphs.

4.5 From Bipartite Graphs to Overlapping Communities, Hypergraphs (B2)

In this class, inputs are bipartite graphs, while outputs are either hypergraphs or lists of node groups representing communities that may or may not overlap. One survey [53] describes this case, in citing [34]. By defining "epistemic communities" in [61], Roth depicts a partial example of this class and then takes it one step further with Galois hierarchies (see below).

4.6 From Bipartite Graphs to Galois Hierarchies (C2)

This class extends another step by attempting to extract communities while preserving knowledge shared in each community. No survey has described this type of method. The inputs are bipartite graphs, and the outputs a Galois hierarchy that reveals communities semantically defined with their common properties or

shared knowledge. Communities are nonempty lattice extents, and the result is a hypergraph whose hyperedges are labeled by lattice intents (i.e., shared knowledge).

However, a Galois hierarchy, which is roughly computed from the hypergraph input (as in the case of Freeman [19]), is not a satisfactory scheme since a significant number of groups may be obtained. Under ideal conditions, reduction methods should be introduced, which at one level cause the loss of some semantic precision, yet on another level add precision, that is, cohesion and reliability inside the extracted communities. Only very few authors have actually addressed this difficulty. Roth [61] found the epistemic communities described above before proposing to retain communities of significant size (extent) and semantics (intent), though with weak justification and validation for the proposed heuristics. The authors in [63] then proposed well-known Galois lattice reduction methods based on the so-called iceberg method as well as the stability method.

The iceberg method from [29] identifies concepts with frequent intents above a set threshold. The authors however point out that some important concepts may be overlooked with this method. Stability methods, as used in [29,32], rely on concept stability. The fewer the number of extent subsets present in child concepts, the greater the concept stability. In [8], it is argued that combining both the iceberg and stability methods yields good results for extracting pertinent communities based on concepts. Two thresholds still need to be set however, as the algorithms computation time may be exponential in the number of objects and attributes (NP complete), and lastly the result presented in the form of a Galois lattice is not easily comprehensible.

4.7 Discussions

In this section, all the major community detection methods have been classified and described. The majority of methods examined lie in cell A1 of Table 2, thus producing a partition scheme of communities, which is a configuration not so well adapted to social networks since individuals may belong to several interest groups. The methods in cell B1 of Table 2 allow extracted communities to overlap. This hypergraph model is better adapted to representing social communities. The methods in cell A2 address the case where individuals are represented with their property knowledge (bipartite graph) yet still provide a partition community scheme. The approaches in cell B2 are more realistic when it comes to representing property sharing communities, although they do require some abstraction. Lastly, the methods in cell C2 are the most accurate, because they extract communities using their precise semantics. Nonetheless, they fall short of giving simple and practical results. It is easy to conclude the lack of perfect methods, as each one presents its pros and cons depending on what the experimenters are seeking. Many methods have been proposed to extract partitioned communities from simple graphs, and this availability of methods is certainly due to the ease of describing this type of problem and drawing a partition in comparison with hypergraphs or Galois lattices.

5 Evaluation Methods for Community Detection

Validation is a key issue: How is it possible to verify that the communities identified are actually the appropriate ones? How is it possible to compare results between two distinct algorithms and declare one better than the other? Several methods may be proposed. One simple sentence from a survey [56] reminds us of the great difficulty involved in evaluating community detection methods: "Now that we have all these ways of detecting communities, what do we do with them?" No evaluation methods are actually given. In his survey, Fortunato [18] provides an effective analysis of evaluation methods in proposing three steps: benchmarks, evaluation measures, and comparative evaluation results. Papadopoulos' survey [53] merges evaluation with various community definitions; like other surveys available, it does not pay great attention to evaluation, except for presenting applications on well-known cases and extensive practical social networks. Most validation methods have been designed for the methods in cell A1 of Tables 1 and 2. As a matter of fact, most detection methods may be compared to "clustering" algorithms as regards evaluation. It is well known that clustering yields results that are difficult to evaluate. This same situation is encountered in the area of community detection. Some standard evaluation methods do however emerge. This section will start by presenting several measures of potential use in evaluating the results of community detection methods. Afterward, evaluation benchmarks will be introduced before reviewing a number of evaluation methods.

5.1 Community Extraction Results' Measures

5.1.1 Referent Graph Versus Expert Panel

One type of evaluation is based on expert validation. Two validation modes may be considered. Either the result is presented to an expert or a panel of experts, who visually decide whether or not each community is valid. Alternatively, referent community allocation schemes, which are perhaps manually designed by an expert or else a users' panel, are available; according to such a scheme, an expert uses the measures described below in order to decide whether or not the computed result is adequate with a given confidence criterion. This method relies on expertise and may vary depending on the assigned expert; it may also depend on the actual viewpoint adopted. Each community can be positively evaluated provided a justifying angle of interpretation can be found. As such, the tasks of expert work and reliability are rendered quite difficult. Some types of referent graphs however do not introduce any doubt, for example, a graph showing civil relations between individuals in a wedding. The karate club example in [77] is famous because it was known that at one time the club divided into two subgroups and moreover any partitioning method would show this result.

5.1.2 F-Measure Based on Recall and Precision Measures

Gregory [22] adapted the measures of recall and precision formulas:

- Recall: the fraction of vertex pairs belonging to the same community within the referent benchmark graph that are also members of the same community in the resulting partition
- Precision: the fraction of vertex pairs that are members of the same community in the resulting partition while also belonging to the same community in the referent benchmark graph

The F-measure is used in this context, that is, the harmonic average of recall and precision. The F-measure provides a useful balanced vision of the community detection algorithm. The dual recall and precision measures may also be introduced, in applying these formulas on the edges instead of the vertices.

These kinds of measures are very practical yet remain difficult to adapt if the community number in the resulting partition scheme is not the same as that in the referent benchmark graph. Certain adapted measures may be implemented for added flexibility on resulting communities.

Girvan-Newman [21] proposed a similar measure: the fraction of correctly assigned community vertices divided by the total size of the graph.

5.1.3 Automatic Graph and Community Generation

As will be seen in greater detail in the following subsections, a number of authors have generated automatic graphs and graph community schemes using several methods. They then compared their community detection algorithms to the communities actually generated. Some generation methods produce random communities, in which case the goal consists of proving that their algorithms are better than randomness. Other generation methods offer a community scheme according to a previously defined goal (e.g., generate five communities), with the authors then attempting to obtain a similar result with their algorithm. This method will be discussed in more precise terms below.

5.1.4 Quality Measure

A community detection scheme may be considered effective or ineffective by using a so-called quality function. This specific measure provides a means for comparing quality across several community detection schemes.

One of these quality measures, "modularity," which was introduced by Newman [46] and has already been mentioned in Sect. 2.4, is very famous and widely used. Modularity expresses the fact that a community has a high density ratio as compared to the same graph without any community structure. This high-density criterion is considered to offer good community detection quality. Some authors have combined

a local quality measure (based on modularity) along with the potential of community communication in order to produce an overall quality ratio (see for example [10]).

Several authors cited in the Fortunato survey [18] (in his Section C2) note that a community detection method yielding strong modularity results is not always the best choice, in arguing that low modularity values could provide greater stability in communities.

5.1.5 Discussion and More Global Measures

Some measures may prove to be contradictory, as indicated in [31]. Moreover, most measures focus on the mathematical properties of a graph. In a social network however, an individual may have different intentions regarding group membership. Real communities may not be optimal with respect to collective modularity. In the real world, communities are determined by their history, which in turn is driven by individual personalities and contingent events. The social optimum at present is not easily able to manage and explain everything. Other methods are needed to take into account these dynamic and human dimensions.

5.2 Standard Benchmarks, Random Generated Graphs

Considered by many researchers as references, several popular real networks are often used as benchmarks. Some of these are cited along with their vertex number v and edge number e: Zachary's Karate Club [77] ($v = 34$, $e = 78$), a social network of dolphins living in Doubtful Sound (New Zealand) [38] ($v = 62$, $e = 159$), college football team games [21] ($v = 115$, $e = 613$), university e-mail network [24] ($v = 1133$, $e = 5{,}451$), and scientific coauthorship in condensed matter physics [43] ($v = 27{,}519$, $e = 116{,}181$). For some of these, a final community list for their graphs exists, while for others only the graph structure is present. Girvan-Newman [21] designed a random computer-generated graph and community partition scheme with four groups and 32 vertices in each group. This setup has since become a standard benchmark. A new enhanced version of this graph has been designed by Brandes [7]. These standard graphs and benchmarks can then be used to compare community results with a particular community detection algorithm by applying the previously defined measures.

5.3 Community Detection Method Evaluation

From an evaluation standpoint, community detection is a complex problem. Each method evaluation turns out to be different, and comparing each method's performance proves to be a real challenge. By using the set of measures presented in Sect. 5.1, a performance evaluation is derived for a given algorithm. A good

solution consists of obtaining results from several methods and maintaining the community detection schemes generated by several methods. This option guarantees good stability of the community detection scheme. Testing a method against the Girvan-Newman benchmark entails calculating the similarity between partitions determined by the method and the natural partition of the graph within the four equal-sized groups. Regarding the two popular real networks with a known community structure, that is, the social networks of Zachary's Karate Club and bottlenose dolphins, the question raised is whether the actual separation into two social groups could have been predicted from the graph topology. Zachary's Karate Club is by far the most widely investigated system. Several algorithms are in fact able to identify the two classes, notwithstanding a few intermediate vertices, which could potentially be misclassified. For example, the so-called Louvain algorithm finds five karate club communities instead of 2. This process however does not guarantee the performance on real networks. Other dimensions, such as semantics and pragmatics, need to be considered as we have argued above. Community detection methods that keep knowledge embedded in the original network, that is, based on Galois hierarchies or hypergraphs, may lead to more accurate results. This evaluation process enhancement has yet to be introduced. Some of the authors from the C2 and C3 cells in Table 2 have addressed this difficult issue, though semantics and pragmatics considerations must still be developed in community detection evaluation methods.

6 Conclusion

The study of networked communities is, in some respects, now quite old, with its origins traced back to sociology, computer science, statistics, and other disciplines. Nevertheless, the expanding field of social networks has focused greater attention on this topic. The present survey has provided a state of the art on existing methods with a new angle: classifying algorithms according to their input and output data schemes. Graphs, hypergraphs, and Galois lattices are proven to be useful in representing the growing complexity of community detection methods. This development has allowed demonstrating how to share knowledge and information among social network users. Further research is still required in this field given the organizing power and commercial interest inherent in knowledge. More methods should be developed, and their associated software tools are expected to follow.

References

1. Ahn, Y., Bagrow, J., Lehmann, S.: Communities and hierarchical organization of links in complex networks. Eprint Phys. **1**, 1–8 (2009)
2. Andrew, A.M.: Information Theory, Inference, and Learning Algorithms, by David J. C. MacKay, Cambridge University Press, Cambridge (2003), hardback, pp. xii + 628, ISBN 0-521-64298-1 (30.00), vol. 22. Cambridge University Press, Cambridge (2004)

3. Bagrow, J.P.: Evaluating local community methods in networks. J. Stat. Mech. Theory Exp. **2008**(05), 8 (2007)
4. Berge, C.: Hypergraphes, Combinatoires des Ensembles Finis. Gauthier-Villars, Paris (1987)
5. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**(10), P10008 (2008)
6. Borgatti, S.P., Everett, M.G., Shirey, P.R.: LS sets, lambda sets and other cohesive subsets. Soc. Netw. **12**(4), 337–357 (1990)
7. Brandes, U., Gaertler, M., Wagner, D.: Experiments on graph clustering algorithms. In: Proceedings of 11th European Symposium on Algorithms (ESA '03), Budapest, pp. 568–579 (2003)
8. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Trans. Knowl. Data Eng. **20**(2), 172–188 (2008)
9. Capocci, A., Servedio, V.D.P., Caldarelli, G., Colaiori, F.: Detecting communities in large networks. Phys. A Stat. Mech. Appl. **352**, 669–676 (2005)
10. Clauset, A.: Finding local community structure in networks. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **72**(2 Pt 2), 7 (2005)
11. Clauset, A., Newman, M., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**(6), 1–6 (2004)
12. Crampes, M., Plantié, M., Julien, B.: Cliques maximales d'un graphe et treillis de Galois. In: MARAMI Conférence sur les Modèles et l'Analyse des Réseaux: Approches Mathématiques et Informatique, Grenoble (2011)
13. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. J. Stat. Mech. Theory Exp. **2005**(09), 10 (2005)
14. David, H., Yehuda, K.: On clustering using random walks. In: Hariharan, R., Mukund, M., Vinay, V. (eds.) FSTTCS 2001, LNCS 2245, pp. 18–41. Springer, Berlin/Heidelberg (2001)
15. Derényi, I., Palla, G., Vicsek, T.: Clique percolation in random networks. Phys. Rev. Lett. **94**(16), 160202 (2005)
16. Evans, T.S., Lambiotte, R.: Line graphs, link partitions and overlapping communities. Phys. Rev. E **80**(1), 9 (2009)
17. Falzon, L.: Determining groups from the clique structure in large social networks. Soc. Netw. **22**(2), 159–172 (2000)
18. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 103 (2009)
19. Freeman, L.C., White, D.R.: Using galois lattices to represent network data. Sociol. Methodol. **23**, 127–146 (1993)
20. Ganter, B., Wille, R.: Formal Concept Analysis: Foundations and Applications. Springer, Berlin (1999)
21. Girvan, M., Newman, M.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**(12), 7821–7826 (2002)
22. Gregory, S.: A fast algorithm to find overlapping communities in networks. Mach. Learn. Knowl. Discov. Databases **5211**, 408–423 (2008)
23. Gregory, S.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**(10), 103018 (2009)
24. Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Phys. Rev. E **68**(6), 1–4 (2003)
25. Guimerà, R., Sales-Pardo, M., Amaral, L.: Module identification in bipartite and directed networks. Phys. Rev. E **76**(3), 036102 (2007)
26. Gulbahce, N., Lehmann, S.: The art of community detection. BioEssays News Rev. Mol. Cell. Dev. Biol. **30**(10), 934–938 (2008)
27. Hughes, D.: Random walks and random environments. Volume 1: random walks. Bull. Math. Biol. **58**(3), 598–599 (1996)
28. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
29. Jay, N., Kohler, F., Napoli, A.: Analysis of Social Communities with Iceberg and Stability-Based Concept Lattices, pp. 258–272. Springer, Berlin/New York (2008)

30. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell. Sys. Tech. J. **49**(2), 291–308 (1970)
31. Kleinberg, J.: An impossibility theorem for clustering. In: Obermayer, K., Becker, S., Thrun, S. (eds.) Advances in Neural Information Processing Systems 15s. MIT, Cambridge (2002)
32. Kuznetsov, S.O.: On stability of a formal concept. Ann. Math. Artif. Intell. **49**(1–4), 101–115 (2007)
33. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. **11**(3), 033015 (2009)
34. Lin, Y.-R., Sun, J., Castro, P., Konuru, R., Sundaram, H., Kelliher, A.: MetaFac: community discovery via relational hypergraph factorization. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 527–536. ACM, New York (2009)
35. Luccio, F., Sami, M.: On the decomposition of networks in minimally interconnected subnetworks. IEEE Trans. Circuit Theory **16**, 184–188 (1969)
36. Luce, R.D.: Connectivity and generalized cliques in sociometric group structure. Psychometrika **15**(2), 169–190 (1950)
37. Luce, R.D., Perry, A.D.: A method of matrix analysis of group structure. Psychometrika **14**(1), 95–116 (1949)
38. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behav. Ecol. Sociobiol. **54**(4), 396–405 (2003)
39. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, L. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
40. Miyagawa, H.: Community extraction in hypergraphs based on adjacent numbers. Oper. Res. **50**, 309–316 (2010)
41. Murata, T.: Detecting communities from tripartite networks. WWW '10. ACM, New York (2010)
42. Murata, T.: Modularity for heterogeneous networks. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia – HT '10, pp. 129. ACM, New York (2010)
43. Newman, M.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98**(2), 7 (2000)
44. Newman, M.: Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**(6), 066133 (2004)
45. Newman, M.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **74**(3 Pt 2), 036104 (2006)
46. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
47. Newman, M., Park, J.: Why social networks are different from other types of networks. Phys. Rev E Stat. Nonlinear Soft Matter Phys. **68**(3 Pt 2), 036122 (2003)
48. Nicolas, N., Klaus, O.: Towards community detection in k-partite k-uniform hypergraphs. In: Proceedings NIPS 2009. Vancouver, B.C., Canada.
49. Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri, M.: Extending the definition of modularity to directed graphs with overlapping communities. J. Stat. Mech. Theory Exp. **2009**(03), P03024 (2009)
50. Noack, A., Rotta, R.: Multi-level algorithms for modularity clustering. Lecture Notes in Computer Science, pp. 257–268. Springer Berlin/Heidelberg (2008)
51. Nordhausen, K.: The elements of statistical learning: data mining, inference, and prediction, second edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. Int. Stat. Rev. **77**(3), 482–482 (2009)
52. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–8 (2005)

53. Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community detection in social media. Data Min. Knowl. Discov. **2460** (June), 1–40 (2011)
54. Plantié, M., Crampes, M.: From photo networks to social networks, creation and use of a social network derived with photos. In: Proceedings of the ACM International Conference on Multimedia, Firenze, October 2010
55. Pons, P.: Détection de communautés dans les grands graphes de terrain. Ph.D. thesis, Paris 7 (2007)
56. Porter, M.A., Onnela, J.P., Mucha, P.J.: Communities in networks. Not. Am. Math. Soc. **56**, 1082–1097 (2009)
57. Pothen, A., Simon, H.D., Liou, K.-P.: Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. **11**(3), 430 (1990)
58. Qian, R., Zhang, W., Yang, B.: Community detection in scale-free networks based on hypergraph model. In: Proceedings of the 2007 Pacific Asia Conference on Intelligence and Security Informatics, PAISI'07, pp. 226–231. Springer, Berlin/Heidelberg (2007)
59. Roth, C.: Compact, evolving community taxonomies using concept lattices. In: Hitzler, P., Schaerfe, H., Ohrstrom, P. (eds.) Contributions to 14th International Conference on Conceptual Structures, pp. 172–187. Aalborg University Press, Aalborg (2006)
60. Roth, C., Bourgine, P.: Binding social and cultural networks: a model. Networks nlin.AO(February), 8 (2003)
61. Roth, C., Bourgine, P.: Epistemic communities: description and hierarchic categorization. Math. Popul. Stud. Int. J. Math Demogr. **12**(2), 107–130 (2005)
62. Roth, C., Bourgine, P.: Lattice-based dynamic and overlapping taxonomies: the case of epistemic communities. Scientometrics **69**(2), 429–447 (2006)
63. Roth, C., Obiedkoy, S., Kourie, D.G.: On succinct representation of knowledge community taxonomies with formal concept analysis. Int. J. Found. Comput. Sci. **19**(2), 383 (2008)
64. Schaeffer, S.: Graph clustering. Comput. Sci. Rev. **1**(1), 27–64 (2007)
65. Seidman, S.B., Foster, B.L.: A graph-theoretic generalization of the clique concept. J. Math. Sociol. **6**(1), 139–154 (1978)
66. Selvakkumaran, N., Karypis, G.: Multiobjective hypergraph-partitioning algorithms for cut and maximum subdomain-degree minimization. Comput.-Aided Des. Integr. Circuits Syst. IEEE Trans. **25**(3), 504–517 (2006)
67. Suzuki, K., Wakita, K.: Extracting multi-facet community structure from bipartite networks. 2009 Int. Conf. Comput. Sci. Eng. **4**, 312–319 (2009)
68. Taramasco, C., Cointet, J.-P., Roth, C.: Academic team formation as evolving hypergraphs. Scientometrics **85**(3), 721–740 (2010)
69. Tyler, J.R., Wilkinson, D.M., Huberman, B.A.: Email as spectroscopy: automated discovery of community structure within organizations. In: Communities and Technologies, pp. 81–96. Kluwer, Norwell (2003)
70. Wan, L., Liao, J., Zhu, X.: CDPM: finding and evaluating community structure in social networks. In: Proceedings of the 4th International Conference on Advanced Data Mining and Applications, ADMA '08, pp. 620–627. Springer, Berlin/Heidelberg (2008)
71. Wan, L., Liao, J., Wang, C., Zhu, X.: JCCM: joint cluster communities on attribute and relationship data in social networks. Adv. Data Min. Appl. **5678**(60525110), 671–679 (2009)
72. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
73. Wu, F.Y.: The potts model. Rev. Mod. Phys. **54**(1), 235–268 (1982)
74. Wu, F., Huberman, B.A.: Finding communities in linear time: a physics approach. Eur. Phys. J. B Condens. Matter **38**(2), 331–338 (2003)
75. Yang, T., Chi, Y., Zhu, S., Gong, Y., Jin, R.: A bayesian approach toward finding communities and their evolutions in dynamic social networks. Work. In: Proceedings of the Ninth SIAM International Conference on Data MiningSociety for Industrial and Applied Mathematics, pp. 990–1001 (2009). Philadelphia, USA
76. Yang, B., Liu, D., Liu, J., Furht, B.: Discovering Communities from Social Networks: Methodologies and Applications. Springer, Boston (2010)

77. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**(4), 452–473 (1977)
78. Zaidi, F., Sallaberry, A., Melancon, G.: Revealing hidden community structures and identifying bridges in complex networks: an application to analyzing contents of web pages. In: IEEE/WIC/ACM International Conference on Web, pp. 198–205. IEEE, Washington, DC (2009)
79. Zhou, H., Lipowsky, R.: Network brownian motion: a new method to measure vertex-vertex proximity and to identify communities and subcommunities. In: International Conference on Computational Science. Lecture Notes in Computer Science, vol. 3038, pp. 1062–1069. Springer, Berlin/New York (2004)

Detecting Multimedia Contents of Social Events in Social Networks

Mohamad Rabbath and Susanne Boll

Abstract Friends visit many social events together, take photos of each other and upload them in several accounts in different types of social media such as Flickr or Facebook. In this chapter we define the different types of events in social media and introduce two approaches in the state of the art of detecting the media content of an event. The first is an ontology-based approach where the metadata is basically used to obtain the content of a well-defined event and the visual features are used to prune the results to finally get the illustrative elements. The second approach exploits a mixture of features (visual, social, structural and metadata) that we also explain in this chapter and fuses them in a probabilistic model to link the media elements of different users and albums to their representative event. We discuss the advantage and disadvantage of each approach and the cases of using each.

1 Introduction

Social media such as Facebook and Flickr became very popular platforms to share multimedia contents, with more than 40 million monthly added photos in Flickr and more than 200 million uploaded photos uploaded per day in Facebook with currently almost 90 billion photos in total [1]. The growing market of smartphones and the resolution of their cameras also has a big effect in the growth of the shared media content. Friends and family members visit the same event together, take photos and videos of each other and upload them each to his account in the social media. If the user wants to collect all the media elements of a specific social event, this task

M. Rabbath (✉)
OFFIS – Institute for Information Technology, Escherweg2, Oldenburg, D-26121, Germany
e-mail: rabbath@offis.de

S. Boll
University of Oldenburg, Oldenburg, Germany
e-mail: susanne.boll@informatik.uni-oldenburg.de

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_5, © Springer-Verlag London 2013

becomes very challenging. Social media differs in treating the stored contents. Some of which preserve the metadata of the contents such as the geo-temporal metadata like the case of Flickr, although much of this metadata is noisy [2]. Others like the case of Facebook strip out most of this metadata included in the EXIF header, although they preserve the information that is manually added by the user to the albums. In these conditions, the content-based analysis remains very important in the task of detecting the media contents that belong to the same event. In this chapter we summarize the state of the art of event detection and clustering approaches. We deal with several types of social media platforms and how do they deal with events. We also show the features used in the multimedia content analysis to support event detection including metadata, visual, people and structure-based features. The state of the art models and approaches are also explained in this chapter. We focus on the ontology-based model and the fusion probabilistic model that are used in several approaches.

2 The State of the Art in Event Clustering

Event detection in general is a very active research area, due to its variety of important applications in multimedia. For example, the work of [3] handled the natural events, where the event model was described in many aspects (temporal, spatial, informational, experiential, structural, causal). The work of [4] studied the effect of personalization and user context in annotating the image to its event represented in four facets (where, when, who, what), where a graph-based model that exploits the WordNet ontology is used. The work of [5, 6] introduced a fusion probabilistic approach that uses camera, temporal and visual metadata to cluster the home-made photos to their representative events.

Regarding the events in social media platforms, [7] extensively analysed the groups in Flickr including the event groups, however based on the textual features only. What is interesting however is the unsupervised probabilistic model introduced in the approach where the topics are hidden parameters and the textual annotations are observed parameters that are used to cluster the groups to their topics using the expectation-maximization algorithm. Closer to the events in Flickr, the work of [8] introduced an ontology-based approach where first the visual features of the photos are used to find the most related photos to the seeds of an event and then an ontology is used to extend the annotations describing the photos to get more related photos in Flickr to the event. Different features were used in event detection and analysis. For example, [8–10] used ontology-based approaches considering well-annotated contents. There is also very few works in the area of event detection based on pure visual features. The work of [2] introduced a parallel computational MapReduce framework, which basically uses visual features for event matching in Flickr. Facebook becomes the largest photo hosting site on the web, and the photos of the same social events are very much distributed among friends. The work of [11] introduced an approach to link the photos of the same event distributed between

friends in Facebook. Global visual, tagged area-, structural- and friendship-based features were studied and evaluated to detect the related events where much of the metadata is missing in Facebook including the geo-temporal information. The work of [12] introduced a probabilistic model to estimate the centric location of an event such as the earthquake in real time. Most of the general models and approaches in the state of the art of event detection in social media are covered in this chapter.

3 Events in Different Social Media

The way of defining the events is different from a social media to another. For example in Flickr, people can add photos and tag these photos with keywords or labels to make it easier to find the photos later. Some of these labels are describing the event, just like the example used in the work of [8] where a dataset of public concerts was used to evaluate the approach. In many of the ontology-based approaches such as [4, 8], the event was formally defined in four facets of Ws:

- *When*: where the time window can differ; it could be a year, month or specific time.
- *Where*: which can be a country or specified place, for example, a house or a public place.
- *What*: for example, birthday and wedding.
- *Who*: the people participating in the event.

Events vary in their geo-temporal window, and defining those facets for each event is a challenging task. Moreover, annotating each photo with those four Ws is fairly hard and impractical. As described in the work of [13], the definition of the event in social media can differ from user to user and not necessarily restricted to four facets. The user can define an event as a trip of several days where several places or countries are visited, and also the users themselves do not agree on annotations. For example, what could be a "vacation in Germany" to some user could be "my friend's visit" to another. Also, the kind of tags differs from one social media to another. The concept of tags in Flickr contains the possible descriptive categories of the photo as shown in Fig. 1.

On the other hand, in Facebook, the concept of tags refers to the annotations added by the users to the faces describing the people within the photo as shown in Fig. 2. In addition to the people tags, Facebook allows adding description, location and time to each album and caption to each photo, and people can comment and like the photos of each other. Most of these textual annotations in Facebook describe the social context and the interaction between friends, rather than the category of the photos as the case of Flickr tags. Unlike Flickr, Facebook strips out the metadata included in the EXIF header of the photo such as the time, camera information and geo-metadata. In the opposite, a large number of photos in Flickr contain metadata, and thus, many datasets from Flickr are used for evaluation. In general, we divide the events containing photos and videos in social media into two types:

Fig. 1 Example of the tags in Flickr where the photo is annotated with its possible categories

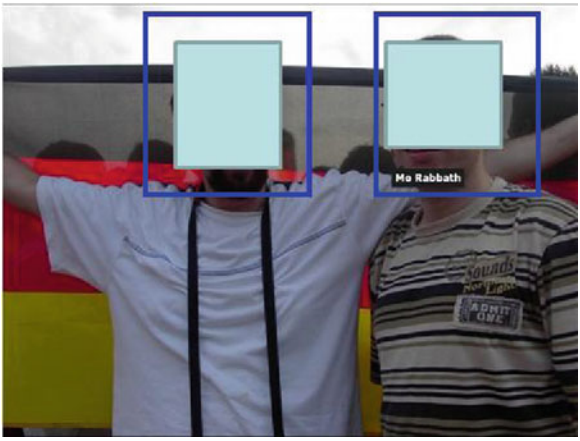


Fig. 2 Example of the tags in Facebook where friends tag each other

Explicitly created events: Users in social media can create a specific event (private, public or other types of privacy settings). Creating an explicit event in Flickr implies adding a hierarchical "event" textual tag with another subcategory tag like "wedding". In Facebook, the explicitly created events are more structured. Once an explicitly created event is created in Facebook, the creator can invite people and add photos directly to the event. People can mark them as "attending", "maybe" or "not attending" as shown in Fig. 3.

Implicit events within albums: The vast majority of the users in social networks such as Facebook add the photos to their albums after attending an event, without explicitly creating an event. Each user can add his collection of photos to an album

Fig. 3 Example of the Facebook explicitly created events

that he creates. This results in many albums of the same event distributed between friends. The maximum allowed number of photos per album changed over time in Facebook, and currently they are 200 photos maximally, so each single user may also need to distribute photos over several albums. Friends can then tag each other, like and comment on each other's photos. Each album may contain full description including the location, time, title and overall textual description. Each individual photo may also contain a caption and title.

Formal definitions: We define the set of explicitly created events e_k created by the user of ID u in some account in the social media as $E_u = \{e_k \mid k : u + 1..u + |E_u|\}$, where each e_k can be defined in terms of its photos $e_k = \{p_i \mid i : k + 1..k + |e_k|\}$. We denote $evt_{p_{i1},p_{i2}}$ as both photos p_{i1}, p_{i2} belong to the same semantic event. If the two photos already belong to an explicitly created event, then by definition they also belong to the same semantic event:

$$p_{i1}, p_{i2} \in e_k \Rightarrow evt_{p_{i1},p_{i2}} \tag{1}$$

In the problem of event detection of the multimedia contents such as photos and videos (which can also be handled as sequences of representative frame photos), our main goal is that given a set of photos P, for each photo $p_i \in P$ finding all photos p_j that belong to the same event (evt_{p_i,p_j}), with p_j, p_i do not belong to the same explicit event set e_k, meaning the photos that are in other album groups (usually added by other friends) but still in the same semantic event.

4 Feature Extraction and Analysis

Photos that belong to the same event have uniform features that can be exploited. They often share global visual similarity. The big advantage of the global visual features is that they are independent of the structure and the preserved metadata in

the social network. However, global visual features alone are not enough to reach efficient detection. A lot of work exploited temporal or geographical metadata to achieve better result, such as the work of [14] where distinct classifications were built from temporal and geoinformation. Also, as explained in the previous section, many photos in Facebook are tagged with people, and therefore depending on the accuracy of the tags, the clothes of the person can be estimated, and this can be very helpful because each person tends to wear the same clothes in the same event. The work of [15], for example, used the upper body to increase the accuracy of face recognition. People in the photos constitute an important feature to recognize different events, for example, [16] introduced a bag of people classification. As previously explained, the structure of the events also differs from a social media platform to another, and the explicit social event structure of Facebook explained in the previous section is a very interesting example to study. In addition to that, the metadata of the media contents such as Flickr textual tags, geo-temporal information and camera metadata if it exists can efficiently increase the accuracy of mapping the multimedia contents to their respective events. In this section we will explain all these features and their effects in retrieving the photos to their respective events.

4.1 Global Visual Features

Unlike the case of object recognition, the low-level visual features evaluated by Strong and Gong [17] play more important role in the problem of event clustering than purely local features, for example, SIFT features [18], because objects change during the event, but low-level features such as the colour histogram usually remain more similar within one-event photos than between different events. For example, a trip event in a garden during the summer most probably may have more green pixels in its photos. Also, the appearance of the same people often implies more edge and colour histogram similarity. In the work of [17], the effectiveness of each visual feature in visual clustering is evaluated. The effectiveness metrics of the work are defined as follows:

If we have N clusters of the photos where a cluster is denoted as S_k, the photo is chosen as centroid c_k of the cluster S_k if it satisfies the following condition:

$$\forall j.j \in S_k \wedge j \neq c_k : \sum\nolimits_{i \in S_k} D(i, c_k) < \sum\nolimits_{i \in S_k} D(i, j)$$

where $D(i, j)$ is the Eucledian distance between two photos on a specific visual feature. Then the closeness of the photos inside the subset S_k is estimated using the function

$$R(S_k) = \frac{1}{\|S_k\|} \sum\nolimits_{i \in S_k} D(i, c_k)$$

Then the average distance to the photos not in the cluster S_k is calculated

$$R(C - S_k) = \frac{1}{\|C - S_k\|} \sum\nolimits_{i \notin S_k} D(i, c_k)$$

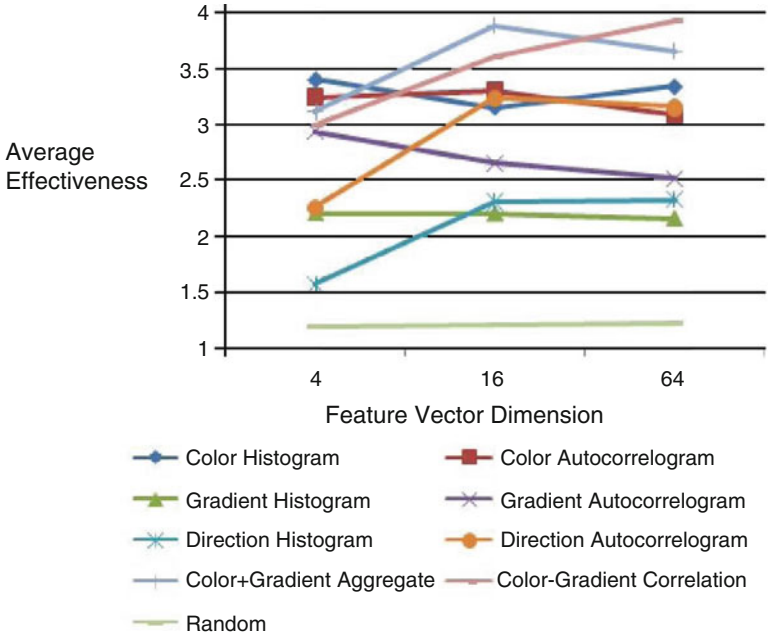


Fig. 4 The evaluation of the features of the colour and directivity described in the work of [17]

where $C = \bigcup_{0<i<=N} S_i$ is the collection of all photos in all the clusters S_i. The effectiveness of a visual feature on the cluster S_k is then calculated as

$$E_k = \frac{R(C - S_k)}{R(S_k)}$$

And finally, the effectiveness is defined as

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i$$

Figure 4 shows the evaluation of the effectiveness of each visual feature with the number of dimensions chosen for each feature vector. As a conclusion, the colour-based features alone generally perform better than the gradient-based features; however, the colour-based features do not necessarily perform better when increasing the dimensionality. A hyper approach between directivity and colour has the best effectiveness, where the colour-gradient correlation performed the best. To benefit most of the visual features, the work of [11] used the colour and directivity histogram (described in [17]). We denote the distances between the colour and directivity histograms of two photos 1, 2 as $\mathrm{CHS}_{1,2}$, $\mathrm{EHS}_{1,2}$, respectively.

Although these features can be quite effective in detecting different events, they totally ignore the localized important features. In the work of [19], a global descriptor that combines the traditional global features with coarsely localized salient regions was introduced. The so-called saliency moments in [19] are extracted from the image of several resolutions. This can be well adjusted to photos in social media where four to five resolutions are offered for each photo in Facebook through the graph api: 2,048 px (if the user uploaded the photo in high quality), 720 px, 180 px, 130 px and 75 px (thumbnail). In the work of [11], the second and third coarser resolutions (720 px, 180 px) are considered because it is both sufficient and computationally efficient to restrict to both resolutions when extracting the saliency moments for the task of visually describing the events. As a summary of extracting the saliency moments as described in [19] and adjusted in [11]:

First, $R_{F_{xy}} = e^{(L_{F_{xy}} + P_{F_{xy}})}$ is computed on the image of both resolutions, where $P_{F_{xy}}$ is the original phase of the signal, and $L_{F_{xy}} = \log\left(A_{F_{xy}}\right) * h_n$ where $A_{F_{xy}}$ is the amplitude of the Fourier spectrum signal amplitude of the image, and h_n is the average filter. This means $L_{F_{xy}}$ is obtained by convoluting $A_{F_{xy}}$ in the logarithmic scale with the average filter. The signal $R_{F_{xy}}$ is multiplied with Gabor filters of 8 orientations, which results in 16 components (eight for each of the two resolutions). A vector is then built by dividing each component into 16 nonoverlapping blogs, and then averaging each block. At the end we obtain a feature vector of 16*16 = 256 dimensions. The feature vector is denoted as $SM = \{x_1, \ldots, x_{256}\}$. To get the distance between two photos on the SM feature, the Euclidean distance gives poor results, because this feature is very sensitive to the movement of the salient objects. In order to calculate the best matches in saliency, [11] applied the following procedure:

1. We order the components in $SM1 =< x_{r1}, \ldots, x_{r256} >$ such that $x_{r1} < x_{r2} < \ldots < x_{r256}$.
2. For each x_{ri} in $SM1$:
   (a) Find x_{ki} in $SM2$ with the min $\|x_{ki} - x_{ri}\|$.
   (b) Remove x_{ki} from $SM2$.
   (c) $V_{SM_{1,2}} = V_{SM_{1,2}} + \|x_{ki} - x_{ri}\|$.
3. $V_{SM_{1,2}}$ will be normalized with a normalization factor ℓ to have a value in [0,1].

The resulting $V_{SM_{1,2}}$ calculated between two photos $1, 2$ can be looked at as the difference between the histograms of the saliency regions in both Photos, and therefore invariant of the positions of these salient regions. We denote the vector of the obtained visual features between two photos $1, 2$ as $Glob_{1,2} =< V_{SM_{1,2}}, \mathrm{CHS}_{1,2}, \mathrm{EHS}_{1,2} >$ where these features cover both the global low-level colour and directivity visual similarity features and coarsely localized salient regions similarity which are most effective in detecting the multimedia contents of an Each photo is divided to 10×10 nonoverlapping areas, and mark the centre of each tag window.

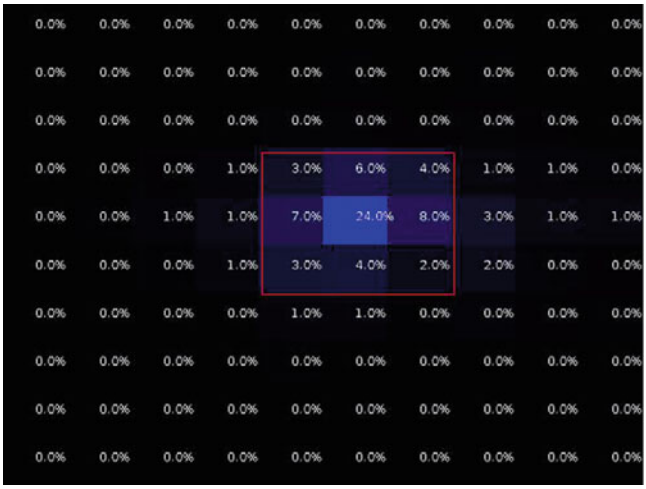


Fig. 5 Tagging accuracy around the face in Facebook according to the work of [11]

4.2 Tagged Area-Based Features

Facebook users can tag each other in a photo p_i which results in tagged areas T_{p_i} of the people appearing in this photo. The positions of the tags are obtainable with the Facebook graph api used by Facebook applications. The tagged areas in a photo is defined as a set of triples where each triple represents the tagged user u and the coordinates of the centre of the tagged area x, y as $T_{p_i} = \{< u, x, y >_j \mid j : p_i + 1..p_i + N_T\}$ where N_T is the number of the tagged users in the photo with id p_i. To estimate how accurately people tag each other, the work of [11] showed a study on the positions of the tags made on a sample of the photos of the users who downloaded the "SmileBooks" Facebook application. The study analysed 9 k tagged photos of few thousands of people. On each photo, a frontal face detection [20] as well as profile face detection in flexible lighting conditions [21] increases the recall of the found faces. If both faces and tags were found, it is assumed that each tag is related to the closest detected face. The red window in Fig. 5 shows the area around the face using a window with the same tag size of Facebook, and the intensity of blue areas shows the percentage of the centre positions of the tags as described in [11]. According to the results, only 4% of the tags are positioned far from the face, while 9% are positioned out of the face but very close, and 61% directly matched the face. The rest of the tags are excluded because no matching faces were detected. This means that around 84% of the tags matching some people are placed accurately on the face or very close. This gives a good indication that the added tags on the photos of Facebook are reliable. In the current time, Facebook added a new feature of suggesting the areas of the tags around the faces which implies even more accuracy in the future.

Due to the accuracy of the added tags, they can be exploited as reliable resources in linking the photos of the same events. For instance, if several photos contain people each of whom wear the same clothes in each photo, this considerably increases the probability that these photos belong to the same event, and the more people who wear the same clothes in the photos are, the higher the probability is that these photos belong to the same event. In the work of [11] for each tagged area, a window 1.5 times bigger than the tag around the face and below it is taken. For this window, the colour and edge histograms ([17]) and texture histogram (using Tamura) are calculated $T_{\mathrm{CHS}}, T_{\mathrm{EHS}}$ and T_{THS}, respectively. We denote $T_{\mathrm{CHS}_{u_i},1,2}, T_{\mathrm{EHS}_{u_i},1,2}$ and $T_{\mathrm{THS}_{u_i},1,2}$ the Euclidean distance of the respective features of the tagged areas in two photos of tagged person u_i, out or N_T tagged people. Because the more tagged people in the photos where they have high similarity under the tagged areas, the higher is the probability for the two photos to be from the same event, the following measures are calculated

$$T_{\mathrm{CHS}_{1,2}} = \prod_{i=1}^{N_T} T_{\mathrm{CHS}_{u_i},1,2} \tag{2}$$

$$T_{\mathrm{EHS}_{1,2}} = \prod_{i=1}^{N_T} T_{\mathrm{EHS}_{u_i},1,2} \tag{3}$$

$$T_{\mathrm{THS}_{1,2}} = \prod_{i=1}^{N_T} T_{\mathrm{THS}_{u_i},1,2} \tag{4}$$

We preserve the vector $T_{1,2} =< T_{\mathrm{CHS}_{1,2}}, T_{\mathrm{EHS}_{1,2}}, T_{\mathrm{THS}_{1,2}} >$. Despite the valuable information of the tagged area features, they do not always work as good as expected. The reason is the overlapping tags, as shown in Fig. 6. Therefore, a filtering process is applied on the tags such that only the features from the nonoverlapping tagged are obtained. Each tagged area which has its below window intersecting with another tagged area will not be considered for feature extraction.

4.3 Friendship-Based Features

People inside the photos constitute a very important feature to detect the content of an event, because usually one specific event contains also the same people. Just like the words in textual retrieval, the importance of the person in a collection of photos depends on how often she appears with the other people in the collection. For example, if a collection of photos contain two people who often appear together (such as a spouse), then it is a less important indicator that this collection is of one event than when these two people rarely appear together. In the work of [11], a measure like the tf-idf in the textual retrieval is introduced for people similarity between two photos. The idf is defined in terms of two people as

Fig. 6 The overlapping tags in Facebook are removed from tagged area-based features extraction according to the work of [11]

$$idf\ (per_1, per_2) = \log \frac{\left|\bigcup A_{per_1}\right| + \left|\bigcup A_{per_2}\right| + \left|E_{per_1}\right| + \left|E_{per_2}\right|}{1 + \left|\bigcup g_l\right| + \left|\bigcup e_k\right| : per_1, per_2 \in \text{People}\,(g_l) \cap \text{People}\,(e_k)} \tag{5}$$

where E_{per} is the set of photos added by or contain person per in all the explicitly created events as explained in Sect. 3, and $\bigcup A_{per}$ is the set of all photos added in the albums of person per. $\left|\bigcup g_l\right| + \left|\bigcup e_k\right|$ is meant to get the size of all photos within the album groups g_l and the explicitly created events e_k where both people $per1, per2$ exist. Obtaining the album groups g_l is explained later in Sect. 5. The similarity between two explicit events or album groups photos in terms of friendship is then calculated as

$$Tf - idf\ (eg_1, eg_2) = \sum_{per_1, per_2 \in \text{People}(eg_1) \cap \text{People}(eg_2)} idf\ (per_1, per_2) * |\text{People}\,(eg_1) \cap \text{People}\,(eg_2)| \tag{6}$$

4.4 Structure-Based Features

As described in Sect. 3, if an explicit event was created in Facebook, for example, the users can mark them as "attending,""maybe attending" or "not attending". Although it might be that no photos are added to the explicit events, and rather photos to other albums are added, but the information if a user recently attended an event might be

quite helpful. Due to the observation of [8], the vast majority of the photos related to a specific event are uploaded directly within 5 days. Therefore, if two groups of photos g_{l1}, g_{l2} are added to two albums of two different people who marked themselves as "attending" or "maybe" attending in an explicitly created event within 5 days from this event, these two groups of photos have higher probability to be from the same event. This feature is highly dependent on the structure of the social network. For example, while the explicitly created social event is more structured in Facebook, it is only annotated with textual tags in Flickr as described in Sect. 3. In the work of [11], the value $Str_{g1,g2}$ between two album groups is calculated. This value represents the number of people tagged in the photos of both album groups who marked themselves as "attending" or "maybe" in the same Facebook event which is explicitly created before maximally 5 days of these two groups.

4.5 Metadata-Based Features

Metadata is very important resources to be exploited. In the work of [5], a mixture of camera (exposure time, aperture and focal) and photo (EXIF) metadata is used in addition to visual features in the task of event clustering. Temporal information proved to be especially efficient and performs better than other features [6, 14]. However, while the temporal metadata τ may be obtainable in some social media like Flickr, they are stripped out in others like Facebook, where both the users and the application have no access to it. Although the EXIF header of the photos is stripped in Facebook, the photo album may contain important metadata, such as the time of creation which can be added manually by the user and geographical information where the user can add the city and the country, in addition to the textual information which is the title and description of the album. Each photo may also contain a caption, and the user's friends can comment on this photo. Flickr photos also contain many textual tags in addition to preserving the geo-temporal information, where the temporal information in Flickr photos is much more common than geographical metadata because almost all digital cameras write the temporal information in the EXIF header of the original photo. But even in the cases where the geo-temporal information are preserved, many of which remain noisy and many photos miss this metadata. For example, in the work of [2], the focus was on the visual features which were analysed using parallel processing framework because much of the geo-temporal metadata was noisy.

5 Clustering Models and Methods

In this section two basic approaches of clustering the media contents to their respective events are explained. The first is an ontology-based model, where the event is characterized in four factual aspects (location, time, description and the

people within). The second approach is a probabilistic-based model, where the probability of the media content to be from the same event is calculated based on the uniform features previously explained.

5.1 Ontology-Based Approach

The ontology model to annotate the photos in the user context or to connect them to their representing events is used in more than little work such as [4, 8]. Each event is connected to four facets of Ws: *where* did it happen, *what*, *when* did it happen and *who* was involved. Practically however, the first three facets are considered where the *Who* dimension is ignored as applied in the work of [8]. In the ontology-based model, most of the process of detecting the media content of the event happens online. The only part that can be executed offline is automatically annotating the photos with external textual information like the work of [4] where a propagation graph-based approach was introduced to annotate the images with four facets due to the user social context. The main steps of this approach can be summarized with the following: querying the event by title, querying the event by geoinformation in the correct respective time window, merging the results and pruning irrelevant media. In the following we explain each step as described in [8], where first the *what* dimension is exploited through query by textual description with filtering by the *when* dimension using the time when the event took place. After that, the *where* dimension is queried using the geotags of the event in the same window of time when the event took place. The visual features are used to remove the irrelevant results in the last step.

5.1.1 Query by Title

As previously described in metadata features, textual information is one of the most important metadata, and the title is one of the most useful information to describe an event. In the work of [8], the full text search of an event is performed using the description extracted from the LODE ontology.[1] The magnitude of the results returned by querying the description is much greater than with geotag queries. But at the same time, the precision is highly affected by the problem of polysemy, where many of the results are irrelevant. Therefore, a pruning step is performed as we will explain in Sect. 5.1.3. In addition to the textual query, the results are filtered to a window time of 5 days of the targeted event, due to the observation of [8] where most of the media elements of an event are uploaded within 5 days as shown in Fig. 7. In the same work, it is also noted that query by title with time filtering results in higher precision in videos from YouTube than photos in Flickr, where more advanced visual-based filtering should be applied as we will describe later.

[1] http://linkedevents.org/ontology/

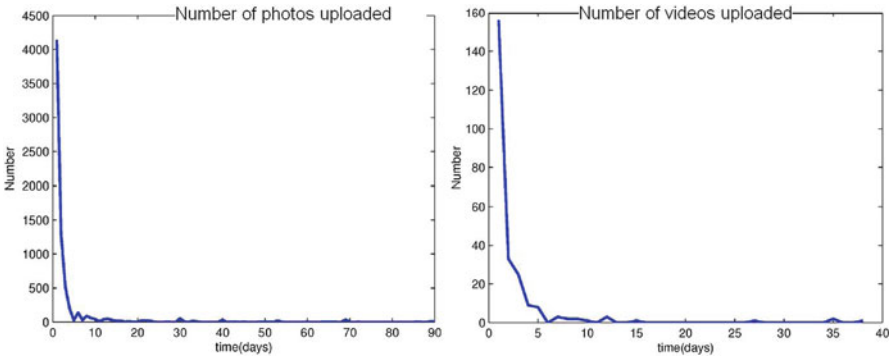


Fig. 7 The uploaded photos and videos in terms of days after a specific event as shown in the work of [8]

5.1.2 Query by Geoinformation in the Correct Time Window

Geotags are very valuable metadata as explained in the metadata features. In the same work of [8], a query by geotags is executed in addition to the query by title to detect the media elements of an event, knowing that each event is defined in four facets as previously described. Just like the case of the query by title, the geo query is extracted from the LODE ontology. The geotags in Flickr usually consist of latitude and longitude coordinates or place names. The query by geotags is also executed in 5 days window of time from the event.

5.1.3 Merging the Results and Pruning the Irrelevant Media

The media elements returned after both queries by geotags and by title are merged after removing the duplicated elements returned from both queries. The results returned from the query by geotags of the event in the respective window time are considered relatively accurate. In contrast, the results returned from the query by title contain a lot of irrelevant contents, especially because of the problem of polysemy. To filter the irrelevant media, a pruning algorithm was introduced in the work of [8]. The visual features play the major role in this phase to filter the very diverse irrelevant media resulted from the title query. First, a testing dataset for each event is composed using the media elements returned from the geotag query, or those who match the Flickr tag of the event and not only the descriptive text. The used visual features are low-level global features due to their effectiveness in the event recognition as described in Sect. 4. The chosen features in the work of [8] are colour moments in Lab space with 225D, Gabor texture with 64D and edge histogram with 73D. For each element in the training test, the nearest neighbours are found using the distance measure $L1$, and the smallest distance is considered as the threshold $Threshold$. After that, for each media element in the results of

Table 1 The algorithm of pruning the irrelevant media from the work of [8]

```
1. INPUT: TrainingSet,TestingSet
2. OUTPUT: PrunedSet
3. foreach img in TrainingSet
     D =[ ]
     foreach imgj in TrainingSet-img
          D.append(dist L1(img,imgj))
     Threshold = min(D)
     foreach imgt in TestingSet
          if dist L1(imgt,img)< Threshold
               PrunedSet.append(imgt)
4. return PrunedSet
```

the title query, if the distance to one element in the training dataset is below the threshold, then this media element is added to the event media. A summary of this simple algorithm is shown in Table 1. Because the threshold is chosen to be the closest visual distance in the neighbourhood of each image in the training dataset, only photos with high certainty to belong to the event are kept. The work of [8] addressed the issue of the high precision, but low recall resulted from the pruning.

5.1.4 Conclusion of This Approach

The explained ontology approach exploits the metadata especially time, geotags and textual description to retrieve the media contents of a well-defined event in the *where*, *when* and *what* dimensions. The pruning phase based on the visual features guarantees a considerably high precision while achieving low recall. This approach works well with well-defined events especially in public event in Flickr. Also due to the high precision and low recall, this approach is very suitable to find the illustrative media of an event which represents this event with high certainty. However, this approach does not work if the metadata such as the geo-temporal information is missing, like the case in Facebook. Even in Flickr, if the event was personal and most or all the photos miss the geoinformation, then the query by geoinformation in the correct window time cannot be applied. Moreover, it is fairly difficult to define each event formally in ontology of four facets. Also in the cases where the user searches for all the media content of an event distributed between friends, then the approach does not perform well due to the low recall. As a conclusion, this approach works at best to find the illustrative media of a public event in a more open social media like Flickr where much of the metadata is preserved, while it does not work fine in detecting the content related to a personal event distributed between friends in less open social networks like Facebook where much of the metadata is stripped out. In the next section we describe a probabilistic approach that exploits a mixture of metadata, social and visual features which are previously explained in Sect. 4.

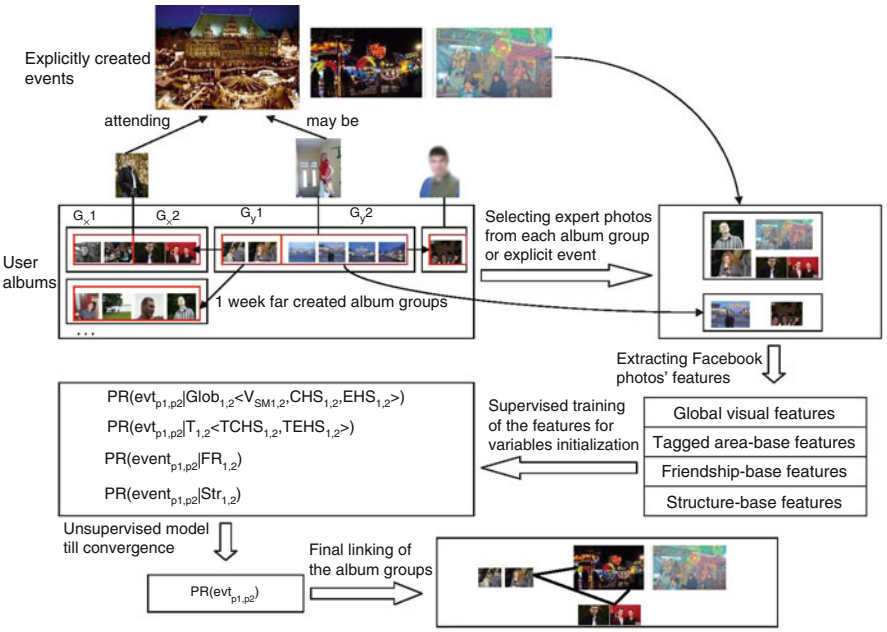


Fig. 8 The steps of the approach introduced by Rabbath et al. [11]

5.2 *Probabilistic Fusion Approach*

In the previous approach, each event is considered to be well defined with at least three facets in ontology as previously explained, and the metadata of the geo-temporal information are the main exploited features, where the visual features are used only in the pruning phase. In this section we introduce a probabilistic model that uses the mixture of features introduced in Sect. 4. Unlike the previous approach, the probabilistic approach performs well even in the social networks where the photo metadata is stripped out like the case in Facebook. The summary of the process is shown by Fig. 8, where the album groups distributed between several users and which belong to the same event are linked together.

5.2.1 Grouping Nearby Created Album Groups

In the work of [11], all the photos within each album are clustered to groups using the mean-shift algorithm based on their upload time. The upload time of the photos within the album in Facebook is obtainable through the graph api, because the original time when the photo was taken is stripped out. The mean shift simply works by calculating

$$m_{h,K} = \frac{\sum_{i=1}^{n} t_i * K \left\| (t - t_i) / h \right\|}{K \left\| (t - t_i) / h \right\|} - t$$

For each timestamp t_i starting from an initial timestamp t using a kernel function K (we use the Gaussian) and bandwidth h, the initial timestamp is then updated as

$$t^{\text{new}} = t^{\text{old}} + m_{h,K}^{\text{old}}$$

The process of updating the mean shift $m_{h,K}$ and the new modes of the time density function t is repeated till convergence. The mean shift can be performed very fast, by adapting the bandwidth h in each iteration for each timestamp t_i data point to be $h = \|t_i - t_{i,k}\|$ where $t_{i,k}$ is the k-nearest neighbour of t_i. The number of clusters is the number of stationary timestamp point t till the convergence. The number of clusters resulting from the mean shift is denoted as Nc. Due to the fact that most of the albums presenting the same event are uploaded in close time window, each album group g_l is linked to other neighbour album groups or photos in the explicitly defined events in both the user account and the friends' accounts. The neighbourhood of a group is defined as the other groups which were created in a time window of one week. This considerably narrows the number of album groups of photos that should be compared.

5.2.2 Selecting Expert Photos

Due to the relatively fewer amount of tagged photos compared to the number of all photos (although the number of tags are growing), not all photos include all features explained in Sect. 4 such as the tagged area-based features. Thus, in the work of [11], only the photos that include tagged areas from each album group are chosen, calling them as *expert photos*. If no expert photo in an album group is found, then a face recognition approach such as that introduced by Stone et al. [22] is applied where the social context is taken into consideration using the already existed tags in the user account and her social contacts as training data. In this case, only photos with recognized faces of more than a high threshold of certainty are taken as expert photos, and the automatic tags are then used for the tagged area-based features.

5.2.3 Extracting the Features of the Pairwise Expert Photos and Performing Offline Supervised Estimation of the Probabilities

The features formally described in Sect. 4 are calculated between the expert photos of each nearby album groups. This process is applied offline when the media is updated with new large amount of contents. The change each of the calculated features between two photos makes to the probability of being from the same event is estimated. The estimation process starts by binning each feature to five clusters using the k-means, where the 0 bin represents the cluster of the highest similarity

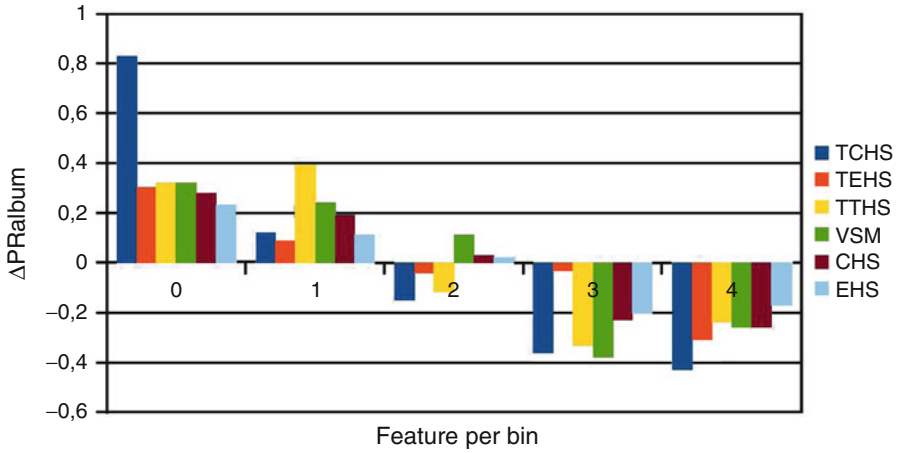


Fig. 9 The ratio change in the probability of being from the same album group (ΔPR_{album}) that each bin of similarity in the global visual and tagged area-based features makes as shown in the work of [11]

values on this feature, and the 4th bin is the one with the lowest values. For example, cluster 0 of feature V_{SM} contains the lowest values (highest similarity) of this feature between two photos. Then the effectiveness of each feature is evaluated in a semi-supervised way as follows:

Global Visual and Tagged Area-Based Features Evaluation: As previously described in Sect. 3, the photos added to explicitly created events are from the same semantic event due to condition 1. Also each album cluster resulting from the first step constitutes one event, and they are better for the case of evaluation because most the photos in Facebook, for example, are added to albums. The work of [11] evaluated how much each bin of these features between two photos changes their probability of being from the same album group (or explicitly created event). This supervised evaluation is performed on around more than 100 k accessible photos in Facebook with around 9 k of them are tagged with at least one person. The photos are accessed through the people who downloaded the SmileBooks application. The photos belong to more than 100 users who downloaded the application and their friendship circle. Figure 9 shows the increased ratio ΔPR_{album} of the probability of being from the same album group per bin per feature. The figure from the work of [11] shows that features of the tagged areas increase/decrease the ratio in a bigger value than global features. For example, bin 0 of the feature $\text{TCHS}_{1,2}$ calculated between two photos 1, 2 increases the probability of them to be in the same album group by more than 80%, while bin 4 decreases it by more than 40%. In some cases, different bins do not make much change, and therefore, they are merged. For example, bins 3, 4 in the feature V_{SM} are merged, as it does not make much difference.

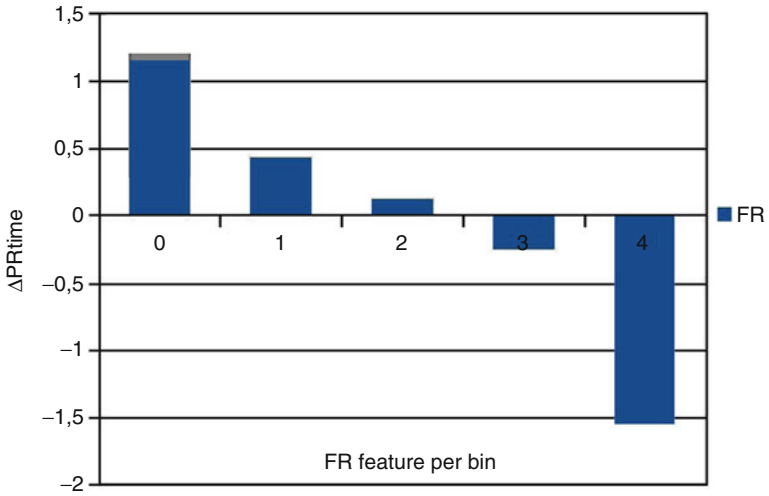


Fig. 10 The ratio of change in the probability of two albums being created in 1 week window of time(ΔPR_{time}) that each bin of similarity in friendship-based feature makes as shown in the work of [11]

Friendship-Based Features Evaluation: The role of the Fr is expected to be very important, but because it is calculated between two groups of album photos or explicitly defined events and not between two photos, it is hard to evaluate using the previous method of intra-album. The work of [11] exploited the observation of [8] where usually photos with the same events are added within 5 days. This feature is evaluated by calculating the increase ratio ΔPR_{time} of the probability of two album groups being created within 1-week window distance from each other. As shown in Fig. 10, bin 0 increases this probability by more than 100%, which indicates the major role of the friendship-based features on linking photos of the same events.

Structure-Based Features Evaluation: The *Str* feature between two album groups is evaluated by calculating the increase ratio of the probability that the creator of one album likes or comments on the album created by the other user. This measure is denoted as ΔPR_{like}. The idea behind this is that usually if two albums in different accounts are connected, it is high likely that the owners of each will like or comment on each other's albums. Although the high similarity in the structure-based features may considerably increase the connectivity between two albums (bin 0 increase ΔPR_{like} by more than 280%), the last three bins representing the low similarity on this feature do not have much negative impact. The reason is that *Str* feature though very helpful but is much less common in Facebook, and the vast majority of the album groups will have low similarity on this feature, even if they belong to the same event. Figure 11 shows the evaluation of the *Str* feature from the work of [11].

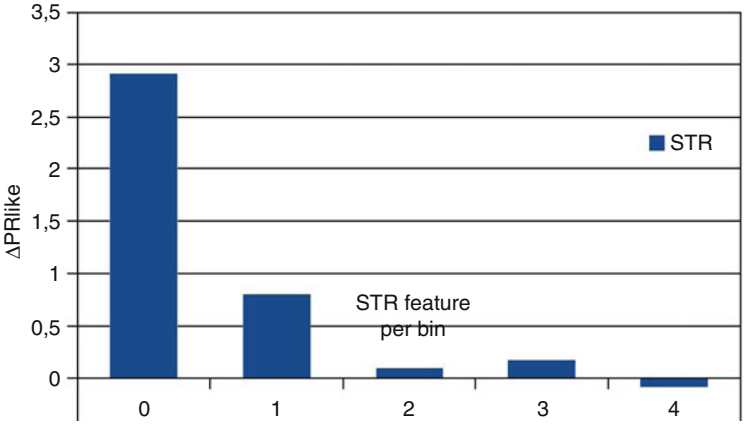


Fig. 11 The ratio of change in the probability of two albums being liked or commented by the creator of each other at least once (ΔPR_{like}) that each bin of similarity in the *Str* feature makes as shown in the work of [11]

Table 2 The obtained look-up table LUT summarizing the initial values of the probability of two photos being from the same event knowing the feature bin

N bin	$PR\,(evt\|T_{\text{CHS}} = N)$	$PR\,(evt\|T_{\text{EHS}} = N)$	$PR\,(evt\|T_{\text{THS}} = N)$
0	$(1+0.83)*pr_{evt}$	$(1+0.3)*pr_{evt}$	$(1+0.37)*pr_{evt}$
1	$(1+0.12)*pr_{evt}$	$(1+0.09)*pr_{evt}$	$(1+0.12)*pr_{evt}$
2	$(1-0.15)*pr_{evt}$	$(1-0.04)*pr_{evt}$	$(1-0.28)*pr_{evt}$
3	$(1-0.36)*pr_{evt}$	$(1-0.31)*pr_{evt}$	
4	$(1-0.43)*pr_{evt}$		
N bin	$PR\,(evt\|V_{sm} = N)$	$PR\,(evt\|\text{CHS} = N)$	$PR\,(evt\|\text{EHS} = N)$
0	$(1+0.32)*pr_{evt}$	$(1+0.28)*pr_{evt}$	$(1+0.23)*pr_{evt}$
1	$(1+0.24)*pr_{evt}$	$(1+0.19)*pr_{evt}$	$(1+0.11)*pr_{evt}$
2	$(1+0.11)*pr_{evt}$	$(1+0.03)*pr_{evt}$	$(1+0.02)*pr_{evt}$
3	$(1-0.29)*pr_{evt}$	$(1-0.25)*pr_{evt}$	$(1-0.18)*pr_{evt}$
N bin	$PR\,(evt\|FR = N)$	$PR\,(evt\|Str = N)$	
0	$(1+1.2)*pr_{evt}$	$(1+2.9)*pr_{evt}$	
1	$(1+0.43)*pr_{evt}$	$(1+0.8)*pr_{evt}$	
2	$(1+0.12)*pr_{evt}$	$(1+0.8)*pr_{evt}$	
3	$(1-0.24)*pr_{evt}$	$(1-0.07)*pr_{evt}$	
4	$(1-1.54)*pr_{evt}$		

The evaluation of each feature helps us in building a supervised look-up table (LUT) to initialize the valued of the model probabilities. The values of the LUT is shown in Table 2. In the work of [11], we assign this random probability of two photos to belong to the same event uniquely to each user as $pr_{evt} = \sum_i \frac{\binom{n_i}{2}}{\binom{N}{2}}$ where n_i is the number of photos in album group i of the user and N is the total

Table 3 Linking the related photos

1. Initialize the posterior probabilities using the look-up table and the previously calculated p_{evt} as explained in the section
2. Apply the **EM**
   E-Step: Use the current values of the probabilities to update 7
   M-Step:
   for $f_l \in F_{ij}$
   foreach bin β 0 **to** N_{bin}
   update $p\,(event|bin\,(f_l) = \beta) = \frac{\sum_{k=1}^{N_\beta} p(event|f_l).f_l=\beta}{N_\beta}$
   update $p\,(event) = \sum p\,(event|f_l) * p\,(f_l)$
3. $RelPhotos = \{\}$
   for $p \in Expert$
   for $E_i \in < E_1, \ldots E_{Ng} >$
   for $p_i \in C_i$
   $Rel_i = getGroup\,(p)$
   if $p\left(event|f_{p,p_i}\right) > 0.5$: add the media contents of the cluster to Rel_i
   if $\|Rel_i\| > 20\% * \|E_i\|$: $Rel = Rel \cup \{Rel_i\}$
6. return Rel

number of photos. As explained in the features evaluation, some of the 5 bins are merged if they do differ much in changing the random probability. In normal cases $pr_{evt} << 1$ except the case where the user has one album group and no other photos at all (not interesting case for event clustering though), but to trust our variables values, technically they are assigned to $Min\,(0.9, PR\,(evt|f))$. The big advantage of the supervised LUT is that it can be directly used by new Facebook applications that do not yet have data and can be updated with the growing contents.

5.2.4 Linking the Photos of the Same Events

The expert photos of each album group or explicitly created event are used to link the content of the same event. The following probability is estimated:

$$PR\left(evt_{p_1,p_2}|F_{1,2} < f_1, \ldots, f_n >\right) = \frac{PR\left(evt_{p_1,p_2}\right) * \prod_{K=1}^{n} PR\left(f_k|evt_{p_1,p_2}\right)}{PR\,(f_1, \ldots, f_n)} \tag{7}$$

where $evt_{p_1,p_2}|$ indicates that both photos p_1 and p_2 belong to the same event and $F_{1,2} < f_1, \ldots, f_n >$ are the explained observed features between the two photos p_1, p_2. The initialization values are read from the LUT where $PR\left(evt_{p_1,p_2}\right)$ is initialized with the value of pr_{evt} and $PR\left(f_k|evt_{p_1,p_2}\right) = \frac{PR\left(evt_{p_1,p_2}|f_k\right) * PR(f_k)}{PR\left(evt_{p_1,p_2}\right)}$ is initialized from the LUT due to the bins of the calculated features. The expectation-maximization is then applied as described in Table 3 to let the posterior probabilities converge to their correct values. The expert photos are then compared against the

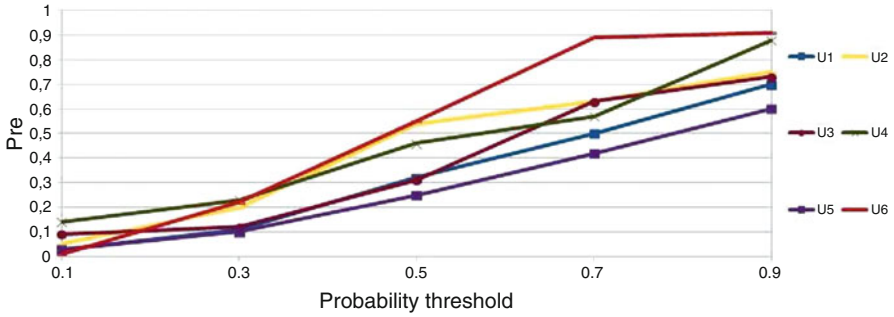


Fig. 12 The precision change with the probability threshold on our evaluation data in the work of [11], where six different datasets were evaluated

expert photos in other album groups and explicitly created events in the friendship circle. As explained in Sects. 5.2.1 and 5.2.2, only the album groups created 1 week far from each other are compared. The threshold to decide that two photos belong to the same event is initially taken as 50%. If at least 20% of the expert photos in a group are marked as related to a photo in another group, then both groups are linked. The whole process is summarized in Table 3, where $getGroup()$ is the album group of photo p, E is the set of all expert photos, N_β is the number of photos with features in bin β and $< E_1, \ldots E_{Ng} >$ is the set of the nearby groups as explained in Sects. 5.2.1 and 5.2.2. The result is a set of all grouped sets where each set represents the content of one event. The work of [11] evaluated the approach over a dataset of several travel events of Facebook users who used a photo book application (the SmileBooks application). Both the precision and recall increased over other unsupervised probabilistic models, such as using the EM model over the Gaussian mixture like the case of [5] where the Gaussian mixture assumption was used for clustering home-made photos or [23] where the Flickr groups were clustered to hidden topics using an EM model. The reason of the considerable improvement is that the posterior probabilities are not randomly initialized and trained with the EM algorithm but after evaluating the effectiveness of each feature in increasing or decreasing the probability of being from the same event. One great advantage of this model is that the number of events should not be determined or estimated in advance, but rather the decision can be done on each two photos if they belong to the same event. In Fig. 12, from the work of [11], the change of precision of the event content detection is shown when changing the threshold probability of two photos being from the same event. Using this observation, the photos of an event can be shown in a "see more" kind of interface as described by Rabbath et al. [11]. The related event photos with a high probability threshold are shown first, and the more "see more" the user clicks, the lower the threshold gets.

5.2.5 Conclusion of This Approach

This approach is especially useful to detect the content of personal event distributed between friends in a social network such as Facebook. Although the photo metadata is missing in Facebook, but a mixture of features can be exploited. Evaluating the effectiveness of each of the visual, tagged area, friendship and structural features makes a big effect in the accuracy of this approach. The look-up table with the initialization values of the posterior probabilities plays a key role in letting the target probability of two photos to belong to the same event converge to its reasonable value. Unlike the previous ontology-based approach, the event should not be defined accurately in the what, where and when facets. The approach can also handle the situations where the metadata is missing. In general, the ontology-based approach can be used in the cases of public events in social media where the event is accurately defined and the metadata such as geo-temporal data is accurate. Because of its high precision but low recall, the ontology-based approach is also suitable in the case where only illustrative subset of the media is looked for, while the probabilistic approach can be used to obtain as much distributed event content as possible, and the threshold of the accepted probability can be configured.

6 Conclusion

In this chapter we introduced the problem of detecting the media content of events in social media such as Flickr and Facebook. We differentiated between the types of events where the event can be created explicitly or its content can be distributed between several friends and accounts. The features that can be exploited in social media to detect the related content of an event are introduced. We described the visual, tagged area, social friendship and structure-based features and studied the effectiveness of each feature in the retrieval process. We introduced two approaches to tackle this problem. The first is the ontology-based approach. In this approach, the metadata such especially geo-temporal data plays the key role in addition to the textual information. The visual features are used to prune the results to remove the noisy and non-related content. This approach is more suitable in the cases where the metadata is available like in Flickr and in well-defined public events such as concerts and big festival where thousands of people add media elements of the event and only illustrative media with high certainty are looked for. The second approach is the probabilistic-based approach which exploits the previously explained feature in a fusion model to group the photos that are likely to be from the same event together. This approach tolerates the missing features such as the geo-temporal information, and therefore more suitable to detect the event content distributed between several friends in a social network such as Facebook, where much of the metadata is stripped out. The probability model also allows controlling the threshold of the probability of two photos belonging to the same

event, such that the user may first retrieve the distributed content of an event with high certainty, and then get the contents with less certainty. In this case, not only the illustrative media elements are retrieved but optimally most of the event content distributed between friends can be retrieved using a small seed of initial set.

References

1. INMAN. [WEBINAR RECORDING] The Photo Economy. http://next.inman.com/2012/04/webinar-recording-the-photo-economy/
2. Trad, M.R., Joly, A., Boujemaa N.: Large scale visual-based event matching. In: ICMR, Trento, p. 53 (2011)
3. Westermann, U., Jain, R.: Toward a common event model for multimedia applications. IEEE Multimed. **14**(1), 19 (2007). doi:http://dx.doi.org/10.1109/MMUL.2007.23
4. Shevade, B., Sundaram, H., Xie, L.: Modeling personal and social network context for event annotation in images. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 127–134. ACM, New York (2007). doi:http://doi.acm.org/10.1145/1255175.1255200
5. Mei, T., Wang, B., Sheng Hua, X., qin Zhou, H., Li, S.: Video booklet: a natural video searching and browsing interface. Multimedia and expo. IEEE Int. Conf. **0**, 1757 (2006). doi:http://doi.ieeecomputersociety.org/10.1109/ICME.2006.262891
6. Platt, J.C., Czerwinski, M., Field, B.A.: Phototoc: automatic clustering for browsing personal photographs. Technical report MSR-TR-2002-17, Microsoft research (2002). citeseer.ist.psu.edu/article/platt02phototoc.html
7. Negoescu, R.A., Gatica-Perez, D.: Analyzing Flickr groups. In: Proceedings of International Conference on Image and Video Retrieval (CIVR), pp. 417–426. ACM, New York (2008). doi:http://doi.acm.org/10.1145/1386352.1386406
8. Liu, X., Troncy, R., Huet, B.: Finding media illustrating events. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11, Trento, pp. 58:1–58:8 (2011). doi:http://doi.acm.org/10.1145/1991996.1992054, http://doi.acm.org/10.1145/1991996.1992054
9. Troncy, R., Malocha, B., Fialho, A.T.S.: Linking events with media. In: Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10, pp. 42:1–42:4. ACM, New York (2010). doi:http://doi.acm.org/10.1145/1839707.1839759, http://doi.acm.org/10.1145/1839707.1839759
10. Gkalelis, N., Mezaris, V., Kompatsiaris, I.: A Joint Content-Event Model for Event-Centric Multimedia Indexing. In: ICSC, Pittsburgh, pp. 79–84. IEEE (2010). http://dblp.uni-trier.de/db/conf/semco/icsc2010.html#GkalelisMK10
11. Rabbath, M., Sandhaus, P., Boll, S.: Analysing Facebook Features to Support Event Detection for Photo-Based Facebook Applications. In: Proceedings of the ACM International Conference on Multimedia Retrieval, ICMR '12, Hong Kong. ACM (2012)
12. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World wide web, WWW '10, pp. 851–860. ACM, New York (2010). doi:10.1145/1772690.1772777, http://doi.acm.org/10.1145/1772690.1772777
13. Rabbath, M., Sandhaus, P., Boll, S.: Automatic creation of photo books from stories in social media. ACM Trans. Multimed. Comput. Commun. Appl. **7S**, 27:1 (2011). http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CB4QFjAA&url=http%3A%2F%2Fdl.acm.org%2Fcitation.cfm%3Fid%3D1878157&ei=iY2BUM2dEMb4sgaN4IGABA&usg=AFQjCNH-G7lRWt_KGulnL8L1TzngS8miCQ"\t"_blank"

14. Pigeau, A., Gelgon, M.: Organizing a personal image collection with statistical model-based ICL clustering on spatio-temporal camera phone meta-data. J. Vis. Commun. Image Represent. **15**(3), 425 (2004). doi:10.1016/j.jvcir.2004.04.002, http://dx.doi.org/10.1016/j.jvcir.2004.04.002
15. Zhao, M., Liu, S.: Automatic person annotation of family photo album. In: Proceedins of International Conference on Image and Video Retrieval (CIVR), Tempe, pp. 163–172 (2006)
16. Kazuya, S., Yujiro, N., Naoko, N., Noboru, B.: Classification based Group Photo Retrieval with Bag of People Features. In: Proceedings of the ACM International Conference on Multimedia Retrieval, ICMR '12, Hong Kong. ACM (2012)
17. Strong, G., Gong, M.: Organizing and browsing photos using different feature vectors and their evaluations. In: Proceedings of the ACM International Conference on Image and Video Retrieval (2009), CIVR '09, Santorini Island, pp. 3:1–3:8. doi:http://doi.acm.org/10.1145/1646396.1646401, http://doi.acm.org/10.1145/1646396.1646401
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91 (2004). doi:10.1023/B:VISI.0000029664.99615.94, http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94
19. Redi, M., Merialdo, B.: Saliency moments for image categorization. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11, Trento, pp. 39:1–39:8 (2011). doi:http://doi.acm.org/10.1145/1991996.1992035, http://doi.acm.org/10.1145/1991996.1992035
20. Viola, P.A., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137 (2004)
21. Gentile, J.E., Bowyer, K.W., Flynn, P.J.: Profile Face Detection: A Subset Multi-Biometric Approach. In: 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, pp. 1–6. IEEE, Piscataway (2008). doi:10.1109/BTAS.2008.4699376, http://dx.doi.org/10.1109/BTAS.2008.4699376
22. Stone, Z., Zickler, T., Darrell, T.: Toward Large-Scale Face Recognition Using Social network Context. Proc. IEEE **98**(8), 1408 (2010)
23. Negoescu, R.A., Gatica-Perez, D.: Analyzing Flickr groups. In: Proceedings of the CIVR, Niagara Falls, pp. 417–426 (2008)

Part II
Tagging of Social Media

Georeferencing in Social Networks

Pascal Kelm, Vanessa Murdock, Sebastian Schmiedeke, Steven Schockaert, Pavel Serdyukov, and Olivier Van Laere

Abstract As technology moves to personalization and mobility, users expect their applications to be location savvy, and relevant to their lives with increasing detail. A person's geographic context includes their current and previous location, the things that surround them, their activity in a given place, as well as their thoughts and feelings in that place. Understanding this context allows us to personalize their experience and refine their interactions with an application, on a hyper-local level. This chapter starts with a discussion of textual modelling of places using Flickr data, following two primary approaches: reliance on a place name gazetteer and with statistical language modelling. We follow with a discussion of the use of the visual content of images in Flickr. This chapter ends with a survey of applications of geographic modelling of this type.

P. Kelm (✉) • S. Schmiedeke
Technische Universität, Berlin, Germany
e-mail: kelm@nue.tu-berlin.de; schmiedeke@nue.tu-berlin.de

V. Murdock
Yahoo! Research, Barcelona, Spain
e-mail: vmurdock@yahoo-inc.com

S. Schockaert
Cardiff University, Cardiff, Wales, UK
e-mail: S.Schockaert@cs.cardiff.ac.uk

P. Serdyukov
Yandex, Moscow, Russia
e-mail: pavser@yandex-team.ru

O. Van Laere
Universiteit Gent, Gent, Belgium
e-mail: Olivier.VanLaere@intec.ugent.be

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_6, © Springer-Verlag London 2013

1 Introduction

Mobile devices are increasingly the primary way people access information on the Web. More and more users are carrying small computing devices with them wherever they go and can access the Internet regardless of their location. Whereas prior to the wide adoption of mobile devices, a user's online life and offline life were partitioned by time and location, now there is less of a division. Users can carry their online friends with them and maintain a constant contact in the form of status updates and uploaded pictures and video, in a running social commentary that blurs the distinction between online and offline interactions. This has lead to a proliferation of data that connects a user's online social network with their offline activities. As a result, the data they create provides a solid link between users, what they are doing offline, and what they are contributing online. The question is how to harness this content.

Because people now have access to information from the Web on the go, their information needs are changing. We propose that the user's location becomes an increasingly important piece of context for understanding their information, need and for interpreting the content they create and upload and share with the community at large. To understand the geographic context of information we need to understand what are the geographic entities in the user's vicinity that may have influenced their interactions with a given application. We leverage the content created by users in a given place to gather the type of information that previously could be obtained only by a geographic surveyor. We consider two primary streams of information: text and visual. The text information on the Web is abundant and relatively easy to access and process. Visual content is also abundant, but it is less obvious how to leverage visual features to understand the geographic context of an image.

Much of the work surveyed in this chapter uses georeferenced images from Flickr.[1] The data is especially valuable because it has textual tags, geographic coordinates (geotags), titles, descriptions, as well as metadata such as the time the image was created, the time it was uploaded, the time each tag was added, and information about the user that uploaded the data. In addition Flickr has a social network, which allows researchers to study the social networking properties of photo sharing. All together, the Flickr data allows people to study the geographic context, the temporal context, and the social context of the user.

Flickr has provided a public API that allows academic researchers and developers to access the data.[2] In addition there are several large public crawls of Flickr data, such as the data provided by the Placing Task, which is part of the MediaEval benchmarking initiative.[3] Another collection of Flickr data was created as part of

[1] www.flickr.com, visited May 2012

[2] http://www.flickr.com/services/api/, visited May 2012

[3] http://www.multimediaeval.org/, visited March 2012

the SAPIR European Project.[4] The CoPhIR test collection[5] consists of 106 million Flickr images, including the complete metadata as well as visual features computed over the entire set. This provides a standard development set for researchers to explore the textual, geographical, social, and visual content of the image data, even if they do not have the computing power at their home institution to process the visual features, which can be computationally expensive. An open source library for computing the visual features is provided by the LIRE Library[6] [38].

Another source of georeferenced data is Twitter.[7] Twitter data consists of a status update, associated with geotags when the user has enabled the capture of geographic coordinates on their device. Users of Twitter 'follow' other users, which is to say they subscribe to the user's stream of status updates. As such the social network is less direct than the explicit reciprocal relationships in the Flickr data because a Twitter user cannot elect to disallow another individual user from following him. Twitter also allows users to 'retweet' the status updates of others, to propagate the information throughout their social network. Because each status update in Twitter is limited to 140 characters, the use of abbreviations is common. In addition, other users are indicated with an '@' symbol preceding their name. Events are frequently identified with hashtags, which is an event name or acronym preceded with a '#' symbol. As a result of the limited length of the status update, URLs are shortened automatically. Twitter data is extremely valuable because people use it to record their thoughts, feelings, and activities throughout the day. In addition, they frequently post links to news articles and images about issues of concern and current events. Unlike Flickr, where the images are typically tagged and uploaded some time later than the time they were created, Twitter status updates are propagated to the user's social network immediately on creation.

Comparing Flickr data and Twitter data in terms of their value specifically for geographic modelling, Twitter differs from Flickr in that a status update in Twitter that is about a location may be about a location other than the user's current location. In Flickr, images about a location are frequently about the user's location at the time the image was created. Twitter contains abbreviations and emoticons and terms in a pseudo-natural language that are frequently only meaningful to the user and his immediate social circle. Flickr tag sets consist largely of nouns describing the content of the image and the context in which it was created. Both data sets are valuable, but they have different properties, and are therefore useful for different purposes.

---

[4]The SAPIR Project was funded by the European Commission under IST FP6, Contract no. 45128.

[5]The CoPhIR data set was built by researchers at the institute of the National Research Council of Italy together with the High Performance Computing and Networked Multimedia Information Systems Laboratories at the Consiglio nazionale Delle Ricerche and the Instituto di Scienza e Tecnologie dell'informazione 'A. Faedo' and is available for download from http://cophir.isti.cnr.it/whatis.html

[6]http://www.semanticmetadata.net/lire/

[7]www.twitter.com, visited March 2012

Other social networks that provide georeferenced user-generated content include Foursquare,[8] Facebook Places (which recently acquired the check-in service Gowalla),[9] and Panoramio.[10] We mention these here for completeness, but we do not study them in this chapter.

One caveat about building systems on user-generated content is that the content itself carries a bias towards users in specific geographic locations and a bias towards users of a socio-economic class that allows them to own devices that enable the content creation. At this writing, a smartphone is still an expensive item owned primarily by the upper middle class and upper classes in the United States and Europe. In addition, the Flickr data contains a large number of images taken by people on vacation, in locations they are not intimately familiar with. This may limit their vocabulary in describing a place and their perception of the place as a whole.

In this chapter, we start with a discussion of textual modelling of places using Flickr data, following two primary approaches: reliance on a place name gazetteer and with statistical language modelling. We follow with a discussion of the use of the visual content of images in Flickr. This chapter ends with a survey of applications of geographic modelling of this type.

2 Textual Location Estimation

Social media enables billions of users to actively participate in the creation of content. Apart from the millions of photos and videos that are uploaded every day, a huge amount of textual data is created in the form of tweets, tags, status updates, news and blog articles, among others. We consider three major categories of textual data, based on the length of a typical message:

- *Articles* such as stories on news websites, blog posts or Wikipedia pages largely correspond to the classical notion of a text document, using full sentences which are structured in paragraphs and sections. Such documents typically discuss one or a few related topics in some level of detail.
- *Microposts* such as Twitter and SMS messages have inherent length restrictions and therefore are mostly limited to a few words. They make wide use of abbreviations and often use short phrases instead of full sentences. Similarly, user comments and Facebook status updates, while not limited in size per se, seldom contain more than a few phrases. In addition to their length, microposts are also characterised by the use of terms that are not found in natural language, including hashtags, emoticons, and (shortened) URLs.

[8] https://foursquare.com/, visited May 2012

[9] http://www.facebook.com, visited May 2012

[10] http://www.panoramio.com/, visited May 2012

– *Tags* are individual terms or keywords which are assigned to a resource, such as a photo on Flickr, a URL on social bookmarking sites such as Delicious, or an artist on Last.fm. Tags can be used to describe the associated resource, either to allow others to find it or to add structure to one's own collections, although tags are used in practice for other purposes as well (e.g. describing actions, such as *toread* in the case of bookmarks).

Textual data often provides cues to its geographic scope, which we can use, for instance, to find out to which user communities a given blog post is likely to be relevant, to find out where a Twitter user is likely located, or to find out where a given photo was taken. There are two basic approaches. The first, and perhaps the most intuitive, is to identify words that are indicative of locations and look them up in a gazetteer. An example of a gazetteer-based system is Yahoo! Placemaker[11] which assigns a geographic context to free text, based on the repository of places, GeoPlanet.[12]

The second basic approach is to estimate a statistical model of the language associated with a place. The idea is to partition the data so as to represent physical geographical boundaries and then associate with each partition all of the textual content associated with the media that was georeferenced in that place. The statistical properties of the language can then be used to predict which partition a particular piece of text came from. This approach essentially turns the problem of place prediction into an information retrieval problem, where the locations are 'documents', the textual metadata associated with each location is analogous to the words in the document, and the text for which a location will be predicted is a 'query'. How the data will be partitioned and how the language models will be estimated vary substantially, depending on the nature of the data. In the following section, we provide an overview of the approaches for the task of assigning a geographic scope to free text, using the three types of data mentioned above. We end this section with an example, comparing the two basic approaches (use of a gazetteer and statistical modelling) described in this section, to locate the mention of a point of interest in a short snippet of text, such as might be associated with an image caption.

2.1 Finding the Geographical Scope of Articles

One of the most natural ideas to map textual data to geographic locations is to identify place names (toponyms) in the text and to look up where the corresponding places are located. For example, a corporate website is likely to contain a contact address, while news stories tend to explicitly mention the places they pertain to.

[11] http://developer.yahoo.com/geo/placemaker/, visited March 2012

[12] http://developer.yahoo.com/geo/geoplanet/, visited March 2012

Once the place names in a text have been identified, *gazetteers* such as Yahoo! GeoPlanet and GeoNames[13] can be used to find the corresponding locations. Such gazetteers typically contain information about administrative place names on Earth (e.g. names of cities, provinces, or countries). Yahoo! GeoPlanet and GeoNames contain information about six and eight million entities, respectively.

One of the most important challenges in this process is to resolve ambiguities. Many place names are also common words in English (e.g. 'Turkey' and 'Nice'). Place names are sometimes used in a nonspatial sense (e.g. 'Brussels' refers to a political entity in a sentence such as *'According to Brussels, the proposed measures have been ineffective'*.). This form of ambiguity can, in principle, be addressed using standard techniques for named entity resolution, although it is a non-trivial problem.

A second form of ambiguity results from the fact that many places on Earth share the same name. GeoNames contains 3,883 entries for *San Antonio*.[14] Such ambiguities are usually handled by (1) preferring locations that are close to other places mentioned in the text and (2) preferring more frequent senses. (All things being equal, *Paris* is more likely to refer to the city in France than to the city in Texas.) An influential example of a system for georeferencing Web pages that follows this strategy is the Web-a-Where system introduced by Amitay et al. [2].

Several techniques improve the accuracy of toponym resolution. For instance, Lieberman et al. [35] look for comma-separated constructions like *Houston, Texas*, which strongly suggest a part of relation between its constituents. Hence, if one of the terms in a comma group is resolved, the remaining terms can be disambiguated more easily. Other techniques include looking for telephone prefixes or postal codes, or identifying the location of the Web server that hosts a page, based on its IP address [40].

Another challenge with using place names is that apart from administrative place names, people frequently use a variety of vernacular place names. The location of these places can often not be found in gazetteers, and they may have inherently vague boundaries. For most cities, for instance, it is not clear precisely where *downtown* is located. Similarly, regions such as *Eastern Europe* or *the Mediterranean* do not have clear-cut boundaries. For this reason, a number of authors have looked at acquiring knowledge about vernacular place names from the Web, in an automated or semiautomated fashion [25, 50, 56]. To cope with the vague nature of region boundaries, these methods represent the spatial extent of a vernacular place name as a probability distribution or a fuzzy set. See Schockaert [49] for a discussion of the similarities and differences between the two representations. Most approaches use some form of social media to collect a set of coordinates that are believed to lie within the region of interest (such as using georeferenced Flickr photos that are tagged with the name of that region) and then estimate a density from the resulting point set.

[13] http://www.geonames.org/, visited March 2012

[14] Visited February 2012

While gazetteer-based methods to georeferencing have been introduced for assigning a geographic scope to Web pages, a number of authors have recently looked at georeferencing social media. In Fink et al. [16], a more or less standard toponym resolution strategy is used to determine the geographic focus of blogs. In their method a named entity tagger finds occurrences of place names. Geographic coordinates are assigned to places occurring in a blog in three steps. First, all toponyms that are sufficiently general are disambiguated directly (e.g. names of continents, countries, capitols, and first-order administrative divisions) as well as those for which only one referent is available in the GeoNames gazetteer. Then the locations of the disambiguated toponyms from the first step are used to disambiguate as many of the other toponyms as possible. Finally, the remaining toponyms are disambiguated by choosing the most populous referent.

Wing and Baldridge [58] propose a supervised learning approach to georeferencing Wikipedia articles. In particular, following Serdyukov et al. [51], they place a grid over the surface of the Earth, and Wikipedia articles with geo-coordinates are associated with their respective grid cells. Both the grid cells and the document to be georeferenced are represented as probability distributions of terms. The probability of a given term in a distribution is proportional to the number of occurrences of that term either in the corresponding document or in the documents associated with the corresponding grid cell. To avoid a zero probability for unseen terms, a form of Good-Turing smoothing is applied to the raw frequency counts. Grid cells are ranked according to the Kullback-Leibler divergence between the document to be georeferenced and the representation of the grid cell. The top ranked grid cell is the predicted location. Their method led to a median prediction error of 11.8 km when estimating the location of Wikipedia articles, without the use of gazetteers to explicitly disambiguate place names. As we will see further, this approach is closer in spirit to methods that have been proposed to georeference tagged resources, such as Flickr photos.

2.2 Finding the Geographical Scope of Microposts

Taken in isolation, microposts are much harder to georeference than full-length texts. To find the location of a tweet, a simple gazetteer look-up is not enough because some inference is needed to understand whether multiple tokens refer to the mention of one place or multiple places. For example, consider the tweet, '9 PM Ginger, Manchester 11 PM'. To a user in Barcelona, it is clear that the tweet mentions two points of interest, as 'Ginger' and 'Manchester' are the names of bars in Barcelona. To people not familiar with Barcelona, it is not clear that Ginger is not a point of interest in the city of Manchester. Ginger could also be another type of named entity, such as a person name. Furthermore, as Twitter lacks the usual syntactic sugar of natural language, there are few, if any, context clues to tell us that '9 PM Ginger' implies 'meet at 9 PM at Ginger'. Specifying 'Manchester, the

bar in Barcelona' would not be necessary because the people for whom the tweet is intended are already aware of the geographical context of the tweet.

For instance, Wing and Baldridge [58] found that when moving from georeferencing Wikipedia pages to georeferencing Twitter messages (tweets), the median error increased from 11.8 to 479 km. Due to their short length, microposts often do not provide enough context for accurately disambiguating place names. Moreover, the abbreviated nature of the phrases in microposts renders techniques such as named entity recognition ineffective. However, microposts are often not posted in isolation. Previous messages from the same user can be exploited as context information. For example, Cheng et al. [7] propose a method to determine the city in which a Twitter user is located (among a pre-selected set of cities). Each city is modelled as a probabilistic language model, which can be used to estimate the probability that a set of tweets was written by a resident of that city. While this baseline model only found the correct city for 10% of the users, substantial improvements were found when using a term selection method to filter out all terms that are not location-relevant, leading to a 49.8% accuracy. Along similar lines, Kinsella et al. [31] train language models over georeferenced Twitter messages and rely on Kullback-Leibler divergence to compare the models of locations with the models of tweets. The results that are reported show that around 65% of the tweets can thus be located within the correct city (among a pre-selected set of cities) and around 20% even within the correct neighbourhood, from among 502 neighbourhoods in New York City identified by GeoPlanet (when restricting to tweets within New York City). To assess the effectiveness of a gazetteer-based method, it was found that passing tweets to Yahoo! Placemaker only classifies 1.5% of the tweets within the correct neighbourhood. This provides further support for the hypothesis that gazetteer-based methods are generally ineffective in dealing with microposts.

The profile information provided by users also can improve the georeferencing of microposts, although the information provided in the profiles is often incomplete or missing entirely. Interestingly, missing values of this field can often be accurately estimated by looking at the location fields of people in the social network of the user. For example, Backstrom et al. [3] show that for users with at least 16 friends whose location field is available, the correct location can be estimated within 25 miles in 69.1% of the cases, as opposed to 57.2% when using IP address georeferencing.

2.3 Finding the Geographical Scope of Tagged Resources

The short length of microposts poses difficulties for traditional gazetteer-based georeferencing methods; however, this is even more true for tagged resources, such as Flickr photos. Since tag sets lack a specific grammar, any form of linguistic processing (such as named entity recognition) is impossible. On the other hand, due to the descriptive nature of tags, the collection of all georeferenced Flickr resources provides a potentially invaluable source of geographic information. As

already mentioned, Flickr photos have been used to find the spatial extents of vernacular regions. Some authors have looked at the related problem of choosing the best term to describe a given region, a choice which is strongly affected by the geographic scale. Indeed, depending on the chosen scale, the same location may best be described by *France*, *Paris*, or *Eiffel tower*. Rattenbury and Naaman [45] use burst analysis techniques to find terms that occur unusually often in a given region at a given scale. In the case of the tag *Paris*, a burst of occurrences will be witnessed at world scale, but not within the city of Paris itself.

Several authors have looked at the problem of estimating the location of a Flickr photo; the Placing Task, part of the MediaEval benchmarking initiative, was proposed to encourage research in this direction.[15] Serdyukov et al. [51] propose to train a probabilistic language model for each cell of a grid-representation of the Earth, where each cell is represented by a language model derived from textual tags associated with images that were geotagged within the cell boundaries. When placing a new image, they assign a photo to the cell whose model is most likely to have generated its tags. Particular emphasis is given to the influence of smoothing, showing that spatially aware forms of smoothing may lead to small, but statistically significant improvements.

In follow-on work, O'Hare and Murdock [42] demonstrate significant improvements by estimating the term distribution with the user frequency rather than the term frequency, where the user frequency is calculated as the number of users applying a given tag in a cell, normalised by all of the user frequencies in a given cell. This solves the problem of near-duplicate tag sets, which occur when a user applies the same basic set of tags to all of the images they upload, with one or two distinguishing tags. In this scenario, the term distribution will be biased towards the tags applied by a single user. When multiple users apply the same tag to a given place, intuitively we would expect the tag to be more representative of that place.

Along similar lines, Van Laere et al. [57] propose a two-step approach to find the location of a Flickr photo. In the first step, language models are used to find the area which is most likely to contain the location where a given photo was taken, although a k-medoids clustering is used instead of a grid. In the second step, similarity search is used to find the photo from the training set that is most similar to the photo to be georeferenced, among all photos that are known to be located within the area that were selected in the first step. To assess the similarity of two photos, their tag sets are compared using the Jaccard measure. The results show that neither of these two steps alone are sufficient for accurate georeferencing. Intuitively, the first step is needed as a form of implicit disambiguation of ambiguous tags, while the second step is needed to escape from the limited granularity of classification-based approaches.

Crandall et al. [10] combine textual, visual, and temporal features to georeference Flickr photos. First, mean-shift clustering is used to cluster the locations of the photos in the training set, yielding a set of highly photographed places. Subsequently, linear support vector machines (SVMs) are trained for each of these places. Their

[15] http://www.multimediaeval.org/mediaeval2011/placing2011/

results show that, depending on the considered scale, combining visual and textual features results in a significant improvement over using only textual features.

As in the case of microposts, using an effective technique to select spatially relevant terms can substantially improve the results. While standard methods such as χ^2 feature selection or selecting the most frequently occurrent tags are sometimes used, these methods are outperformed using a term selection method proposed by Hauff et al. [23]. Their *geographic spread* term selection is based on the idea that spatially relevant tags are those tags that occur only around a few clusters of locations, while still favouring tags that occur often.

Training location-specific language models from georeferenced Flickr photos have uses beyond the task of georeferencing other Flickr resources. For example, De Rouck et al. [14] show how Wikipedia pages about places can be georeferenced using language models trained on Flickr photos. They found that for 15.4% of the pages, coordinates were found within 1 km of their true location, as opposed to 4.2% in the case of Yahoo! Placemaker. This suggests that implicit geographic information can indeed be extracted from the social Web, in this case Flickr data, which can be used to complement or even replace gazetteers. As another example of the use of location-specific language models, several authors have looked at extending topic models [5] with a spatial component. For example, Sizov et al. in [54] propose such a method, called GeoFolk, and show its benefits on tasks such as tag recommendation. Yin et al. [59] propose a geographically informed topic model to analyse the geographic influence on the popularity of different topics. Somewhat related, Eisenstein et al. [15] use topic models trained from georeferenced Twitter messages to analyse lexical variation across different parts of the US.

2.4 An Example: Identifying the Location of a Point of Interest

We apply several of the approaches described above to show a comparison of the benefits and drawbacks of each approach, using as an example the task of determining the location of a point of interest (POI). A point of interest is a geographical entity, such as a landmark, business, school, or public building. They include such places as airports, university campuses, and parks and thus may represent a wide geographical area; however, they are typically smaller than a neighbourhood. They are difficult to localise because the mentions of points of interest vary greatly, depending on the text, for example, the official name of the University at Buffalo is ‘The University at Buffalo, The State University of New York’, but is commonly referred to as ‘UB’ or it may be ‘#UBuffalo’ in Twitter. Many of them are ephemeral such as restaurants which may open and go out of business within a year, while others are permanent, but have regular name changes, such as the ‘Hollywood and Highland Center’ which was formerly known as the ‘Kodak Theater’. Points of interest are often mentioned in image captions or news articles, where their location is implied by the geographic context of the surrounding

text. Knowing their geographic coordinates allows us to place them on a map and determine other geographic entities that are near them.

We take the point of interest in isolation, separate from its surround textual context. This solves the problem of identifying the boundaries of the mention. For example, in the tweet '10 PM Ginger, Manchester 11 PM', we take as given that we can identify 'Ginger' as one POI and 'Manchester' as another. We abstract the problem in this way to focus on the georeferencing, rather than on the entity extraction, but we acknowledge that the entity extraction problem is non-trivial. Also, in abstracting the problem in this way, we potentially hobble the gazetteer-based approach because we remove some of the linguistic context that may have helped the system identify the place.

The points of interest in this example are listed in the Appendix of O'Hare and Murdock [42]. They originally appeared in the 'location' field in the metadata associated with a sample of images from the Getty stock agency.[16] There are 74 unique mentions of points of interest, although several of the POIs may refer to the same location. It is also important to note that the POI mentions may contain other locations that are not directly related to the POI itself.

Table 1 shows the result of a gazetteer approach in the row labelled 'gazetteer' in median metres from the true location. This result is obtained by querying the Placemaker system with the point of interest and taking the centroid of the most likely location returned by Placemaker as the predicted location. When the point of interest is listed in the gazetteer, the results are extremely accurate: In this example, the Placemaker service is a median distance of 230 m from the true location for the points of interest it can localise, but it cannot localise every POI. Some POIs are not listed in the gazetteer that Placemaker searches, or Placemaker does not recognise them when they are mentioned in a non-standard way.

The language modelling approach described in O'Hare and Murdock [42] is presented in the row labelled 'LM: cell'. The data is quantised along latitude/longitude lines, rather than clustered. The language modelling approach described in Van Laere et al. [57] is presented in the row labelled 'LM: cluster'. The differences may be accounted for by the size of the cells. In O'Hare and Murdock, the cells are one-kilometre square at the equator (and smaller towards the poles). In Van Laere et al., the dimensions of the clusters are determined by the data and may be larger than one kilometre. Furthermore, the term estimates in O'Hare and Murdock are estimated by the user frequency, rather than the term frequency as in Van Laere et al., and this may account for its better performance.

An advantage of the approach of Van Laere et al. is that the system estimates the location boundaries, which is useful for determining the locations and extents of neighbourhoods and vague or colloquial regions which are not mentioned in any gazetteer. Applying the geographic coordinates to the most similar photo within the region shows a substantial improvement (the row labelled 'LM + SIM'). This is particularly effective for points of interest because well-known landmarks are well-represented in the Flickr data. Interestingly, applying the geographic coordinates

[16] www.gettyimages.com, visited March 2012

Table 1 Median distance (in metres) of predicted geographic coordinates from the true location of points of interest from news images

	Median distance (in metres)
Gazetteer	1,752
LM: cell	469
LM (cell) + gazetteer	322
LM: cluster	1,903
SIM	718
LM (cluster) + SIM	457

of the most similar photo directly does not perform well on its own (the row labelled 'SIM'), but using the language model to restrict the choices of similar image improves the results over either approach independently.

As the gazetteer approach is very precise for points of interest that it can localise, it makes sense to use that information when it is available and falls back to the language modelling approach when the gazetteer fails. The result of this is in the row labelled 'LM + gazetteer'. In this result, the language modelling approach used is the system shown in the row labelled 'LM: cell'. This shows the clear benefits of combining the knowledge encoded in a gazetteer approach, with the statistical properties of language modelling. Both language modelling approaches demonstrate that although the data that the model is built upon consists of Flickr tag sets, they can be applied effectively to other types of data that better represent natural language.

The systems described above rely on textual metadata from images, which may not always be available. The next section discusses the use of visual information for georeferencing images. Whether textual metadata is available or not, the visual information may provide additional information about the content and the context of the image.

3 Visual Location Estimation

Visual georeferencing of media items such as photographs or video recordings is an important problem since not all of them have user-contributed tags, and georeferencing can provide coarse estimations of concepts in the media, which may be suitable for certain applications (e.g. travel recommendation). Georeferencing is also useful for other vision tasks by narrowing down the possibilities for further processing. For example, object recognition tasks, such as the detection of landmarks in images, can be made more efficient (and possibly more effective) by restricting the candidate landmarks to those that are located within the most likely geographic area of the image. This section gives an overview of the different approaches for estimating locations based on visual similarity and recognition of landmarks.

3.1 Georeferencing via Scene Matching

The estimation of geographic locations from visual content is related to the field of content-based image retrieval (CBIR). In CBIR visually similar media items are returned for a given query image. CBIR and scene matching systems are put in place to retrieve images showing variety of views of a particular scene. The composition of the image is typically determined by the type of scene, such as whether it is a shot of a city or a nature shot. CBIR and scene matching overlap considerably, and scene matching can be considered a first step for visual georeferencing.

The basic technique for georeferencing media using their visual content is as follows. Imagine a query image depicting a Mediterranean coastal scene for which the geo-coordinates are not known. A CBIR system retrieves images that capture the gist of the query image and returns similar georeferenced images of Mediterranean costal scenes, which can be used to predict the location of the query image. Most of the approaches described below adopt this approach, based on the assumption that similar scenes in images (or video sequences) correlate to similar geographic areas.

The easiest way to identify visually similar media is to compare their colour space. Torralba et al. [55] demonstrate that tiny images contain enough information to recognise the scene and even objects within the scene. The *tiny image descriptor* represents the colour information of each pixel as feature vector; typically the RGB colour space is used, but other more perceptual-adapted colour spaces, like the CIE L*a*b* [9], might also be employed. Torralba et al. propose the use of 32×32 pixel images, but smaller versions (16×16) are also used [19, 24]. This descriptor can be compared by Euclidean distance [24] or by other metrics such as cosine similarity [19].

Colour histograms are commonly used in image retrieval, in which the colour distribution is presented in quantised colour component bins. As with the tiny image descriptor, several colour spaces are employed. More sophisticated geo-related approaches consistently choose the CIE L*a*b* space, which was designed to approximate human vision. The distances in L*a*b* colour space correlate with the perception of changes in colour. Histogram functions such as L1-norm or χ^2 distance [19, 24] are used as the distance (or similarity) function.

Edge and texture histograms are another class of commonly applied descriptor in scene matching. Edge features help distinguish between natural scenes and scenes containing man-made structures, while texture features might help to discriminate properties such as different terrain types. Hays and Efros [24] noticed that man-made structures in scenes contain a higher frequency of straight lines, while natural scenes contain more undirected lines. These properties are encoded in either histograms based on edge angles or texton histograms. The term ‘texton’ was introduced by Julesz [27] as a unit of preattentive human texture perception. Malik et al. [39] define textons as cluster of responses of oriented Gaussian derivative filters. So the texton dictionary is determined by clustering, for example, by the centroids of a k-means algorithm. The size of the dictionary (and therefore the histogram) differs in recent georeferencing work (120 in [19], 512 in [24]). While

edge histograms are generated using quantised angles, texton histograms are created by mapping each pixel to the nearest texton according to its filter response.

The *gist descriptor* was proposed for retrieving structurally similar images by Oliva and Torralba [43]. This descriptor was inspired by the ability of humans to recognise scenes rapidly. This low-level descriptor does not carry high-level semantics, but structurally similar images can be interpreted as semantically similar if their distance is low enough. This descriptor has shown good performance in scene classification and is used in different variations for georeferencing as a derivative of scene matching. It encodes the structural information of the image by superimposing a square grid and extracting the filter response for each tile separately. Recent work in georeferencing [19, 24] uses a filter response at six orientations and four scales averaged for each 5×5 tile. This descriptor can also be enriched with colour information, as with the tiny image descriptor, but for each tile. Euclidean distance is employed as matching metric. Results based on gist descriptor will only be reliable if the matching is applied on a large data set; according to Torralba et al. [55], this will only work when the neighbourhood of a query image is densely covered (more than 100,000 media items).

These proposed features have been shown to be effective for scene matching [43] and [55]. Scene matching is relevant for georeferencing because certain scene characteristics correlate with locations. For example, a photo depicting the Mediterranean coast could have originated from many places in the world, but many candidate locations can be excluded, such as midland areas or Arctic regions.

Georeferencing by matching media items may not accurately predict the location, but it can provide an estimation which reduces the number of candidates that must be searched. The simplest approach looks for the nearest neighbour of an item by comparing relative low-dimensional feature vectors, which is faster than performing a sophisticated object recognition algorithm that has to compute over many object models. The geographic estimation using the nearest neighbour to a queried item requires a large data set that covers the entire world. At the beginning of 2012, the image and video hosting website Flickr[17] hosts almost 170 million georeferenced items under a Creative Commons licence. Since these resources are publicly available, the primary obstacle to use them is the required computational power to describe these media items and to match them against new items.

There is a bias in the Flickr images towards North American and Europe and popular tourist destinations due to the large number of photos taken in these areas. This limits the ability of approaches based on visual similarity to georeference locations that look very different. A result of this bias is that images from regions that are not well covered in the Flickr data will be mistakenly georeferenced to visually similar locations that happen to have better coverage in the data. Another limitation of this approach is that many regions look very similar and are visually hard to distinguish. For these reasons, image content is usually used as one feature in a georeferencing system, but not as a stand-alone approach.

[17] www.flickr.com, visited March 2012

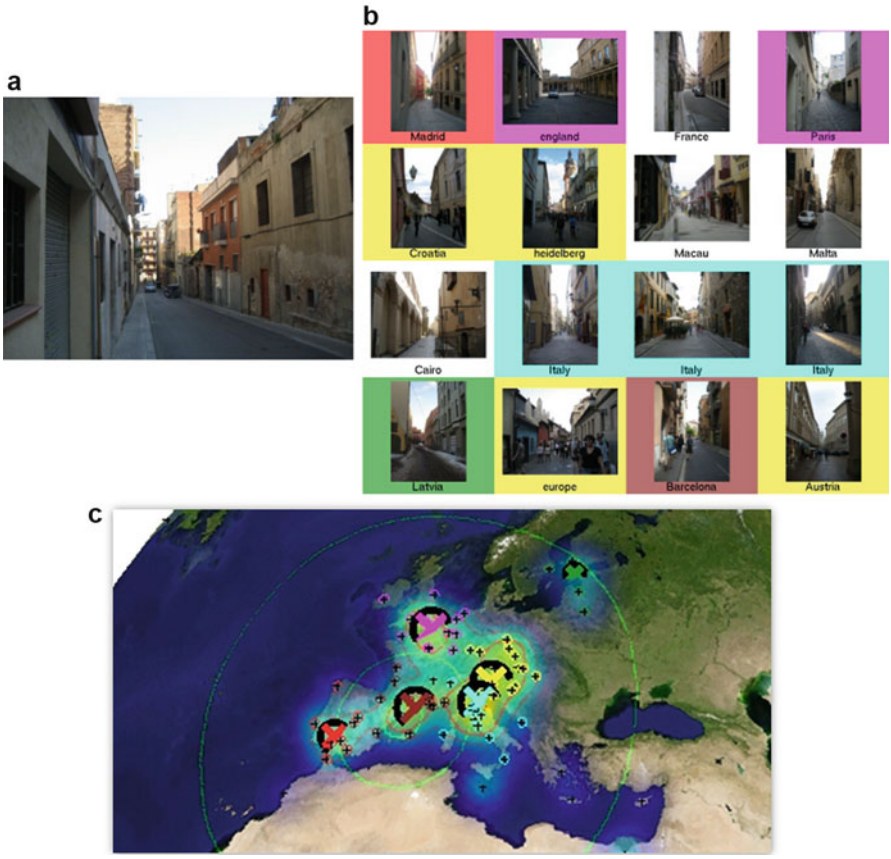


Fig. 1 Illustration of geolocation estimation by Hays and Efros [24]: given a query image (**a**), a set of georeferenced nearest neighbours (**b**) is returned, which are mapped on the globe. The different geographical clusters of neighbours are shown in different colours. The resulting probability map (**c**) is shown as overlay on the world map

The simplest approach described above applies geographic coordinates to photos by taking the geographic information from their nearest neighbour. This strategy is quite effective, but it is not robust against outliers in the database. A more advanced approach by Hays and Efros [24] takes a *set* of visually similar neighbours which implicitly forms a probability map of likely geographic areas. If the query image has enough distinguishing visual characteristics, the majority of neighbours will lie in a dense area that corresponds to the true location. Figure 1 shows the geolocation probability map for a query image depicting a typical street scene in Barcelona. Although the query image lacks visual characteristics, its location is estimated to be in Western Europe. The areas formed by the set of nearest neighbours are determined by clustering, with the biggest cluster centroid predicting the true location. Hays and Efros use kernel-based mean-shift clustering to identify the most

likely geographic area, but simpler distortion-based clustering has been shown to be effective [30]. A similar procedure is applied to georeference video sequences. Keyframes are extracted from the video, and then each keyframe is processed as the images described above. If the video was shot at a single location, sequences with distinctive visual content will yield a set of neighbours in a dense area. If the video was shot at multiple locations, multiple dense area are detected, projected by their nearest neighbours. Much of the video data publicly available is shot at a single location; therefore, the entire sequence can be successfully matched by selecting the nearest neighbour that has the smallest mean distance to all frames, respectively, as reported in [29].

The approach by Hays and Efros [24] yields 16% correctly assigned photos within a radius of 200 km. Accuracy is not expected to improve much with a much larger database, as the accuracy is limited by the huge portion of generic-looking images that lack distinctive visual characteristics such as landmarks. Even for humans it would be hard to determine the location of a generic coastal scene in the absence of additional context. Gallagher et al. [19] use the image tags as additional information, in an approach similar to that of Hays and Efros. Although their approach to tag similarity is simple, they demonstrate that using both visual information and textual tags improves over either visual or textual information alone. In accordance with Gallagher et al., Kelm et al. [29] also use tags to achieve a more precise georeferencing of video sequences. A common geographical area boundary is determined for all tags of a video recording by querying the GeoNames.org gazetteer. Within this boundary the image with the smallest mean distance to all keyframes is returned together with its coordinates. This extension yields 69% of the videos correctly assigned within a radius of 200 km, in contrast to 20% correct decisions achieved with the visually nearest neighbours only. A purely data-driven extension to Hays' approach was proposed by Kalogerakis et al. [28]. They note that many geo-related photos are shot during vacations, and multiple images represent the same location, often a well-known tourist destination. Consequently, a person's photos can be assigned to the same location if they are shot within a narrow time window. Thus, otherwise ambiguous photos can be geo-located by their temporal association to other images by the same photographer. For example, if a set of photos by a single photographer shows a Mediterranean coastal scene, the assignment will be ambiguous because Mediterranean beaches look very similar to each other and to other beaches. If the same series of images show a certain landmark (such as the Rock of Gibraltar), it will be clear to people where this entire set of images was taken. Kalogerakis et al. [28] show improved precision by superimposing each single image's probability map. Even photos with few visual cues can be georeferenced, as long as they were taken together with at least one unambiguous image, such as a photo depicting a landmark. The primary contribution of their work is the modelling of the probabilities of geographical distances between images taken in the same temporal window.

3.2 Georeferencing via Landmark Recognition

The recognition of a landmark or a well-known building is usually easy for a person. It is clear that the Statue of Liberty is in New York and the Taj Mahal in Agra. Even when the image only contains some cultural features, building materials, and a typical architecture, it is possible to locate the image. In addition to the scene matching approaches, landmark recognition approaches aim to recognise the precise place depicted in a query image using a database of images annotated with geolocation information. Landmark recognition is a challenging task, as the query image and images available in the database might show the same place depicted at a different scale, from different viewpoints or under different illumination conditions. For this reason landmark recognition requires a large data set of photographs of objects, such as buildings and works of art to extract key points. The data set should contain no irrelevant information, such as indoor party pictures or photographs of pets.

Quack et al. [44] present an unsupervised labelling approach for objects to collect this type of high-quality data set. They crawl community photo collections to cluster the content referring to an object or event based on visual similarity. They label the clusters with the metadata from the photographs and crawl-related content. The recognition of an object photographed from multiple viewpoints is improved by having a large number of images in each cluster. The system makes use of *Frequent Itemset Mining* [22] in the textual information in order to assign labels to the clusters. Furthermore, the text query was submitted to Wikipedia for assigning articles and associated images to the clusters. The images found in the Wikipedia articles are used for verification of the cluster link. One limitation of the approach of Quack et al. is a possibly insufficient number of images for a given topic. Gammeter et al. [20] modify this approach by inverting the process. Here, each location in a Wikipedia article is used to find Flickr photographs related to this point.

Another approach uses map-based collections of street side images provided by Google Street View or Microsoft StreetSide to enable image-based place recognition, as in Yu et al. [60] who use a data set of 300,000 street view images of Manhattan in New York. In the street view data, each geographical location contains six surrounding views separated by 45°. This information is used for a mobile search application and offers a method for georeferencing independent of GPS or network-based localisation. Estimation of the camera location is related to a 3D reconstruction of a city, as in Agarwal et al. [1]. The recognised location is often more precise in terms that viewing directions are included and neither network infrastructure nor line of sight to satellites is needed.

To solve the problem of matching images with a different scale or a different viewpoint, local features like scale-invariant feature transform (SIFT) [36] or speeded up robust features (SURF) [4] are used. These algorithms provide key points for regions of interest. For any landmark in a photograph, interesting points on the object can be extracted to provide a feature description of the landmark. This description is stable across a range of image transformations such as scaling, rotation, and perspective distortion.

One of the first benchmarks in landmark recognition was the 'Where am I?' task at the ICCV 2005. The goal was to find a corresponding image with GPS information in the training set. For this contest Zhang and Kosecka [61] built a system for image-based localisation in an urban environment using three steps: coarse location recognition, camera motion estimation, and position triangulation. Coarse location recognition uses SIFT features to select the closest views in the training database. To eliminate the outliers, a camera motion model is estimated between the query view and the reference views, assuming that there is sufficient overlap between the views to calculate the top five closest views. The final result is specified by the triangulation based on planar homographies of the queried views.

In an extension of this work, Li et al. [34] use a large set of 30 million georeferenced Flickr images, of which two million were labelled into one of 500 most popular landmarks on Flickr. For these landmarks a multiclass support vector machine of vector-quantised SIFT features is trained. The most severe limitation of this kind of approach is the huge number of training images required, which is problematic because a non-linear SVM scales between $O(N^2)$ and $O(N^3)$ (where N is the number of training images). Linear SVMs are more efficient because the training cost is $O(N)$.

Shrivastava et al. [53] extend the work of Hays et al. [24] to estimate a location using visual similarity for matching images across visually different domains, such as hand-drawn sketches and photographs in different seasons. The idea behind their approach is to find a good visual similarity function that exhibits high 'uniqueness' for features which discriminate the image (the positive sample) against the rest of the data (the negative samples). Each image is represented as a grid of Histogram of Oriented Gradients (HoG) [12] features. The system adds extra positive data points by appending images with the same object, thus generating a sample that is robust to small errors due to image misalignments. The negative examples are millions of sub-images extracted by 10,000 random Flickr photographs. An SVM classifier learns the best weighting for the visual similarity function. A sliding window with HoG features is applied to evaluate the sub-windows in the image. This approach shows good results for matching paintings or sketches to real photographs, but the computation is too expensive for practical applications such as *Paintings2GPS* [53].

Vajda et al. [26] adopt a graph-based duplicate object detection system for the propagation of georeference using a spatial neighbourhood graph for a given object. The graph is based on the connections between the neighbours of a key point in the query image and the corresponding feature neighbours in the model. To avoid erroneous connections, only those neighbours at a distance similar to the object size will be connected. This approach is robust for 3D landmark recognition even if only a few images are used for training.

One challenging problem in landmark recognition is that objects are frequently occluded by other objects such as cars, trees, and road markings. Knopp et al. [32] resolved this problem assuming that an image of a particular place will not match other images at faraway locations. The system automatically detects spatially localised groups of meaningful features and regions with less distinguishing information. The approach applies quantised SURF from a database of 2,942

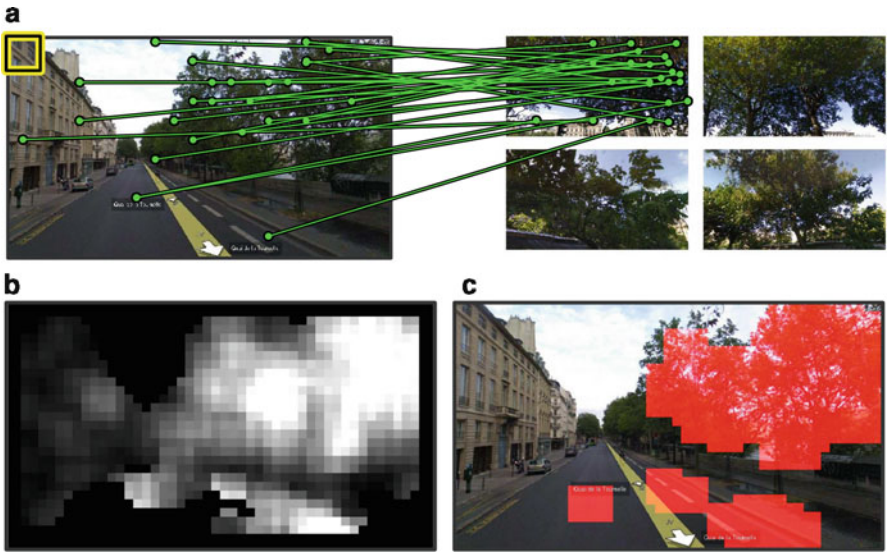


Fig. 2 Detection of spatially localised groups of features without meaningful characteristics: (**a**) key point matching of similar images with different locations. (**b**) Similarity score: the brightness corresponds to the similarity to image segments of faraway places. (**c**) Masking of regions with less distinguishing information (*red* areas) [32]

georeferenced images with approximately six million features. This location-specific obfuscation process is shown in Fig. 2.

4 Applications

In previous sections we discussed approaches to georeferencing based on textual data (or metadata) and visual characteristics of images. In this section we survey applications of georeferencing technologies. The ability to determine the places to which an image or a piece of text refer allows us insight into a user's geographic context: where the user is, what is around him, and what he may be doing in a given location.

Understanding geographic context is critical for applications such as local search, advertising, and delivering content through applications on a mobile device. Most companies that develop these applications rely on licensed data from companies such as Navteq.[18] The data is expensive to license and maintain because it is collected by human surveyors who travel to a location and document the geographic entities in that place. Furthermore, because of its reliance on human geographers,

[18] www.navteq.com

the types of points of interest they focus on are more permanent structures such as public buildings, landmarks, stadiums, and parks. More ephemeral places, such as restaurants and businesses, are not as well covered. Social media data has the potential to supplement the licensed data to improve its coverage and reduce its bias. Furthermore, people often use colloquial names for places they are familiar with, and these names are present in the content they upload and share with others, while the colloquial names may not be present in licensed data derived from official sources.

The opportunity to browse personal photo sets, tweet streams, or place visits over geographic locations is important for users to better keep track of their personal memories. However, there are only a few attempts to build useful applications using the huge amounts of personal travel histories accumulated at popular social networks and easily accessible via their public APIs. This section overviews several major directions of research aiming at prototyping and testing such applications. While local search and targeted advertising are important applications of geographic information, they are well understood and well studied. In this section, we focus on other types of initiatives.

4.1 Recommending Links in Social Networks

Predicting and recommending the connections in social networks is a well-known fundamental problem. As it was shown in a number of studies, the knowledge about users' locations has some potential to improve link prediction algorithms and to help measure the strength of ties between users more precisely.

Crandall et al. [11] assumed that two people A and B 'co-occurred', if both A and B took georeferenced photos within the bounds of a certain region and within a certain time period. They further tried to find a correlation between the co-occurrence itself and the existence of a social tie between A and B. They discovered that two people have almost a 60% chance of having a social tie on Flickr when they have five co-occurrences at a temporal range of a day in distinct Earth grid cells of side length equal to one latitude-longitude degree (about 80 km on a side at the mid-latitudes). In other words, their research suggested that when two people exhibit multiple spatio-temporal co-occurrences, this is a strong indicator of a social tie. Still, only 0.1% of the friendships met such conditions that allowed them to be predicted with such confidence based only on co-occurrences.

Sadilek et al. [46] suggested an approach that is able to predict over 90% of friendships in Twitter with confidence above 80%. Based on the fair assumption that a significant share of friendships cannot be explained by location alone, they demonstrate that locations can be augmented with text and structural features to infer social ties with high accuracy. At the same time, locations played a major role in discovering 'strong ties' as it was noticed that very close friends (with at least eight common friends) spend most of their time in a relatively small area.

Zhuang et al. [62] presented a kernel-based learning to rank framework for inferring the strength of social ties between Flickr users, which involves a kernel target alignment algorithm to integrate the heterogeneous data into a holistic similarity space and a pairwise learning to rank approach to estimate the social strength. One of the kernels was built using the number of locations where two given users took photos (the other kernels were built using the number of mutual friends, textual similarity of the uploaded photo's descriptions, their visual similarity, and the numbers of commonly favourited photos and groups). Although the location-based kernel was not the most useful for the task of top-k friend recommendations, it outperformed the kernels based on other features, such as the visual similarity of uploaded photos, the number of common contacts, and the number of common favourited photos.

In location-based social systems such as Foursquare or Gowalla, co-occurrence in the same location appears to be a much stronger predictor of social ties, than in networks focusing on content sharing, such as Flickr and Twitter. Scellato et al. [48] found that in Gowalla, in particular, while about 50% of new links appear among friends-of-friends, more than 30% of new friends are added among place-friends (those users that are not connected to each other but have checked in in at least one common place). By using only location-based features calculated for each pair of users and a classifier based on the random forests algorithm, they were able to predict up to 72% of friendship relationships with nearly 100% precision. Chang and Sun [6] also examined the correlation between check-in data and friendship at Facebook Places. They found that each additional check-in by two users to the same place increases the likelihood that they are friends by approximately 2.3%. Their logistic regression classifier with features characterising users' co-locations was able to predict a friendship relationship for any two users with 72% accuracy.

4.2 Travel Recommendation

The ability of social network applications to trace users' locations via georeferenced user-generated content led to a number of proposals to build place (landmark, trip, route) recommenders. Typically, such applications assume that the collection of each user's georeferenced objects (photos, videos, or tweets) can be viewed as a sequence of visited locations at specific times. They often start from assigning georeferenced objects to so-called destinations (landmarks, POIs) either in an unsupervised way, by making spatial clusters of georeferenced objects [8, 33, 37], or in a supervised way, by assigning georeferenced objects to the known POIs (by matching their textual descriptions and locations) [13, 52]. In addition, they estimate the average stay time at each such destination and the average travel time between destinations using timestamps of georeferenced objects. Relying on such travel patterns mined from user location histories, they propose different ways to assist tourists in exploring geographic regions.

Kurashima et al. [33] assumed that the user's preference for visiting each next destination in a city can be estimated by analysing travel behaviour of similar users. First, they inferred the probability of a user's general interest to the destination under study by means of a collaborative filtering method based on probabilistic latent semantic analysis. Second, they inferred the probability for an average user staying at the current destination to visit the next destination under study. Finally, they used a simple greedy search algorithm to find the most probable route given the above-mentioned probabilities of visiting destinations and the amount of time available for travelling.

Lu et al. [37] also regarded trip planning as a problem of finding the optimal path within a given time, but the user's preference for a destination was estimated differently. It was represented as a linear combination of the destination's all-time popularity among all users, its popularity during the current seasons and the similarity of its description to the description explicitly given by the user. The probability of transition between two destinations was also estimated as the probability of such transition for an average user. The problem of finding the optimal route was solved using a dynamic programming algorithm. De Choudhury et al. [13] formulated the problem of finding popular travel itineraries of Flickr users in a very similar way, but used a quasi-polynomial recursive greedy approximation algorithm for orienteering to find the optimal solution. Their algorithm is parameterised by the number of POIs the user is willing to visit in a day. They suggested filtering out city residents and focusing specifically on the travel patterns of tourists (those Flickr users who stayed in the city no more than 21 days and visited at least two POIs).

Clements et al. [8] proposed not to recommend the most popular travel trajectories, but to suggest a list of landmarks worth visiting ranked in a personalised way. They proposed that a user's favourite landmarks in a previously unvisited city can be predicted by reranking the most popular locations based on users with similar travel preference. They found that a statistical improvement over all users is hard to achieve, but for users with a clear travel preference, very accurate predictions can be made. Shi et al. [52] proposed to build a similar application using a category-regularised matrix factorisation approach to recommend landmarks to individual users based on both user-landmark preference information and category-based landmark similarity (based on matching POIs to Wikipedia articles and, hence, their categories). They observed that personalised landmark recommendation must go beyond recommending the most popular, widely visited landmarks, and instead focus on less frequently visited 'off-the-beaten-track' landmarks that are well fit to users' individual tastes. They demonstrated the advantage of their method with respect to a popularity-based baseline for the task of recommending such non obvious landmarks.

4.3 Social Sensing

Users of social networks have the potential to act as social sensors, serving a similar purpose to physical sensors by observing and posting about events rapidly emerging at specific locations (such as earthquake tremors or the spread of disease). Such an opportunity is currently under-explored, but some studies have already confirmed that it may bear significant social importance in certain situations and societies.

Sakaki et al. [47] proposed an algorithm for prompt detection of critical events, such as earthquakes, using georeferenced tweets reporting sensed ground shaking. To detect an earthquake, they first select those tweets that are likely to mention something about an earthquake using a text-based classifier. Subsequently, they build a probabilistic spatiotemporal model of the earthquake event to find its centre and its trajectory. They claimed that the notification of users about an upcoming earthquake is delivered much faster if earthquakes are detected using Twitter, than the announcements that are broadcast by the Japanese officials. Similar to the work of Sakaki et al., Fontugne et al. [17] proposed to analyse georeferenced Flickr photos of disaster outbreaks. Their implementation takes advantage Google Street View to allow the user to compare the scene captured by Flickr photos during the disaster with the one captured by the Google cameras in normal conditions. Gomide et al. [21] demonstrated the potential of Twitter data for dengue fever surveillance in Brazil. Specifically, they analysed how users refer to dengue in Twitter with sentiment analysis and used the result to focus only on tweets that somehow express personal experience about dengue. They constructed a linear regression model for predicting the actual number of dengue fever cases in any city using the proportion of tweets originating from this city and expressing personal experience about dengue.

While social networks may be used for spreading information about current events, they are also used for spreading misinformation about events. A study by Mendoza et al. [41], of the 2010 earthquake in Chile discovered that while people spread misinformation and rumours via Twitter, users questioned rumours more than actual news, and this allowed the misinformation to be detected automatically by analysing tweets in the aggregate.

5 Conclusion

Applications such as local search, targeted advertising, mapping, and location-based social networking applications such as Foursquare and Facebook Places require an understanding of human-centric geographies. They need to know the colloquial names people assign to a place as well as the points of interest in and around a given location. Understanding human-centric geographies allows an application to make sense of a user's information need, and interaction with the device. This in turn allows the system to provide more relevant information to the user.

Learning information about places from the content provided by users allows us to supplement information that would normally be licensed by a surveying organisation. This provides a richer representation of the place, additional coverage of locations that are not normally included in licensed data, and reduces our reliance on official sources and proprietary geographic information.

In the future many more devices will determine the location of a media item automatically at the time of recording, using GPS hardware or cellular tower proximity. Even today, these devices provide geographic information while requiring very little data analysis. However, visual georeferencing of photographs and video recordings is still an important problem. For several reasons (such as maintaining the privacy of users), many recordings have been and will be captured without annotations. Visual georeferencing also provides knowledge about which location information can be extracted from the media item and has been shown to provide additional useful context. So far, of all the content on the internet, it is estimated that only about 3% of the media items have been georeferenced (see [18]). Even partially successful automatic location estimation would extend the reach of georeferencing to many use cases where location information has not been collected or cannot be collected since not all media items are associated with geo-relevant tags or any tags at all.

Building models that extract geographic meaning from media content allows us to leverage the vast amount of data provided by users in the course of their every day lives. The content users provide gives us a window into how people use a space, what they are thinking about and doing in that place, and how they perceive it. Ultimately this allows us to build systems that enrich a person's experience of their world with their mobile device.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building rome in a day. In: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, pp. 72–79, 29–Oct 2 2009. IEEE, Piscataway (2009)
2. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 273–280. ACM, New York, NY, USA (2004). doi:10.1145/1008992.1009040. http://doi.acm.org/10.1145/1008992.1009040
3. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International Conference on World Wide Web, pp. 61–70. ACM, New York, NY, USA (2010). doi:10.1145/1772690.1772698. http://doi.acm.org/10.1145/1772690.1772698
4. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) Computer Vision - ECCV 2006. Lecture Notes in Computer Science, vol. 3951, pp. 404–417. Springer, Berlin/Heidelberg (2006) 10.1007/1174402332
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Chang, J., Sun, E.: Location3: how users share and respond to location-based data on social. In: Adamic, L.A., Baeza-Yates, R.A., Counts, S. (eds.) Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, July 17–21, 2011. AAAI, Menlo Park (2011)

7. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating Twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759–768. ACM, New York (2010)
8. Clements, M., Serdyukov, P., de Vries, A.P., Reinders, M.J.: Using flickr geotags to predict user travel behaviour. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, pp. 851–852. ACM, New York (2010)
9. Commission Internationale de L'Eclairage (CIE): Colorimetry. CIE Publication, vol. 15.2, 2nd edn. Central Bureau of the CIE, Vienna (1986). ISBN 3-900-734-00-3
10. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th International Conference on World Wide Web, Madrid, pp. 761–770. ACM, New York (2009)
11. Crandall, D.J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J.: Inferring social ties from geographic coincidences. Proc. Natl. Acad. Sci. U.S.A. **107**(52), 22436–22441 (2010)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE, Los Alamitos (2005)
13. De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., Yu, C.: Automatic construction of travel itineraries using social breadcrumbs. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, HT '10, pp. 35–44. ACM, New York (2010)
14. De Rouck, C., Van Laere, O., Schockaert, S., Dhoedt, B.: Georeferencing Wikipedia pages using language models from Flickr. In: Proceedings of the Terra Cognita 2011 Workshop, Bonn, pp. 3–10 (2011)
15. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, pp. 1277–1287 (2010)
16. Fink, C., Piatko, C., Mayfield, J., Chou, D., Finin, T., Martineau, J.: The geolocation of web logs from textual clues. In: Proceedings of the 2009 International Conference on Computational Science and Engineering, pp. 1088–1092. IEEE, Los Alamitos (2009)
17. Fontugne, R., Cho, K., Won, Y., Fukuda, K.: Disasters seen through flickr cameras. In: Proceedings of the Special Workshop on Internet and Disasters, SWID '11, p. 5:1–5:10. ACM, New York (2011)
18. Friedland, G., Sommer, R.: Cybercasing the joint: on the privacy implications of geo-tagging. In: USENIX Workshop on Hot Topics in Security, Washington, DC (2010)
19. Gallagher, A., Joshi, D., Yu, J., Luo, J.: Geo-location inference from image content and user tags. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, pp. 55–62. IEEE, Piscataway (2009)
20. Gammeter, S., Bossard, L., Quack, T., Gool, L.V.: I know what you did last summer: object-level auto-annotation of holiday snaps. In: Proceedings of the IEEE 12th International Conference on Computer Vision, pp. 614–621, 29–Oct 2 2009. IEEE, Piscataway (2009)
21. Gomide, J., Veloso, A., Meira, W., Jr., Almeida, V., Benevenuto, F., Ferraz, F., Teixeira, M.: Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In: Proceedings of the ACM SIGWEB Web Science Conference (WebSci), Koblenz (2011)
22. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. SIGMOD Rec. **29**, 1–12 (2000)
23. Hauff, C., Houben, G.-J.: WISTUD at MediaEval 2011: Placing task. In: Proceedings of the Working Notes of the MediaEval Workshop, Pisa (2011)
24. Hays, J., Efros, A.: IM2GPS: estimating geographic information from a single image. In: Proceedings of the Computer Vision and Pattern Recognitio, Anchorage, pp. 1–8 (2008)
25. Hollenstein, L.: Capturing vernacular geography from georeferenced tags. Master's thesis, University of Zurich (2008)

26. Ivanov, I., Vajda, P., Lee, J.-S., Goldmann, L., Ebrahimi, T.: Geotag propagation in social networks based on user trust model. Multimed. Tools Appl. **56**, 155–177 (2012). doi:10.1007/s11042-010-0570-7
27. Julesz, B.: Textons, the elements of texture perception, and their interactions. Nature **290**, 3 (1981)
28. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A., Hertzmann, A.: Image sequence geolocation with human travel priors. In: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, pp. 253–260, 29–Oct 2 2009. IEEE, Piscataway (2009)
29. Kelm, P., Schmiedeke, S., Sikora, T.: A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs. In: Proceedings of the 2011 ACM workshop on Social and Behavioural Networked Media Access, SBNMA '11, pp. 15–20. ACM, New York (2011)
30. Kelm, P., Schmiedeke, S., Sikora, T.: Multi-modal, multi-resource methods for placing flickr videos on the map. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11, pp. 52:1–52:8. ACM, New York (2011)
31. Kinsella, S., Murdock, V., O'Hare, N.: "I'm eating a sandwich in Glasgow": modeling locations with tweets. In: Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, pp. 61–68 (2011)
32. Knopp, J., Sivic, J., Pajdla, T.: Avoiding confusing features in place recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision at ECCV 2010. Lecture Notes in Computer Science, vol. 6311, pp. 748–761. Springer, Berlin/Heidelberg (2010)
33. Kurashima, T., Iwata, T., Irie, G., Fujimura, K.: Travel route recommendation using geotags in photo sharing sites. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, pp. 579–588. ACM, New York (2010)
34. Li, Y., Crandall, D., Huttenlocher, D.: Landmark classification in large-scale image collections. In: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, pp. 1957–1964, 29–Oct 2 2009. IEEE, Piscataway (2009)
35. Lieberman, M.D., Samet, H., Sankaranayananan, J.: Geotagging: using proximity, sibling, and prominence clues to understand comma groups. In: Proceedings of the 6th Workshop on Geographic Information Retrieval, pp. 6:1–6:8. ACM, New York (2010)
36. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004). doi:10.1023/B:VISI.0000029664.99615.94
37. Lu, X., Wang, C., Yang, J.-M., Pang, Y., Zhang, L.: Photo2trip: generating travel routes from geo-tagged photos for trip planning. In: Proceedings of the International Conference on multimedia, MM '10, pp. 143–152. ACM, New York (2010)
38. Lux, M., Chatzichristofis, S.A.: LIRe: Lucene image retrieval – an extensible java CBIR library. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 1085–1088. ACM, New York (2008)
39. Malik, J., Belongie, S., Shi, J., Leung, T.: Textons, contours and regions: cue integration in image segmentation. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 918–925. IEEE, Los Alamitos (1999)
40. McCurley, K.S.: Geospatial mapping and navigation of the web. In: Proceedings of the 10th International Conference on World Wide Web, pp. 221–229. ACM, New York (2001)
41. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: can we trust what we rt? In: Proceedings of the Workshop on Social Media Analytics (KDD). ACM, New York (2010)
42. O'Hare, N., Murdock, V.: Modeling locations with social media. J. Inf. Retr. (2012, In press)
43. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**, 145–175 (2001)
44. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, CIVR '08, pp. 47–56. ACM, New York (2008)
45. Rattenbury, T., Naaman, M.: Methods for extracting place semantics from flickr tags. ACM Trans. Web **3**, 1:1–1:30 (2009)

46. Sadilek, A., Kautz, H., Bigham, J.P. Finding your friends and following them to where you are. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, pp. 723–732. ACM, New York (2012)
47. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, pp. 851–860. ACM, New York (2010)
48. Scellato, S., Noulas, A., Mascolo, C.: Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, KDD '11, pp. 1046–1054. ACM, New York (2011)
49. Schockaert, S.: Vague regions in geographic information retrieval. SIGSPATIAL Spec. **3**, 24–28 (2011)
50. Schockaert, S., De Cock, M.: Neighborhood restrictions in geographic ir. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 167–174. ACM, New York (2007)
51. Serdyukov, P., Murdock, V., van Zwol, R.: Placing flickr photos on a map. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 484–491. ACM, New York (2009)
52. Shi, Y., Serdyukov, P., Hanjalic, A., Larson, M.: Personalized landmark recommendation based on geotags from photo sharing sites. In: Adamic, L.A., Baeza-Yates, R.A., Counts, S. (eds.) Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, July 17–21, 2011. AAAI, Menlo Park (2011)
53. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. ACM Trans. Graph. **30**(6), 154:1–154:10 (2011)
54. Sizov, S.: Geofolk: latent spatial semantics in web 2.0 social media. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pp. 281–290. ACM, New York (2010)
55. Torralba, A., Fergus, R., Freeman, W.: 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1958–1970 (2008)
56. Twaroch, F.A., Jones, C.B., Abdelmoty, A.I.: Acquisition of a vernacular gazetteer from web sources. In: Proceedings of the 1st International Workshop on Location and the Web, pp. 61–64. ACM, New York (2008)
57. Van Laere, O., Schockaert, S., Dhoedt, B.: Finding locations of flickr resources using language models and similarity search. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pp. 48:1–48:8. ACM, New York (2011)
58. Wing, B.P., Baldridge, J.: Simple supervised document geolocation with geodesic grids. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 955–964 (2011)
59. Yin, Z., Cao, L., Han, J., Zhai, C., Huang, T.: Geographical topic discovery and comparison. In: Proceedings of the 20th International Conference on World Wide Web, Portland, pp. 247–256 (2011)
60. Yu, F.X., Ji, R., Chang, S.-F.: Active query sensing for mobile location search. In: Proceedings of the 19th ACM International Conference on Multimedia, MM '11, pp. 3–12. ACM, New York (2011)
61. Zhang, W., Kosecka, J.: Image based localization in urban environments. In: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission, pp. 33–40. IEEE, Los Alamitos (2006)
62. Zhuang, J., Mei, T., Hoi, S.C., Hua, X.-S., Li, S.: Modeling social strength in social media community via kernel-based learning. In: Proceedings of the 19th ACM International Conference on Multimedia, MM '11, pp. 113–122. ACM, New York (2011)

Predicting User Tags in Social Media Repositories Using Semantic Expansion and Visual Analysis

Tomas Piatrik, Qianni Zhang, Xavier Sevillano, and Ebroul Izquierdo

Abstract Manually annotating large scale content such as Internet videos is an expensive and consuming process. Furthermore, community-provided tags lack consistency and present numerous irregularities. This chapter aims to provide a forum for the state-of-the-art research in this emerging field, with particular focus on mechanisms capable of exploiting the full range of information available online to predict user tags automatically. The exploited information covers both semantic metadata including complementary information in external resources and embedded low-level features within the multimedia content. Furthermore, this chapter presents a framework for predicting general tags from the associated textual metadata and visual features. The goal of this framework is to simplify and improve the process of tagging online videos, which are unbounded to any particular domain. In this framework, the first step is to extract named entities exploiting complementary textual resources such as Wikipedia and WordNet. To facilitate the extraction of semantically meaningful tags from a largely unstructured textual corpus, this framework employs GATE natural language processing tools. Extending the functionalities of the built-in GATE named entities, the framework also integrates a bag-of-articles algorithm for effectively extracting relevant articles from the Wikipedia articles. Experiments were conducted for validation of the framework against MediaEval 2010 Wild Wild Web dataset for the tagging task.

T. Piatrik (✉) • Q. Zhang • E. Izquierdo
School of EE and CS, Queen Mary University London, Mile End Road, E1 4NS, London, UK
e-mail: tomas.piatrik@eecs.qmul.ac.uk; qianni.zhang@eecs.qmul.ac.uk;
ebroul.izquierdo@eecs.qmul.ac.uk

X. Sevillano
La Salle - Universitat Ramon Lull, Spain
e-mail: xavis@salle.url.edu

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_7, © Springer-Verlag London 2013

1 Motivation and Challenges

With the advances in computer technologies and the evolution of social networks, there has been an explosion in the amount and complexity of digital media that is being generated, stored, transmitted and accessed through the Internet. Much of this information is multimedia in nature, including digital images, video, audio, graphics and textual data. Large-scale social media repositories enable users to creatively share thoughts among a much wider audience. As a consequence, every online user has been transformed into the role of a broadcaster. In efforts to be heard, there is an increasing interest in associating these media items with free text annotations. The disadvantages of manual textual annotation, and in particular of tagging, have been studied over the years, and the three main problems associated with it include (1) manual labour, (2) differences in the interpretation of the media items and (3) inconsistency of the keyword assignments among tags. Due to these disadvantages, recently there has been large amount of research focusing on automatically generating reliable and useful tags for multimedia content in social networks. In other words, there is currently great interest in the development of techniques that are able to take advantage of the characteristics of Internet multimedia that sets it apart from multimedia in more conventional environments in order to generate effective and useful annotations.

To tackle these problems, recently there has been a lot of research focusing on automatically generating reliable and useful tags for multimedia content in the Internet. Such systems usually rely on textual or low-level features, as well as some predefined knowledge focusing on particular domains. Therefore, one aim of this chapter is to provide a survey on the state-of-the-art research in this emerging field and to address the growing interests in automatic tagging of Internet multimedia. In particular, this survey concentrates on mechanisms capable of exploiting the full range of information available online to predict user tags automatically, with specific focuses on technologies related to query expansion, exploitation of complementary resources and visual-based approaches.

Despite of the large amount of research work done on multimedia tagging in social network repositories, the tagging of online multimedia resources is particularly challenged by the fact that these are unbounded to any particular domain. This makes users' requirements for tagging and indexing both too general and specific. On one hand, it is ideal to have a system that 'works for everything'. The universal context is very broad, while the usable resources are limited. Therefore, the task of tagging in a general context is very difficult and often intractable. On the other hand, the systems designed for a specific area can exploit the rich domain knowledge, but they are restricted to the domain and thus may not be useful in an irrelevant context. Therefore, the challenge is how to derive rich and correct tags in a general context using the limited metadata and at the same time can be easily adapted for more specific applications.

Addressing this challenge, in this chapter we also present a framework that aims at predicting user tags of online videos from the associated textual metadata. Despite significant research developments in the area of semantic tagging, much

of these techniques are bounded to the a priori knowledge of their domains. Since by nature, Internet videos are not bounded to anything particular, we considered textual metadata to provide a more reliable source of information that does not require training based on a priori knowledge. To extend the limited information available in the textual metadata, this framework is able to exploit complementary resources such as Wikipedia and WordNet in order to extract more semantically meaningful tags from a largely textual resource. The proposed framework has been tested in a social network tagging scenario using Flickr videos and images. A very important feature of the proposed framework is that it relies only on existing features associated to the multimedia content and general complementary resources which are available to anyone through the Internet. Without relying on domain specific knowledge, the proposed framework can be used for any general purposes. However, if specific application is required, the framework is flexible enough to be adapted for the domain of concern, using available complementary context in that domain.

Based on the survey on related research and on our experiments using the proposed framework, at the end of this chapter we also identify some potential research directions towards a future user tag-prediction systems. The focus of these identified future research directions is on their capability of handling large-scale social network media repositories.

2 Related Research in Social Multimedia Tagging

Nowadays, large-scale online multimedia repositories have become available through various Web 2.0 applications, such as Flickr,[1] Wikipedia,[2] YouTube,[3] Facebook,[4] Second Life[5] and Twitter,[6] providing access to tremendous amount of multimedia data which are mostly created by users. For example, Flickr has been providing access to over five billion images by September 2010, and there are over 3,000 uploads every minute to the website. YouTube has stored 400 million videos by 2010, and in every minute around 20 h videos are being uploaded to the website. The number of images on Facebook has exceeded 60 billion by the end of 2010, and around 138 MB of new content is being uploaded every minute. This user-uploaded and user-generated audio-visual content belongs to the established concept of user-generated content (UGC). UGC includes all kinds of data that comes from regular people who voluntarily contribute with data, information or media that then appears before others in a useful or entertaining way. All digital media technologies can be

[1] http://www.flickr.com/

[2] www.wikipedia.org/

[3] www.youtube.com/

[4] http://www.facebook.com/

[5] secondlife.com/

[6] twitter.com/

related to UGC, such as question-answer databases, digital video, blogging, podcasting, forums, review sites, social networking, mobile phone photography and wikis.

Among all kinds of user-generated data, digital audio-visual content is certainly the one receiving most public interests, and the one generating most technological challenges compared to the others. For example, automatic tagging and search for multimedia content has been a tremendous challenge, particularly in uncontrolled environments such as UGC applications. Collaborative tagging has been a typical and promising approach for tagging of user-generated multimedia content [37]. This kind of approach enables a process where users add and share tags to other shared items. Collaborative tagging is an organisational method. Its most important contribution is the concept of folksonomy, which will be further elaborated in Sect. 2.2. Still, it faces some serious limitations that restrict its usability, such as the nonstructured tags, tags validation, spamming detection and removal, redundancy and subjectivity in tags.

In this section, we present a survey of technologies related to the multimedia content tagging in a large-scale online repositories. First, an overview of the related works on multimedia tagging in general is presented. Then, the survey is focused on some specific topics in social media tagging, including approaches using query expansion, folksonomies, complementary resources, visual analysis techniques and some other related works.

2.1 Multimedia Tagging

Indexing and retrieval of multimedia content in the large scale online repositories has become an increasingly active field. Annotation and tagging have been recognised as a very important and essential mechanism to enable the effective organisation and sharing of large scale of multimedia information. However, manual annotation on large multimedia datasets is extremely labour intensive and time-consuming. Therefore, efficient automatic tagging methods are highly desirable. This interdisciplinary research direction has attracted various attentions and resulted in many algorithmic and methodological developments. There has been a significant amount of research on automatic video indexing based on textual and visual analysis [5, 10, 12, 16, 23].

In general, such approaches for automatic labelling or tagging can be classified in two types, 'open-set tagging' and 'closed-set tagging' [21]. The first type of approaches 'extract' appropriate labels for items from the words or phrases already associated with item content or metadata. In this case, the tags to be assigned are not known in advance. In comparison, the second type of approaches 'assign' tags in a known set of labels to multimedia content. The tagging problem can be posed as a classification problem to be solved either using a series of binary classifiers, one for each tag, or a multi-class classifier [8]. Another approach to close-set tagging relies on multimedia search and retrieval systems for assigning tags to the items, where each tag is treated as a query [16]. In this approach, conventional query expansion methods in information retrieval can be used to expand the tags into

appropriately enriched queries. Such approach often applies a certain threshold in the list of retrieved multimedia items and assigns the queried tag to all items above the threshold.

In [77], authors have tested three different techniques, namely, language modelling, query expansion and maximum entropy, for tagging videos based solely on the video abstracts. Another approach for video tagging based only on the use of associated metadata is discussed in [28]. In [29], tags are predicted for bookmarked URLs using page text, anchor text, linked websites and tags of other URLs. In [56], different sources of information have successfully been integrated in factorisation models to predict the tags that a user will assign to an item. A very important group of research employs query expansion. In the following two subsections, a list of such research is reviewed. Our proposed framework shows that using other metadata resources and complementary information improves the quality of assigned tags.

2.2 Query Expansion and Folksonomy

The associated textual information in social networks is identified as a rich source of information for extracting high-level semantics for collaborative tagging systems. However, in order to effectively index these media items, the free text description needs to be analysed, and corresponding tags with semantic meaning should be extracted.

Most research in this field has so far focused on nonstatistical approaches, particularly on the lexico-syntactic patterns (Hearst patterns) first introduced in [27]. While purely statistical approaches such as latent semantic indexing (LSI) are prevalent in other fields of natural language processing, until recently they were only suitable for discovering symmetrical relations between words. The closest task to hypernym discovery mentioned in the seminal text book on statistical natural language processing [46] is unsupervised disambiguation, in which k meanings of a term are determined automatically. This approach has however the limitation that meaning is not represented by a single word (term) but by a context. Recent research [6] introduced one of the first statistical methods to hypernym discovery. Their work utilis principal component analysis (PCA) for discovering term taxonomies (hierarchies of hypernyms). The algorithm presented here is closest to the research of Cimiano et al. [13], who use lexico-syntactic patterns, also codified in a JAPE transducer grammar. The focus is however different, as their Text2Onto framework tries to learn the whole ontology, while the work presented here tries to discover only hypernyms for the given query.

Query expansion is probably the most typical application of hypernym (taxonomy) discovery. Query expansion is a method for improving recall and possibly the precision of information retrieval by expanding the query with other terms related to the original query. These terms are usually weighted. Query expansion has not been found to provide any significant objective improvement, although it is perceived positively by the users [52, 60]. Generally, query expansion comprises two basic steps: expand the initial queries using new words and term re-weighting

in the set of the expansion queries. Currently, five query expansion techniques have extensively applied, namely, query expansion based on global document analysis [17, 78], query expansion based on local analysis [42, 76], query expansion based on query log analysis [36, 79], query expansion based on association rules [18, 83] and query expansion based on complementary semantic resources [25, 54]. Xu et al. [42] proposed a local context analysis method, which selects expansion terms based on cooccurrence with the query terms in the top-ranked documents. The method produces more effective and robust query expansion than traditional global and local techniques. However, the main drawback of this method is that it may lead to irrelevant addition of terms. In global analysis methods, new terms are added to an original query before searching. This method needs external resources such as thesaurus and WordNet [78]. Cui [15] proposed a query expansion model based on user logs. By mining user logs, a probability method is used to optimise the query. Some researchers have also worked on the ontology-based expansion but they have been static in their approach [84]. To improve this method, authors in [49] propose an approach called dynamic document analysis considering thesaurus analysis as well as dynamic documents.

Social networks and social resource sharing systems use the lightweight knowledge representation, called folksonomy. The term 'folksonomy', first proposed by Thomas Vander Wal in a mailing list [3], is combination of 'folk' and 'taxonomy' to describe the social classification phenomenon. Folksonomy provides user-created metadata rather than professional-created and author-created metadata. As discussed in [47], the tags, which constitute the core of folksomony, can be seen as good keywords for describing the respective web pages from various aspects. The folksonomy tags have the keyword property which may convey the topics of web pages from various aspects. Al-Khalifa and Davis [2] analysed the semantic value of social tags and concluded that the folksonomy tags are semantically richer than keywords extracted using a major search engine extraction services. X. Wu et al. [80] explored machine understandable semantics from social annotations in a statistical way and applied the derived emergent semantics to discover and search shared web bookmarks. In [31], authors proposed Adapted PageRank and FolkRank to find communities within the folksonomy. Bao et al. [4] proposed to measure the similarity and popularity of web pages from web users' perspective by calculating SocialSimRank and SocialPageRank. In [82], a personalised search framework to utilise folksonomy for personalised search has been proposed.

2.3 Query Expansion Using Complementary Resources

A gold standard dataset for training and testing hypernym discovery algorithms is WordNet (e.g. [24, 64]). WordNet is a lexical database developed by Princeton University to model the lexical knowledge of a native speaker of English [20]. Sets of synonym terms called synsets constitute its basic organisation. Several types of relations between synsets are recorded in WordNet, including hypernymy/hyponymy

(is-a relation) and meronymy/holonym (part-of relation). In addition, each synset has a gloss that defines the synset. WordNet is one of the most important lexical semantic resources in information retrieval. Faced with the defects of traditional query expansion methods by choosing similar terms to query terms based on some criterion, a query expansion method based on concepts has been proposed in [55]. In this approach, terms with a common sense are chosen as one of the candidate terms for expansion. To improve this approach, WordNet has been used to expand queries using the well-defined synonyms [73]. But in this work, query terms were deemed independent from each other and only synonyms were selected as term candidates for expansion. In other work, Smeaton [57] tried to perform query expansion using various strategies of weighting expansion terms, along with manual and automatic word sense disambiguation techniques, but it proved not able to improve the performance of retrieval. Hoeber manually constructed a concept network based on which terms are selected to perform conceptual query expansion [43]. The performance of this method depends highly on the quality of the concept network. In contrast, Liu et al. [30] proposed automatically generating expanded query terms by WordNet. Once original query terms' concepts are determined, their synonyms, hyponyms and the like are considered to be the expanded terms. But in their work, queries to be expanded are confined to noun phrases. The main drawback of this technique is that it does not take term relationships into consideration. In [84], the word sense disambiguation is utilised to recover the sense of a word in the given query context. Based on the extracted concepts, similar terms in the corresponding synset are extracted from WordNet. Then through combining the newly chosen terms, the candidate expanded query set is generated, from which final expanded queries are selected.

Although WordNet contains general knowledge of a wide range of fields, it is difficult to instantly add new knowledge, particularly proper nouns, to these general ontologies. Therefore, Wikipedia has been used as a useful corpus for knowledge extraction because it is a free and large-scale online encyclopedia that continues to be actively developed. Wikipedia presents a much larger data resource for named entity extraction such as people, places, organisation and events to name a few. There have been many attempts to combine web search and Wikipedia article titles and hyperlinks for extraction of instances of arbitrary relations [7]. In [66], authors used the Wikipedia category system for the purpose of ontology learning. Kliegr et al. [34] found the first section of Wikipedia articles as particularly suitable for hypernym discovery and use it as the sole source of information. However, making judgements about the semantic relatedness of different terms in Wikipedia articles are yet a deceptively complex task. Any attempt to compute semantic relatedness automatically must also consult external sources of knowledge. Some techniques use statistical analysis of large corpora while some others use hand-crafted lexical structures such as taxonomies and thesauri. In either case, it is the background knowledge that is the limiting factor limited in scope and scalability. These limitations are the motivation behind several new techniques which infer semantic relatedness from the structure and content of Wikipedia. Strube and Ponzetto [65] were the first to compute measures of semantic relatedness using Wikipedia. Their

approach 'WikiRelate' took familiar techniques that had previously been applied to WordNet and modified them to suit Wikipedia. In another work, authors achieved extremely accurate results with ESA, a technique that is somewhat reminiscent of the vector space model widely used in information retrieval [22]. Instead of comparing vectors of term weights to evaluate the similarity between queries and documents, they compare weighted vectors of the Wikipedia articles related to each term. The difference to this approach is the use of Wikipedia's hyperlink structure to define relatedness [48]. This approach offers a measure that is both cheaper and more accurate than ESA: cheaper, because Wikipedia's extensive textual content can largely be ignored, and more accurate, because it is more closely tied to the manually defined semantics of the resource.

2.4 Tagging Using Visual Analysis Approaches

Content-based tagging and search for multimedia content has been a most important approach in parallel to the textual features-based approach. Therefore, in this subsection, we give an overview on the important works in this direction. In the state-of-the-art research, many automatic tagging methods use visual content analysis together with text features in order to predict tag assignments. These visual-based approaches borrow many concepts and techniques from the content-based image retrieval field, a comprehensive survey of which can be found in [62].

One of the first approaches to tagging using visual analysis was based on machine translation [19]. The rationale was annotating image regions with words. To that end, the regions an image was segmented into were categorised using a taxonomy of region types. Subsequently, an EM-based learning approach is used for mapping region types and keywords, thus captioning the image.

Latent space models (namely, latent semantic analysis and probabilistic latent semantic analysis) were applied to image annotation for discovering the links between visual features and words in an unsupervised fashion, propagating tags from the most similar images in the latent space [51].

The work by Li and Wang [38] introduced a fully automatic and high speed system for annotating online pictures called ALIPR (Automatic Linguistic Indexing of Pictures – Real Time). It was based on the use of generative models for learning the joint distributions of visual features and vocabulary subsets, thus characterising each image by a statistical distribution. By exploiting statistical relationships between images and words, tagging could be conducted in realtime without the need of recognising individual objects in the images.

According to [44], the availability of training data required by most approaches to tagging limits their performance and scalability. This is one of the motivations of the dual cross-media relevance model for automatic image tagging proposed by Liu et al., which estimates the joint probability by the expectation over words in a predefined lexicon. To do so, the proposed model considers two types of relations in image annotation: word-to-image relations and word-to-word relations, which are estimated by using search techniques on Web data as well as available training data.

In [1], visual features were mapped to semantic categories by designing a dedicated feature space for each image category. To that end, a two-layer ensemble learning system called Supervised Annotation by Descriptor Ensemble (SADE) was proposed. In a nutshell, the proposal was based on an initial extraction of multiple low level visual descriptors from the image, each one of which is separately fed into a learning machine in the first layer. Finally, the meta-layer classifier is trained on the output of the first layer classifiers, and the images are annotated by using the decision of the meta-layer classifier.

The analysis of visual contents is coupled with the exploitation of collaboratively annotated image databases in [41]. The proposed approach applied two techniques based on image analysis: an SVM classifier annotated images with a controlled vocabulary, while a tag propagation module exploited user-generated, folksonomic annotations from Flickr, thus being able to deal with an unlimited vocabulary.

It is a commonplace that the tags associated with images in social media repositories are a source of valuable information source for superior multimedia retrieval experiences [67]. For this reason, it is necessary to evaluate the descriptive power (or relevance) of user-generated tags. However, users tag images with uncontrolled and often personalised and ambiguous terms. This is the motivation behind the work of Sun and Bhowmick [67], who proposed a measure called Normalized Image Tag Clarity (NITC) – a version of the clarity score proposed for query performance prediction in classic information retrieval – for evaluating the descriptiveness of a tag with respect to the visual contents of the image it is attached to. To that end, images are represented using a bag of visual words scheme, which allows to build a collection language model upon which the NITC evaluation measure is computed.

Focusing also on the tag relevance evaluation problem, Li et al. proposed a scalable algorithm for computing tag relevance values from visually similar neighbours [39]. In a subsequent work, Li et al. [40] used an extended version of their previous work for automatic image tagging. Broadly speaking, the proposal consisted in annotating an untagged image with the most relevant tags attached to its visual neighbours, retrieved from a large user-tagged image database. However, the validity of this approach suffered from the unreliability and sparsity of user tagging, so a joint-modality tag relevance estimation method based on textual and visual clues was introduced to mitigate their effect.

This idea of exploiting the nearest neighbours for annotating an untagged image was also explored in [26]. The proposed model (called TagProp), though, was based on a discriminatively trained nearest neighbour model in which neighbours were weighted according to their rank. The TagProp model included a word specific sigmoidal modulation of the weighted neighbour tag predictions to boost the recall of rare words. Moreover, it allowed to combine several visual similarity metrics in order to consider simultaneously local and global aspects of image contents.

The power of groups of images uploaded to online repositories like Flickr was exploited by Ulges et al. in [72]. Their approach was based on the realistic assumption that Flickr users group their pictures into batches (e.g. all snapshots taken over the same holiday trip) and that the images within a batch are likely to

have a common tagging style. Therefore, these batches are matched with categories learned from Flickr groups, and leveraged for accurate context-specific image annotation.

A problem related to image tagging is tag recommendation, which tries to avoid both the noise inherent to user tags and also semantic noise. In [81], a multimodal tag recommendation algorithm was introduced. In there, tag recommendation was posed as a learning problem that was tackled using tag and visual correlations. Each modality was used to generate a ranking feature, and the optimal ranking features' combination from different modalities was learnt by means of the RankBoost algorithm.

Another related problem is the creation of visual tags dictionaries, which was the goal of Wang et al. [75]. The main idea is describing textual tags by means of visual words related to a bag of visual words' representation of images. With the proposed method, the visual tags dictionary is built in a fully automatic manner by harnessing tagged images available online. Once the dictionary is created, a connection between textual tags and visual words is established, which can be exploited for image annotation.

The tagging of online video resources has also attracted the attention from researchers in the last years. At least two main trends coexist in this area. The first one is based on annotating the video using concept detectors that describe objects, locations or activities appearing in it [63]. In order to alleviate the problem caused by the little availability of large-scale collections of annotated videos for training tagging systems, the work by Ulges et al. [71] proposes training concept detectors on videos available in online repositories such as YouTube. This allows exploiting existing user tags, besides scaling concept detection up to thousands of concepts with need of no manual labour at all.

An alternative strategy to video tagging is based on exploiting the redundancy of its content [58, 61]. The underlying rationale is based on the existence of a large amount of videos with overlapped or duplicated content on YouTube. Thus, this can be harnessed in order to obtain useful information about connections between videos, which are revealed by means of robust content-based video analysis techniques thus allowing to generate new tag assignments using tag propagation methods.

2.5 Other Related Research

Another interesting field of multimedia tagging is music annotation. Indeed, songs can be tagged with highly semantic concepts related to their mood, usage, instrumental contents, among others, which are of interest for building music recommendation systems and large scale music discovery engines.

In [70], Turnbull et al. presented a computer audition system capable of annotating novel audio tracks with semantically meaningful words. They posed the problem as a supervised multiclass, multilabel problem in which the joint probability of acoustic features and words was modelled. Using a dataset of human-generated

annotations that describe popular music tracks, a Gaussian mixture model was trained over an acoustic feature space for each word in the vocabulary, obtaining music annotations comparable with the performance of humans on the same task.

More recently, a larger dataset comprising 10,870 annotated songs was collected in order to develop a novel music tagging system [68]. The novelty of this approach was that it considered both genre tags as well as 'acoustically-objective' tags, the main feature of the latter being that they can be consistently applied to songs by expert musicologists. Another interesting aspect of this work was the analysis of the tagging performance of two novel content-based audio features related to timbre and mid-level acoustic parameters.

However, the obtainment of accurate and reliable tags for annotating multimedia resources is a great challenge. This is due to the fact that harnessing user tags of publicly available videos and images may lead to unreliable results, whereas manual annotation is expensive though more accurate in general. For this reason, some researchers have devised collaborative strategies for motivating users to manually annotate multimedia resources, particularly by means of gaming.

One of the earliest attempts to do so in the image field was the work by von Ahn and Dabbish [74]. Their motivation was to take advantage of the people's desire to be entertained to make them do the work that computers are unable to do well enough due to the shortcomings of computer vision techniques. The proposed game, called ESP, encouraged players to tag a given image with the same strings (i.e. a *think like each other* type of game), as the strings two players agree on turned out to be good labels for the image. The authors estimated that if the proposed game was played as much as popular online games, most images on the Web could be labelled in a few months.

More recently, a new gaming approach to gaming-based image annotation was proposed in [59]. Its main features were the fact that it takes into account the social aspects of human-based computation, as it aimed at what millions of individual gamers are enthusiastic to do, to enjoy themselves within a social competitive environment. This goal was achieved by setting the focus of the system on the social aspects of the gaming environment, which involved a widely distributed network of human players. Furthermore, the proposed framework integrated a number of different algorithms commonly found in image processing and game theoretic approaches to obtain an accurate label. As a result, the framework was able to assign accurate tags for images besides being able to detect and eliminate annotations made by cheater players.

A less gaming-oriented approach is the one presented by Moehrmann et al. [50] that introduces an image labelling interface based on self-organising maps (SOM) for optimising its usability.

As for the manual tagging of music based on gaming, a parallel road has been followed. For instance, Mandel and Ellis [45] designed a web-based game to collect descriptions of musical excerpts. Their goal was to make this task fun and easy for users, besides obtaining useful and objective tags. They apply the same idea than in [74], as the goal of players is to describe song clips using the same tags as other participants.

Another example of game-based music tagging is an online multiplayer game called Listen Game, aimed to measure the semantic relationship between music and words [69]. The game has two playing modes: in the normal mode, the player is prompted to select the best and worst words (describing semantic music concepts such as instruments, emotions, song usages and genres) to describe a song. In the freestyle mode, the player is asked to suggest a new word that describes the music, receiving feedback of other players' answers.

3 Predicting Tags Using Semantic Expansion and Visual Analysis

In this section, we present a framework for predicting user tags, by jointly exploiting the associated textual metadata, the expanded query terms and their complementary resources, as well as the visual features embedded in content. The visual features we employed in the proposed system are MPEG-7 colour layout and edge histogram features [32].

The proposed framework consists of two stages. The first stage is the tag preprocessing where each tag from the list of all tags is processed and further expanded if needed. The algorithmic workflow is presented in Fig. 1. As tags in general can contain any keyword which the author might consider as relevant, it was important to contextualise the tags. To this end, the preprocessing framework developed is aimed at categorising the tags into two general categories, namely, (1) common tags and (2) named entity tags. Common tags are those which correspond to either an action, country or as depicted in the figure have a synset associated to it in WordNet. On the other hand, named entity tags are those tags which do not have a WordNet synset and depend on external resources to contextualise them. The objective of this preprocessing is to ensure that named entity tags are disambiguated enough to enable a match semantic similarity search.

An overview of the second stage of processing is presented in Fig. 2. As we considered the metadata (i.e. video title, video description, automatic speech

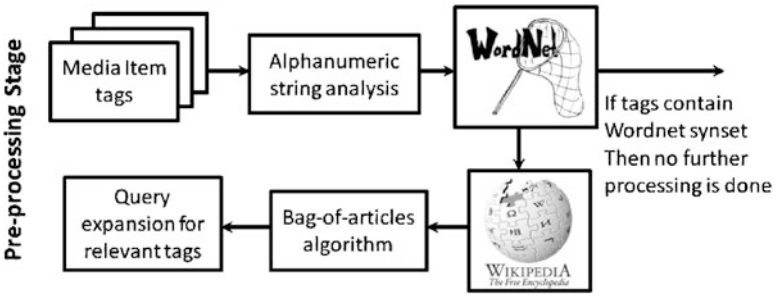


Fig. 1 Overview of the tag preprocessing phase

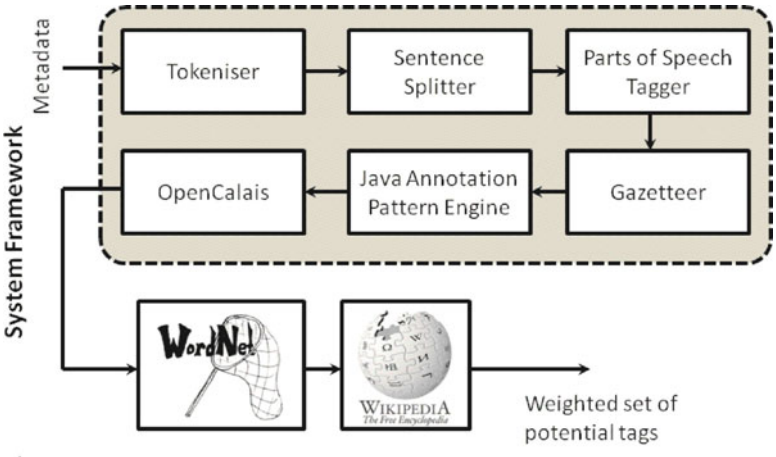


Fig. 2 Overview of the proposed system

recognition (ASR) transcripts) to be of value in determining the nature of tags, we first processed the metadata with GATE[7] NLP framework. The framework includes a tokeniser, sentence splitter, and part-of-speech (POS) tagger. In addition to the basic text components, we also included a gazetteer in order to identify entity names in the text based on lists of predefined words. Also, for extraction of additional semantic information, we included the Java Annotation Pattern Engine (JAPE) to extract hypernyms from Wikipedia. Finally, we also included the OpenCalais[8] plugin for extraction of named entities from the textual metadata.

One of the significant contributions of this framework is the integration of Bag-of-Articles (BOA) algorithm as an extension to GATE NLP tools. Briefly, the module locates a Wikipedia article using the unlabelled entity through media wiki API. The similarity measure for determining the article's relevance to the tag is obtained through text relevance with popularity of the articles [34]. From the selected article, a JAPE implementation of Hearst patterns was used to extract a hypernym. This hypernym was then looked up in WordNet, thus establishing a link between the entity and a WordNet synset.

3.1 Wikipedia as the Source of Knowledge

WorldNet has a structured nature, and its general coverage makes it a good choice for general disambiguation tasks. The focus of work presented here is on specialised domain, which makes the use of WordNet less appealing. Most existing lexical

[7] http://gate.ac.uk/

[8] http://www.opencalais.com/

resources including WordNet will have difficulty finding hypernyms for specialised search queries such as the name of a footballer or football arena. In experiments with automatically learned rather than hand-crafted lexico-syntactic patterns [64], using TREC dataset and Wikipedia as the training corpus gave a significant improvement to the best WordNet classifier (F-Measure from 0.2339 to 0.3592).

Our previous work relied on WordNet thesaurus [53], but it turned not to be exhaustive enough, and we decided to search for another source of information. In this sense Wikipedia turned out to be convenient as we needed a closed corpus of texts where the duplicity of articles describing the distinctive semantic category of the given word is minimal. In this regard, the general web cannot serve as a good source while Wikipedia tries to cover most of the semantic meanings using only limited number of pages (usually only one page). Therefore, we found the first section of Wikipedia articles as particularly suitable for hypernym discovery and use it as the sole source of information.

3.2 Bag-of-Articles Classifier

As previously mentioned, Wikipedia presents a much larger data resource compared to WordNet for named entity extraction such as people, places, organisation and events to name a few. In order to exploit Wikipedia resources, the BOA classifier has been developed. The proposed BOA is an extension of the well-known bag-of-words (BOW) approach [33]. The input for the BOA classifier is the classified entity represented as a noun chunk and a set of class entities, represented with a Wikipedia page title. For unlabelled entities, the BOA classifier locates articles in Wikipedia that might define the entity and selects one of them using a disambiguation function. Subsequently, it uses link analysis to try to identify related articles falling into the same semantic category, and then creates a BOA term-weight vector by aggregating their BOW's vectors. The class is assigned by choosing the closest class entity, also a BOA term weight vector, with cosine similarity or other suitable metric.

Formally, the input of a BOA classifier is a set of t labelled instances (titles of Wikipedia articles) C and a set of u unlabelled instances (noun phrases) E. Wikipedia article titles provide an unanimous mapping between the labelled instance and a Wikipedia article. We use symbol W to denote a collection of all pages in Wikipedia at a given time. Each article is described by its title, term-weight vector, outbound links, a list of categories it belongs to and type (article page, disambiguation page, category page, ...). The BOA representation, as proposed here, does not process Wikipedia infoboxes.

For an unlabelled instance $e_x \in E$, it is first necessary to determine the articles that may be defining its various senses. The ranking function ρ maps it onto the vector of its n possible senses $s_x = \rho(e_x, W) = \langle s_{x,1} \dots s_{x,l} \dots s_{x,n} \rangle$. The senses – titles of Wikipedia *article pages* – are sorted in the vector in the decreasing order of relevance. The sense l of an unlabelled instance e_x is represented by article title $s_{x,l}$. The fact that there are multiple senses for the unlabeled instance gives space

for disambiguation function δ. In the base scenario, we use disambiguation function δ_{mfs}, which assigns the most frequent sense:

$$\delta_{mfs}(s_x) = s_{x,1}. \tag{1}$$

Now, both a disambiguated unlabelled instance and a labelled instance is a Wikipedia article title and can be mapped to a Wikipedia article. In the following, we will use the variable a to refer to a Wikipedia article to which an instance (labelled or unlabelled) is mapped. The bag of articles $\beta(a)$ is constructed by aggregating related article across the set of modalities M with the help of the modality membership function μ, article term-weighting function τ and recursive term-weight aggregation function θ.

Modality Membership μ

Modality membership function $\mu(a, a_r) \mapsto \{0, 1\}$ expresses if article a_r is considered related to a ($\mu = 1$) or not ($\mu = 0$). Several modality membership functions are suggested below. Article a is evaluated as related to a_r ($a \neq a_r$) if

- $\mu_{outlink}(a, a_r) = 1$ iff a links to a_r.
- $\mu_{backlink}(a, a_r) = 1$ iff a_r links to a.
- $\mu_{related\ outlink}(a, a_r) = 1$ iff a links to a_r and there is an article a_c linking to a and a_r, $a_r \neq a \neq a_c$.
- $\mu_{backlinking\ outlink-firstpara}(a, a_r) = 1$ iff a links to a_r, a_r links to a and the link from a to a_r is contained in the first paragraph of a.
- $\mu_{shared\ category\ outlink}(a, a_r) = 1$ iff a links to a_r and a and a_r share the same category.

Other modality membership function definitions are also possible and various have been in fact suggested in the literature, albeit under a different name. This applies, for example, to $\mu_{backlinking\ outlink-firstpara}$ [14] or $\mu_{related\ outlink}$, which is used in the Lucene-search Mediawiki extension (refer to Sect. 3.3). We use the symbol $A^a_{\mu_m}$ to denote the set of all articles a_r that are related to a with respect to modality membership function μ_m:

$$A^a_{\mu_m} = \{a_r | a_r \in W, \mu_m(a, a_r) = 1\}. \tag{2}$$

The bag of articles might contain articles related according to multiple modalities.

Article Term-Weighting τ

The weight function $\tau(a) \mapsto R^n$ represents the article a as a vector of term weights. The parameter $w_{m,d}$ is a weight assigned to term vectors $\tau(a)$ in modality m and depth d. The term weight functions considered are

- Term frequency (TF)
- Term frequency – inverse document frequency (TF-IDF) computed over entire Wikipedia
- Term frequency – inverse document frequency computed over articles included in bag of articles of labelled instances C
- Term frequency with first paragraph[9] boost

Other term-weight function definitions can be also considered.

Recursive Term-Weight Aggregation θ

The function $\theta_m(a, d, maxd_m) \rightarrow R^n$ recursively aggregates term-weight vectors of articles related to a according to the modality membership function μ_m:

$$\theta_m = \begin{cases} \sum_{a_r \in A^a_{\mu_m}} [w_{m,d}\ \tau(a_r) + \\ \quad \theta_m(a_r, d+1, maxd_m)] \text{ if } d < maxd_m \\ 0 \text{ if } d = maxd_m. \end{cases} \tag{3}$$

Bag of Articles β

Function $\beta(a) \mapsto R^n$ creates the bag of articles for article a:

$$\beta(a) = \tau(a) + \sum_{m \in M} \theta_m(a, 1, max_d). \tag{4}$$

The formula aggregates the term-weight vector for article a with term-weight vectors of articles recursively related to it up to level $maxd_m, maxd_m \in N$. The articles (directly) related to it have level 1.

The classification is done by comparing the BOA vector of the unlabelled instance $\beta(a_x)$ with BOA-term vectors of labelled instances $\beta(a_c)$ with the similarity metrics sim and selecting the class with the highest similarity:

$$BOAclass(a_x) = \underset{c}{arg\ max}\ sim(\beta(a_x), \beta(a_c)). \tag{5}$$

A BOA classifier implementation needs to make decisions as of the selection of the ranking function ρ, modality membership functions μ_m, term-weighting function τ and the BOA similarity function sim. The weights $w_{m,d}$ and the maximum depth $maxd_m$ for gathering related pages in modality m are externally set. Except for the function sim, all these settings are made separately for labelled and unlabelled instances.

[9]The first paragraph of a Wikipedia article contains usually the definition of the article subject, it can be therefore expected to contain more relevant words than the rest of the text.

3.3 Implementation of BOA Classifier

This section describes an experimental implementation of the BOA-based classification system. As the ranking function ρ, the implementation uses a composite metric, which combines text-based similarity between the noun chunk and article text and article popularity as measured by the number of backlinks. As modality membership function μ_m, there is one option – *outlinks*, implementation of *backlinks* is in progress. For the term-weighting function τ, there is a TF and TF-IDF support. As the BOA similarity metrics *sim*, the implementation uses cosine similarity.

A BOA classifier requires a Wikipedia index containing the following pieces of information about each article:

- Term vectors with term frequencies
- Outlinks
- Popularity ranking (for most frequent sense relevance ranking)

Given the current size of English Wikipedia and the fact that it is constantly updated, meeting these data acquisition requirements results in a considerate engineering effort, and in fact, a reimplementation of an existing software as these functions are from the most part performed by the existing *Lucene-search* Mediawiki extension.[10] This *Lucene*[11]-based Mediawiki search engine indexes the Mediawiki article database and creates five Lucene indexes: the main index, the links index, the related index, the headlines index and spellcheck index. For the BOA classifier, the main index containing term vectors and the links index containing links leading out of each article are the most important. This extension provides two additional vital functions for the BOA classifier – parsing of wikitext and prospectively the ability to perform incremental updates.

The main `wiki` index contains the following important fields: `title`, `key` with a numeric article identifier, the term vectors are saved in the `contents` field, `category` stores article's categories, `related` stores titles of articles that were determined as related during indexing.[12] The `wiki.links` index contains the following fields: `Article key` containing concatenated article title, `Article PageID` with a unique numeric identifier that binds the entry with the main index `key` field, `links` with a list of article titles to which the article links. The index differentiates between different types of links (article/image) using a namespace (prefix), `redirect` contains the title of the article to which the current article is redirected, `rank` contains the number of backlinking articles. In the BOA classifier implementation, these indexes are exploited as follows.

[10] http://www.mediawiki.org/wiki/Extension:Lucene-search

[11] http://lucene.apache.org

[12] A is said to be related to B, if A links to B, and there is some C that links to both A and B (source: Lucene-Search Extension documentation).

Term vectors Indexed Wikipedia articles are stored in the `wiki.main` index, however the Lucene-seach extension does not store term vectors. For the purpose of the BOA classifier, it was necessary to modify the extension with code for storing the term vectors.

Outlinks This information can be obtained from the `links` field of the article entry in the `wiki.links` index.

Popularity ranking The Lucene-search extension contains a search engine, which uses sophisticated relevance ranking involving the number of backlinks. The BOA implementation uses the first-ranked article as the MFS baseline.

The Lucene Mediawiki indexer as used in the BOA classifier system has several changes in code, the most marked one is the extension of the index with stored term vectors. The term vector computations are done with a *sparse matrix toolkit* java library.[13]

3.4 WordNet-Based Classification

To expand known entities using WordNet, we perform a similarity matching function by constructing TF/IDF matrix. We used the Lin similarity metric between the WordNet synsets representing an entity with each of the target tags. The Lin similarity measure has sound theoretical foundation stated in the similarity theorem [9] and is defined as

$$sim_L(c_1, c_2) = \frac{2 * log\ p(lso(c_1, c_2))}{log\ p(c_1) + log\ p(c_2)} \tag{6}$$

The function lso returns the lowest common subsumer from the hierarchy, and the value $-log(p(c))$ is called information content (IC). The value $p(c)$ denotes the probability of encountering an instance of concept c, which is estimated from frequencies from a large corpus. More details of the method can be found in [11].

3.5 Filename-Based Classification

The filename-based approach exploits the human reasoning behind naming video files and is aimed at transforming the user behaviour towards predicting user tags. In addition, the video file name contains intrinsic semantic information, in particular when multiple file names starting with or containing a major portion of the file name. This approach is based on the implementation of a filename-based classifier

[13] http://code.google.com/p/matrix-toolkits-java/

Table 1 Close set annotation results in MAP

	Methods	MAP (%)
Proposed approach	All videos (1,727)	30
	Videos with tags (1,671)	43
	Filename-based approach	17
MediaEval2010 tagging task competition	DCU team	0.16
	TUD team	0.27

for which the development set from MediaEval 2010 dataset was used as a training set. The filename-based classifier was developed based on the Weka statistical signal processing library.

3.6 Experiments and Evaluation

In this section, we present an overview of the evaluation methodology we adopted for the evaluation of the proposed framework on a user tagging task.

The evaluation consists of two parts, namely, 'closed-set annotation' and 'open-set annotation'. On one hand, the objective of closed-set annotation is to predict user tags only from a list of tags provided. Although it should be noted that there are no restrictions on the data domain. On the other hand, in the 'open-set annotation,' there are no restrictions assigned to the list of tags that could be associated with the media items.

3.6.1 Closed-Set Annotation

For the closed-set annotation, the evaluation was treated as a retrieval problem, and using the TRECVID evaluation tool, we obtained MAP measure for predicted tags. Although the dataset contained 1,727 videos, we extracted tags only for 1,671 videos. This was due to either the absence of title and/or description or the absence of named entities from these textual resources. In summary, using our proposed framework we achieved 30 % MAP for all 1,727 videos and 43 % MAP against 1,671 videos for which we found any tags. It is worth noting the filename-based approach has been responsible for 17 % MAP of correctly detected tags. Overall, our proposed framework performed the best among all participants who submitted their results to the MediaEval2010 Tagging Task competition. Our method has been compared to other techniques: DCU team achieving 0.16 % MAP and TUD team achieving 0.27 % MAP. More details about approaches proposed by other teams can be found at [35]. These results are more clearly presented in Table 1.

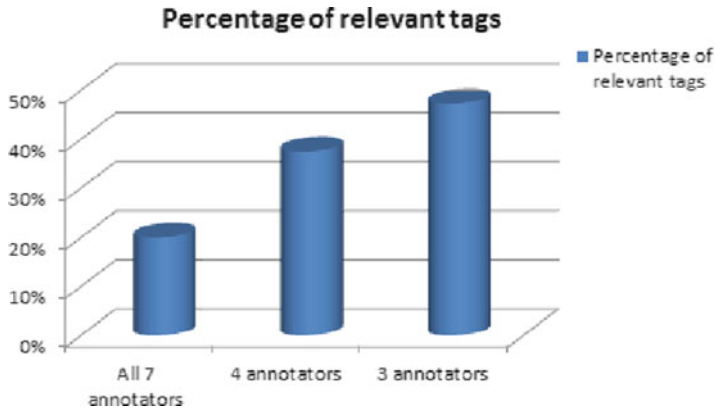


Fig. 3 Open-set annotation results

3.6.2 Open-Set Annotation

We were the only team participating to the MediaEval2010 open-set annotation task. In order to provide a fair evaluation on the open-set annotation,we randomly selected 40 videos and had seven annotators to manually label if the tags associated to each video are 'relevant' or 'irrelevant'. As a measure of relevance, we considered the 'inter-annotator' agreement [28] among any three or more annotators. A total of 296 tags were generated for the 40 videos considered for the evaluation and among them, 35.8 % of generated tags were considered to be irrelevant and 20 % tags relevant by all annotators. Considering a tag with more than 3 inter-annotator agreement, then 47.3 % of the tags generated were considered to be relevant and with four inter-annotator agreement, the percentage drops to 37.5 %. For the total dataset of 1,727 videos, we obtained 6,095 unique tags. These results are presented in Fig. 3.

In summary, the performance analysis of the results for closed-set annotation shows the benefit from exploiting complementary textual resources such as Wikipedia, WordNet and considering filenames as another strong tag predictor. Proposed framework proved successful also on the open-set annotation with almost 40 % generated tags being considered relevant by 4 out of 7 manual annotators.

4 Future Research Directions

One of the most relevant future research directions in the use of visual analysis for tagging is the exploitation of online multimedia repositories as substitutes of hard-to-collect training datasets. Although already a reality in image and video tagging applications, a boost in performance could be achieved if the group and hypergroup structures of sites like Flickr or YouTube were explored [72]. However, this issue still remains a challenge in the area of music annotation.

Another promising issue resides in the integration of multiple annotation techniques under a single framework. An interesting idea is the combination of tagging models with different scalabilities, so that good performance can be obtained regardless of the datasets size [72]. In a similar sense, another way of extending tagging approaches would consist in taking into account the relationships link between different resources such as videos, pictures or text found in different sites, which may be of help for extracting additional information for improving tagging accuracy [61].

Moreover, a very interesting direction for future research, specially in the music annotation field, is the construction of user-specific models that allow to reduce the influence of subjectivity, thus making it possible to model each user's concept of audio semantics [70].

Another relevant issue is the analysis and generation of the so-called *deep tags* (i.e. tags linked to a small part of a larger media resource (e.g. a segment of a video [61], a region of an image, or an audio sample)).

References

1. Akbas, E., Yarman Vural, F.T.: Automatic image annotation by ensemble of visual descriptors. In: CVPR, Minneapolis, pp. 1–8 (2007)
2. Al-Khalifa, H.S., Davis, H.C.: Exploring the value of folksonomies for creating semantic metadata. IJSWIS **3**(1), 13–39 (2007)
3. Atomiq, G.S.: Folksonomy: social classification. http://atomiq.org/archives/2004/08/folksonomysocialclassification.html. Accessed August 2004
4. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: Proceedings of WWW2007, pp. 501–510. ACM, New York (2007)
5. Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. J. Mach. Learn. Res. **3**, 1107–1135 (2003)
6. Bast, H., Dupret, G., Majumdar, D., Piwowarski, B.: Discovering a term taxonomy from term similarities using principal component analysis. In: Semantic Web Mining. Springer, Berlin/New York (2006)
7. Blohm, S., Cimiano, P.: Using the web to reduce data sparseness in pattern-based information extraction. In: PKDD. Lecture Notes in Computer Science, vol. 4702, pp. 18–29. Springer, Berlin/New York (2007)
8. Brezeale, D., Cook, D.J.: Automatic video classification: a survey of the literature. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **38**(3), 416–430 (2008)
9. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)
10. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 394–410 (2007)
11. Chandramouli, K., Kliegr, T., Svatek, V., Izquierdo, E.: Towards semantic tagging in collaborative environments. In: 16th International Conference on Digital Signal Processing 2009, pp. 1–6. IEEE, Piscataway (2009)
12. Chang, E., Goh, K., Sychay, G., Wu, G.: Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Trans. Circuits Syst. Video Technol. **13**(1), 26–38 (2003)

13. Cimiano, P., Voelker, J.: Text2onto – a framework for ontology learning and data-driven change discovery. In: NLDB 2005, Alicante (2005)
14. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pp. 708–716 (2007)
15. Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y.: Query expansion by mining user logs. IEEE Trans. Knowl. Data Eng. **15**(4), 829–839 (2003)
16. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. (CSUR) **40**(2), 5 (2008)
17. Deerwester, D.S., Fumas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. ACM Trans. Inf. Syst. **41**(6), 391–408 (2000)
18. Ding, G., Bai, S., Wang, B.: Local co-occurrence based query expansion for information retrieval. J. Chin. Inf. Process. **20**, 84–91 (2006)
19. Duygulu, P., Barnard, K., de Freitas, J., Forsyth, D.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: ECCV 2002, Copenhagen, pp. 349–354 (2002)
20. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT, Cambridge/London/England (1998)
21. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G.: Domain-specific keyphrase extraction. In: Proceeding of 16th International Joint Conference on Artificial Intelligence, Stockholm, pp. 668–673 (1999)
22. Gabrilovich, E., Markovich, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 07), Hyderabad (2007)
23. Gao, Y., Fan, J., Xue, X., Jain, R.: Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 901–910. ACM, New York (2006)
24. Gong, Z., Cheang, C.W., Hou, U.L.: Web query expansion by wordnet. In: DEXA 2005, Copenhagen. LNCS, vol. 3588, pp. 166–175 (2002)
25. Grootjen, T.P.: Conceptual query expansion. Data Knowl. Eng. **56**, 174–193 (2005)
26. Guillaumin, M., Mensink, T., Verbeek, J.: TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV, Kyoto, pp. 309–316 (2009)
27. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Fourteenth International Conference on Computational Linguistics, Nantes, pp. 539–545 (1992)
28. Hernández-Aranda, D., Granados, R., Cigarran, J., Rodrigo, A., Fresno, V., Garcıa-Serrano, A.: UNED at mediaeval 2010: exploiting text metadata for automatic video tagging. In: MediaEval 2010 Workshop, Pisa (2010)
29. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 531–538. ACM, New York (2008)
30. Hoeber, O., Yang, X.-D., Yao, Y.: Conceptual query expansion. In: Proceedings of the Atlantic Web Intelligence Conference, Lodz (2005)
31. Hotho, A., Jaschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: search and ranking. In: Proceedings of ESWC 2006, Budva, pp. 411–426 (2006)
32. http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm
33. Kliegr, T.: Entity classification by bag of wikipedia articles. In: Proceedings of the 3rd Workshop on Ph.D. Students in Information and Knowledge Management, pp. 67–74. ACM, New York (2010)
34. Kliegr, T., Chandramouli, K., Nemrava, J., Svátek, V., Izquierdo, E.: Combining captions and visual analysis for image concept classification. In: MDM/KDD'08: Proceedings of the 9th International Workshop on Multimedia Data Mining. ACM, New York (2008)
35. Larson, M., Soleymani, M., Serdyukov, P., Murdock, V., Jones, G. (eds.): In: Working Notes Proceedings of the MediaEval 2010 Workshop, Pisa (2010)

36. Li, D., Cai, D.: A study of query extension based on query log analysis. In: Proceedings of the Fourth National Student Conference on Computational Linguistics (SWCL-2008). Knowledge Engineering Center, Shenyang Institute of Aeronautical Engineering, Shenyang, Limning, 110034 (2008)
37. Li, Q., Lu, S.C.Y.: Collaborative tagging applications and approaches. IEEE Multimed. **15**(3), pp. 14–21 (2008)
38. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. In: MM, Santa Barbara, pp. 911–920 (2006)
39. Li, X., Snoek, C.G.M., Worring, M.: Learning tag relevance by neighbor voting for social image retrieval. In: MIR, Vancouver, pp. 180–187 (2008)
40. Li, X., Snoek, C.G.M., Worring, M.: Annotating images by harnessing worldwide user-tagged photos. In: ICASSP, Taipei, pp. 3717–3720 (2009)
41. Lindstaedt, S., Mörzinger, R., Sorschag, R., Pammer, V., Thallinger, G.: Automatic image annotation using visual content and folksonomies. Multimed. Tools Appl. **42**(1), 97–113 (2009)
42. Liu, X., Bruce Croft, W.: Cluster-based retrieval using language models. In: The 2004 ACM 1-58113-881-4/04/0007, New York, NY, USA, 25–29 July 2004
43. Liu, S., Liu, F., Yu, C., Meng, W.: An effective approach to document retrieval via utilizing wordNet and recognizing phrases. In: Proceedings of the 27th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Sheffield (2004)
44. Liu, J., Wang, B., Li, M., Li, Z., Ma, W.Y., Lu, H., Ma, S.: Dual cross-media relevance model for image annotation. In: MM, Augsburg, pp. 605–614 (2007)
45. Mandel, M., Ellis, D.: A web-based game for collecting music metadata. In: ISMIR, Vienna (2007)
46. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT, Cambridge (1999)
47. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Position paper, tagging, taxonomy, flickr, article, toRead. In: Proceedings of the 17th Conference on Hypertext and Hypermedia, Odense, pp. 31–40. ACM, New York (2006)
48. Milne, D., Witten, I.H.: Witten An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: David, M., Ian, H. Advancement of Artificial Intelligence, Chicago, USA (2008)
49. Mittal, N., Nayak, R., Govil, M.C., Jain, K.C.: Dynamic query expansion for efficient information retrieval. In: The Proceedings of International Conference on Web Information Systems and Mining, Sanya (2010)
50. Moehrmann, J., Bernstein, S., Schlegel, T., Werner, G., Heidemann, G.: Improving the usability of hierarchical representations for interactively labeling large image data sets. In: Jacko, J. (ed.) Human-Computer Interaction, Design and Development Approaches. Lecture Notes in Computer Science, vol. 6761, pp. 618–627. Springer, Berlin/New York (2011)
51. Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: MM, Berkeley, pp. 275–278 (2003)
52. Nemeth, Y., Shapira, B., Taeib-Maimon, M.: Evaluation of the real and perceived value of automatic and interactive query expansion. In: SIGIR '04, Sheffield, pp. 526–527 (2006)
53. Nemrava, J.: Refining search queries using wordnet glosses. In: EKAW 2006, Podebrady, pp. 2–6 (2006)
54. Paltoglou, G.: A study of information retrieval weighting schemes for sentiment analysis. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, vol. 11–16, pp. 1386–1395 (2010)
55. Qiu, Y., Frei, H.-P.: Concept based query expansion. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 160–169. ACM, Pittsburgh (1993)
56. Rendle, S., Schmidt-Thieme, L.: Pairwise interaction tensor factorization for personalized tag recommendation. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 81–90. ACM, New York (2010)

57. Richardson, R., Smeaton, A.F.: Using wordNet in a knowledge-based approach to information retrieval. In: Proceedings of the BCS-IRSG Colloquium, Crewe (1995)
58. San Pedro, J., Siersdorfer, S., Sanderson, M.: Content redundancy in YouTube and its application to video Tagging. ACM Trans. Inf. Syst. **29**(3), 13:1–13:31 (2011)
59. Seneviratne, L., Izquierdo, E.: An interactive framework for image annotation through gaming. In: MIR, Philadelphia, pp. 517–526 (2010)
60. Shapira, B., Taieb-Maimon, M., Nemeth, Y.: Subjective and objective evaluation of interactive and automatic query expansion. In: Online Information Review, pp. 374–390. Emerald, Bradford (2005)
61. Siersdorfer, S., San Pedro, J., Sanderson, M.: Automatic video tagging using content redundancy. In: SIGIR 2009, Boston, pp. 395–402 (2009)
62. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
63. Snoek, C.G.M., Worring, M.: Concept-based video retrieval. Found. Trends Inf. Retr. **2**(4), 215–322 (2008)
64. Snow, R., Jurafsky, D., Ng, A.: Learning syntactic patterns for automatic hypernym discovery. In: NIPS. Morgan Kaufmann, San Mateo (2005)
65. Strube, M., Ponzetto, S.P.: WikiRelate! computing semantic relatedness using wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), Boston, pp. 1419–1424 (2006)
66. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW 2007: 16th International World Wide Web Conference. ACM, New York (2007)
67. Sun, A., Bhowmick, S.S.: Image tag clarity: in search of visual-representative tags for social images. In: WSM, Beijing, pp. 19–26 (2009)
68. Tingle, D., Kim, Y.E., Turnbull, D.: Exploring automatic music annotation with acoustically-objective tags. In: MIR, Philadelphia, pp. 55–62 (2010)
69. Turnbull, D., Liu, R., Barrington, L., Lanckriet, G.: A game-based approach for collecting semantic annotations of music. In: ISMIR, Vienna (2007)
70. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Semantic annotation and retrieval of music and sound effects. IEEE Trans. Audio Speech Lang. Process. **2**(16), 467–476 (2008)
71. Ulges, A., Schulze, C., Koch, M., Breuel, T.M.: Learning automatic concept detectors from online video. Comput. Vis. Image Underst. **114**(4), 429–438 (2010)
72. Ulges, A., Worring, M., Breuel, T.: Learning visual contexts for image annotation from flickr groups. IEEE Trans. Multimed. **13**(2), 330–341 (2011)
73. Varelas, G., Voutsakis, E., Raftopoulou, P.: Semantic similarity methods in wordNet and their application to information retrieval on the web. In: 7th ACM International Workshop on Web Information and Data Management, Bremen (2005)
74. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: CHI, Vienna, pp. 319–326 (2004)
75. Wang, M., Yang, K., Hua, X.S., Zhang, H.J.: Visual tag dictionary: interpreting tags with visual words. In: WSCM, New York, NY, USA, pp. 1–8 (2009)
76. Wang, Z., Li, X., Xu, R.: Multi-keywords query expansion with OLCA based concept tree pruning. Comput. Sci. **37**(4), 132 (2010)
77. Wartena, C.: Using a divergence model for mediaeval tagging task. In: MediaEval 2010 Workshop, Pisa (2010)
78. Wen, N.J., Zhang, H.J.: Clustering user queries of a search engine. In: Proceedings of the 10th International World Wide Web Conference (WWW10), Hong Kong (2001)
79. Wen, J., Cui, H., Li, M.: A statistical query expansion model based on query logs. J. Softw. **14**(9), 1593–1599 (2003)
80. Wu, X., Zhang, L., Yu, Y.: Exploring social annotations for the semantic web. In: Proceedings of WWW06, Edinburgh, pp. 417–426 (2006)
81. Wu, L., Yang, L., Hua, X.S., Yu, N.: Learning to tag. In: WWW, Madrid, pp. 361–370 (2009)

82. Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y.: Exploring folksonomy for personalized search. In: Proceedings of ACM SIGIR, Singapore, pp. 155–162 (2008)
83. Yan, X., Huang, M., Zhang, S.: Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining. Inst. Softw. Chin. Acad. Sci. **20**, 1854–1865 (2009)
84. Zhang, J., Deng, B., Li, X.: Concept based query expansion using wordNet. In: AST '09 Proceedings of the 2009 International e-Conference on Advanced Science and Technology, Daejeon, pp 52–55 (2009)

A Rule-Based Flickr Tag Recommendation System

Luca Cagliero, Alessandro Fiori, and Luigi Grimaudo

Abstract Personalized tag recommendation focuses on helping users find desirable keywords (tags) to annotate Web resources based on both user profiles and main resource characteristics. Flickr is a popular online photo service whose resource sharing system significantly relies on annotations. However, recommending tags to a Flickr user who is annotating a photo is a challenging task as the lack of a controlled tag vocabulary makes the annotation history collection very sparse.

This chapter presents a novel rule-based personalized tag recommendation system to suggest additional pertinent tags to partially annotated resources. Rules represent potentially valuable correlations among tag sets. Intuitively, the system should recommend tags highly correlated with the previously annotated tags. Unlike previous rule-based approaches, a WordNet taxonomy is used to drive the rule mining process and discover rules, called generalized rules, that may contain either single tags or their semantically meaningful aggregations. The use of generalized rules in tag recommendation makes the system (1) more robust to data sparsity and (2) able to capture different viewpoints of the analyzed data. Experiments demonstrate the usefulness of generalized rules in recommending additional tags for real photos published on Flickr.

1 Introduction

Social networks and online communities allow creating and managing annotations to categorize and index the resources published by the community users. Tags are keywords that provide meaningful descriptors of the Web resources (e.g., bookmarks, photo, academic articles). Their usefulness has been recently demonstrated in a number of research contexts, e.g., Web content indexing [3],

L. Cagliero (✉) • A. Fiori • L. Grimaudo
Politecnico di Torino. Corso Duca degli Abruzzi, 24 10129 Torino, Italy
e-mail: luca.cagliero@polito.it; alessandro.fiori@polito.it; luigi.grimaudo@polito.it

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_8, © Springer-Verlag London 2013

multimedia data retrieval [10], and enterprise Web searches [12]. When dealing with multimedia content (e.g., images or videos) for which the amount of available textual content is limited, tags play an even more important role in Web searching and browsing.

The popularity of Web tagging sites (e.g., Delicious [11], Flickr [14], and Zooomr [36]) has prompted the need of novel and more effective recommendation systems to support users in resource annotation by suggesting novel and pertinent tags. Recommendations may be either *personalized*, i.e., dependent on the user who is tagging the resource, or *collective* (user-independent). A significant research effort has been devoted to addressing personalized tag recommendation for Flickr photos [15,25,29]. However, the lack of a controlled vocabulary from which tags could be selected during the annotation process makes the set of previously assigned annotations very sparse. Thus, results provided by the mostly used information retrieval or data mining techniques may become unreliable. Indeed, the problem of personalized Flickr tag recommendation is a challenging task.

This chapter presents a novel personalized Flickr tag recommendation system that suggests additional tags to partially annotated Flickr photos. To this aim, it discovers strong generalized association rules from the personal and the community-based sets of past annotations and exploits them to support the process of tag recommendation. An association rule [2] is an implication $A \Rightarrow B$, where A and B are itemsets (sets of items) named, respectively, as rule antecedent and consequent. In the context of tag recommendations, any item is associated with a distinct tag assigned by a user, while each transaction belonging to the source data is associated with a specific annotation made by user to a given photo and is composed of a set of tags. A rule $A \Rightarrow B$ may be used to recommend one or more tags contained in B if the given photo has already been annotated with the tags in A. A WordNet taxonomy is used to define a hierarchy of is-a or is-part-of aggregations built over the tags occurring in the source data. For instance, based on the WordNet taxonomy, a tag (e.g., *Rome*) may be generalized as its corresponding geographical aggregation *Italy*, while *Europe* may be considered a generalization of both of them. Items relative to aggregated values (e.g., *Italy*) are also called generalized items and denote high-level tag generalizations. The generalization level of a node (i.e., a tag or an aggregation) indicates the length of the path on the taxonomy from the node to a leaf. For instance, in the above-mentioned example, *Italy* has generalization level 1 while *Europe* has level 2. This chapter proposes to exploit WordNet taxonomies to drive the rule mining process and discover rules $A \Rightarrow B$, called generalized rules [30], in which itemsets A and B may include either single tags (items) or their aggregations (generalized items). To make the generalized rule extraction problem tractable in real-life cases, only a subset of all the possible generalized rules is usually extracted. Selected generalized rules $A \Rightarrow B$ are characterized by the following properties: (a) the observed frequency of occurrence of the itemset $A \cup B$ in the analyzed data, called support, is above a given threshold and (b) the conditional probability of occurrence $P(B|A)$ in the source data, called confidence, is above a given threshold. A (generalized) rule that satisfies (a) is said to be *frequent*. Differently, a (generalized) rule that satisfies both (a) and (b) is said to be *strong*. The use of

generalized rules may allow effective coping with sparse data collections as well as provides different viewpoints of the analyzed data, as shown in the following with the help of a toy example.

Consider, as running example, a photo of the Colosseum, the famous Roman amphitheater situated in the center of the city of Rome (Italy). An example of not generalized association rule may be *Rome* → *Colosseum*, where *Rome* and *Colosseum* are tag examples. If the user has already annotated the photo with *Rome*, *Colosseum* is an example of subsequent tag to recommend. However, if the collection is very sparse, the rule is likely to be infrequent in the collection of the past annotations, and thus it is not extracted. The use of a taxonomy that generalizes the tag *Rome* as the corresponding state *Italy* may allow the extraction of a generalized rule *Italy* → *Colosseum* that suggests the same annotation while considering a higher-level viewpoint the latter tag correlation.

To select the most relevant tags to recommend, two distinct rule sets are generated: (1) a personalized rule set, which includes the generalized rules extracted from the past annotations made by the user to which the recommendation is targeted, and (2) a community-based rule set, which includes the generalized rules mined from the past annotations made by the community. Tags contained in the consequents of the selected rules are ranked based on the confidence value of the corresponding rules. The ranking process is driven by a newly proposed metrics that weighs differently the rules extracted from the personal and community-based collections. Experiments, reported in Sect. 4, show that the best tag recommendation performance were achieved when giving higher importance to the confidence of the rules extracted from the personal annotations than those mined from the community-based annotations.

The effectiveness of the proposed system has been validated on a real photo collection retrieved from Flickr. The use of generalized rules allows improving the performance of the state-of-the-art approaches.

This chapter is organized as follows. Section 2 overviews the most relevant related works concerning tag recommendation and generalization rule mining. Section 3 presents the framework of the proposed recommender system and describes its main blocks. Section 4 assesses the effectiveness of the system in providing personalized tag recommendations, while Sect. 5 draws conclusions and presents future developments of this work.

2 Previous Work

A recommender system helps users find desirable products or services by analyzing user interests and behaviors. Overviews of the most recently proposed recommendation systems are given in [1, 21, 26]. In the last years, a relevant effort has been devoted to the development of novel and more effective tag recommendation systems. This chapter specifically addresses the issue of personalized tag recommendation by means of generalized association rules. In the following, we present

and compare the main state-of-the-art works concerning tag recommendation (see Sect. 2.1) and generalized association rule mining (see Sect. 2.2) with the proposed approach.

2.1 Tag Recommendation

The popularity of social networks and online communities (e.g., Delicious [11], Flickr [14], and Zooomr [36]) has increased the attention to the problem of recommending Web resource annotations, i.e., the tags. More specifically, tag recommendation is focused on suggesting pertinent tags to users who are annotating a Web resource. The suggestion may be either personalized, i.e., dependent on the user who is annotating the resources or not, i.e., exclusively based on the collective knowledge.

Several approaches have been proposed to address personalized tag recommendation. For instance, content-based filtering methods [9,22] focus on recommending tags that are similar to those that a user annotated in the past (or is annotating in the present). They commonly analyze the characteristics of the recommended tags to generate detailed user profiles. Tag relevance and user similarity are commonly evaluated by exploiting information retrieval or data mining techniques. In [22], tags for Delicious bookmarks are recommended by evaluating the cosine similarity among tags and by considering both the cases in which prior tag information is available or not. Differently, in [9], the authors present an application for large-scale automatic generation of personalized annotation tags. They propose an algorithm, named P-TAG, that automatically extracts personalized keywords as tags from bookmarked Web page contents to generate personalized Web document recommendations. Tags are selected based on their relevance to the textual content of the target Web page as well as to the documents residing on the surfer's Desktop.

The use of collaborative filtering approaches in personalized tag recommendation has been addressed in [20, 23, 28]. They collect and analyze a large amount of information on user behaviors, activities, or preferences to predict what users will like based on their similarity to other user features. To this aim, they commonly rely on the assumption that similar users share similar tastes. For instance, in [23], the authors address post tag recommendation by combining, similar to [28], a collaborative filtering method with information retrieval techniques for evaluating the similarities between posts, users, and tags. A hybrid approach that combines a collaborative filtering method with a content-based analysis is proposed in [20]. This system tunes its parameters based on the user feedback to better suit the user preferences. Differently, the combined usage of collaborative filtering and graph-based indexing algorithms is addressed in [18, 33]. In particular, in [18], a user-resource-tag (URT) graph is analyzed by means of an ad hoc indexing strategy derived from the popular PageRank algorithm [7], while in [33], singular value decomposition (SVD) methods are applied to reduce the sparsity of the generated graphs. In [15], an interactive approach to Flickr tag recommendation is proposed.

Suggested tags are first selected according to the set of previously assigned tags based on co-occurrence measures. Next, based on the suggestion, the candidate set is narrowed down to make the suggestion more specific. To overcome challenges of co-occurrence and graph-based measures due to the sparsity of the analyzed data, this chapter proposes to exploit associations at different abstraction levels.

A parallel research issue has been devoted to addressing the problem of collective tag recommendation [17,19,29]. The most commonly used approaches are based on co-occurrence measures. For instance, authors in [29] propose a Flickr tag recommendation system that analyzes tag co-occurrences in the collective past annotation collection to suggest additional tags to partially annotated resources. Authors in [25] extend the previous approach to the context of personalized recommendation by combining the knowledge coming from different contextual layers (i.e., personal, collective, and group levels). Differently, the system presented in [19] specifically tackles the cold start problem, i.e., the annotation of not previously annotated resources, by using latent dirichlet allocation (LDA). Unlike [19], this chapter specifically addresses, similar to [25,29], the task of tag recommendation to partially annotated Flickr photos by using generalized rules instead of traditional rules or co-occurrence measures. Authors in [17] reformulate the task of content-based tag recommendation as a (supervised) classification problem. Using page text, anchor text, surrounding hosts, and available tag information as training data, they build a classifier for each tag they want to predict. The main drawback of the proposed approach is that the overall training time may become very high when the cardinality of the considered tags increases. In the same work, the use of association rules in tag recommendation has been also proposed. Unlike [17], this chapter proposes to overcome the limitations of traditional association rules in coping with sparse data collections by aggregating tags at different abstraction levels according to the given generalization hierarchies.

2.2 Generalized Association Rule Mining

Association rule mining is a widely used exploratory data mining technique, introduced in [2] in the context of market basket analysis, to discover valuable correlations among data. To focus on rules that are relatively strong, i.e., the ones that frequently occur in the source data and hold in most cases, the mining phase is commonly driven by two main rule quality indexes, i.e., the support and the confidence indexes. However, in some cases, this approach is not effective in discovering relevant data recurrences due to the excessive level of detail of the hidden information. Generalized rules have been first introduced in [30] to address rule mining in the presence of taxonomy. By evaluating a taxonomy built over the data items, items are aggregated into higher-level (generalized) concepts. Each generalized itemset is a high-level representation of a set of "lower level" itemsets according to the given taxonomy. The first generalized association rule mining algorithm [30] generates itemsets by considering, for each item, all its parents in

a taxonomy. Hence, candidate frequent itemsets are generated by exhaustively evaluating the taxonomy. To reduce the mining complexity, several optimizations have been proposed (e.g., [4, 16, 24, 31, 32]). Furthermore, the application of generalized itemsets or rules in different application contexts has been recently investigated as well (e.g., network traffic analysis [4, 6], context-aware systems [5, 8]). Unlike any previously mentioned approaches, this chapter proposes to exploit generalized rules to accomplish the personalized tag recommendation task.

3 The Rule-Based Recommendation System

This chapter presents a novel personalized tag recommendation system. Given a photo and a set of user-defined tags, the system proposes novel pertinent tags based on both the personal user preferences, i.e., the tags already annotated by the same user, and the community-based knowledge, i.e., the annotations provided by the other users. Its main architectural blocks are shown in Fig. 1. A brief description of each block follows.

Tag set data representation. This block aims at making the history collection of the previously assigned tags suitable for the rule mining process. The tag set is tailored to a transactional data format, where each transaction corresponds to an

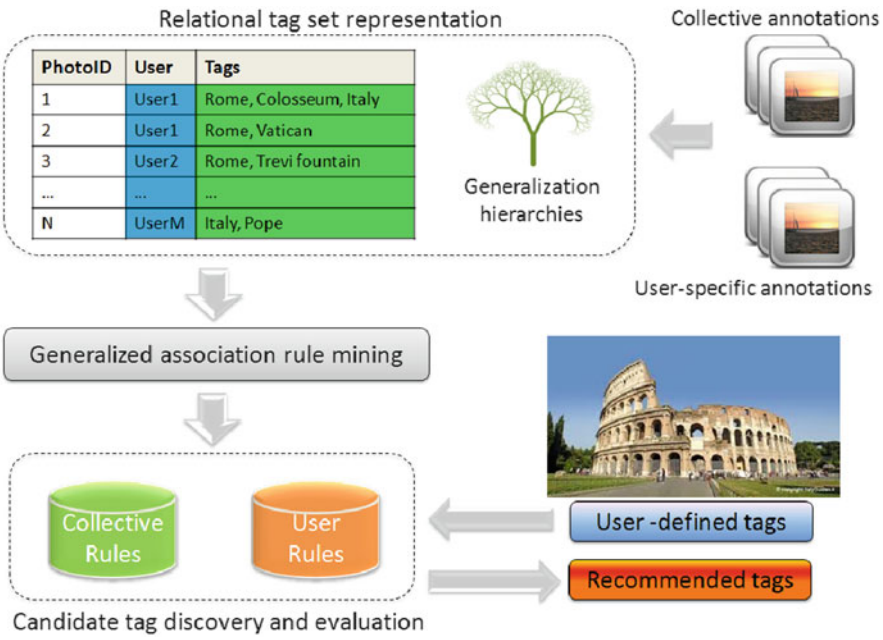


Fig. 1 The recommendation system architecture

annotation performed by a user to a given photo and includes the corresponding set of assigned tags. To allow generalizing tags as the corresponding high-level categories, a set of hierarchies of aggregations (i.e., the generalization hierarchies) built over the analyzed tags is derived from the WordNet [30] lexical database.

Generalized association rule mining. This block focuses on discovering high-level correlations, in the form of generalized association rules, from the transactional representation of the past annotation collection. The tag generalization hierarchies are exploited to aggregate tags at higher abstraction levels. Two distinct generalized rule sets are generated: (1) a personalized rule set, which includes the generalized rules extracted from the past annotations made by the user to which the recommendation is targeted and (2) a community-based rule set, which includes the generalized rules mined from the collective knowledge (i.e., the photo annotations of the other users).

Candidate tag discovery and evaluation. Given a photo and a set of tags already assigned by the user, this block aims at generating a ranked list of suggested tags. To this aim, the selection of the tags pertinent to the previously annotated ones is driven by the personalized and community-based rules.

This section is organized as follows. Section 3.1 formally states the recommendation task addressed by this chapter, while Sects. 3.2, 3.3, and 3.4 thoroughly describe the main blocks of the recommendation system separately.

3.1 Problem Statement

Given a set of photos P, a set of tags T, and a set of users U, the ternary relation $X = P \times T \times U$ represents the user-specific assignments of tags in T to photos in P. We denote as $\mathcal{T}(p_i,u_j) \subseteq T$ the set of tags assigned by the user u_j to an arbitrary photo u_j, where $u_j \in U$ and $p_i \in P$. It could be defined as follows:

$$\mathcal{T}(\mathrm{p_i}, \mathrm{u_j}) = \pi_{\mathrm{t}} \sigma_{\mathrm{p_i,u_j}} X \tag{1}$$

where π and σ are the commonly used projection and selection primitive operators of the relational algebra [13].

To discriminate between past assignments made by user u_j and those made by the other users, the ternary relation X may be partitioned as follows:

$$X(u_j) = \pi_t \sigma_{u_j} X \tag{2}$$

$$X(\neg u_j) = \pi_t \sigma_{U \setminus u_j} X \tag{3}$$

Given a set $\tau(p_i,u_j)$ of user-defined tags and the personal and community-based knowledge $X(u_j)$ and $X(\neg u_j)$, the Flickr personalized tag recommendation task addressed by this chapter focuses on suggesting to user u_j new tags in $T \setminus \tau(p_i,u_j)$ for a photo p_i.

3.2 Tag Set Data Representation

Flickr [14] is an online photo-sharing system whose resources are commonly annotated by the system users. The use of association rule mining techniques is particularly suitable for discovering correlations hidden in large real-life tag annotation collections [17]. This chapter investigates the use of an established data mining technique, i.e., generalized association rule mining [30], in recommending tags to partially annotated Flickr photos. Since, in real cases, photo annotations are commonly unsuitable for being directly analyzed by means of data mining algorithms, a preprocessing step is needed.

The collection of past Flickr photo annotations is tailored to a transactional data format. A transactional dataset is a set of transactions, where each transaction is a set of items of arbitrary size. To map the tag set to a transactional data format, the annotations made by a user to a given photo are considered as a transaction composed of the set of (not repeated) assigned tags. More formally, given a ternary relation X (Cf. Sect. 3.1), any tag set $\tau(p_i, u_j)$ generated from X is considered as a transaction. For instance, if the user u_j assigns to the photo p_i the tags *Colosseum* and *Rome*, the corresponding transaction becomes {*Colosseum, Rome*}. The transactional dataset D is the set of all distinct $\tau(p_i, u_j)$ occurring in X, i.e., the full list of the past annotations.

Given a user u_j to which the personalized tag recommendation is targeted, the transactional dataset D is partitioned between the annotations made by u_j and not denoted as $D(u_j)$ and $D(\neg u_j)$ are generated. The separate analysis of $D(u_j)$ and $D(\neg u_j)$ allows the discovery of both user-specific and collective recurrences, in the form of generalized rules.

To enable the process of generalized rule mining process, a WordNet taxonomy composed of set of hierarchies of aggregations (generalizations) over the tag set T is built. The WordNet lexical database [30] is queried to retrieve the most relevant semantic relationships holding between a tag in T and any other term. More specifically, hyponyms (i.e., is-a-subtype-of relationships) and meronyms (is-part-of relationships) are considered. All the terms that belong to these relationships are generalizations of the original tag. For instance, consider again the tag *Rome*. If the following semantic relationship is retrieved from the WordNet database

```
<Rome> <is-part-of> <Italy>
```

then the term *Italy* is selected as the upper level generalization (aggregation) of the tag *Rome*. Tag generalizations may be further aggregated into high-level categories. For instance, the semantic relationship $< Italy > < is-part-of > < Europe >$ prompts the selection of *Europe* as generalization of *Italy* and *Rome*. The generalization level of a tag (or a tag aggregation) is defined as the length of the path on the taxonomy hierarchy from the corresponding node to a leaf. Recalling the previous example, *Europe* has generalization level 2 as the path from *Europe* to *Rome* has length two. Differently, *Rome* has level 0 because it is already a leaf of the taxonomy. Notice that the generalization relationship of two tags (or tag sets)

holds even if they are not characterized by consecutive generalization levels. For instance, Europe is considered one of the possible generalizations of Rome as well. The generalization of an itemset (i.e., a tag set) is defined as the maximum among the levels of its items (tags). For instance, since {*Colosseum*, *Italy*} is composed of one item with level 0 and one item of level 1, its generalization level is 1.

3.3 Generalized Association Rule Mining

This block focuses on discovering high-level correlations among tags, in the form of generalized association rules, from the transactional representations of the tag sets $D(u_j)$ and $D(\neg u_j)$. Strong association rules represent implications among tags or tag sets that frequently occur and almost hold in the source data [2]. More specifically, an association rule is an implication $A \rightarrow B$, where A and B are itemsets (i.e., sets of data items). In the transactional representation of the tag set, items are tags in T associated with any photo included in the collection.

In the context of tag recommendation, generalized association rules [30] are association rules that may include either tags or their high-level aggregations, also denoted as generalized items. By considering the taxonomy built over the tag set (Cf. Sect. 3.2), any concept that aggregates one or more tags in T at a higher level of generalization is considered as a semantically meaningful tag aggregation. For instance, consider again the semantic relationship <*Rome*> <*is-part-of*> <*Italy*>. If *Rome* is a tag that occurs in the analyzed data, *Italy* is an example of tag aggregation (generalized item). Similarly, generalized itemsets are itemsets including at most one aggregation (e.g., {*Colosseum, Italy*}). Generalized itemsets are characterized by a notable quality index, i.e., the support, which is defined in terms of the itemset coverage with respect to the analyzed data. A generalized itemset I covers a given transaction $d \in D$ if all its (possibly generalized) items $x \in I$ are either included in d or ancestors (generalizations) of items $i \in d$. Given a transaction dataset D and a (generalized) itemset I, the support of I is given by the ratio between the number of transactions $d \in D$ covered by I and the cardinality of D.

The concept of generalized association rule extends the traditional association rules to the case in which they may include either generalized or not generalized itemsets. A generalized association rule is represented in the form $A \rightarrow B$, where A and B are two (generalized) itemsets that are named, respectively, as the body and the head of the rule. Similarly, A and B are also denoted as rule antecedent and consequent. Generalized association rule extraction is driven by rule support and confidence quality indexes. The support of a generalized rule is defined as the observed frequency of occurrence of $A \cup B$ in the source dataset. The confidence of a rule $A \rightarrow B$ is the conditional probability of occurrence of the generalized itemset B given A and represents the strength of the implication. For instance, the generalized association rule {*Colosseum* $\rightarrow$ *Italy*} characterized by support equal to 10 % and confidence equal to 88 % states that the tag *Colosseum* co-occurs with

the tag generalization *Italy* in 10 % of the transactions (photo annotations) of the collection and the implication holds in 88 % of the cases.

The generalized association rule mining task is usually accomplished by means of a two-step process [30]: (1) generalized itemset mining, driven by a minimum support threshold *minsup* and (2) generalized association rule generation, starting from the set of previously extracted itemsets, driven by a minimum confidence threshold *minconf*. A generalized association rule is said to be *strong* if it satisfies both *minsup* and *minconf* [2].

Given a set of generalization hierarchies built over the tags in T, a minimum support threshold *minsup*, and a minimum confidence threshold *minconf*, the generalized rule mining process is performed on $D(u_j)$ and $D(\neg u_j)$ separately. More specifically, given a photo p_i, a user u_j, and a set of user-specific tags $\tau(p_i,u_j)$, the main idea behind our approach is to treat strong high-level correlations related to annotations made by the user u_j differently from that made by the other users. To this aim, two distinct rule sets are generated: (1) a personalized rule set $R_{D(u_j)}$, which includes the strong generalized rules extracted from the past annotations made by the user to which the recommendation is targeted and (2) a community-based rule set $R_{D(\neg u_j)}$, which includes all the strong generalized rules mined from the past annotations made by the other users.

To perform generalized rule mining from the tag history collections, we exploit our more efficient implementation of the Cumulate algorithm [30]. However, different algorithms may be easily integrated as well.

3.4 Candidate Tag Discovery and Evaluation

Given a photo p_i, a set of user-defined tags $\tau(p_i,u_j)$ already assigned by user u_j, and the sets $R_{D(u_j)}$ and $R_{D(\neg u_j)}$ of generalized rules mined, respectively, from $D(u_j)$ and $D(\neg u_j)$, this block entails the selection and the ranking of the tags to recommend to u_j for p_i. In the following, we discuss how to tackle the candidate tag selection and ranking problems separately.

3.4.1 Candidate Tag Selection

The selection step focuses on identifying additional tags, pertinent to the user-specified tag set $\tau(p_i,u_j)$, based on the previously generated rule sets. To guarantee the pertinence of the candidate tags, for each photo p_i only the subset of the personalized and community-based rules including tags in $\tau(p_i,u_j)$ in their ancedent are considered. More specifically, the candidate tag selection step exclusively considers the strong generalized rules in $R_{D(u_j)}$ and $R_{D(\neg u_j)}$ whose (1) rule antecedent exactly covers, at any level of abstraction, the tag set $\tau(p_i,u_j)$ (or any of its subsets) and (2) rule consequent includes an arbitrary set of not generalized

Table 1 Generalized rules used for recommending to user u_j tags subsequent to *Rome*

ID	Generalized rule	Support (%)	Confidence (%)
Annotations made by user u_j ($R_{D(u_j)}$)			
1	$\{Rome\} \rightarrow \{Colosseum\}$	25	100
2	$\{Rome\} \rightarrow \{History\}$	17	85
Annotations made by the other users ($R_{D(\neg u_j)}$)			
3	$\{Rome\} \rightarrow \{Colosseum, Gladiator\}$	15	95
4	$\{Italy\} \rightarrow \{Roman\ Age\}$	23	80

items (tags). The coverage of a tag in $\tau(p_i,u_j)$ may be due to the presence in the rule antecedent of either an exact matching (i.e., the same tag) or one of its higher-level generalizations. Any rule that does not fulfill the above-mentioned constraints is not considered in the subsequent ranking process. The set of tags that occur in the selected rule consequents is chosen as set of candidate recommendable tags.

Consider, for instance, a photo p_i annotated by user u_j with the tag *Rome*. In Table 1 is reported the selection of generalized rules, fulfilling the above-mentioned constraints, that has been taken from the set of rules mined from the personalized collection $D(u_j)$ and community-based one $D(\neg u_j)$. In this example, we exploit the generalization hierarchies described in Sect. 3.2, and we enforce, respectively, a minimum support threshold equal to 15 % and a minimum confidence threshold equal to 50 %.

Readers could notice that any selected rule must have (1) as rule antecedent, either the user-specified tag *Rome* or its generalization *Italy*, and (2) as rule consequent, an arbitrary set of (not generalized) tags. Tags occurring in the rule consequents include the potentially relevant tags to recommend. Recalling the previous example, the set C of candidate tags is {*Colosseum*, *History*, *Gladiator*, *Roman Age*}. Notice that a single rule may include one or more candidate tags (e.g., *Colosseum* and *Gladiator* co-occur in the consequent of the rule (3)).

The generalization process prevents the discarding of potentially relevant knowledge. In fact, it allows also recommending tags contained in the consequent of rules having as antecedent a generalization of the user-defined tags. For instance, the generalized rule $Italy \rightarrow Roman\ Age$ mined from $R_{D(\neg u_j)}$ suggests to recommend the tag *Roman Age* subsequently to *Rome*. Indeed, even if the rule $Rome \rightarrow Roman\ Age$ is infrequent with respect to the minimum support threshold in $R_{D(\neg u_j)}$ (possibly because of the sparsity of the personal annotation collection), the co-occurrence between *Roman Age* and *Rome* does not remain hidden.

Consider now the case in which the set of user-specified tags $\tau(p_i,u_j)$ is {*Rome*, *Roman Empire*}. Rules including as antecedent the tag set {*Rome*, *Roman Empire*} or any of its subsets belonging to any abstraction level (e.g., *Rome*, *Italy*) are deemed worth considering in the selection of the candidate tags. For instance, $Italy, Roman\ Empire \rightarrow Roman\ Age$ may be considered to recommend the tag *Roman Age* as well.

3.4.2 Candidate Tag Ranking

The last but not the least task in tag recommendation is the ranking of the candidate tags in C to recommend to u_j for p_i. The tag ranking should reflect (1) their significance with respect to the user-defined tags, (2) their relevance according to the personal user preferences, and (3) their relevance based on the collective knowledge.

To take the correlation with the previously annotated tags into account, we propose a ranking strategy that evaluates the candidate tags in terms of the interestingness of the rules in $R_{D(u_j)}$ and $R_{D(\neg u_j)}$ from which they have been selected. Generalized rule interestingness is evaluated in terms of its confidence index value [2], i.e., the rule strength in the analyzed dataset (see Sect. 3.3). Based on the assumption that personal recommendations frequently assigned by user u_j might be weighted differently from that made by the other users, we evaluate the contribution of each rule set separately and then we properly combine the resulting scores.

More formally, let $c \in C$ be an arbitrary candidate tag and $R^c_{D(u_j)} \subseteq R_{D(u_j)}$ and $R^c_{D(\neg u_j)} \subseteq R_{D(\neg u_j)}$ be, respectively, the subsets of rules in $R_{D(u_j)}$ and $R_{D(\neg u_j)}$ whose antecedent covers c (at any level of abstraction). The ranking score of c is defined as follows:

$$\text{rankscore(c)} = \lambda \cdot \frac{\sum_{r_{u_j} \in R^c_{D(u_j)}} \text{conf}(r_{u_j})}{|R^c_{D(u_j)}|} + (1-\lambda) \cdot \frac{\sum_{r_{\neg u_j} \in R^c_{D(\neg u_j)}} \text{conf}(r_{\neg u_j})}{|R^c_{D(\neg u_j)}|} \quad (4)$$

where $\lambda \in [0,1]$ is a user-provided algorithm parameter.

Relatively speaking, when $\lambda > 0.5$ the impact of the confidence value of the rules occurring in $R^c_{D(u_j)}$ is higher than that in $R^c_{D(\neg u_j)}$, i.e., preferences given by user u_j are deemed more significant than those given by the other users. Oppositely, in case $\lambda < 0.5$ user u_j preferences are averagely penalized. An analysis of the impact of λ on the performance of the proposed recommendation system is reported in Sect. 4.

The recommendation system returns the set C of selected tags sorted by the ranking score reported in Eq. 4.

4 Experimental Results

We performed a set of experiments addressing the following issues: (1) a performance comparison between our system and a set of recently proposed approach, (2) the impact of the generalization process on the recommendation process, and (3) the analysis of the recommendation system parameters.

This section is organized as follows. Section 4.1 describes the characteristics of the photo collection exploited in the experimental evaluation. Section 4.2 describes the experimental design and introduces the evaluation metrics adopted for the performance evaluation. Section 4.3 compares the results achieved by our system

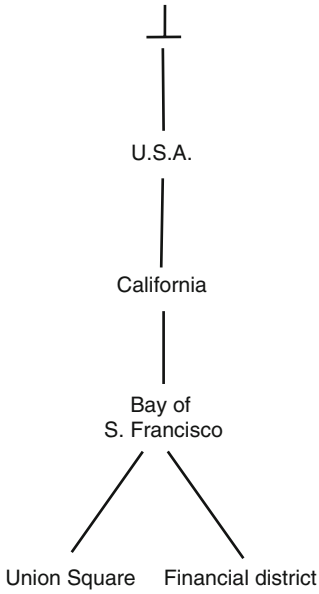


Fig. 2 Portion of an example generalization hierarchy built over the photo collection tags

with that achieved by the system presented in [25] and a baseline version of our approach that does not exploit generalized rules. Finally, Sect. 4.4 analyzes the impact of the system parameters on the recommendation performance.

4.1 Photo Collection

To evaluate the performance of our approach we retrieved, by means of the Flickr APIs, 2,300 real photos, each one annotated with at least 5 tags by a set of 30 users. The selected photos were chosen based on a series of high-level geographical topics, i.e., *New York*, *San Francisco*, *London*, and *Vancouver*. By following the strategy described in Sect. 3.2, a set of generalization hierarchies is derived from the WordNet lexical database over the collected photo tags. A portion of one of the generated generalization hierarchies is reported in Fig. 2.

To evaluate the effectiveness of our system in coping with heterogeneous photo annotations, the considered photo collection is ensured to be unevenly distributed among the analyzed upper level tag categories.

4.2 Experimental Design

Since our system retrieves a ranked list of pertinent additional tags based on the extracted frequent generalized rules, we defined the tag recommendation task as a ranking problem. Given a photo p_i and a set of user-defined tags $\tau(p_i,u_j)$,

the system has to recommend tags that describe the photo based on both the user-specific and the collective past annotations. To perform personalized recommendation, from the whole photo collection, the user-specific annotations made by 10 users who annotated at least 15 photos are considered separately. The above selection allows making the statistical evaluation of our recommendation system reliable. Once a user-specific annotation subset is selected, the rest of the collection is considered as the collective set. For each analyzed user collection, the evaluation process performs an hold-out train-test validation, i.e., the user-specific collection is partitioned in a training set, including the 75 % of the whole annotations, whereas the remaining part is chosen as test set. To evaluate the additional tag recommendation performance of our system, for each test photo, two random tags are selected as initial (user-specified) tag set and the recommended tag list is compared with the held-out test tags. A recommended tag is judged as correct if it is present in the held-out set. Since the held-out tags need not be the only tags that could be assigned to the photo, the evaluation method actually gives a lower bound on the system performance.

To evaluate the performance of both our recommendation system and its considered competitors, we exploited three standard information retrieval metrics, previously adopted in [25, 29] in the context of additional Flickr tag recommendation. The selected measures are deemed suitable for evaluating the system performance at different aspects. Let Q be the set of relevant tags, i.e. the tags really assigned by the user to the test photo, and C the tag set recommended by the system under evaluation. The adopted evaluation measures are defined as follows.

Mean reciprocal rank (MRR). This measure captures the ability of the system to return a relevant tag (i.e., a held-out tag) at the top of the ranking. The measure is averaged over all the photos in the testing collection and is computed by:

$$\text{MRR} = \max_{q \in Q} \frac{1}{c_q} \tag{5}$$

where c_q is the rank achieved by the relevant tag q.

Success at rank k (S@k). This measure evaluates the probability of finding a relevant tag among the top-k recommended tags. It is averaged over all the test photos and is defined as follows:

$$S@k = \begin{cases} 1 & \text{if } q \in C_k, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $q \in Q$ is a relevant tag and C_k is the set of the top-k recommended tags.

Precision at rank k (P@k). This metric evaluates the percentage of relevant tags among the set of retrieved ones. The measure, averaged over all test photos, is defined as follows:

$$P@k = \frac{|Q \cap C_k|}{|Q|} \tag{7}$$

For any evaluated measure, the estimates on each test photo are averaged over ten runs, where, within each run, a different (randomly generated) held-out tag set ranking is considered.

4.3 Performance Comparison

The aim of this section is twofold. Firstly, it experimentally demonstrates the effectiveness of our system against a state-of-the-art approach. Secondly, it evaluates the impact of the generalization process on the recommendation performance. To achieve these goals, we compared the performance of our system, in terms of the evaluation metrics described in Sect. 4.2, with (1) a recently proposed personalized Flickr tag recommendation system [25] and (2) a baseline version of our approach, which does not exploit generalized rules.

The system presented in [25] is a personalized recommender system that proposes additional photo tags pertinent to a number of different user contexts, among which the personal and the collective ones. The system generates a list of recommendable tags based on a probabilistic co-occurrence measure for each context. Then, it aggregates the results achieved within each context in a final recommended list by exploiting the Borda count group consensus function [34]. To the best of our knowledge, it is the most recent work proposed on the topic of personalized additional tag recommendation. To perform a fair comparison, we evaluated the performance of our implementation of the approach presented in [25] (denoted as *probabilistic prediction* in the following) when coping with the combination of the collective and the personalized contexts.

To demonstrate the usefulness of generalized rules in tag recommendation, we also compared the performance of our system with that of a baseline version, which exploits traditional (not generalized) association rules [2] solely. More specifically, the baseline method performs the same steps of the proposed approach, while disregarding the use of tag generalizations in discovering significant tag associations (see Sect. 3.4.1).

To test the performance of our approach, we consider as standard configuration the following setting: minimum support threshold $minsup = 30\%$, minimum confidence threshold $minconf = 40\,\%$, and $\lambda = 0.75$. A more detailed analysis of the impact of these parameters on the recommendation system performance is reported in Sect. 4.4. Even for the baseline version of our system we tested several support and confidence threshold values. For the sake of brevity, in the following we select as representative the configuration that achieved the best results in terms of MMR measure, i.e., minimum support threshold equal to 30 % and minimum confidence threshold equal to 40 %.

Table 2 Performance comparison in terms of S@1, P@1, S@5, P@5, and MRR metrics

System	S@1–P@1	S@5	P@5	MRR
GR-TAG	*0.7392*	*0.8695*	*0.5696*	*0.7935*
Baseline rule-based	0.7174*	0.8261*	0.4957*	0.7572*
Probabilistic prediction	0.6956*	0.8478	0.4435*	0.7681*

Statistically relevant worsening in the comparisons between our system and the other approaches is starred

The achieved results are summarized in Table 2. In particular, the success and the precision at ranks 1 and 5 (i.e., S@1, P@1, S@5, and P@5, respectively) as well as the mean reciprocal rank (MRR) achieved by both our system (named *GR-TAG*) and all the tested competitors are reported. The selected ranks (k) for the precision at rank k and the success at rank k are chosen analogously to what was previously done in [25,29]. To validate the statistical significance of the achieved performance improvements, the student t-test has been adopted [27] by using as p-value 0.05. Significant worsening in the comparisons between our system and the other tested competitors are starred in Table 2. For each tested measure, the result(s) of the best system(s) is written in boldface.

Our recommendation system outperforms both its baseline version and *probabilistic prediction* in terms of all the tested measures. The performance improvement with respect to the baseline version is always statistically significant, while, for *probabilistic prediction*, is significant for MRR, S@1, P@1, P@5. To have a more deep insight into the achieved results, in Fig. 3a and b we also plot the variation of, respectively, the precision and the success by varying k in the range [1,10]. Results show that, when increasing the rank value, our system and all the other competitors worsen their performance in terms of precision at rank k, while averagely perform better in terms of success until reaching a steady state value. Our approach performs best for any value of k in terms of precision (see Fig. 3b) and for $k = 1,3,4,5$ in terms of success (see Fig. 3a), while it performs as good as *Probabilistic prediction* in terms of success for the other values of k.

In summary, results show that our approach averagely selects the most suitable recommendable tags at the top of the ranking and precisely identify the potential user interests.

4.4 Parameter Analysis

We also analyzed the impact of the system parameters on the performance of the tag recommendation process. In Fig. 4, we plot the average MRR estimate achieved by our GR-TAG system by (1) varying the support threshold and by setting the minimum confidence threshold $minconf$ to 40 % and λ to 0.75 (see Fig. 4a), (2) varying the confidence threshold and by setting the minimum support threshold

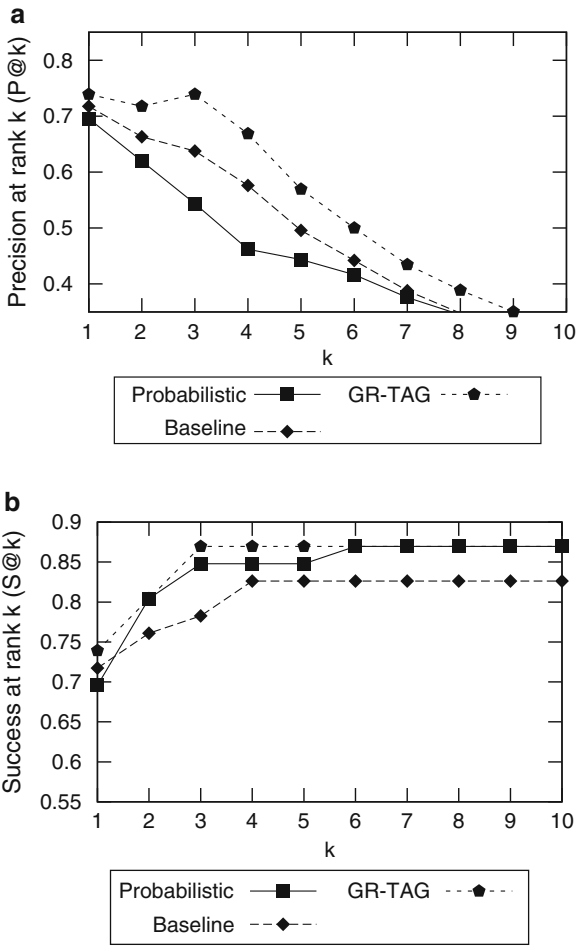


Fig. 3 Performance comparison by varying the reference rank k. (**a**) Precision at rank k (P@k) and (**b** success at rank k (S@k))

$minsup$ to 30 % and λ to 0.75 (see Fig. 4b), and (3) varying the lamba parameter by setting the minimum support threshold $minsup$ to 30 % and the minimum confidence threshold $minconf$ to 40 % (see Fig. 4c).

The support threshold relevantly affects the quality of the tag recommendation. When higher support thresholds (e.g., 70 %) are enforced, the percentage of not generalized rules is quite limited (e.g., around 18 % of the user-specific rule set mined from the training photo collection described in Sect. 4.1) and many informative rules (generalized and not) are discarded. Oppositely, when low-support thresholds (e.g., 20 %) are enforced, many low-level tag associations (e.g., around 3.5 % of the user-specific rule set from the same training data) become frequent and, thus, are extracted by our system. However, the high sparsity of the analyzed tag collections still left some of the most peculiar associations hidden. Aggregating tags into high-level categories allows achieving the best balancing between specialization and generalization of the discovered associations.

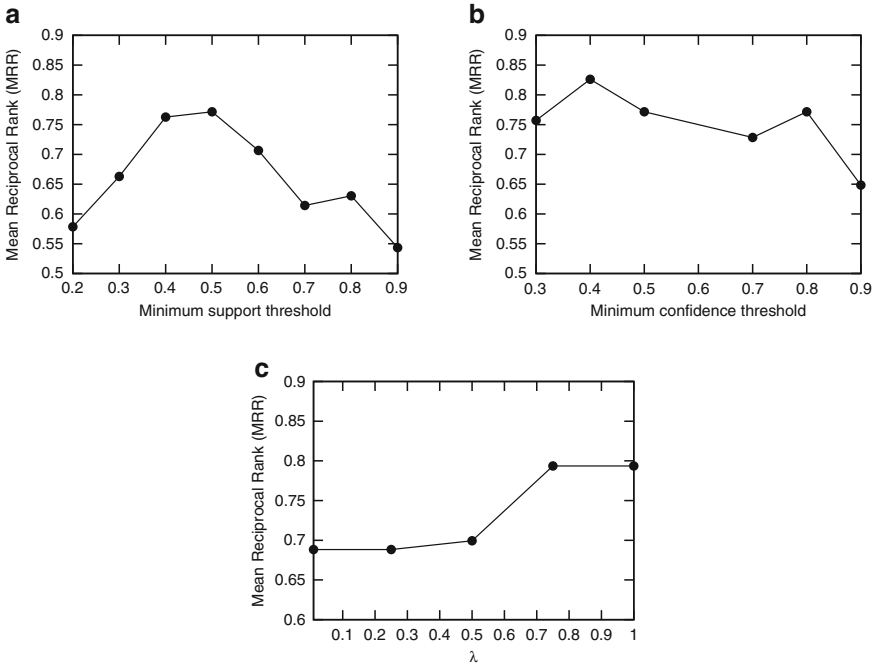


Fig. 4 GR-TAG performance analysis. (**a**) Impact of the support threshold on MRR estimate. $minconf = 40\,\%$. $\lambda = 0.75$, (**b**) impact of the confidence threshold on P@1 and S@1 estimates. $minsup = 30\,\%$. $\lambda = 0.75$, and (**c**) impact of λ on the MRR estimate. $minsup = 30\,\%$. $minconf = 40\,\%$

The confidence threshold may also significantly affect the system performance. By enforcing very low confidence threshold values (e.g., 20 %), a large amount of (possibly misleading) low-confidence rules is selected. Indeed, the quality of the rule-based model at the top of which the recommendation system is built worsens. Differently, when increasing the confidence threshold, a more selective pruning of the low-quality rules may allow significantly enhancing the system performance. Finally, when enforcing very high confidence thresholds (e.g., 90 %), the rule pruning selectivity becomes too high to allow dealing with a considerable amount of interesting rules.

Finally, we also analyzed the impact of the parameter λ on the achieved MRR. Similarly trends were achieved by using the other tested measures. The value of λ discriminates between the contribution of personal and collective knowledge. More specifically, when $\lambda < 0.5$, rules extracted from the community-based history of past annotations are deemed more significant than that discovered from the personal annotation set. Indeed, the recommendation process becomes less personalized, and the knowledge about the personal user interests is partially ignored. Differently, when setting $\lambda > 0.5$, tags mainly referable to the personalized rule set are deemed the most relevant ones for tag recommendation. Results show that, as expected, the

proposed system performs significantly better when the recommendation is more personalized, i.e., when user preferences are considered more relevant than the community-based annotations.

5 Conclusions and Future Work

In this chapter, we addressed the issue of recommending additional tags to partially annotated Flickr photos by exploiting both the personalized and the collective knowledge. We propose a rule-based recommendation system that also considers associations at higher abstraction levels, i.e., the generalized rules, to counteract the effect of data sparsity on the recommendation performance. A set of experiments performed on a real Flickr photo collection show the effectiveness of the proposed approach.

As pointed out by our work, the integration of the personalized and the collective annotations may effectively improve the quality of the recommended tags. To enrich the background knowledge related to the users and the community, we plan to integrate the analysis of the user-generated content coming from social networks and online communities in the tag recommendation system. Furthermore, we will also address, as future work, the integration in the proposed system of efficient disk-based indexing strategies to store and retrieve very large pattern collections.

References

1. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Proceedings of the 2008 ACM conference on Recommender Systems, RecSys '08, pp. 335–336. ACM, New York (2008). doi:http://doi.acm.org/10.1145/1454008.1454068, http://doi.acm.org/10.1145/1454008.1454068
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD Record, vol. 22, pp. 207–216. ACM, New York (1993)
3. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: Proceedings of the 16th International Conference on World Wide Web, WWW '07, pp. 501–510. ACM, New York (2007). DOI http://doi.acm.org/10.1145/1242572.1242640, http://doi.acm.org/10.1145/1242572.1242640
4. Baralis, E., Cagliero, L., Cerquitelli, T., D'Elia, V., Garza, P.: Support driven opportunistic aggregation for generalized itemset extraction. In: IEEE Conference of Intelligent Systems, pp. 102–107 (2010)
5. Baralis, E., Cagliero, L., Cerquitelli, T., Garza, P., Marchetti, M.: Cas-mine: providing personalized services in context-aware applications by means of generalized rules. Knowl. Inf. Syst. **28**(2), 283–310 (2011)
6. Baralis, E., Cagliero, L., Cerquitelli, T., Garza, P.: Generalized association rule mining with constraints. Inf. Sci. **194**, 68–84 (2012)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the Seventh International Conference on World Wide Web 7, pp. 107–117. Elsevier Science Publishers B. V. Amsterdam, The Netherlands (1998)

8. Cagliero, L.: Discovering temporal change patterns in the presence of taxonomies. IEEE Trans. Knowl. Data Eng. **99** (2011, Preprints). doi:http://doi.ieeecomputersociety.org/10.1109/TKDE.2011.233
9. Chirita, P.A., Costache, S., Nejdl, W., Handschuh, S.: P-tag: large scale automatic generation of personalized annotation tags for the web. In: Proceedings of the 16th International Conference on World Wide Web, WWW '07, pp. 845–854. ACM, New York (2007). doi:http://doi.acm.org/10.1145/1242572.1242686, http://doi.acm.org/10.1145/1242572.1242686
10. Datta, R., Ge, W., Li, J., Wang, J.Z.: Toward bridging the annotation-retrieval gap in image search. IEEE MultiMed. **14**, 24–35 (2007). doi:10.1109/MMUL.2007.67, http://dl.acm.org/citation.cfm?id=1435658.1436725
11. Delicious: Delicious. Website (2012). http://delicious.com. Accessed 28 Oct 2012
12. Dmitriev, P.A., Eiron, N., Fontoura, M., Shekita, E.: Using annotations in enterprise search. In: Proceedings of the 15th International Conference on World Wide Web, WWW '06, pp. 811–817. ACM, New York (2006). doi:http://doi.acm.org/10.1145/1135777.1135900, http://doi.acm.org/10.1145/1135777.1135900
13. Elmasri, R., Navathe, S.B.: Fundamentals of Database Systems, 5th edn. Addison Wesley, Boston (2006). http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20&path=ASIN/0321369572
14. Flickr: Flickr Website (2012). http://www.flickr.com. Accessed 28 Oct 2012
15. Garg, N., Weber, I.: Personalized, interactive tag recommendation for flickr. In: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys '08, pp. 67–74. ACM, New York (2008). doi:http://doi.acm.org/10.1145/1454008.1454020. http://doi.acm.org/10.1145/1454008.1454020
16. Han, J., Fu, Y.: Mining multiple-level association rules in large databases. IEEE Trans. Knowl. Data Eng. **11**(5), 798–805 (2002)
17. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pp. 531–538. ACM, New York (2008). doi:http://doi.acm.org/10.1145/1390334.1390425. http://doi.acm.org/10.1145/1390334.1390425
18. Jäschke, R., Marinho, L., Hotho, A., Lars, S.T., Gerd, S.: Tag recommendations in folksonomies. In: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2007, pp. 506–514. Springer, Berlin/Heidelberg (2007). doi:http://dx.doi.org/10.1007/978-3-540-74976-9_52, http://dx.doi.org/10.1007/978-3-540-74976-9_52
19. Krestel, R., Fankhauser, P., Nejdl, W.: Latent dirichlet allocation for tag recommendation. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pp. 61–68. ACM, New York (2009). doi:http://doi.acm.org/10.1145/1639714.1639726, http://doi.acm.org/10.1145/1639714.1639726
20. Lipczak, M., Milios, E.: Efficient tag recommendation for real-life data. ACM Trans. Intell. Syst. Technol. **3**(1), 2:1–2:21 (2011). doi:10.1145/2036264.2036266, http://doi.acm.org/10.1145/2036264.2036266
21. Lops, P., Gemmis, M., Semeraro, G.: Content-based recommender systems: state of the art and trends. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 73–105. Springer, New York/London (2011)
22. Lu, Y.T., Yu, S.I., Chang, T.C., Hsu, J.Y.j.: A content-based method to enhance tag recommendation. In: Proceedings of the 21st International Joint Conference on Artifical Intelligence, IJCAI'09, pp. 2064–2069. Morgan Kaufmann Publishers, San Francisco (2009). http://dl.acm.org/citation.cfm?id=1661445.1661775
23. Mishne, G.: Autotag: a collaborative approach to automated tag assignment for weblog posts. In: Proceedings of the 15th International Conference on World Wide Web, WWW '06, pp. 953–954. ACM, New York (2006). doi:http://doi.acm.org/10.1145/1135777.1135961, http://doi.acm.org/10.1145/1135777.1135961
24. Pramudiono, I., Kitsuregawa, M.: Fp-tax: tree structure based generalized association rule mining. In: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, p. 63. ACM, New York (2004)

25. Rae, A., Sigurbjörnsson, B., van Zwol, R.: Improving tag recommendation using social networks. In: Adaptivity, Personalization and Fusion of Heterogeneous Information, RIAO '10, pp. 92–99. Le centre de Haute Etudes Internationales d'Informatique Documentaire, Paris (2010). http://dl.acm.org/citation.cfm?id=1937055.1937077
26. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.): Recommender Systems Handbook. Springer, New York/London (2011)
27. Sanderson, M., Zobel, J.: Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, pp. 162–169. ACM, New York (2005). doi:http://doi.acm.org/10.1145/1076034.1076064, http://doi.acm.org/10.1145/1076034.1076064
28. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, WWW '01, pp. 285–295. ACM, New York (2001). doi:http://doi.acm.org/10.1145/371920.372071. http://doi.acm.org/10.1145/371920.372071
29. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th International Conference on World Wide Web, WWW '08, pp. 327–336. ACM, New York (2008). doi:http://doi.acm.org/10.1145/1367497.1367542. http://doi.acm.org/10.1145/1367497.1367542
30. Srikant, R., Agrawal, R.: Mining generalized association rules. In: Proceedings of the International Conference on Very Large Data Bases, pp. 407–419. Morgan Kaufmann, San Fransisco (1995)
31. Srikant, R., Vu, Q., Agrawal, R.: Mining association rules with item constraints. In: Proceedings of the Conference on Knowledge Discovery and Data Mining, vol. 97, pp. 67–73. AAAI Press, Palo Alto, California (1997)
32. Sriphaew, K., Theeramunkong, T.: A new method for finding generalized frequent itemsets in generalized association rule mining. In: Seventh International Symposium on Computers and Communications, pp. 1040–1045. IEEE, Washington, DC (2002)
33. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Tag recommendations based on tensor dimensionality reduction. In: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys '08, pp. 43–50. ACM, New York (2008). doi:http://doi.acm.org/10.1145/1454008.1454017, http://doi.acm.org/10.1145/1454008.1454017
34. van Erp, M., Schomaker, L.: Variants of the borda count method for combining ranked classifier hypotheses. In: Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition, Amsterdam, pp. 443–452 (2000)
35. WordNet: WordNet Lexical Database (2012). http://wordnet.princeton.edu. Accessed 28 Oct 2012
36. Zooomr: Zooomr Website (2012). http://www.zooomr.com. Accessed 28 Oct 2012

Sentic Computing for Social Media Analysis, Representation, and Retrieval

Erik Cambria, Marco Grassi, Soujanya Poria, and Amir Hussain

Abstract As the web is rapidly evolving, web users are evolving with it. In the era of social colonisation, people are getting more and more enthusiastic about interacting, sharing and collaborating through social networks, online communities, blogs, wikis and other online collaborative media. In recent years, this collective intelligence has spread to many different areas in the web, with particular focus on fields related to our everyday life such as commerce, tourism, education, and health. These online social data, however, remain hardly accessible to computers, as they are specifically meant for human consumption. To overcome such obstacle, we need to explore more concept-level approaches that rely more on the implicit semantic texture of natural language, rather than its explicit syntactic structure. To this end, we further develop and apply sentic computing tools and techniques to the development of a novel unified framework for social media analysis, representation and retrieval. The proposed system extracts semantics from natural language text by applying graph mining and multidimensionality reduction techniques on an affective common sense knowledge base and makes use of them for inferring the cognitive and affective information associated with social media.

E. Cambria (✉)
Temasek Laboratories, National University of Singapore, Singapore, 117411, Singapore
e-mail: cambria@nus.edu.sg

M. Grassi
Department of Information Engineering, Universitá Politecnica delle Marche, Ancona, 60131, Italy
e-mail: m.grassi@univpm.it

S. Poria
Department of Computer Science and Engineering, Jadavpur University, Kolkata, 700 032, India
e-mail: soujanya.poria@gmail.com

A. Hussain
Department of Computing Science and Mathematics, University of Stirling, Scotland, FK9 4LA, UK
e-mail: ahu@cs.stir.ac.uk

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_9, © Springer-Verlag London 2013

1 Introduction

Web 2.0 has changed the ways people communicate, collaborate, express their opinions and sentiments. The distillation of knowledge from this huge amount of unstructured information is an extremely difficult task as today web contents are perfectly suitable for human consumption, but they remain hardly accessible to machines. The web, in fact, mostly owes its success to the development of search engines like Google and Yahoo, which represent the starting point for information retrieval. Such engines, which base their searches on keyword-based algorithms relying on the textual representation of the web page, are very good in retrieving texts, splitting them into parts, checking the spelling and counting their words. But when it comes to interpreting sentences and extracting useful information for users, their capabilities result still very limited.

Current attempts to perform automatic understanding of text, for example, textual entailment and machine reading, still suffer from numerous problems including inconsistencies, synonymy, polysemy, entity duplication and more, as they focus on a mere syntactical analysis of text. To bridge the cognitive and affective gap between word-level natural language data and the concept-level opinions and sentiments conveyed by them, we need intelligent user interfaces able to learn new affective common sense knowledge and to perform reasoning on it, in order to semantically and affectively analyse natural language text. In human cognition, thinking and feeling are mutually present: emotions are often the product of our thoughts as well as our reflections are often the product of our affective states. Emotions, in fact, are intrinsically part of our mental activity and play a key role in decision-making processes: they are special states, shaped by natural selection, to adjust various aspects of our organism to make it better face particular situations, for example, anger evolved for reaction, fear evolved for protection and affection evolved for reproduction [1].

For these reasons, we cannot prescind from emotions in the development of intelligent systems: if we want computers to be really intelligent, not just have the veneer of intelligence, we need to give them the ability to recognise, understand and express emotions. In this work, we further develop and apply AI tools and techniques (Sect. 2) to the development of a novel unified framework for analysing (Sect. 3), representing (Sect. 4) and retrieving (Sect. 5) social media. The developed system (Fig. 1), in particular, consists of four main modules:

1. NLP module: This module is in charge of preprocessing the input text by using the affective valence indicators that are usually contained in opinionated text such as special punctuation, complete upper-case words, onomatopoeic repetitions, exclamation words, degree adverbs and emoticons.
2. Semantic parsing module: This module deconstructs the text into concepts using a lexicon contains several n-grams extracted from different semantic resources. The input text is deconstructed into several small bags of concepts (SBoCs), which are used as inputs to the ConceptNet module and AffectiveSpace module to infer their relative cognitive and affective information, respectively.

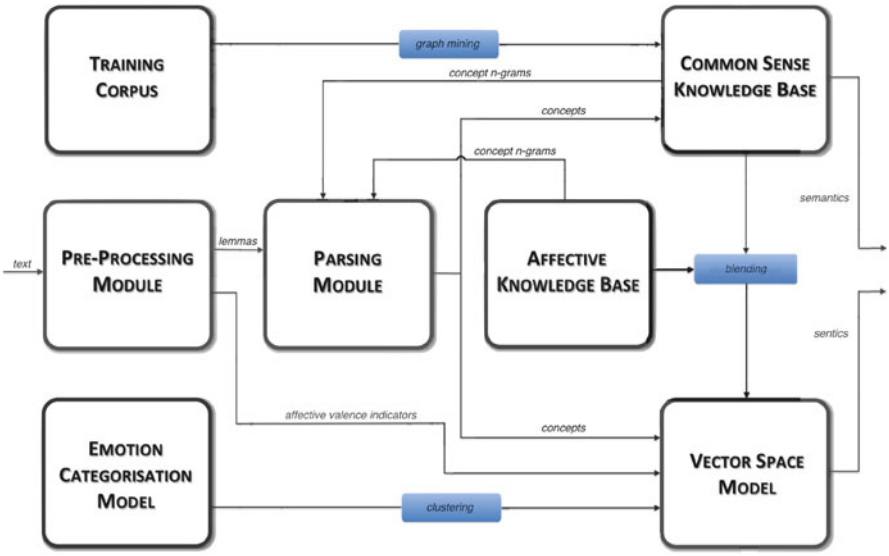


Fig. 1 System architecture. Graph mining and dimensionality reduction techniques are employed on two knowledge bases for open-domain sentiment analysis

3. ConceptNet module: This module exploits the graph representation of a common sense knowledge base to detect semantics. The concepts of each SBoC obtained from the output of the semantic parser are projected on the matrix resulting from many steps of spreading activation in order to calculate their semantic relatedness to each seed concept and, hence, their degree of belonging to each different class.
4. AffectiveSpace module: The concepts of each SBoC are projected into a vector space of affective common sense knowledge and clustered according to their coordinates in such space. This module assigns a score to each concept of the SBoC which defines the affinity of a concept belonging to a particular affective cluster.

2 Methodology

Sentic computing is a multidisciplinary approach to sentiment analysis, recently proposed by Cambria and Hussain [2], at the crossroads between affective computing and common sense computing. In the field of opinion mining, in fact, not only common sense knowledge but also emotional knowledge is important to grasp both the cognitive and affective information (termed semantics and sentics) associated with natural language opinions and sentiments.

Although scientific research in the area of emotion stretches back to the nineteenth century when Charles Darwin and William James proposed theories of

emotion that continue to influence thinking today [3, 4], the injection of affect into computer technologies is much more recent. During most of the last century, research on emotions was conducted by philosophers and psychologists, whose work was based on a small set of emotion theories that continue to underpin research in this area. The first researchers to try linking text to emotions were actually social psychologists and anthropologists who tried to find similarities on how people from different cultures communicate [5]. This research was also triggered by a dissatisfaction with the dominant cognitive view centred around humans as 'information processors' [6]. Later on, in the 1980s, researchers such as Turkle [7] began to speculate about how computers might be used to study emotions.

Systematic research programmes along this front began to emerge in the early 1990s. For example, Scherer [8] implemented a computational model of emotion as an expert system. A few years later, Picard's landmark book Affective Computing [9] prompted a wave of interest among computer scientists and engineers looking for ways to improve human–computer interfaces by coordinating emotions and cognition with task constraints and demands. Picard described three types of affective computing applications:

1. Systems that detect the emotions of the user
2. Systems that express what a human would perceive as an emotion
3. Systems that actually 'feel' an emotion

Although touching upon HCI [10] and affective modelling [11, 12], sentic computing primarily focuses on affect detection from text. Affect detection is critical because an affect-sensitive interface can never respond to users' affective states if it cannot sense their affective states. Affect detection need not be perfect but must be approximately on target. Affect detection is, however, a very challenging problem because emotions are constructs (i.e., conceptual quantities that cannot be directly measured) with fuzzy boundaries and with substantial individual difference variations in expression and experience.

To overcome such problem, sentic computing builds upon a brain-inspired and psychologically motivated affective categorisation model, proposed by Cambria et al. [13], that can potentially describe the full range of emotional experiences in terms of four independent but concomitant dimensions, whose different levels of activation make up the total emotional state of the mind. In sentic computing, whose term derives from the Latin *sentire* (root of words such as sentiment and sentience) and *sensus* (intended both as capability of feeling and as common sense), the analysis of natural language is based on affective ontologies and common sense reasoning tools, which enable the analysis of text not only at document, page or paragraph level but also at sentence and clause level.

In particular, sentic computing involves the use of AI and Semantic Web techniques, for knowledge representation and inference; mathematics, for carrying out tasks such as graph mining and multidimensionality reduction; linguistics, for discourse analysis and pragmatics; psychology, for cognitive and affective modelling; sociology, for understanding social network dynamics and social influence; and finally ethics, for understanding related issues about the nature of mind and

the creation of emotional machines. In this work, we exploit sentic computing tools and techniques to extract the semantics and sentics (i.e., the cognitive and affective information) associated with social media and, hence, bridge the gap between unstructured natural language data and structured machine-processable data. In particular, for the extraction of semantics, we use the following sentic computing tools and techniques:

1. A directed graph representation of common sense knowledge (Sect. 2.1)
2. A statistical method for the identification of common semantics (Sect. 2.2)
3. A technique that expands semantics through spreading activation (Sect. 2.3)

In turn, for the extraction of sentics, we use:

1. A language visualisation and analysis system (Sect. 2.4)
2. A novel emotion categorization model (Sect. 2.5)
3. A technique for clustering sentics (Sect. 2.6)

2.1 ConceptNet

ConceptNet [14] is a semantic resource structurally similar to WordNet, but whose scope of contents is general world knowledge, in the same vein as Cyc [15].Instead of insisting on formalising common sense reasoning using mathematical logic [16], ConceptNet uses a new approach: it represents data in the form of a semantic network and makes it available to be used in natural language processing. The prerogative of ConceptNet, in fact, is contextual common sense reasoning: while WordNet is optimised for lexical categorization and word-similarity determination, and Cyc is optimised for formalised logical reasoning, ConceptNet is optimised for making practical context-based inferences over real-world texts.

In ConceptNet, WordNet's notion of node in the semantic network is extended from purely lexical items (words and simple phrases with atomic meaning) to include higher-order compound concepts, for example, 'satisfy hunger' and 'follow recipe', to represent knowledge around a greater range of concepts found in everyday life (see Table 1). Moreover, WordNet's repertoire of semantic relations is extended from the triplet of synonym, is-a and part-of, to a repertoire of twenty semantic relations including, for example, EffectOf (causality), SubeventOf (event hierarchy), CapableOf (agent's ability), MotivationOf (affect), PropertyOf and LocationOf. ConceptNet's knowledge is also of a more informal, defeasible and practically valued nature. For example, WordNet has formal taxonomic knowledge that 'dog' is a 'canine', which is a 'carnivore', which is a 'placental mammal'; but it cannot make the practically oriented member-to-set association that 'dog' is a 'pet'. ConceptNet also contains a lot of knowledge that is defeasible, that is, it describes something that is often true but not always, for example, EffectOf ('fall off bicycle', 'get hurt'), which is something we cannot leave aside in common sense reasoning.

Table 1 Comparing WordNet and ConceptNet: while WordNet synsets contain vocabulary knowledge associated with concepts, ConceptNet assertions convey generic knowledge about what such concepts are used for

Term	WordNet hypernyms	ConceptNet assertions
Cat	Feline; felid; adult male; man; gossip; gossiper; gossipmonger; rumormonger; rumourmonger; newsmonger; woman; adult female; stimulant; stimulant drug; excitant; tracked vehicle	Cats can hunt mice; cats have whiskers; cats can eat mice; cats have fur; cats have claws; cats can eat meat; cats are cute
Dog	Canine; canid; unpleasant woman; disagreeable woman; chap; fellow; feller; lad; gent; fella; scoundrel; sausage; follow	Dogs are mammals; a dog can be a pet; a dog can guard a house; you are likely to find a dog in kennel; an activity a dog can do is run; a dog is a loyal friend; a dog has fur
Language	Communication; auditory communication; word; higher cognitive process; faculty; mental faculty; module; text; textual matter	English is a language; French is a language; language is used for communication; music is a language; a word is part of language
iPhone	N/A	An iPhone is a kind of a telephone; an iPhone is a kind of computer; an iPhone can display your position on a map; an iPhone can send and receive emails; an iPhone can display the time
Birthday gift	Present	Card is birthday gift; present is birthday gift; buying something for a loved one is for a birthday gift

Most of the facts interrelating ConceptNet's semantic network are dedicated to making rather generic connections between concepts. This type of knowledge can be brought back to Minsky's K-lines as it increases the connectivity of the semantic network and makes it more likely that concepts parsed out of a text document can be mapped into ConceptNet. ConceptNet is produced by an automatic process, which first applies a set of extraction rules to the semistructured English sentences of the Open Mind Common Sense (OMCS) corpus and then applies an additional set of 'relaxation' procedures, that is, filling in and smoothing over network gaps, to optimise the connectivity of the semantic network (Fig. 2). In ConceptNet version 2.0, a new system for weighting knowledge was implemented, which scores each binary assertion based on how many times it was uttered in the OMCS corpus and on how well it can be inferred indirectly from other facts in ConceptNet. In ConceptNet version 3.0 [17], users can also participate in the process of refining knowledge by evaluating existing statements on Open Mind Commons [18], the new interface for collecting common sense knowledge from users over the web.

By giving the user many forms of feedback and using inferences by analogy to find appropriate questions to ask, Open Mind Commons can learn well-connected structures of common sense knowledge, refine its existing knowledge and build analogies that lead to even more powerful inferences. The pieces of common

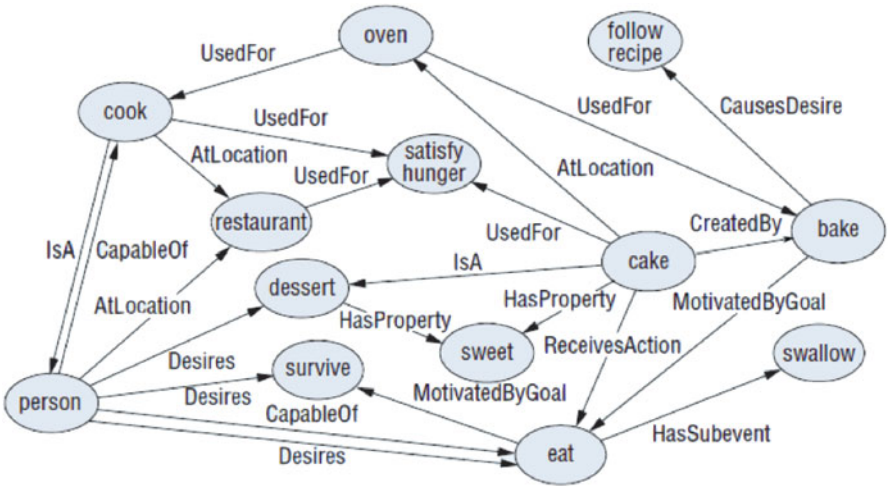


Fig. 2 ConceptNet represents the information in the Open Mind corpus as a directed graph where nodes are concepts and labelled edges are assertions of common sense that interconnect them

sense knowledge acquired through this interface are made publicly available in ConceptNet, which is released periodically both as an SQL database and through an API.

2.2 CF-IOF Weighting

CF-IOF (concept frequency – inverse opinion frequency) [19] is a technique that identifies domain-dependent semantics, using an approach similar to TF-IDF weighting, in order to evaluate how important a concept is to a set of opinions concerning the same topic. Firstly, the frequency of a concept c for a given domain d is calculated by counting the occurrences of the concept c in the set of available d-tagged opinions and dividing the result by the sum of number of occurrences of all concepts in the set of opinions concerning d. This frequency is then multiplied by the logarithm of the inverse frequency of the concept in the whole collection of opinions, that is:

$$\text{CF-IOF}_{c,d} = \frac{n_{c,d}}{\sum_k n_{k,d}} \log \sum_k \frac{n_k}{n_c}$$

where $n_{c,d}$ is the number of occurrences of concept c in the set of opinions tagged as d, n_k is the total number of concept occurrences and n_c is the number of occurrences of c in the whole set of opinions. A high weight in CF-IOF is reached by a high concept frequency in a given domain and a low frequency of the concept in the whole collection of opinions. Therefore, thanks to CF-IOF weights, it is possible to filter out common concepts and detect relevant domain-dependent semantics.

2.3 Spectral Association

Spectral association [20] is a technique that involves assigning values, or activations, to 'seed concepts' and applying an operation that spreads their values across the ConceptNet graph. This operation, an approximation of many steps of spreading activation, transfers the most activation to concepts that are connected to the key concepts by short paths or many different paths in common sense knowledge. In particular, we build a matrix C that relates concepts to other concepts, instead of their features, and add up the scores over all relations that relate one concept to another, disregarding direction. Applying C to a vector containing a single concept spreads that concept's value to its connected concepts. Applying C^2 spreads that value to concepts connected by two links (including back to the concept itself). But what we would really like is to spread the activation through any number of links, with diminishing returns, so the operator we want is:

$$1 + C + \frac{C^2}{2!} + \frac{C^3}{3!} + \ldots = e^C$$

We can calculate this odd operator, e^C, because we can factor C. C is already symmetric, so instead of applying Lanczos' method to CC^T and getting the singular value decomposition (SVD), we can apply it directly to C and get the spectral decomposition $C = V\Lambda V^T$. As before, we can raise this expression to any power and cancel everything but the power of Λ. Therefore, $e^C \approx Ve^{\Lambda}V^T$. This simple twist on the SVD lets us calculate spreading activation over the whole matrix instantly. As with the SVD, we can truncate these matrices to k axes and therefore save space while generalising from similar concepts. We can also rescale the matrix so that activation values have a maximum of 1 and do not tend to collect in highly connected concepts such as 'person', by normalising the truncated rows of $Ve^{\Lambda/2}$ to unit vectors, and multiplying that matrix by its transpose to get a rescaled version of $Ve^{\Lambda}V^T$.

2.4 AffectiveSpace

AffectiveSpace is a multidimensional vector space built by 'blending' [21] ConceptNet with WordNet-Affect (WNA) [22], a linguistic resource for the lexical representation of affective knowledge. Blending is a technique that performs inference over multiple sources of data simultaneously, taking advantage of the overlap between them. It basically combines two sparse matrices linearly into a single matrix in which the information between the two initial sources is shared. When we perform SVD on a blended matrix, the result is that new connections are made in each source matrix taking into account information and connections present in the other matrix, originating from the information that overlaps. The alignment operation operated over ConceptNet and WNA yields a new matrix, A, in which

common sense and affective knowledge coexist, that is, a matrix $14{,}301 \times 117{,}365$ whose rows are concepts (e.g., 'dog' or 'bake cake'), whose columns are either common sense or affective features (e.g., 'isA-pet' or 'hasEmotion-joy') and whose values indicate truth values of assertions. Therefore, in A, each concept is represented by a vector in the space of possible features whose values are positive for features that produce an assertion of positive valence (e.g., 'a penguin is a bird'), negative for features that produce an assertion of negative valence (e.g., 'a penguin cannot fly') and zero when nothing is known about the assertion.

The degree of similarity between two concepts, then, is the dot product between their rows in A. The value of such a dot product increases whenever two concepts are described with the same feature and decreases when they are described by features that are negations of each other. In particular, we use truncated singular value decomposition (TSVD) [23] in order to obtain a new matrix containing both hierarchical affective knowledge and common sense. The resulting matrix has the form $\tilde{A} = U_k * \Sigma_k * V_k^T$ and is a low-rank approximation of A, the original data. This approximation is based on minimising the Frobenius norm of the difference between A and $\tilde{A}$ under the constraint $\text{rank}(\tilde{A}) = k$. For the Eckart–Young theorem [24], it represents the best approximation of A in the mean-square sense, in fact

$$\min_{\tilde{A}|\text{rank}(\tilde{A})=k} |A - \tilde{A}| = \min_{\tilde{A}|\text{rank}(\tilde{A})=k} |\Sigma - U^* \tilde{A} V| = \min_{\tilde{A}|\text{rank}(\tilde{A})=k} |\Sigma - S|$$

assuming that $\tilde{A}$ has the form $\tilde{A} = \text{USV}^*$, where S is diagonal. From the rank constraint, i.e., S has k non-zero diagonal entries, the minimum of the above statement is obtained as follows:

$$\min_{s_i} \sqrt{\sum_{i=1}^{n} (\sigma_i - s_i)^2} = \min_{s_i} \sqrt{\sum_{i=1}^{k} (\sigma_i - s_i)^2 + \sum_{i=k+1}^{n} \sigma_i^2} = \sqrt{\sum_{i=k+1}^{n} \sigma_i^2}$$

Therefore, $\tilde{A}$ of rank k is the best approximation of A in the Frobenius norm sense when $\sigma_i = s_i$ $(i = 1, \ldots, k)$, and the corresponding singular vectors are the same as those of A. If we choose to discard all but the first k principal components, common sense concepts and emotions are represented by vectors of k coordinates: these coordinates can be seen as describing concepts in terms of 'eigenmoods' that form the axes of AffectiveSpace, that is, the basis $e_0, \ldots, e_{k-1}$ of the vector space (Fig. 3). For example, the most significant eigenmood, e_0, represents concepts with positive affective valence. That is, the larger a concept's component in the e_0 direction is, the more affectively positive it is likely to be. Concepts with negative e_0 components, then, are likely to have negative affective valence. Thus, by exploiting the information sharing property of TSVD, concepts with the same affective valence are likely to have similar features – that is, tend to fall near each other in AffectiveSpace. Concept similarity does not depend on their absolute positions in the vector space, but rather on the angle they make with the origin. For example, we can find concepts such as 'beautiful day', 'birthday

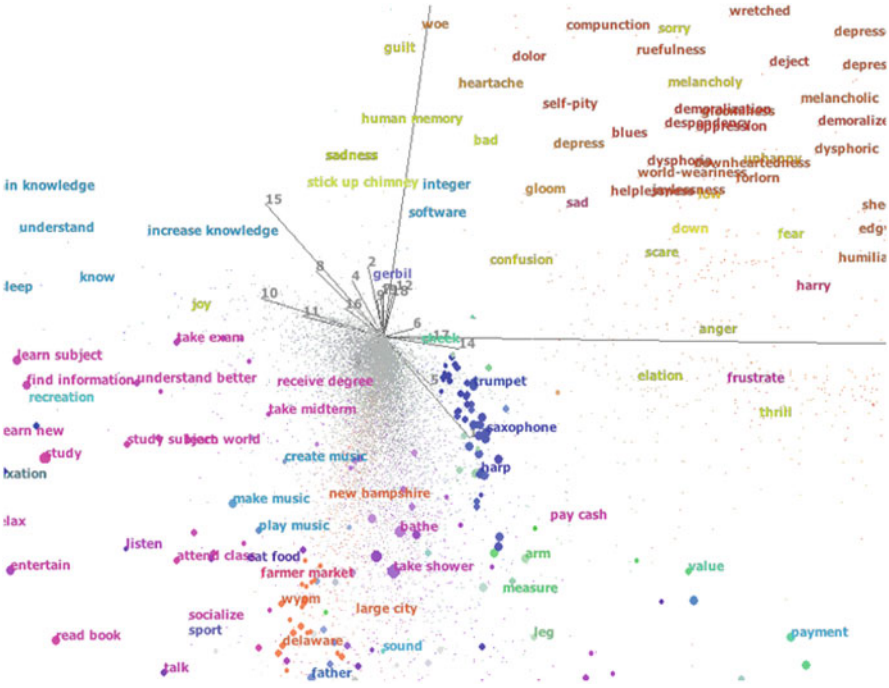
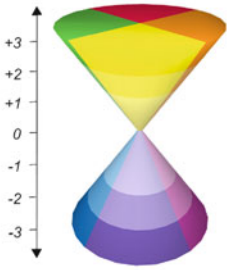


Fig. 3 Affectively positive (*bottom-left* corner) and affectively negative (*up-right* corner) common sense concepts in AffectiveSpace

party', 'laugh' and 'make person happy' very close in direction in the vector space, while concepts like 'sick', 'feel guilty', 'be laid off' and 'shed tear' are found in a completely different direction.

2.5 The Hourglass of Emotions

This model is a variant of Plutchik's emotion categorization [25] and constitutes an attempt to emulate Marvin Minsky's theories on human emotions. Minsky sees the mind as made up of thousands of different resources and believes that our emotional states result from turning one set of these resources on and turning another set of them off [1]. Each such selection changes how we think by changing our brain's activities: the state of anger, for example, appears to select a set of resources that help us react with more speed and strength while also suppressing some other resources that usually make us act prudently [44]. The Hourglass of Emotions (Fig. 4) is specifically designed to recognise, understand and express emotions in the context of human–computer interaction (HCI). In the model, in fact, affective



	Pleasantness	Attention	Sensitivity	Aptitude
+3	ecstasy	vigilance	rage	admiration
+2	joy	anticipation	anger	trust
+1	serenity	interest	annoyance	acceptance
0	—	—	—	—
-1	pensiveness	distraction	apprehension	boredom
-2	sadness	surprise	fear	disgust
-3	grief	amazement	terror	loathing

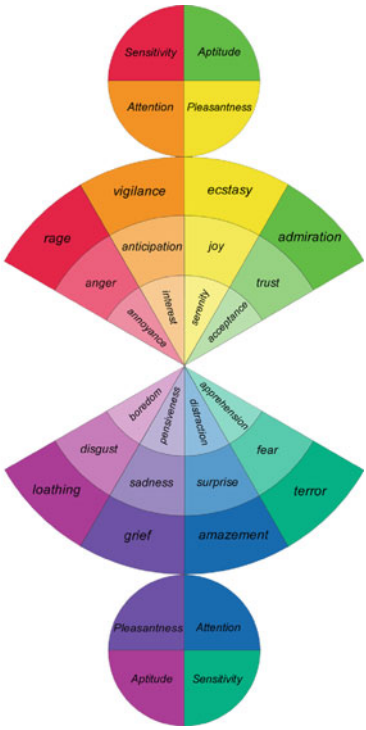


Fig. 4 The 3D model and the net of the Hourglass of Emotions. Dimensional and discrete forms of the different sentic levels are summarised in the proposed emotion categorization table

states are not classified, as often happens in the field of emotion analysis, into basic emotional categories, but rather into four concomitant but independent dimensions in order to understand how much respectively:

1. The user is happy with the service provided (pleasantness).
2. The user is interested in the information supplied (attention).
3. The user is comfortable with the interface (sensitivity).
4. The user is disposed to use the application (aptitude).

Each affective dimension is characterised by six levels of activation, called 'sentic levels', which determine the intensity of the expressed/perceived emotion as an int $\in [-3,+3]$. These levels are also labelled as a set of 24 basic emotions (six for each of the affective dimensions) in a way that allows the model to specify the affective information associated with text both in a dimensional and in a discrete form. The dimensional form, in particular, is called 'sentic vector', and it is a four-dimensional *float* vector that can potentially express any human emotion in terms of pleasantness, attention, sensitivity and aptitude. Some particular sets of sentic vectors have special names as they specify well-known compound emotions. For

example, the set of sentic vectors with a level of pleasantness ∈ (+1,+2] ('joy'), a null attention, a null sensitivity and a level of aptitude ∈ (+1,+2] ('trust') are called 'love sentic vectors' since they specify the compound emotion of 'love'.

2.6 Sentic Medoids

Sentic medoids [26] is a clustering technique that adopts a k-medoids approach [27] to partition affective common sense concepts in AffectiveSpace into k clusters around as many centroids, trying to minimise a given cost function. Differently from the k-means algorithm [28], which does not pose constraints on centroids, k-medoids do assume that centroids must coincide with k observed points. The k-means approach finds the k centroids, where the coordinate of each centroid is the mean of the coordinates of the objects in the cluster and assigns every object to the nearest centroid. Unfortunately, k-means clustering is sensitive to the outliers, and a set of objects closest to a centroid may be empty, in which case centroids cannot be updated. For this reason, k-medoids are sometimes used, where representative objects are considered instead of centroids. In many clustering problems, in fact, one is interested in the characterisation of the clusters by means of typical objects, which represent the various structural features of objects under investigation. Because it uses the most centrally located object in a cluster, k-medoids clustering is less sensitive to outliers compared with k-means.

Among many algorithms for k-medoids clustering, partitioning around medoids (PAM) is one of the most widely used. The algorithm, proposed by Kaufman and Rousseeuw [27], first computes k representative objects, called medoids. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. PAM determines a medoid for each cluster selecting the most centrally located centroid within the cluster. After selection of medoids, clusters are rearranged so that each point is grouped with the closest medoid. Compared to k-means, PAM operates on the dissimilarity matrix of the given data set. It is more robust, because it minimises a sum of dissimilarities instead of a sum of squared Euclidean distances. A particularly nice property is that PAM allows clustering with respect to any specified distance metric. In addition, the medoids are robust representations of the cluster centres, which is particularly important in the common context that many elements do not belong well to any cluster. However, PAM works inefficiently for large data sets due to its complexity.

To this end, a modified version of the algorithm recently proposed by Park and Jun [29] was used, which runs similarly to the k-means clustering algorithm. This has shown to have similar performance when compared to PAM algorithm while taking a significantly reduced computational time. In particular, we have N concepts ($N = 14,301$) encoded as points $x \in \mathbb{R}^p (p = 50)$. We want to group them into k clusters, and, in our case, we can fix $k = 24$ as we are looking for one cluster for each sentic level s of the Hourglass model. Generally, the initialization of clusters for clustering algorithms is a problematic task as the process often risks to get

stuck into local optimum points, depending on the initial choice of centroids [30]. However, we decide to use as initial centroids the concepts that are currently used as centroids for clusters, as they specify the emotional categories we want to organise AffectiveSpace into. For this reason, what is usually seen as a limitation of the algorithm can be seen as advantage for this approach, since we are not looking for the 24 centroids leading to the best 24 clusters but indeed for the 24 centroids identifying the required 24 sentic levels (i.e., the centroids should not be 'too far' from the ones currently used). In particular, as the Hourglass affective dimensions are independent but concomitant, we need to cluster AffectiveSpace four times, once for each dimension. According to the Hourglass categorization model, in fact, each concept can convey, at the same time, more than one emotion (which is why we get compound emotions), and this information can be expressed via a sentic vector specifying the concept's affective valence in terms of pleasantness, attention, sensitivity and aptitude. Therefore, given that the distance between two points in AffectiveSpace is defined as $D(a,b) = \sqrt{\sum_{i=1}^{p}(a_i - b_i)^2}$ (note that the choice of Euclidean distance is arbitrary), the used algorithm, applied for each of the four affective dimensions, can be summarised as follows:

1. Each centroid $C_n \in \mathbb{R}^{50}\,(n = 1, 2, \ldots, k)$ is set as one of the six concepts corresponding to each s in the current affective dimension.
2. Assign each record x to a cluster Ξ so that $x_i \in \Xi_n$ if $D(x_i, C_n) \leq D(x_i, C_m)$ $m = 1, 2, \ldots, k$.
3. Find a new centroid C for each cluster Ξ so that $C_j = x_i$ if $\sum_{x_m \in \Xi_j} D(x_i, x_m) \leq \sum_{x_m \in \Xi_j} D(x_h, x_m) \quad \forall x_h \in \Xi_j$.
4. Repeat steps 2 and 3 until no changes on centroids are observed.

Note that condition posed on steps 2 and 3 may occasionally lead to more than one solution. Should this happen, our model will randomly choose one of them. This clusterization of AffectiveSpace allows to calculate, for each common sense concept x, a four-dimensional sentic vector that defines its affective valence in terms of a degree of fitness $\mathbf{f}(x)$ where $f_a = D(x, C_j) \quad C_j | D(x, C_j) \leq D(x, C_k)$ $a = 1, 2, 3, 4 \quad k = 6a\text{-}5, 6a\text{-}4, \ldots, 6a$.

3 System Architecture

In order to effectively mine and analyse opinions and sentiments, it is necessary to bridge the gap between unstructured natural language data and structured machine-processable data. To this end, an intelligent software engine has been proposed by Cambria et al. [31] that aims to extract the semantics and sentics, that is, the cognitive and affective information, associated with natural language text, in a way that the opinions and sentiments contained in it can be more easily aggregated and interpreted. The engine exploits graph mining and multidimensionality reduction techniques on ConceptNet, and it is based on the Hourglass model (Fig. 5).

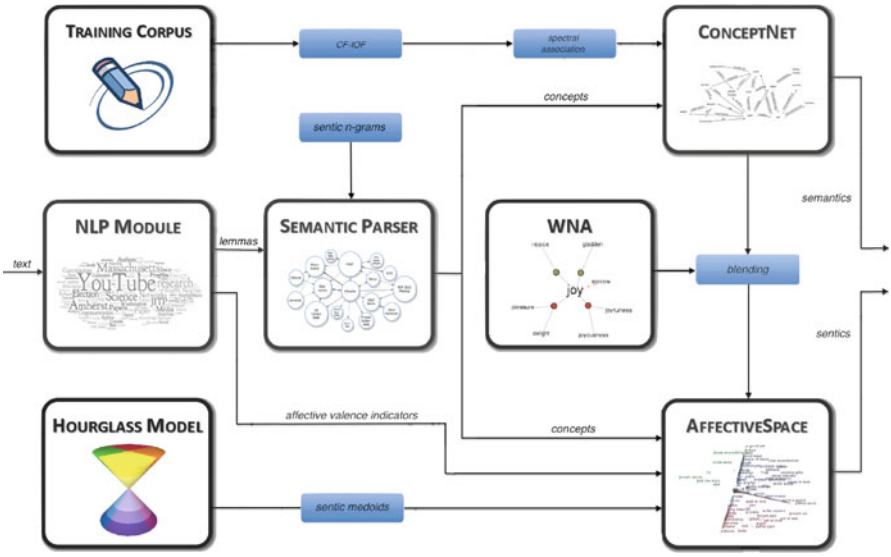


Fig. 5 Opinion mining engine *block diagram*. After performing a first skim of the input text, the engine extracts concepts from it and, hence, infers related semantics and sentics

Table 2 An overview of recent model-based affect recognition and sentiment analysis systems. Studies are divided by techniques applied, number of categories of the model adopted, corpora and knowledge base used

Study	Techniques	Model	Corpora	Knowledge base
[32]	NB, SVM	2 categories	Political articles	None
[33]	LSA, MLP, NB, KNN	3 categories	Dialogue turns	ITS interaction
[34]	Cohesion indices	4 categories	Dialogue logs	ITS interaction
[35]	VSM, NB, SVM	5 categories	ISEAR	ConceptNet
[36]	WN presence, LSA	6 categories	News stories	WNA
[37]	WN presence	6 categories	Chat logs	WNA
[38]	Winnow linear, C4.5	7 categories	Children stories	None
[39]	VSM, KNN	24 categories	LiveJournal	ConceptNet, WNA
[31]	VSM, k-means	24 categories	YouTube, LiveJournal	ConceptNet, WNA, HEO
[40]	VSM, k-means	24 categories	LiveJournal, Patient Opinion	ConceptNet, WNA
[41]	VSM, k-medoids	24 categories	Twitter, LiveJournal, Patient Opinion	ConceptNet, Probase

Several other affect recognition and sentiment analysis systems [32–38] are based on different emotion categorisation models, which generally comprise a relatively small set of categories (Table 2). The Hourglass of Emotions, in turn, allows the opinion mining engine to classify affective information both in a categorical way (according to a wider number of emotion categories) and in a

dimensional format (which facilitates comparison and aggregation). Such engine, in particular, consists of four main components: an NLP module, which performs a first skim of the opinion (Sect. 3.1); a semantic parser, whose aim is to extract concepts from the opinionated text (Sect. 3.2); the ConceptNet module, for inferring the semantics associated with the given concepts (Sect. 3.3); and the AffectiveSpace module, for the extraction of sentics (Sect. 3.4). Eventually, this section illustrates an output example of the engine, given a short natural language sentence as input (Sect. 3.5).

3.1 NLP Module

This preprocessing module firstly interprets all the affective valence indicators usually contained in opinionated text such as special punctuation, complete upper-case words, onomatopoeic repetitions, exclamation words, degree adverbs and emoticons. Secondly, the module detects negation and spreads it in a way that it can be accordingly associated to concepts during the parsing phase. Handling negation is an important concern in opinion- and sentiment-related analysis, as it can reverse the meaning of a statement.

Such task, however, is not trivial as not all appearances of explicit negation terms reverse the polarity of the enclosing sentence and that negation can often be expressed in rather subtle ways, for example, sarcasm and irony, which are quite difficult to detect. Lastly, the module converts text to lower case and, after lemmatising it, splits the opinion into single clauses according to grammatical conjunctions and punctuation.

3.2 Semantic Parser

The semantic parser deconstructs text into concepts using a lexicon based on sequences of lexemes that represent multiple-word concepts extracted from ConceptNet and WordNet. These n-grams are not used blindly as fixed word patterns but exploited as reference for the module, in order to extract multiple-word concepts from information-rich sentences. So, differently from other shallow parsers, the module can recognise complex concepts also when irregular verbs are used or when these are interspersed with adjective and adverbs, for example, the concept 'buy christmas present' in the sentence 'I bought a lot of very nice Christmas presents'. For each clause, the module outputs a small bag of concepts (SBoC), which is later on analysed separately by the ConcepNet and AffectiveSpace modules to infer the cognitive and affective information associated with the input text, respectively. In case any of the detected concepts is found more than once in the vector space (that is, any of the concepts has multiple senses), all the SBoC concepts are exploited

for a context-dependent coarse sense disambiguation. In particular, to represent the expected semantic value of the clause as a whole, the vectors corresponding to all concepts in the clause (in their ambiguous form) can be averaged together. The resulting vector does not represent a single meaning but the 'ad hoc category' of meanings that are similar to the various possible meanings of concepts in the clause [42]. Then, to assign the correct sense to the ambiguous concept, the concept sense with the highest dot product (and thus the strongest similarity) with the clause vector is searched.

3.3 ConceptNet Module

Once natural language text is deconstructed into concepts, these are given as input to both the ConceptNet and the AffectiveSpace modules. While the former exploits the graph representation of the affective common sense knowledge base to detect semantics, the latter exploits the vector space representation of ConceptNet to infer sentics. In particular, the ConceptNet module applies spectral association for assigning activation to key concepts, that is, nodes of the semantic network, which are used as seeds or centroids for classification. Such seeds can simply be the concepts corresponding to the class labels of interest plus their available synonyms and antonyms, if any.

As shown in Sect. 2.3, seeds can also be found by applying CF-IOF on a training corpus (when available), in order to perform a classification that is more relevant to the data under analysis. After seeds concepts are identified, the module spreads their values across the ConceptNet graph. This operation, an approximation of many steps of spreading activation, transfers the most activation to concepts that are connected to the seed concepts by short paths or many different paths in affective common sense knowledge. Therefore, the concepts of each SBoC provided by the semantic parser are projected on the matrix resulting from spectral association in order to calculate their semantic relatedness to each seed concept and, hence, their degree of belonging to each different class. Such classification measure is directly proportional to the degree of connectivity between the nodes representing the retrieved concepts and the seed concepts in the affective common sense knowledge graph.

3.4 AffectiveSpace Module

In the ConceptNet module, graph-mining techniques are exploited to extract semantics from the concepts retrieved by the semantic parser. Such concepts are also given as input to the AffectiveSpace module, which, in turn, exploits dimensionality reduction techniques to infer the affective information associated with them. To this

end, the concepts of each SBoC are projected into AffectiveSpace, and, according to their position in the vector space representation of affective common sense knowledge, they are assigned to an affective class defined through the sentic medoids technique.

As well as in the ConceptNet module, the categorisation does not consist in simply labelling each concept but also in assigning a confidence score to each emotional label, which is directly proportional to the degree of belonging to a specific affective cluster (dot product between the given concept and the relative sentic medoid). Such affective information can also be exploited to calculate a polarity value associated with each SBoC provided by the semantic parser as well as to detect the overall polarity associated with the opinionated text.

3.5 Output Example

As an example of how the software engine works, intermediate and final outputs obtained when a natural language opinion is given as input to the system can be examined. The following tweet was chosen: 'I think iPhone4 is the top of the heap! OK, the speaker is not the best i hv ever seen bt touchscreen really puts me on cloud 9… camera looks pretty good too!'. After the preprocessing and semantic parsing operations, the following SBoCs are obtained:

SBoC#1:
<Concept: 'think'>
<Concept: 'iphone4'>
<Concept: 'top heap'>
SBoC#2:
<Concept: 'ok'>
<Concept: 'speaker'>
<Concept: !*'good'++>*
<Concept: 'see'>
SBoC#3:
<Concept: 'touchscreen'>
<Concept: 'put cloud nine'++>
SBoC#4:
<Concept: 'camera'>
<Concept: 'look good'−−>

These are then concurrently processed by the ConceptNet and the AffectiveSpace modules, which output the cognitive and affective information associated with each SBoC, both in a discrete way, with one or more labels, and in a dimensional way, with a polarity value $\in [-1,+1]$ (Table 3).

Table 3 Structured output example of opinion mining engine. For each clause, the engine detects the opinion target, the category it belongs to and the affective information associated with it

Opinion target	Category	Moods	Polarity
'iphone4'	'phones', 'electronics'	'ecstasy', 'interest'	+0.71
'speaker'	'electronics', 'music'	'annoyance'	−0.34
'touchscreen'	'electronics'	'ecstasy', 'anticipation'	+0.82
'camera'	'photography', 'electronics'	'acceptance'	+0.56

4 Data Model

Our framework for social media representation and analysis aims to be applicable to most of online resources (videos, images, text) coming from different sources, for example, online video sharing services, blogs and social networks.

To such purpose, it is necessary to standardise as much as possible the descriptors used in encoding the information about multimedia resources and people to which the text refer (considering that every website uses its own vocabulary) in order to make it univocally interpretable and suitable to feed other applications. To achieve this purpose, Semantic Web techniques are exploited.

The Semantic Web initiative by W3C[1] tackles this problem through an appropriate representation of information in the web page, able to univocally identify resources and encode the meaning of their description. In particular, the Semantic Web uses uniform resource identifiers (URIs) to univocally identify entities available on the web as documents or images but not as concepts or properties and RDF data model to describe such resources in univocally interpretable format, whose basic building block is an object-attribute-value triple, that is, a statement.

Resources may be authors, books, publishers, places, people, hotels, rooms, search queries, etc., while properties describe relations between resources such as 'writtenBy', 'age', and 'title'. Statements assert the properties of resources, and their values can be either resources or literals (strings). To provide machine-accessible and machine-processable representations, it is usual to encode RDF triples using XML syntax. Each triple can also be seen as a directed graph with labelled nodes and arcs, where the arcs are directed from the resource (the subject of the statement) to the value (the object of the statement). Each statement describes the graph node or connects it to other nodes, linking together multiple data from different sources without pre-existing schema. It is according to this representation that indeed the Semantic Web in its whole can be envisioned as a Giant Global Graph of Linked Data. RDF, however, does not make assumptions about any particular application domain, nor does it defines the semantics of any domain. For this purpose, it is necessary to introduce ontologies.

[1] http://w3.org

Ontologies basically deal with knowledge representation and can be defined as formal explicit descriptions of concepts in a domain of discourse (named classes or concepts), properties of each concept describing various features and attributes of the concept (roles or properties), and restrictions on property (role restrictions). Ontologies make possible the sharing of common understanding about the structure of information among people or software agents. In addition, ontologies make possible reasoning, that is, it is possible, starting from the data and the additional information expressed in the form of ontology, to infer new relationships between data. Different languages have been developed for the design of ontologies, among the most popular there are RDFS (RDF Schema) and OWL (Ontology Web Language). RDFS can be seen as a RDF vocabulary and a primitive ontology language. It offers certain modelling primitives with fixed meaning.

Key concepts of RDF are class, subclass relations, property, sub-property relations and domain and range restrictions. OWL is a language more specifically conceived for ontologies creation. It builds upon RDF and RDFS and a XML-based RDF syntax is used. Instances are defined using RDF descriptions, and most RDFS modelling primitives are used. Moreover, OWL introduces a number of features that are missing in RDFS such as local scope of property, disjointness of classes, Boolean combination of classes (like union, intersection and complement), cardinality restriction and special characteristics of properties (like transitive, unique or inverse). The proposed framework for opinions and affective information description aims to be applicable to most of online resources (videos, images, text) coming from different sources, for example, online video sharing services, blogs and social networks. To such purpose, it is necessary to standardise as much as possible the descriptors used in encoding the information about multimedia resources and people to which the opinions refer (considering that every website uses its own vocabulary) in order to make it univocally interpretable and suitable to feed other applications.

To this end, we encode the cognitive and affective information associated with multimedia resources and people using the descriptors provided by OMR (Ontology for Media Resources), FOAF (Friend of a friend ontology), HEO (Human Emotion Ontology) [43], and WNA (WordNet-Affect) [22]. OMR represents an important effort to help circumventing the current proliferation of audio/video meta-data formats, currently carried on by the W3C Media Annotations Working Group. It offers a core vocabulary to describe media resources on the web, introducing descriptors such as ‘title’, ‘creator’, ‘publisher’, ‘createDate’ and ‘rating’. It defines semantic-preserving mappings between elements from existing formats. This ontology is supposed to foster the interoperability among various kinds of meta-data formats currently used to describe media resources on the web. FOAF represents a recognised standard in describing people, providing information such as their names, birthdays, pictures, blogs and especially other people they know, which makes it particularly suitable for representing data that appears on social networks and communities. OMR and FOAF together supply most of the vocabulary needed for describing media and people and other descriptors are added only when necessary. For example, OMR, at least in the current realisation, does not supply vocabulary for describing comments, which are analysed to extract the

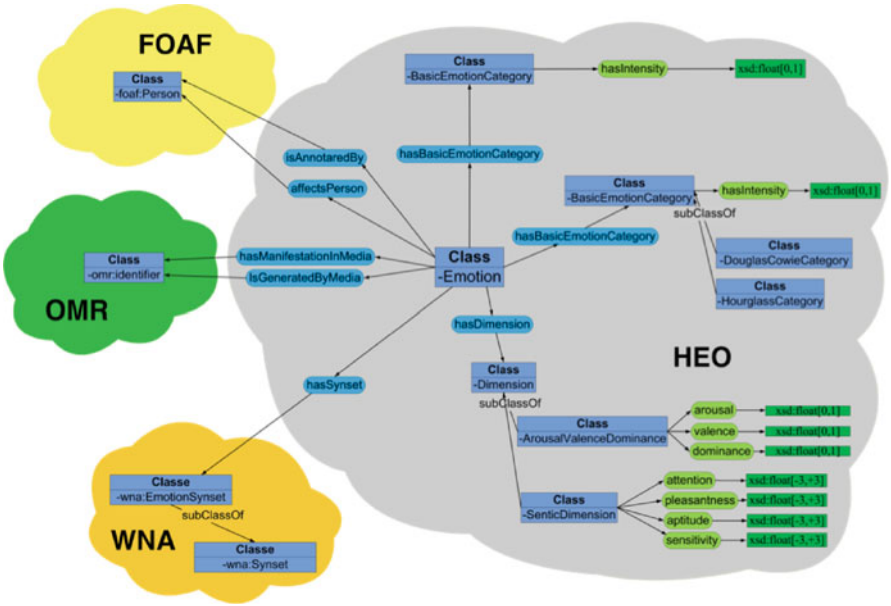


Fig. 6 Social media representation. *HEO*, *WNA*, *OMR* and *FOAF* are accordingly merged for effectively representing the cognitive and affective information associated with social media

affective information relative to media. This ontology is extended by introducing the 'Comment' class and by defining for it the 'author', 'text' and 'publicationDate' properties.

HEO is a high-level ontology for human emotions that supplies the most significant concepts and properties which constitute the centrepiece for the description of every human emotion. The main purpose of HEO is to create a description framework that could grant at the same time enough flexibility, by allowing the use of a wide and extensible set of emotion feature descriptors, and interoperability, by allowing to map concepts and properties belonging to different emotion representation models. In HEO, we introduce properties to link emotions to multimedia resources and people. In particular, we have defined the 'hasManifestationInMedia' and 'isGeneratedByMedia', to describe emotions that respectively occur and are generated in media, and the property 'affectPerson' to connect emotions to people.

Moreover, to improve the hierarchical organisation of emotions in HEO, we exploit WNA, a linguistic resource for the lexical representation of affective knowledge, built by assigning to a number of WordNet synsets one or more affective labels (a-labels) and then by extending the core with the relations defined in WordNet. Thus, the combination of HEO with WNA, OMR and FOAF provides a complete framework to describe not only multimedia contents and the users that have created, uploaded or interacted with them but also the opinions and the affective content carried by the media and the way they are perceived by people (Fig. 6).

5 User Interface

As remarked above, due to the way they are created and maintained, community-contributed multimedia resources are very different from standard web data. One fundamental aspect is constituted by the collaborative way in which such data is created, uploaded and annotated. A deep interconnection emerges in the nature of these data and meta-data, allowing, for example, to associate videos of completely different genre but uploaded by the same user, or different users, even living in opposite sides of the world, who have appreciated the same pictures.

Such interdependence can be exploited, for example, to find similar patterns in customer reviews of commercial products and hence to gather useful information for marketing, sales, public relations and customer service. To visualise the cognitive and affective information associated to social media, we exploit the multifaceted categorization paradigm. Faceted classification allows the assignment of multiple categories to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, predetermined, taxonomic order. This makes possible to perform searches combining the textual approach with the navigational one. Faceted search, in fact, enables users to navigate a multidimensional information space by concurrently writing queries in a text box and progressively narrowing choices in each dimension.

For our framework, we use SIMILE Exhibit API, a set of JavaScript files that allows to easily create rich interactive web pages including maps, timelines and galleries, with very detailed client-side filtering. Exhibit pages use the multifaceted classification paradigm to display semantically structured data stored in a Semantic Web aware format, for example, RDF or JavaScript Object Notation (JSON). One of the most relevant aspects of Exhibit is that, once the page is loaded, the web browser also loads the entire data set in a lightweight database and performs all the computations (sorting, filtering, etc.) locally on the client-side, providing high performances.

We encode the cognitive and affective information associated with social media in RDF/XML, using the descriptors defined by HEO, WNA, OMR and FOAF, and store it in a Sesame triple store, a purpose-built database for the storage and retrieval of RDF meta-data. Sesame can be embedded in applications and used to conduct a wide range of inferences on the information stored, based on RDFS and OWL-type relations between data. In addition, it can also be used in a standalone server mode, much like a traditional database with multiple applications connecting to it (Fig. 7).

In this way, all the knowledge stored inside Sesame can be queried, and the results can also be retrieved in a semantic aware format and used for other applications. We export all the information contained in the triplestore into a JSON file to feed the Exhibit interface, in order to make it available for being browsed as a unique knowledge base. We choose to use Exhibit in our framework due to the ease with which it allows to create rich and interactive web pages. Social media are displayed in a dynamic gallery that can be ordered according to different parameters and the cognitive and affective information associated with them. Using faceted

Fig. 7 Social media retrieval. The faceted classification interface allows multimodal social media retrieval according to the semantics and sentics associated with them

menus, it is possible to explore such information both using the search box, to perform keyword-based queries, and filtering the results using the faceted menus, that is, by adding or removing constraints on the facet properties. One of the most relevant aspects of Exhibit is that, once the page is loaded, the web browser also loads the entire data set in a lightweight database and performs all the computations (sorting, filtering, etc.) locally on the client-side, providing high performances.

6 Conclusions

With the advent of the social web, the way people express their views and opinions has dramatically changed. They can now post reviews of products at merchant sites and express their views on almost anything in Internet forums, discussion groups, and blogs. Such online word-of-mouth behaviour represents new and measurable sources of information with many practical applications.

However, finding opinion sources and monitoring them can be a formidable task because there are a large number of diverse sources, and each source may also have a huge volume of opinionated text. In many cases, in fact, opinions are hidden in long forum posts and blogs. It is extremely time consuming for a human reader to find relevant sources, extract related sentences with opinions, read them, summarise them and organise them into usable forms. Thus, automated opinion discovery and summarisation systems are needed. Sentiment analysis, also known as opinion mining, grows out of this need. It is a challenging NLP or text mining problem. Due to its tremendous value for practical applications, there has been an explosive growth of both research in academia and applications in the industry.

Due to many challenging research problems and a wide variety of practical applications, opinion mining has been a very active research area in recent years. All the sentiment analysis tasks, however, are very challenging. Our understanding and knowledge of the problem and its solution are still limited. The main reason is that it is a NLP task, and NLP has no easy problems. Another reason may be due to our popular ways of doing research. So far, in fact, researchers have probably relied too much on machine learning algorithms. Some of the most effective machine learning algorithms, for example, SVM and CRF, in fact, produce no human understandable results such that, although they may achieve improved accuracy, little about how and why is known, apart from some superficial knowledge gained in the manual feature engineering process.

All such approaches, moreover, rely on syntactical structure of text, which is far from the way human mind processes natural language. In this work, common sense computing techniques were further developed and applied to bridge the semantic gap between word-level natural language data and the concept-level opinions conveyed by these. In particular, the ensemble application of graph mining and multi-dimensionality reduction techniques was exploited on a common sense knowledge base to develop a novel intelligent engine for open-domain opinion mining and sentiment analysis. The proposed framework performs a clause-level semantic analysis of text, which allows the inference of both the conceptual and emotional information associated with natural language opinions and, hence, a more efficient passage from (unstructured) textual information to (structured) machine-processable data.

References

1. Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon and Schuster, New York (2006)
2. Cambria, E., Hussain, A.: Sentic Computing: Techniques, Tools, and Applications. Dordrecht, Netherlands: Springer (2012)
3. Charles, D.: The Expression of the Emotions in Man and Animals. John Murray, London (1872)
4. James, W.: What is an emotion? Mind **34**, 188–205 (1884)
5. Osgood, C., May, W., Miron, M.: Cross-cultural universals of affective meaning. University of Illinois, Urbana (1975)
6. Lutz, C., White, G.: The anthropology of emotions. Ann. Rev. Anthropol. **15**, 405–436 (1986)
7. Turkle, S.: The Second Self: Computers and the Human Spirit. Simon and Schuster, New York (1984)
8. Scherer, K.: Studying the emotion-antecedent appraisal process: an expert system approach. Cognit. Emot. **7**, 325–355 (1993)
9. Picard, R.: Affective computing. MIT, Boston (1997)
10. Cambria, E., Hupont, I., Hussain, A., Cerezo, E., Baldassarri, S.: Sentic avatar: multimodal affective conversational agent with common sense. In: Esposito, A., Hussain, A., Faundez-Zanuy, M., Martone, R., Melone, N. (eds.) Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues. Lecture Notes in Computer Science, vol. 6456, pp. 82–96. Springer, Berlin/Heidelberg (2011)

11. Cambria, E., Olsher, D., Kwok, K.: Sentic activation: a two-level affective common sense reasoning framework. In: Proceedings of the AAAI, Toronto, pp. 186–192 (2012)
12. Cambria, E., Olsher, D., Kwok, K.: Sentic panalogy: swapping affective common sense reasoning strategies and foci. In: Proceedings of the CogSci, Sapporo, pp. 174–179 (2012)
13. Cambria, E., Hussain, A., Havasi, C., Eckl, C.: SenticSpace: visualizing opinions and sentiments in a multi-dimensional vector space. In: Setchi, R., Jordanov, I., Howlett, R., Jain, L. (eds.) Knowledge-Based and Intelligent Information and Engineering Systems. Lecture Notes in Artificial Intelligence, vol. 6279, pp. 385–393. Springer, Berlin (2010)
14. Liu, H., Singh, P.: ConceptNet: a practical commonsense reasoning toolkit. BT Technol. J. **22**(4), 211–226 (2004)
15. Lenat, D., Guha, R.: Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Boston (1989)
16. Mueller, E.: Commonsense Reasoning. Morgan Kaufmann, Amsterdam/Boston (2006)
17. Havasi, C., Speer, R., Alonso, J.: ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In: Proceedings of the RANLP, Borovets (2007)
18. Speer, R.: Open mind commons: an inquisitive approach to learning common sense. In: Proceedings of the Workshop on Common Sense and Interactive Applications, Honolulu (2007)
19. Cambria, E., Hussain, A., Durrani, T., Havasi, C., Eckl, C., Munro, J.: Sentic computing for patient centered application. In: Proceedings of the IEEE ICSP, Beijing, pp. 1279–1282 (2010)
20. Havasi, C., Speer, R., Holmgren, J.: Automated color selection using semantic knowledge. In: Proceedings of the AAAI CSK, Arlington (2010)
21. Havasi, C., Speer, R., Pustejovsky, J., Lieberman, H.: Digital intuition: applying common sense using dimensionality reduction. IEEE Intell. Syst. **24**(4), 24–35 (2009)
22. Strapparava, C., Valitutti, A.: WordNet-affect: an affective extension of WordNet. In: Proceedings of the LREC, Lisbon (2004)
23. Wall, M., Rechtsteiner, A., Rocha, L.: Singular value decomposition and principal component analysis. In: Berrar, D., Dubitzky, W., Granzow, M. (eds.) A Practical Approach to Microarray Data Analysis, pp. 91–109. Kluwer Academic Publishers, Boston (2003)
24. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika **1**(3), 211–218 (1936)
25. Plutchik, R.: The nature of emotions. Am. Sci. **89**(4), 344–350 (2001)
26. Cambria, E., Mazzocco, T., Hussain, A., Eckl, C.: Sentic medoids: organizing affective common sense knowledge in a multi-dimensional vector space. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) Advances in Neural Networks. Lecture Notes in Computer Science, vol. 6677, pp. 601–610. Springer, Berlin (2011)
27. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
28. Hartigan, J., Wong, M.: Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc. **28**(1), 100–108 (1979)
29. Park, H., Jun, C.: A simple and fast algorithm for k-medoids clustering. Expert Syst. Appl. **36**(2), 3336–3341 (2009)
30. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
31. Cambria, E., Grassi, M., Hussain, A., Havasi, C.: Sentic computing for social media marketing. Multimed. Tools Appl. **59**(2), 557–577 (2012)
32. Lin, W., Wilson, T., Wiebe, J., Hauptmann, A.: Which side are you on? Identifying perspectives at the document and sentence levels. In: Proceedings of the Conference on Natural Language Learning, New York, pp. 109–116 (2006)
33. D'Mello, S., Craig, S., Sullins, J., Graesser, A.: Predicting affective states expressed through an emote-aloud procedure from autotutor's mixed-initiative dialogue. Int. J. Artif. Intell. Educ. **16**, 3–28 (2006)
34. D'Mello, S., Dowell, N., Graesser, A.: Cohesion relationships in tutorial dialogue as predictors of affective states. In: Proceedings of the Conference Artificial Intelligence in Education, pp. 9–16. Springer, New York (2009)

35. Danisman, T., Alpkocak, A.: Feeler: emotion classification of text using vector space model. In: Proceedings of the AISB, Aberdeen (2008)
36. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the ACM Symposium Applied Computing, pp. 1556–1560. ACM, New York (2008)
37. Ma, C., Osherenko, A., Prendinger, H., Ishizuka, M.: A chat system based on emotion estimation from text and embodied conversational messengers. In: Proceedings of the International Conference Active Media Technology, pp. 546–548. IEEE, Piscataway (2005)
38. Alm, C., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the HLT/EMNLP, pp. 347–354. Association for Computing Linguistics, Morristown (2005)
39. Grassi, M., Cambria, E., Hussain, A., Piazza, F.: Sentic web: a new paradigm for managing social media affective information. Cognit. Comput. **3**(3), 480–489 (2011)
40. Cambria, E., Benson, T., Eckl, C., Hussain, A.: Sentic PROMs: application of sentic computing to the development of a novel unified framework for measuring health-care quality. Expert Syst. Appl. **39**(12), 10533–10543 (2012)
41. Cambria, E., Song, Y., Wang, H., Howard, N.: Semantic Multi-Dimensional Scaling for Open-Domain Sentiment Analysis. In press: IEEE Intelligent Systems (2013)
42. Havasi, C., Speer, R., Pustejovsky, J.: Coarse word-sense disambiguation using common sense. In: Proceedings of the AAAI CSK, Arlington (2010)
43. Grassi, M.: Developing HEO human emotions ontology. Lecture Notes in Computer Science, vol. 5707, pp. 244–251. Springer, Berlin/Heidelberg (2009)
44. Cambria, E., Livingstone, A., Hussain, A.: The Hourglass of Emotions. LNCS, vol. 7403, pp. 144–157. Springer, Heidelberg (2012)

Highlight Detection in Movie Scenes Through Inter-users, Physiological Linkage

Christophe Chênes, Guillaume Chanel, Mohammad Soleymani, and Thierry Pun

Abstract Automatic summarization techniques facilitate multimedia indexing and access by reducing the content of a given item to its essential parts. However, novel approaches for summarization should be developed since existing methods cannot offer a general and unobtrusive solution. Considering that the consumption of multimedia data is more and more social, we propose to use a physiological index of social interaction, namely, physiological linkage, to determine general highlights of videos. The proposed method detects highlights which are relevant to the majority of viewers without requiring them any conscious effort. Experimental testing has demonstrated the validity of the proposed system which obtained a classification accuracy of up to 78.2%.

1 Introduction

The rapid rate of multimedia creation motivates the development of novel multimedia management and indexing methods. These methods are facing many challenges such as the semantic gap and the complexity of multimedia data. In this context, tagging is an essential step for an effective indexing of multimedia content. In

C. Chênes (✉) • M. Soleymani • T. Pun
Computer Science Department, Computer Vision and Multimedia Laboratory (CVML)

University of Geneva, Battelle Campus, Building A, 7 route de Drize, CH-1227 Carouge, Geneva, Switzerland
e-mail: christophe.chenes@gmail.com; mohammad.soleymani@unige.ch; thierry.pun@unige.ch

G. Chanel
CVML and Swiss Center for Affective Sciences, 7 rue des Battoirs, CH-1205, Geneva, Switzerland
e-mail: guillaume.chanel@unige.ch

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_10, © Springer-Verlag London 2013

addition, since tags are usually given by users, this approach is likely to reduce the semantic gap. Summarization techniques are fundamental to manage large-scale multimedia collections, since they allow to reduce the content to its essential parts. Summarization can thus be considered as a form of tagging with video sequences being "tagged" as a highlight (i.e., a relevant and essential part of the video) or a non-highlight.

Major improvements have been made in the multimedia indexing and retrieval field over the last years thanks to user-generated tags and social networks. Some of these tagging methods have brought a new approach, human-centered tagging, by which the data are tagged by the users. An improvement to these methods is presented in the previous chapter with the use of implicit tagging which aims at reliably extracting tags from nonverbal behaviors of users who are confronted to multimedia data [1]. These new tags, such as emotional keywords, are supposed to be more robust, more useful, and more convenient for content recommendation or retrieval as they are unobtrusive and uncontrollable.

In the perspective of highlight detection in movie scenes, emotions play an essential role. The movie structure, the plot, and the narrative are carefully designed by directors to elicit strong emotions from the spectators. This is the reason why movies have often been used for emotion elicitation studies [2]. The emotional moments of movies can then be considered as strong highlights. As a consequence, the spectators' emotional responses are reliable indicators for highlight detection. Moreover, as emotions enhance the memory [3], highlights corresponding to the spectators' emotional response peaks are potentially the most memorable ones, which is also a desirable property of detected highlights.

Automatic detection of emotions is one of the main goals of affective computing [4]. This field of research is thus of high interest for implicit tagging and emotional highlights detection. Researchers have proposed to use several modalities to automatically detect human emotions [5–7]. These include facial expressions, speech, and also physiological signals. The physiological signals have certain advantages, for example, when measured with the appropriate sensors, physiological signals are difficult to hide or fake. So far, affective computing has mostly focused on the detection of a single person's emotional experience. However, emotions often emerge during social interactions, and it would thus be valuable to switch the unit of analysis from the person to the group. This can, for instance, be achieved by computing physiological linkage which measures the extent to which the physiological signals of two people are dependent of each other [8].

Physiological linkage can occur during any social interaction, and also due to the common interpretation and perception of a stimulus. For example, the spectators, feeling empathy with the movie characters, having similar emotional reactions to the movie content and experiencing social-emotional contagion in the case of multi viewers' show, are expected to have synchronized and similar physiological reactions. It is expected that this form of linkage increases

during moments of a movie when most of the spectators are strongly moved and engaged in the movie. Physiological linkage can then be considered as a reliable indicator of highlights when most spectators are emotionally and similarly impacted [9].

This chapter proposes a new approach for highlight detection in videos. The proposed approach is based on the analysis of several spectators' physiological signals. It is thus a social approach that provides a user-independent and content-independent or general (as opposed to personalized and content dependent) summarization of the videos by aggregating the subjective experiences of all spectators. The use of physiological measures also has the advantage of not obstructing the watching of the video in comparison with methods that require users to explicitly provide feedback. One of the central ideas of the proposed work, linked with the social media, is thus to record several spectators' physiological responses and to consider high physiological linkage sequences [10] as highlights. This technique enables us to obtain an accurate and user-independent summary of a given video. It is important to clarify that this method does not require emotion recognition, since it is sufficient to compute linkage directly between spectators' physiological signals. However, it is expected that linkage is a result of the synchronization of spectators' emotional responses.

Therefore, this human-centered method uses the viewers' physiological responses to tag a video sequence either as a highlight (the positive class) or a non-highlight (the negative class). Based on the literature, we expect that movies elicit strong and synchronous emotional responses in spectators that can be detected by measuring their physiological signals. Consequently, the system we propose for highlight detection is based on the following hypothesis: *the moments when viewers' physiological responses are highly linked correspond to the highlight sequences of the scene*.

To answer this hypothesis, this chapter is structured as follows. Firstly, a summary of the work related to the proposed method is given. It focuses on the importance of social considerations for affective tagging in indexing and retrieval processes and also describes physiological linkage. It then thoroughly explains the existing video summarization techniques, both internal and external with a particular focus on the closest techniques to the proposed work. Secondly, the user-independent system developed is described. This includes explanations of the concept, the signal processing, the physiological linkage method applied, the classification methods used, and the stimuli-reaction delay management. Thirdly, the proposed system is evaluated based on the physiological data collected in a previous experiment. In this phase particular attention was given to the collection of a reliable groundtruth. The results validating the proposed approaches are then presented and discussed. Finally, this chapter ends with a summary of the study, the remaining issues, and the future work.

2 Related Work

2.1 Social and Affective Tagging

The act of tagging is a well-known and largely applied method which has been used very often to simplify the description of complex data by experts in their domain. With the development of the Web 2.0, this method has become collaborative [11]. The action of associating a tag to a given content is no longer done by a single expert but by a large number of users who are free to use their own terms to describe the content of the data they are rating. When there is a very large amount of data to process, such as on the Web, the collaborative tagging process is the most effective [12]. However, since any user is free to tag with any term, this approach is not fully reliable. Users tend to tag data for personal motivations (selfishness) and social motivations (reputation) [13, 14]. Therefore, applications based on this process suffer from lack of objectivity and reliability. These limitations also demonstrate the importance of social factors in tagging processes.

A solution to improve tagging reliability is to rely on the implicit cues given by multimedia consumers rather than on their explicit evaluation of the multimedia content. This human-centered method is known as implicit tagging [1] and is discussed in the previous chapter. Most of the studies on implicit tagging have focused on affective tags since affect is a highly relevant criterion for multimedia indexing and retrieval based on preferences. Affective tags are also useful to partially bridge the semantic gap. The tags used in this method are linked with spontaneous reactions of the users to the multimedia content they are watching/listening; they can be based on the user's facial expression [15] or physiological reactions [16–18].

Affective implicit tagging is possible thanks to the advancement of affective computing [4]. Affective computing has two major goals:

1. The detection and recognition of emotions: Computers should be able to detect users' emotional changes based on different signals recorded through specific sensors; they should also be able to recognize users' emotions.
2. Computer affective reactions: Computers should be able to synthesize emotions and empathetic reactions and to improve and create a naturalistic human-computer interaction; this branch is motivated by the Turing test [19].

The research in this innovative field has lead to several models for the detection and recognition of emotions. It has also described three main channels for affective sensing: visual, audio, and physiological. Within the human-computer interaction (HCI) development, affective computing applications range from e-health service [20] to video games [21, 22]. Independently from the topic, every study in affective computing has helped to better understand the user's experience, including those employing physiological signals [23–25]. Different physiological signals from both the central and the peripheral nervous system have been used for the purpose of emotion assessment [6, 26, 27]. This includes electroencephalography (EEG), which measures brain electrical potentials; electrodermal activity (EDA), which measures

the resistance of the skin (an index of perspiration), blood pressure (BP), heart rate (HR), respiration, and temperature; and electromyography (EMG), which measures the electrical activity originated from muscular activity.

Studies on affective computing, so far, were mostly focused on emotional experience of a single user. However, multimedia is produced and consumed socially. This is, for instance, the case of music and movies which are very often produced by bands and teams that interact during the whole creation phase. This social aspect of creation has, for instance, encouraged the creation of social musical instruments [28]. Furthermore, people enjoy sharing and exchanging media, as well as the experiences they felt during their consumption. Finally, they also get together in theaters and festivals to watch movies and listen to music. There is now clear evidence that the emotional expression of a person can shape the emotions of their peers [29, 30]. The experience of users should thus not be considered independently of each other, and some measures of joint-mediated social interactions and associations should be determined. This is in line with the work done in the field of social signal processing [31] which aims at improving the social abilities and social intelligence [32] of computers. In this context, a measure of social interaction can help to better understand and quantify social processes [33] but can also be used to provide new multimedia tags. In the range of social interactions, we include any social association, affective bond, and affinities that can exist between two people experiencing similar content (e.g., if spectators have similar preferences and thus react in the same way to the movie or if spectators are friends and share a common ground).

2.2 Social Interaction and Physiological Linkage

We use social signals in our daily interactions to send messages and mediate our social behaviors. The signals given from one communicator to another are mainly gesture, posture, facial expression, and voice prosody. However, physiological signals also carry relevant social information since they are modulated by affective reactions [34]. Social effects do not only occur in a classical one-to-one communication; interactions also happen when experiencing shared material, such as music and movies [28]. Participants are socially linked when watching such multimedia content together, they feel the others' presence, and they share similar emotions, such as joy through smiling or laughter and fear through screaming. This social context induces emotional contagion [29] which is the emotional convergence of participants who are influenced by others' reactions. This fact represents the importance of social context on physiological responses since it emphasizes the linkage between participants' reactions. From another perspective, in the case of movies, viewers can feel empathy with a character. Depending on their involvement in the movie, they can experience social presence by identifying themselves with a movie character or imagining themselves in the character's situation [33, 35].

As proposed in [33], physiological linkage is an interesting measure to study mediated social interactions. Physiological linkage has been originally proposed by Gottman [8] and measures to what extent physiological signals of two people depend on each other. The most straightforward method to infer physiological linkage is to compute correlation between two physiological signals. However, several other methods can be used [33]: coherence measures allow to determine if signals oscillate at the same frequency by switching from the temporal to the frequency domain; Granger causality and similar methods based on linear autoregressive models account for nonindependence of time series samples; and methods from nonlinear dynamical systems, such as synchronization measures [36], are able to deal with nonlinear signals.

Physiological linkage has been shown to be related to several social processes. In [37] the authors observed a higher level of physiological linkage during conflicting interactions compared to low-intensity nonconflicting interactions of married couples. Levenson et al. [38] also demonstrated that physiological linkage is related to the accuracy of rating others' feelings which can be considered as an indicator of empathy. More recently Henning et al. [39, 40] have analyzed physiological linkage in the context of collaborative processes. They discovered that linkage is associated with team performance [39] when playing on a collaborative video game. However, these results were contradicted by a later study [40] that concluded physiological linkage was inversely proportional to self-reported team productivity, quality of communication, and ability to work together. While the studies cited above focus on the use of peripheral physiological signals, it is also possible to compute linkage and synchronization on brain signals. For instance, it has been demonstrated that inter-person brain signal synchronization occurs during imitation of the other [41]. Concerning movies, it has been shown in [42] that synchronization can also occur during natural visualization of films and it is argued that interbrain synchronization can be understood as supporting the coordination of actions and the common understanding of the environment [43]. Finally, the analysis of synchronicity and linkage is not limited to physiological signals and can also be used to measure synchronous social behaviors [28] between musicians. All these results clearly demonstrate that physiological linkage is associated to several types of measurements and situations that involve social interactions and associations.

2.3 Automatic Highlight Detection in Videos

The automation of highlight detection or extraction from video is a far-reaching subject. With the exponential growth of video media, it becomes essential to have some algorithms able to summarize video content. The aim of video abstraction is, in general, to put together the highlights of the video. The first step of video summarization techniques is then the detection of highlight sequences which are considered as relevant by the majority of viewers. The second step of a complete

summarization method includes the cut of the highlight sequences from the video stream, respecting the video structure (scenes, clusters, shots) [44, 45] and the reconstruction of the summary with the selected highlights.

The vast quantity of videos induces a huge variation of genres. The differences between genres can be significant and, most important, the definition of highlights can be extremely diverging. For example, a soccer match video summary will be mainly composed by goal actions, while the highlight of a drama scene will be a wave of sadness. It is then obvious that the goal of a general method raises great difficulties.

In the current literature, two categories of video abstraction methods are reported [46, 47] as well as the combination of these two categories:

Internal summarization techniques are based on the analysis of low-level features present within the video stream, such as color or speech. Many interesting techniques have been developed with promising results. For example, a method based on the grass proportion in the video, the slow-motion effect, and the cinematic of the scene was implemented for soccer match summarization to detect goals, slow motions, and referee actions [48]. Its global correct classification rate reaches 89%. The recall of the method is impressive, more than 90%, but the precision is about 40%. However, internal techniques face a certain number of remaining challenges, in particular the semantic gap and their highly domain-specific design which makes the goal of a general method utopian, though they appear to be very popular and effective techniques for the sport videos [48–50] due to the structures of sport games and the interferences of the spectators.

External summarization techniques are based on features entirely external to the video, such as user's description of the video. The information used by these techniques to summarize the video can be of two types:

1. User-based, sourced directly from the user, such as from the facial expression of the user [51]
2. Contextual, sourced from the context of the user but not the user himself, such as the video recording GPS position [52]

These techniques try to summarize videos with a higher level of abstraction. They reflect better the comprehension of the user and thus reduce the semantic gap. In spite of this significant advantage over the internal summarization techniques, such systems have been rarely implemented [46, 47].

In survey written by Money et al. [46], three key external techniques are reviewed. The first one is contextual and unobtrusive as it does not require the user to give any information explicitly. It consists fixed-position video cameras and integrated pressure-based floor sensors tracking the location of user activity in the home during the shooting stage. The video can then be summarized according to the characters' positions in the house and their footsteps analysis. The two other reported studies are both obtrusive since based on manual annotations of the users [53, 54]. These descriptions are detailed and domain specific, in this case for baseball and soccer. Moreover, two out of the three techniques reported

Table 1 ELVIS results compared to random

	Comedy		Horror/comedy		Horror	
	Random	ELVIS	Random	ELVIS	Random	ELVIS
Mean on 20 users	30.37%	45.96%	31.34%	43.77%	28.38%	41.74%

produce personal summary and are therefore user-dependant. Although external summarization techniques might reach a general method, existing approaches are too costly to the end user.

Another existing external technique is the detection of personal highlights through facial expression [51]. This method tracks motion vector of 12 points on the participant's face. An experiment on ten participants was conducted. The participants watched eight video clips of different genres and reported highlight annotations at the end of the clip. The best precision result is 40%, and the authors reported that the best point on the face varies for each participant. Consequently, this method cannot be extended to a user-independent technique.

In the perspective of what has been done, the analysis of user's physiological responses for video summarization, a new user-based external summarization technique, can dramatically reduce the user effort and improve the external summarization techniques. This is an almost unexplored branch. The most and only advanced study, to our best knowledge, is *ELVIS* [47]. It is a complete system of personal video summarization based on five physiological signals (electrodermal activity, heart rate, blood volume pulse, respiration rate, and respiration amplitude). It is motivated almost by the same reasons than this project : video content elicits strong physiological responses from the user, highlight sequences elicit stronger responses, and these responses can be detected by recording physiological signals. The authors explore whether user's physiological responses can serve as new information to produce personalized video summaries. The result is a user personalized video summary corresponding to the favorite segments of the user. The signals from a single user are processed to obtain high and low values and are then combined to form a physiological significance index. Highlights are identified as the most significant segments. The system is precisely evaluated thanks to a large-scale experiment conducted over 60 participants. Not only physiological signals were recorded during these sessions but participants also reported what was the highlight sequences for them. This allowed a thorough analysis of their system (cf. Table 1). *ELVIS* has showed the usability of users' physiological responses as information for video summarization. However, they used single-user signals in order to compute personalized summary and therefore cannot propose a user-independent approach.

As a summary, most of the reported internal techniques focus on sports highlights which reduce their interest since they are not extendable to other genres of videos. One of the disadvantages of existing summarization techniques is their lack of users' self-reported highlights for ground truth construction. The evaluation of the techniques is then not based on the opinion of media consumers but on the evaluations of a unique experimenter. Finally, the few external summarization

techniques reported are designed to produce personalized summarization for a single user. This lack of generalization over both the type of videos (sports, drama, horror, comedy, etc.) and the population is the main drawback of existing techniques. In order to achieve a user-independent and content-independent video summarization technique, it is necessary to compute the summary based on the information gathered from several users and to evaluate such a system on a strong reference obtained through users' self-reports.

3 System Definition

The new user- and content-independent, unobtrusive external video summarization technique proposed in this chapter is defined in the following section.

3.1 Concept

The system aims to provide a novel highlight detection method in order to compute user-independent summaries of videos. The major challenges are the difficulty to reach a general method – that is, a method which detects sequences considered as relevant for a majority of viewers and content independent – and the difficulty to have an unobtrusive approach, that is, an approach which will not require any effort to the end user. Considering all the facts reported in the previous section about tagging, the right way forward seems to be a human-centered implicit tagging. This approach is the most likely to bridge the semantic gap since it relies on the user's comprehension of the multimedia content.

With the evolution of affective computing, recording of emotional reactions through physiological signals allows to detect user's affective changes. Using modern sensors, these parameters can be obtained without any conscious effort from the user. This represents therefore an unobtrusive technique able to provide reliable tags. Since emotional reactions are common to everyone, they can lead to user independent tags. The system is designed as a user-based external summarization technique which tags every instant of the video as highlight or non-highlight on the basis of the participants' physiological responses.

Although emotions make sense to everyone, not everyone has exactly the same reactions to movies; thus, to guarantee a user-independent method, it is necessary to combine the physiological signals of several participants together to obtain the user-independent highlight sequences. This is the main difference with the existing summarization techniques based on single-user physiological responses, which produces personalized summaries [47].

The concept of the system is to compute inter-users, physiological linkage to detect user-independent highlight sequences without any effort required from users. With this design, the system is likely to detect user-independent highlights.

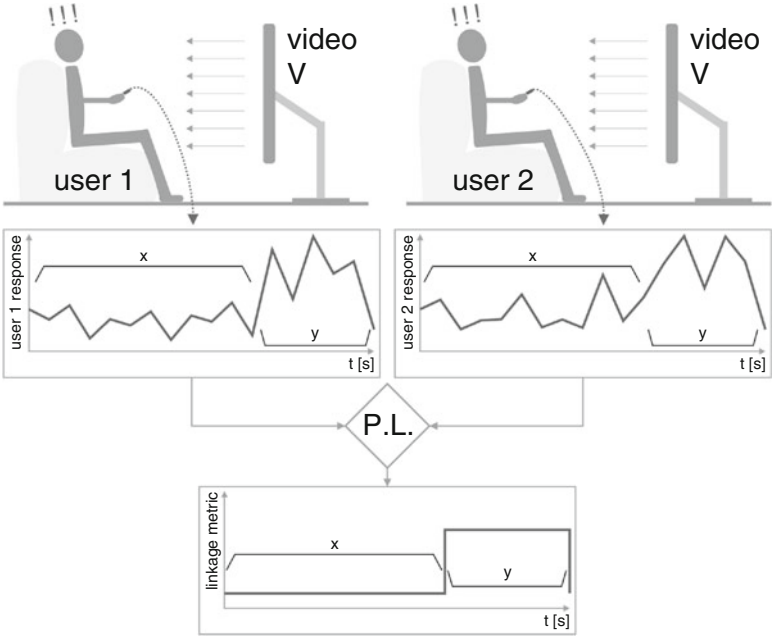


Fig. 1 Physiological linkage (P.L.) system concept

Moreover, with its social approach, it allows to detect social interaction, such as emotional contagion, which should enhance the generalizability of the algorithm. As the emotions improve the memory [3], the highlights detected with this system should be the most memorable ones. The method belongs to the category of external summarization techniques but only concerns the detection of highlights and does not include the summary reconstruction.

3.2 Physiological Linkage

The system concept is based on the linkage between several participants' emotional responses. These responses are represented by participants' physiological signals which are recorded during the viewing of the videos and stored as time series. Figure 1 presents the scheme of the physiological linkage concept with two participants (*user 1* and *user 2*). One of their physiological signals is recorded; while they are watching *video V*, the corresponding time series is presented for each participant. These two signals are then given to the physiological linkage (*P.L.*) unit which computes the physiological linkage and gives a metric vector of the linkage as a result. Highlight sequences can then be extracted from this metric as the highest values. On the figure, two periods can be identified. The *x* period corresponds

to the signals randomly oscillating around their mean and corresponding to a low physiological linkage (non-highlight). The *y* period corresponds to an event in *video V* inducing some coupling in the physiological activity and resulting in high physiological linkage (highlight). In this example, there is only one highlight sequence, but the system is designed to detect as many highlights as the scene contains.

The physiological linkage computation is realized with a sliding-window linear correlation between every possible pairs of participants. This metric was chosen for its simplicity and efficiency and because it is used in other works on physiological linkage [33, 39, 40]. This continuous linkage metric is computed on a window time frame which is shifted of half the window length at each iteration. When more than two participants' physiological signals are available, it results in as many linkage metric vectors as number of pairs. These vectors form then a matrix of linkage metrics, and the final vector is computed as the mean of all values at each time interval.

3.3 Sequence Classification

Once the global linkage metric vector is obtained for a video, the highlights are extracted by classification using the best limit on the metric. Therefore, several highlights can be detected in a scene. This detection is performed by a support vector machine (SVM) classifier in its quadratic mode, which was previously trained with the datasets presented in the *system evaluation* section. Two approaches are experimented: the first one uses only a single signal to extract the highlight and the second one combines multi-signals to extract the highlights. This second method aims to obtain better results by increasing the system dimensions and by taking advantage of physiological signal combination.

4 System Evaluation

The proposed system was tested on a database of physiological signals and its results are thoroughly analyzed to determine whether the user-independent, content-independent, and unobtrusive system goal is reached

4.1 Physiological Signals Recording and Preprocessing

The physiological signals used to test the proposed system were recorded in an experiment previously conducted for the purpose of emotional implicit tagging [55]. This dataset is composed of 64 scenes extracted from eight movies at the rate of eight scenes by movie. A subset of 26 scenes out of the 64 was created

Table 2 Physiological signal preprocessing

EMG	High-pass 10 Hz, absolute value, running average window of size 0.5 s, logarithm
BPM (from BVP)	Low-pass 5 Hz, beat detection, beat correction
EDA	Low-pass 3 Hz, logarithm
Skin temperature	Low-pass 1 Hz

by experimentators for this work. The selection is distributed among four major genres:

- Action: `Saving Private Ryan` and `Kill Bill, Vol.1`
- Drama: `Hotel Rwanda` and `The Pianist`
- Comedy: `Mr. Bean's Holiday` and `Love Actually`
- Horror: `28 Days Later` and `Ringu`

A selection of various movie genres is necessary to test the generality of the system. Each scene lasts about 2 min and, based on the experimentators' judgment, contains an emotional event.

The physiological signals were recorded for eight healthy participants, three females and five males from 22 to 40 years old. Six physiological signals were recorded, but only five are used in this work:

- *Electromyogram (EMG) from right zygomaticus major*: measure of the activation of this muscle, involved in smile and laughter
- *Electromyogram (EMG) from right frontalis*: measure of the activation of this muscle, involved in attention and surprise
- *Blood volume pulse (BVP)*, using a plethysmograph: measure of the relative change of blood pressure on the top of the thumb
- *Electrodermal activity (EDA)*: measure of the skin resistance between the index and the middle fingers
- *Skin temperature*: measure of the skin temperature with a temperature sensor placed on the top of the little finger

The preprocessing of each kind of signal was done according to the guidelines found in the literature [56–58] (see Table 2). For all the signal a high-pass filter was applied to remove drifts. Some signals were also corrected by computing their logarithm in order to normalize their range across participants. The envelope of the EMG signals was computed by taking the absolute value of the filtered EMG and applying a running average window. Heart beats were detected from the BVP signals by identification of local maximums.

4.2 Highlights in Scenes Reference

The demonstration of the system generality goes through a thorough evaluation which requires a highlight ground truth in the used scenes. This information is

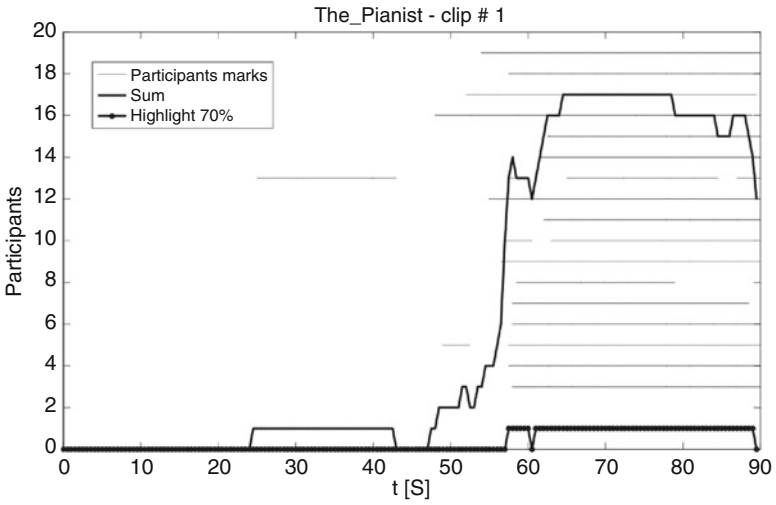


Fig. 2 Highlights ground truth example

compulsory to train and test the classifier and to compute statistical results. It was obtained during another experiment, conducted on 18 participants who did not participate in the physiological signal recording experiment. The large number of participants enabled to obtain the user-independent definition of highlights in the selected scenes.

Participants had to watch a scene once entirely and then were asked to mark the sequences they judged as highlights through a simple, original interface. The whole process was repeatedly done for each scene. The scenes were randomly presented to each participant. From these experiments, a vector containing the sum of highlights marked at each instant by the 18 participants was obtained for each scene. Each instant corresponds to each 0.5 s of the scene. An instant was finally judged as a highlight when 70% of the participants marked it. Figure 2 shows an example of result for a scene of `The Pianist` for which participants agreed on the highlight. The marked moment corresponds to a bombing near the main character after a peaceful sequence of piano. The horizontal thin lines represent each participant's marks, the curve is the sum, and the thick line is the final highlight definition used for the system evaluation.

As participants did not agree about highlight annotations for each scene, the coherence among participants' annotations was computed with the Fleiss' kappa. This was done to remove scenes on which the participants did not clearly agreed on the presence of highlights. This metric, ranging from 0 to 1, informs on the coherence, and thus the generality, of the highlights marked by the participants. Figure 3 shows the coherence distribution for the 26 scenes. It appeared clearly that the distribution was bimodal which conducted to the creation of a subset of scenes. This second dataset contains 13 scenes for which the kappa was over the median of the distribution and the highlights are general. The system was evaluated on both the 13 scene and the 26 scene datasets.

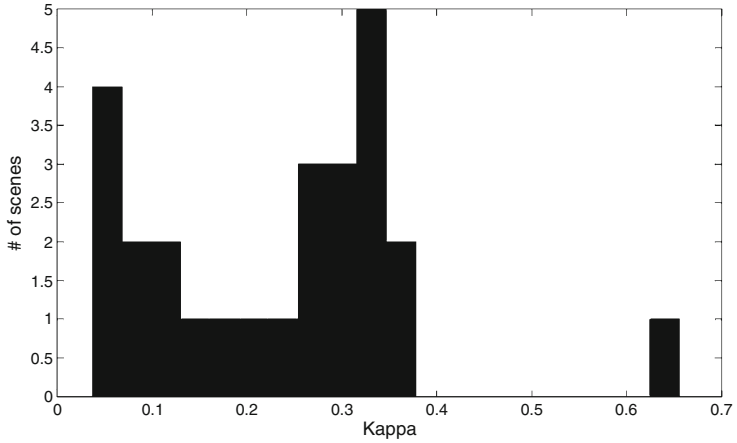


Fig. 3 Coherence distribution among the participants' annotations

4.3 Specifications for System Parameters

The physiological linkage unit of the system is based on the sliding-window linear correlation, which has two main parameters: the window length and the overlap between two consecutive windows. Considering that the scenes last about 2 min a window length of 10 s with a time lag of half the window was chosen.

The use of a 10-s window with a 5-s shift is optimal for the detection of events but can artificially create a delay. In addition, some signals have a long reaction time, especially the skin temperature. These two facts induce a delay between the stimuli – that is, the event in the video – and the reaction, that is, the change of the physiological linkage index, which has an impact on the system. Figure 4 presents an example of clear delay between the stimuli – the plain curve, and the reaction, the dashed curve. This figure also demonstrates that highlights can be detected in the physiological response. Indeed, the two highlight sequences with ground truth reported on the figure correspond in the video (a `28 Days Later` scene), respectively, to a very short movement of a dead character after a distressing period of calm which is surprising but not extremely relevant and to the assault of the main character by a zombie, which is an emotionally intense moment and therefore induces a strong physiological linkage. Several delays were then tested depending on the signal to realign the curves. Their choice is discussed in the next section.

4.4 Results

The system was evaluated on two setups, a single-signal analysis in which only one kind of signal at a time was used and a multi-signal analysis in which all signals

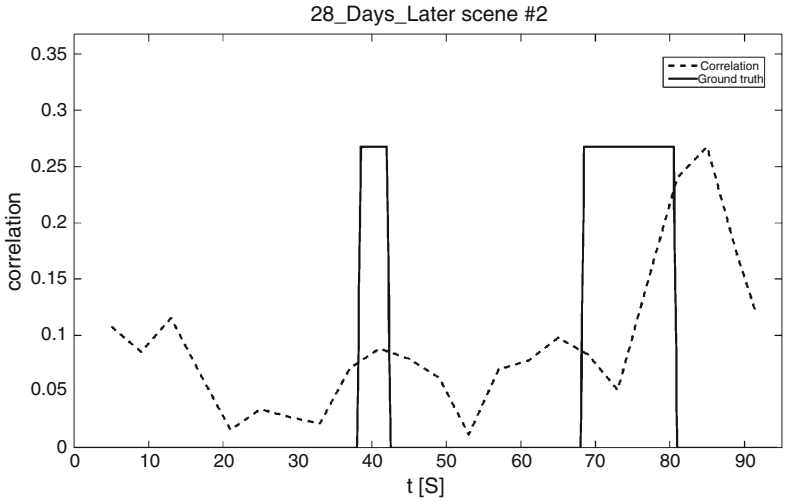


Fig. 4 Stimuli-reaction delay example, high value for ground truth represents highlight sequence

Table 3 Highlight detection results for the *skin temperature* signal with the delay influence

Delay	−10 [s]	*−8 [s]*	−6 [s]	−4 [s]	−2 [s]	0 [s]
ROC area	0.692	*0.701*	0.692	0.674	0.647	0.619
F1-score	0.547	0.549	*0.567*	0.549	0.549	0.515
Accuracy	0.766	*0.775*	0.760	0.745	0.723	0.720

were combined together. The results were analyzed through the *receiver operating characteristic (ROC)* curve, characterized by the area under the curve, the accuracy, and the *precision-recall (P-R)*, characterized by the F1-score.

The single-signal analysis allowed to identify the best signal for the highlight detection process as the *skin temperature* signal and also proved the significant, positive influence of the stimuli-reaction delay on the results. Highlight detection results for the skin temperature are presented in Table 3 in which the delay influence is clearly visible and the best delay results are displayed in bold. The *ROC* and *P-R* curves are displayed on Fig. 5 for the *skin temperature* signal with the best delay of 8 s. In this configuration, the system returned a correct classification rate of *77%*. In comparison, a random classifier obtains 50% and a classifier returning always the majority class (in this case non-highlight) obtains 72% of accuracy. The first and foremost observation is thus that the proposed system is able to identify user- and content- independent highlights in video scenes.

The second analysis which combined all the signals together aimed to improve the results by increasing the number of dimensions, which could lead to a better separation between the two classes, and by disclosing physiological dependencies among the signals. The classification was performed by a *support vector machine (SVM)* using its quadratic mode. The training/testing phases were conducted using

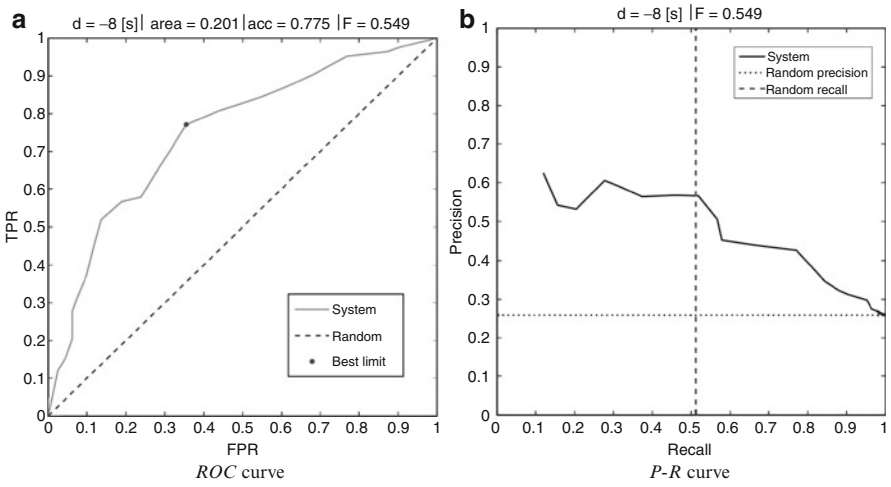


Fig. 5 *ROC* and *P-R* curves for the *skin temperature* signal with the 8-s delay (*d*). On the *top* of the figures, the area under the *ROC* curve (*area*), the accuracy (*acc*), and the F1-score (*F*) are displayed. (**a**) *ROC* curve and (**b**) *P-R* curve

a *leave-one-out* cross-validation technique. This second method did improve the results, obtaining *78.2%* of correct classification, but had only a limited influence. Nevertheless, it was proven that the system works even better by increasing the number of physiological signals involved in the highlight detection process.

What is surprising about the single-signal result is that it is the *skin temperature* signal which returned the best correct classification rate, whereas this signal was not expected to be the most informative for the detection process because of its very slow response time. On the contrary, the *EDA* signal was considered as the most promising signal because of its correlation with felt arousal and finally returned results inferior to the *skin temperature* signal, that is, 75% of correct classification. Even more surprisingly, the two *EMG* signals turned out to be totally uninformative; their results were not even better than what a random classifier would obtain. What the single-signal approach teaches is that using only one signal, a satisfactory, general, highlights detection method is obtained. The original system which is proposed in this chapter has then been proved to be workable.

Even though it has been shown that the multi-signal approach improved the results in comparison to the single-signal approach, does this approach really improve the system? Considering that four additional signals were necessary to obtain 1.2% of better classification, it may represent too much work for such a small benefit. If the design was applied on a larger scale, such as at home through the Web for online video summarization, the single-signal method, especially with the use of the *skin temperature* signal, is clearly advantageous since it only requires a cheap, simple sensor to record it.

As the viewing of a movie should elicit smiles and frowns, reactions driven by *EMGs* should be informative. This lack of emotional reaction demonstrated through the obtained results could be explained by the context of the physiological recording experiment: the participants were alone in front of a monitor in a laboratory environment, and short scenes were randomly submitted to their viewing. This is extremely different from a multi-person ecological situation in a theater or at home. The emotional contagion occurring when watching a film surrounded by several people in a theater was missing, and the participants could feel not very at ease in the laboratory and thus could limit their facial expressions. In addition, as there was no narrative context and the scenes duration were short, the participants were not allowed to be in the mood of the scene, and the randomly submitted scenes from various genres could induce residual affective states: a participant's emotional reaction to a funny scene differs depending on whether the previous scene was a horror scene or a comedy scene [2]. Consequently, the context of the physiological recording experiment appears to be a key point.

The system provides better results than the closest work, *ELVIS*, which only obtains between 40 and 45% of correct classification on three genres of videos [47] and than the work using facial expressions which obtains an F-score of only 0.150 for highlight detection [59]. In addition, though many internal summarization techniques obtain better result in their own context, for example, 86.4% of accuracy for sports highlights correctly detected in [49], they are certainly not capable to detect highlights for other kinds of videos. However, the proposed system still suffers from two drawbacks: too many false positives were retrieved and the delays were chosen without cross-validation. These issues must be resolved through future work.

5 Conclusion

This project aimed to provide a user-independent and content-independent video highlight detection method based on inter-users, physiological linkage. Confidence was granted to this approach on the basis of its novel use of unobtrusive implicit human-centered tagging able to take advantage of the social interaction occurring between spectators and their common emotional interpretation of a movie. The system uses correlation on inter-users, physiological signals to compute the physiological linkage during the scene. It then tags the video sequences as highlight or non-highlight depending on the physiological linkage metric. The very interesting results obtained, 77% of correct classification with only the *skin temperature* signal and 78.2% of correct classification with all the signals combined together, demonstrate that the system works better than existing external summarization techniques and is context independent.

However, several limitations have been identified, such as the physiological recording context, the low precision, and the empirical choice of the delays. The

following future work is thus necessary. Firstly, the ecological and physiological recording should be conducted on complete movies with several participants at the same time in an actual cinema. Of course, this raises many difficulties, but the results and the hypothetical improvements obtained will be extremely interesting. Secondly, the system precision as long as the system results could benefit from the use of another physiological linkage unit. Since the correlation is a very simple, linear method, a more complex, nonlinear mathematical tool, such as the synchronization likelihood, is likely to improve the results. Finally, the delay choice must be asserted through cross-validation.

The implemented system led to results which enabled the verification of the research hypothesis: the moments when viewers' physiological responses are highly linked correspond to highlight sequences of the video. The feasibility of a user-independent and content-independent, unobtrusive method for highlight detection in videos using inter-users, physiological linkage has been demonstrated. Even if a great amount of work should still be accomplished to reach a complete summarization tool, it represents a progress in external video summarization method, which could be, in the near future, widely used.

References

1. Pantic, M., Vinciarelli, A.: Implicit human centered tagging. IEEE Signal Process. Mag. **26**, 11 (2009)
2. Gross, J.J., Levenson, R.W.: Emotion elicitation using films. Cognit. Emot. **9**(1), 87–108 (1995)
3. Hamann, S.: Cognitive and neural mechanisms of emotional memory. Trends Cognit. Sci. **5**(9), 394–400 (2001)
4. Picard, R.W.: Affective computing. M.I.T media laboratory perceptual computing section technical report, 321 (1995)
5. Zeng, Z.H., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
6. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. Int. J. Hum. Comput. Stud. **67**(8), 607–627 (2009)
7. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. Proc. IEEE **91**(9), 1370–1390 (2003)
8. Gottman, J.M.: Detecting cyclicity in social-interaction. Psychol. Bull. **86**(2), 338–348 (1979)
9. Money, A.G., Agius, H.: Analysing user physiological responses for affective video summarisation. Displays **30**(2), 59–70 (2009)
10. Levenson, R.W., Gottman, J.M.: Marital interaction: physiological linkage and affective exchange. J. Personal. Soc. Psychol. **45**(3), 587–597 (1983)
11. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32**, 198–208 (2006)
12. Golder, S., Huberman, B.A.: Huberman: Usage Patterns of Collaborative Tagging Systems. J. Inf. Sci. **32**(2), 198–208 (2006)
13. Nov, O., Naaman, M., Ye, C.: What drives content tagging: the case of photos on flickr. In: Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pp. 1097–1100. ACM, New York(2008)

14. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07, pp. 971–980. ACM, New York (2007)
15. Jiao, J., Pantic, M.: Implicit image tagging via facial information. In: Proceedings of the 2nd International Workshop on Social Signal Processing, SSPW '10, pp. 59–64. ACM, New York (2010)
16. Koelstra, S., Mühl, C., Patras, I.: Eeg analysis for implicit tagging of video data. In: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 27–32. IEEE, Los Alamitos (2009)
17. Yazdani, A., Lee, J.S., Ebrahimi, T.: Implicit emotional tagging of multimedia using eeg signals and brain computer interface. In: Proceedings of the First SIGMM Workshop on Social Media, WSM '09, pp. 81–88. ACM, New York (2009)
18. Soleymani, M., Chanel, G., Kierkels, J.J.M., Pun, T.: Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In: Proceedings of the Tenth IEEE International Symposium on Multimedia, pp. 228–235, 15–17 Dec (2008)
19. Turing, A.M.: Computing Machinery and Intelligence, pp. 11–35. MIT, Cambridge (1995)
20. Lisetti, C., Lerouge, C.: Affective computing and tele-home health. In: Proceedings of the 37th Hawaii International Conference on System Sciences, pp. 148–155, Orlando, FL, USA, 5–8 Jan. (2004)
21. Paiva, A., Prada, R., Chaves, R., Vala, M., Bullock, A., Andersson, G., Höök, K.: Towards tangibility in gameplay: building a tangible affective interface for a computer game. In: Proceedings of the 5th International Conference on Multimodal Interfaces, ICMI '03, pp. 60–67. ACM, New York (2003)
22. Chanel, G., Rebetez, C., Betrancourt, M., Pun, T.: Emotion assessment from physiological signals for adaptation of games difficulty. IEEE Trans. Syst. Man Cybern. Part A **41**(6), 1052–1063 (2011)
23. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1175–1191 (2001)
24. Kim, K., Bang, S., Kim, S.: Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. **42**, 419–427 (2004). doi:10.1007/BF02344719
25. Scheirer, J., Fernandez, R., Klein, J., Picard, R.W.: Frustrating the user on purpose: a step toward building an affective computer. Interact. Comput. **14**(2), 93–118 (2002)
26. Katsis, C.D., Katertsidis, N., Ganiatras, G., Fotiadis, D.I.: Toward emotion recognition in car racing drivers: a biosignal processing approach. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **38**(3), 502–512 (2008)
27. Rani, P., Liu, C., Sarkar, N.: An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Anal. Appl. **9**, 58–69 (2006)
28. Varni, G., Volpe, G., Camurri, A.: A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Trans. Multimed. **12**(6), 576–590 (2010)
29. Hatfield, E., Cacioppo, J.T., Rapson, R.L.: Emotional contagion. Curr. Dir. Psychol. Sci. **2**, 96–100 (1993)
30. Jakobs, E., Fischer, A.H., Manstead, A.S.R.: Emotional experience as a function of social context: the role of the other. J. Nonverbal Behav. **21**(2), 103–130 (1997)
31. Pentland, A.: Social signal processing [exploratory dsp]. Signal Process. Mag. IEEE **24**(4), 108–111 (2007)
32. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009). Visual and multimodal analysis of human spontaneous behaviour
33. Ekman, I., Chanel, G., Kivikangas, J.M., Salminen, M., Järvelä, S., Ravaja, N.: Social interaction in games: measuring physiological linkage and social presence. Simul. Gaming **43**(3), 321–338 (2012)

34. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic nervous-system activity distinguishes among emotions. Science **221**(4616), 1208–1210 (1983)
35. Coplan, A.: Catching characters' emotions: emotional contagion responses to narrative fiction film. Film Stud. **8**, 26–38 (2006)
36. Stam, C.J.: Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. Clin. Neurophysiol. **116**(10), 2266–2301 (2005)
37. Levenson, R.W., Gottman, J.M.: Marital interaction: physiological linkage and affective exchange. J. Personal. Soc. Psychol. **45**(3), 587–597 (1983)
38. Levenson, R.W., Ruef, A.M.: Empathy: a physiological substrate. J. Personal. Soc. Psychol. **63**(2), 234–246 (1992)
39. Henning, R.A., Boucsein, W., Gil, M.C.: Social-physiological compliance as a determinant of team performance. Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol. **40**(3), 221–232 (2001)
40. Henning, R.A., Armstead, A.G., Ferris, J.K.: Social psychophysiological compliance in a four-person research team. Appl. Ergon. **40**(6), 1004–1010 (2009)
41. Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., Garnero, L.: Inter-brain synchronization during social interaction. PloS one **5**(8), 1–10 (2010)
42. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R.: Intersubject synchronization of cortical activity during natural vision. Science (New York) **303**(5664), 1634–1640 (2004)
43. Hasson, U., Ghazanfar, A.A., Galantucci, B., Garrod, S., Keysers, C.: Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends in Cognit. Sci. **16**, 114–121 (2012)
44. Ngo, C.-W., Ma, Y.-F., Zhang, H.-J.: Automatic video summarization by graph modeling. IEEE Int. Conf. Comput. Vis. **1**, 104 (2003)
45. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE Trans. Multimed. **7**, 1097–1105 (2005)
46. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. J. Vis. Commun. Image Represent. **19**(2), 121–143 (2008)
47. Money, A.G., Agius, H.: Elvis: entertainment-led video summaries. ACM Trans. Multimed. Comput. Commun. Appl. **6**, 17:1–17:30 (2010)
48. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Trans. Image Process. **12**(7), 796–807 (2003)
49. Wang, J., Xu, C., Chng, E., Tian, Q.: Sports highlight detection from keyword sequences using hmm. In: Proceedings of the IEEE ICME, Taipei, pp. 27–30 (2004)
50. Li, J., Wang, T., Hu, W., Sun, M., Zhang, Y.: Soccer highlight detection using two-dependence bayesian network. IEEE Int. Conf. Multimed. Expo **0**, 1625–1628 (2006)
51. Joho, H., Staiano, J., Sebe, N., Jose, J.M.: Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. Multimed. Tools Appl. **51**, 505–523 (2011)
52. Aizawa, K., Tancharoen, D., Kawasaki, S., Yamasaki, T.: Efficient retrieval of life log based on context and content. In: Proceedings of the the 1st ACM workshop on Continuous Archival and Retrieval of Personal Experiences, CARPE'04, pp. 22–31. ACM, New York (2004)
53. Jaimes, A., Jaimes, R., Echigo, T., Teraguchi, M., Satoh, F.: Learning personalized video highlights from detailed mpeg-7 metadata. In: MPEG-7 Metadata, in Proceedings of the ICIP, p. 2002. IEEE, Piscataway (2002)
54. Takahashi, Y., Nitta, N., Babaguchi, N.: Video summarization for large sports video archives. In: Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2005, Amsterdam, pp. 1170–1173 (2005)
55. Soleymani, M., Chanel, G., Kierkels, J.J.M., Pun, T.: Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In: Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia, ISM '08, pp. 228–235. IEEE, Washington, DC (2008)
56. Fridlund, A.J., Cacioppo, J.T.: Guidelines for human electromyographic research. Psychophysiology **23**(5), 567–589 (1986)

57. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G.: Handbook of Psychophysiology. Cambridge University Press, Cambridge/New York (2000)
58. Dawson, M.E., Schell, A.M., Filion, D.L.: The electrodermal response system. In: Cacioppo, J.T., Tassinary, L.G. (eds.) Principles of Psychophysiology: Physical, Social and Inferential Elements, pp. 295–324. Cambridge University Press, Cambridge (1990)
59. Joho, H., Jose, J.M., Valenti, R., Sebe, N.: Exploiting facial expressions for affective video summarisation. In: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09, pp. 31:1–31:8. ACM, New York (2009)

Toward Emotional Annotation of Multimedia Contents

Ashkan Yazdani, Jong-Seok Lee, and Touradj Ebrahimi

Abstract By annotating multimedia contents, users of a web resource can associate a word or a phrase (tag) with that resource such that other users can retrieve it by means of searching. Nowadays, tags play an important role in search and retrieval process in multimedia content sharing social networks. Explicit tagging refers to assigning tags directly in an explicit way such as typing. Implicit tagging, however, refers to assigning tags by observing users' behaviors during exposure to multimedia contents. Among various kinds of information that can be obtained for the purpose of implicit tagging, emotional information about a given content is of great interest. In this chapter, we discuss various means of emotion recognition and emotional characterization, which can be used as tools for emotional tagging. A P300-based brain-computer interface system is proposed for the purpose of emotional tagging of multimedia content. We show that this system can successfully perform emotional tagging and naive users who have not participated in the training of the system can also use it efficiently. Furthermore, we present emotional annotating systems using multimedia content analysis and electroencephalogram signal processing and will compare them. Finally, a road map for developing a practical multimodal system for implicit emotional annotation of multimedia contents will be sketched out.

This work is performed in the framework of European Community's Seventh Framework Program (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia) and the Swiss National Foundation for Scientific Research. The authors would also like to thank Krista Kappeler for her contribution on emotional characterization using MCA.

A. Yazdani (✉) • T. Ebrahimi
Multimedia Signal Processing Group (MMSPG), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
e-mail: ashkan.yazdani@epfl.ch; touradj.ebrahimi@epfl.ch

J.-S. Lee
School of Integrated Technology, Yonsei University, Incheon 406-840, Korea
e-mail: jong-seok.lee@yonsei.ac.kr

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_11, © Springer-Verlag London 2013

1 Introduction

With the rapid evolution of digital media, services such as social networks, web-based multimedia archives, and search engines are becoming increasingly popular, and, consequently, the amount of multimedia data on the Internet is escalating exponentially. Therefore, developing appropriate schemes for efficiently performing search and retrieval in immense multimedia databases becomes significantly important. Numerous efforts have been made to automatically analyze multimedia content for generating relevant content labels or descriptors that can be used for search and retrieval [2, 44]. However, automatic multimedia content analysis (MCA) approach has not yet satisfied users' expectations, which could be mainly attributable to the so-called semantic gap. MCA is incapable of adequately incorporating the process of human interpretation of multimedia content. A prevalent trend to solve this problem is to attach short nonhierarchical textual annotations describing the content, which are known as tags. In other words, a tag is a form of metadata that provides users of a multimedia content with information about it and also facilitates search retrieval processes for that content. Nowadays, tagging is increasingly performed in many social networks and web pages that provide multimedia contents and is an important feature of many Web 2.0-based services.

In general, there are two approaches to assign tags to a given content, namely, explicit and implicit tagging. The former refers to a user's explicit action of manually entering appropriate keywords associated with the content, whereas in the latter approach, users do not necessarily input tags and automatic analysis of the users' behavior is used to generate tags for the content. Explicit tagging is used in most of the currently available social network-based systems such as YouTube and Flickr. Nevertheless, it cannot be considered as the ultimate solution for assigning tags to multimedia contents due to the following reasons: First, manual explicit annotating of the enormous amount of multimedia data on the Internet is clearly infeasible, and hence, a large portion of the multimedia data either remains untagged or is annotated with unuseful information, which in turn imposes limits on the search and retrieval process and deteriorates the performance of multimedia search engines. Secondly, users who annotate multimedia contents do not necessarily aim at improving the performance of the current retrieval systems. In fact, personal and social motivations often underlie the annotation [4]: A personal need-driven tag may be meaningless to other users, e.g., my lovely granddaughter in my place. Furthermore, it is also possible that some users generate tags to increase their reputation, e.g., spam tags for advertisement. Therefore, complementary and/or alternative annotation methods are needed to overcome the aforementioned shortcomings of explicit tagging. Implicit tagging can be considered as one such solution. Implicit tagging allows the annotation of a multimedia content, each time a user interacts with it based on his/her reactions to the multimedia content (e.g., laughter when seeing a funny video) [32].

Among various kinds of information that can be obtained for the purpose of tagging, emotional information about a given content is of great interest. Emotional

information plays an important role for personalized as well as social content delivery [11]. For example, users may prefer to watch video clips containing funny contents when they feel sad or depressed in order to improve their mood. Another example would be that some users might not want to watch video clips containing scary or violent scenes. Finally, users may want to have multimedia content recommendations, which match their current or desired affective state. In such cases, emotional information about the content can be used effectively in search and retrieval [17].

In this chapter we discuss systems, which can extract emotional values of multimedia contents and/or emotions induced in people while consuming multimedia contents. The rest of this chapter is organized as follows: Sect. 2 provides the emotion assessment methodology and a review of state of the art on the means of emotion recognition. Section 3 presents the result of our work on development of a BCI system which can be used to emotionally annotate multimedia contents. In Sect. 4, the results of emotion recognition using and electroencephalogram (EEG) signal processing during the watching of music video clips and emotional characterization using MCA will be presented, and finally, Sect. 5 presents a road map for developing a multimodal implicit multimedia annotator for social media retrieval.

2 Emotion Recognition

An appropriate modeling for emotion must be developed and used, in order to assess users' affective state. How to represent and model emotions is, however, a challenging task. Until today, numerous theorists and researchers have conducted research on this subject, and consequently a large amount of literature exists today with sometimes very different solutions [29]. Generally, there are two different families of emotion models: the categorical models and the dimensional models. The rational for the categorical models is to have discrete basic categories of emotions from which every other emotion can be built by combining these basic emotions. The dimensional models, on the other hand, describe the components of emotions and are often represented as a two- or three-dimensional space, where the emotions are presented as points in the coordinate space of these dimensions. The goal of the dimensional model is not to find a finite set of emotions as in the categorical model but to find a finite set of underlying components of emotions [35, 38]. In this work, Russell's arousal-valence dimensional model of emotion will be used to quantitatively analyze the emotions. The dimension "valence" provides information about the degree of pleasantness of the content and ranges from pleasant (positive) to unpleasant (negative). The dimension "arousal" represents the inner activation and ranges from energized to calm. In these scales, each emotional state can be placed on a two-dimensional plane with arousal and valence as the horizontal and vertical axes. While arousal and valence explain most of the variation in emotional states, a third dimension of dominance may be also included in the

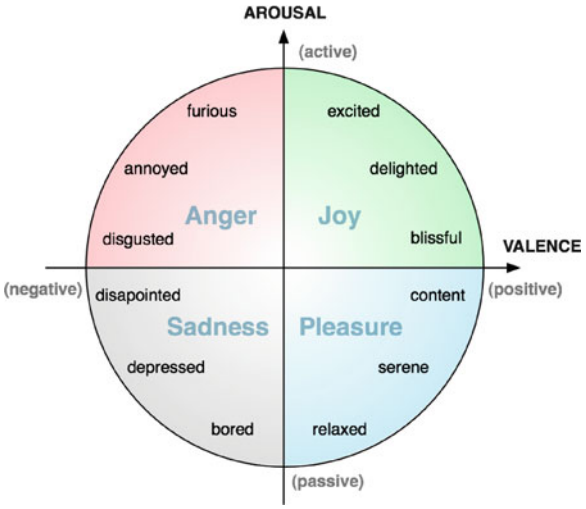


Fig. 1 Arousal-valence space and basic emotions

model [38], which helps to distinguish between "grief" and "rage" and goes from no control to full control. Figure 1 illustrates the arousal-valence space and the distribution of basic emotions on this space [18].

Until today, many research studies have investigated human facial expressions [15, 46], speech prosody [43], and the fusion of different modalities [8, 10, 40] to extract information about users' affective states. Physiological signals originating from the peripheral nervous system (PNS) are also known to convey traces of emotion, and they have been studied for the aim of emotion recognition [12, 18, 23]. Kim and André [18] investigates the potential of physiological signals as reliable channels for emotion recognition during music listening, and they showed that for four emotional states of three users, an average recognition accuracy of 95% can be achieved. Kim et al. [19] presented a physiological signal-based emotion recognition system induced by combination of photos and music. They showed that an average correct classification ratios of 78.4 and 61.8% can be achieved for recognition of three and four categories, respectively. Lisetti and Nasoz [28] investigated physiological changes during watching movie scenes, and they showed that a high recognition rate of 84% for the recognition of six emotions can be achieved.

Furthermore, a few studies have examined the feasibility of analyzing EEG signal, emanating from the central nervous system (CNS), for extracting information about human emotion [3, 14, 21, 22]. Chanel et al. [7] studied the changes in EEG signals during watching emotion evocative images and showed that an accuracy of 58% for classification of three emotions can be achieved. Schaaff and Schultz [39] used a headband EEG acquisition interface and support vector machines (SVMs) to recognize three emotional categories induced by images. In [27], power spectrum density of different EEG sub-bands were extracted as features during different emotions induced during listening to music, and a correct classification of 82%

for four emotions was achieved. Petrantonakis and Hadjileontiadis [33] studied the changes in the EEG signal of users when presented with images of faces expressing six basic emotions. They showed that a classification accuracy of 83% can be achieved using features based on higher-order crossings and support vector machine classification.

3 BCI for Emotional Tagging

In this section, we propose a novel application of EEG-based brain-computer interface (BCI) for emotional tagging of multimedia contents. It is worthy of mention that, unlike the previous EEG-based emotion recognition research studies mentioned above, the proposed system obtains the emotional information of a user in an indirect way in order to achieve high accuracy. It is shown that our system can successfully perform emotional tagging for naive users who had never experienced training sessions before. Moreover, we introduce a measure of easiness of tagging for a given content, namely, emotional taggability (ET), which is obtained from another set of subjective assessments. This measure is used to analyze the recognition performance of the system. It is shown that the ET measure and the system performance have a correlation in that, for a content with a low ET value (i.e., a high ambiguity of expected emotional reaction), the recognition performance is relatively poor, and vice versa.

3.1 Experimental Protocol

Users were positioned in front of a desktop monitor and were asked to watch several video clips. These video clips were collected from YouTube for our experiments. Four clips were chosen for each of the six basic emotional categories defined by Paul Ekman (i.e., joy, sadness, surprise, disgust, fear, and anger), and thus, 24 clips were used in total for the experiment. In order to ensure that the duration of the test session remains reasonable, we mostly chose relatively short video clips. The minimum, mean, and maximum lengths of the clips were 15, 58, and 161 s, respectively. Immediately after each video was finished, six images were displayed on the screen. These images were happy, sad, surprised, disgusted, afraid, and angry faces representing the six basic emotions. They were flashed in a pseudorandom order, one image at a time. During each flash, one image was intensified for 100 ms followed by 300 ms during which none of the images were intensified so that an interstimulus interval of 400 ms was achieved. The EEG signals were acquired at a 2,048-Hz sampling rate from 32 electrodes that were placed on the scalp of the users according to the 10–20 international electrode positioning system. A Biosemi Active Two amplifier was used for amplification and analog-to-digital conversion of

the recorded EEG signal. Signal processing and pattern recognition algorithms used in this study were implemented in Matlab. These algorithms and interfaces were implemented and tested initially for P300-based environment control BCI [13].

We trained the BCI system with eight healthy users, Ph.D. students recruited by our laboratory (all male, age 29 ± 3.4). None of them had any known neurological deficiencies. After training a general classifier, we tested the system with four other users (all male, age 29 ± 1.5), who had never used the system. In the training phase, each user was asked to complete four training sessions for recording the EEG signals. The first two sessions were performed during 1 day and the remaining two sessions on another day within 2 weeks after the first session. Each session consisted of six runs, one run for each image. During each run, the images of the GUI were flashed in a pseudorandom order. The users were asked to perform a covert task, i.e., silently count how many times a prescribed image was intensified (e.g., now please count how often the sad face flashes). After this massage, the six images were shown on the screen and a warning beep was issued. The images then started to flash according to a random sequence starting 4 ms after the preparation beep, and simultaneously the EEG signals were recorded. The sequence of images to be flashed was block-randomized. In other words, after each block (6 flashes), each image was flashed one time, and after two blocks (12 flashes), each image was flashed twice and so forth. The number of blocks inside each run was selected randomly between 20 and 25. Therefore, for instance, a sequence might have included 23 blocks, which provided 23 target (P300) trials together with $23 \times 5 = 115$ non-target (non-P300) trials. At the end of each run, the users were asked to report the result of their counting. This number was then compared to the actual number of blocks to monitor the performance of the user and also to know whether he/she was concentrated throughout the test. Based on the data recorded in the first session, a simple classifier was built, and at the end of each run during the second, third, and fourth sessions, the image inferred by the classification algorithm was flashed five times so that the users could have a feedback of their performance. The duration of one run was approximately 1 min, for a mean value of 22.5 blocks and six image intensification of 0.4 s long each inside a block, and the 4 s preparation time, i.e., $(22.5 \times 6 \times 0.4) + 4 = 58$ s. Each session lasted approximately 30 min, including the setup of electrodes and short breaks between runs, and comprised on average of 810 trials. The whole data gathered for each user consisted on average of 3,240 trials.

3.2 Data Processing

The recorded EEG data was needed to be preprocessed, and some features had to be extracted from each trial in order to perform classification and to train the discriminating functions. In this section, the preprocessing and feature extraction methods are described. During the signal acquisition, two electrodes were placed on the mastoids of the user. The average signal of these two electrodes was used for referencing. In the next step, a sixth-order forward-backward Butterworth band-pass

filter with zero phase shift was used to filter the data. The cutoff frequencies of the band-pass filter were set to 1.0 Hz and 12 Hz. The data was then downsampled from 2,048 to 32 Hz. To extract the single trials from the whole EEG data acquired during each run, windows of the duration 1,000 ms were extracted from the data. The stimulus presentation interface provided the exact system clock of the stimulus onset, which was used for windowing the signal. Single trials started at the stimulus onset, i.e., at the beginning of the intensification of an image, and ended 1,000 ms after the stimulus onset. It is worthy of mention that the last 600 ms of each single trial overlapped the 600 ms of its subsequent single trial, due to the fact that the interstimulus interval was set to 400 ms. After the data was broken down into single trials and downsampled, it needs to be purified from artifacts. During standard EEG acquisition, eye blinks, eye movements, muscle activity, or user movement can cause large-amplitude outliers in the recorded signal. As the sources of these peaks are not directly related to brain activities, they might influence the results of the classification in that they can simply be mistaken as P300 peaks. To reduce the effects of such outliers, the signal from each electrode was windsorized in the following manner. For the samples of each electrode, the 10th percentile and the 90th percentile were computed. Amplitude values that fall below the 10th percentile or above the 90th percentile were then replaced by the 10th and 90th percentiles, respectively. In the next step, the samples of each electrode were scaled to the range $[-1, 1]$, and finally these normalized samples were concatenated to constitute the feature vectors. Considering that the number of electrodes was 32 and the number of decimated temporal samples was 32 for each single trial, the dimensionality of each feature vector representing each single trial was $32 \times 32 = 1,024$. For performing the classification, the Bayesian linear discriminant analysis (BLDA) was used. BLDA can be seen as an extension of Fisher's linear discriminant analysis (FLDA). In contrast to FLDA, BLDA uses regularization to prevent overfitting to high-dimensional and possibly noisy datasets. Through a Bayesian analysis, the degree of regularization can be estimated automatically and quickly from training data without time-consuming cross validation. Algorithms that are closely related to this method are the Bayesian least-squares support vector machine and the algorithm for Bayesian nonlinear discriminant analysis [6]. BLDA is also closely related to the so-called evidence framework [5].

3.3 Results

As explained in the previous section, we have developed a general classifier using the training data gathered from eight users and tested this classifier with four additional naive users who had never been through the training phase. More precisely, in the test experiment, each of the four users was asked to watch 24 video clips. One run of BCI was performed immediately after each video ended. The total duration of each test session, including the setup of the EEG signal acquisition equipment, was around 90 min. Before the beginning of the test experiment, the

Table 1 Rates of the correctly annotated video clips using the proposed BCI system for the test users

User1	User2	User3	User4	Average
79.17%	91.67%	70.83%	79.17%	80.19%

users were asked to select one image on the screen using their brainwaves, so that an appropriate number of blocks (B) for each user can be defined. In this manner, a proper value of B was chosen for each test user separately, which varied between 4 and 10. For each run in the test session, the single trials corresponding to the B blocks were extracted using the processing techniques before. In the next step, the single trials were classified using the BLDA classifier. The classifications of single trials resulted in B blocks of outputs so that each block consisted of six classifier outputs, one output for each image on the display. In order to make the final decision about the selected image, i.e., to recognize which image was selected by the user, the classifier outputs were summed over the B blocks for each image, and finally the image with the maximum summed classifier output value was selected. Table 1 shows the performance of the four users for annotation of 24 video clips. On average, we obtained about 80% of tagging accuracy. While the users were asked to choose only one emotional category for each video clip in our system, such a task may be difficult for some video clips due to ambiguity of the messages conveyed in the clips or simultaneous elicitation of multiple emotions. This difficulty would deteriorate the accuracy of the recognition in our system in that the EEG signal recording and the image presentation GUI are already running but the user still hesitates to choose one among the six emotional categories, and consequently he/she is unable to generate P300 patterns properly. Therefore, we conducted an additional subjective experiment for further analysis of the relationship between the recognition performance of our system and the difficulty of tagging for the users. More precisely, we asked another group of users (18 people) to rate the video clips by assigning an integer value from 0 to 10 for each of the six emotional categories. For example, if a user did not feel sad at all for a clip, he/she entered "0" for sadness; if he/she felt that the video clip was very sad, he/she entered "10" for sadness. Thus, for each video clip and for each user, six integer numbers corresponding to the six emotional categories (e.g., {3 6 0 8 10 2}) were obtained. To perform this subjective test, a web page containing all the 24 video clips used in the test was created. From the rating values obtained in the subjective test, we defined a measure of emotional taggability for each video clip. This measure indicates the feasibility and easiness of assigning an emotional tag to the given content when using the proposed system. It is defined in a way that, if the six integer values are low and similar, the ET measure should also be low. On the other hand, we would rather have a large ET value when only one of the six values is dominantly large compared to the others. Defining a measure for this purpose has been already addressed for classifier fusion in the field of audiovisual speech recognition, where it is necessary to compare the reliability of different classifiers based on their outputs for the target classes; a classifier showing

large difference between the probabilities of the given input for different classes is considered as a reliable one due to its large discriminability between the classes. Several candidates for measuring the reliability have been proposed, e.g., entropy, variance, dispersion, and average differences [36]. However, it has been shown that many of them are not appropriate due to their intrinsic errors. For example, the variance measure cannot distinguish between {10, 0, 0, 0, 0, 0} and {10, 10, 10, 10, 10, 0}, which are in fact completely different cases. On the other hand, the following definition has shown to be effective [25]:

$$\mathrm{ET1} = \frac{\sum_{i=1}^{M}(\max_{1\le j\le M}(e_j) - e_i)}{(M-1)E_{\max}} \tag{1}$$

where $E_{\max}$ is the maximum rating value ($E_{\max} = 10$ in our case) and e_i the rating value of the ith emotion among the M emotional categories ($M = 6$ in our case). This measure has minimum and maximum values of 0 and 1, respectively, due to the normalization factor in the denominator. Nevertheless, the above measure is still insufficient for our purpose of measuring ET because it produces the same value for the two cases {10, 5, 5, 5, 5, 5} and {5, 0, 0, 0, 0, 0}, while a higher ET value would be more suitable for the first case. Therefore, the measure given above is weighted by the maximum value among e_is, and, thus, we finally define the ET of a video clip as

$$\mathrm{ET2} = \frac{\max_{1\le j\le M}(e_j)\sum_{i=1}^{M}(\max_{1\le j\le M}(e_j) - e_i)}{(M-1)E_{\max}^2} \tag{2}$$

Again, note that this is a normalized measure so that it ranges between 0 and 1. Figure 2 illustrates the average rating values for the 18 users, sorted in an ascending order with respect to their ET values. The bright and the dark pixels represent the high and low rating values, respectively. It can be seen that a video clip with a high ET value induces only one dominant emotion. For instance, when we compare video number 10 and number 21, the ratings are higher for video number 10 on average, but video number 21 has a higher ET value as only one rating is dominant and the others are nearly zero. Figure 3 shows the relationship between the ET measure and the emotional tagging performance for the two different definitions of ET presented above. The tagging accuracy was measured among the four test users of the system. For example, an accuracy of 75% means that three among the four users could tag the video successfully. It is observed that the proposed measure, ET2, has a relatively higher correlation with the accuracy when compared with ET1, which supports our analysis of the recognition performance based on ET. It appears that, if the ET value for a given video clip is sufficiently large (e.g., 0.04), the proposed BCI system can be used to tag the clip with a high accuracy. As it can be seen in Fig. 4b, the video that has the second highest ET value resulted in a recognition accuracy of 75%. This video was the 21st video during the test session, and the user who made the error while tagging this video reported that he was too tired to fully concentrate, since the presentation of this video and the annotation occurred almost 80 min after the beginning of the session.

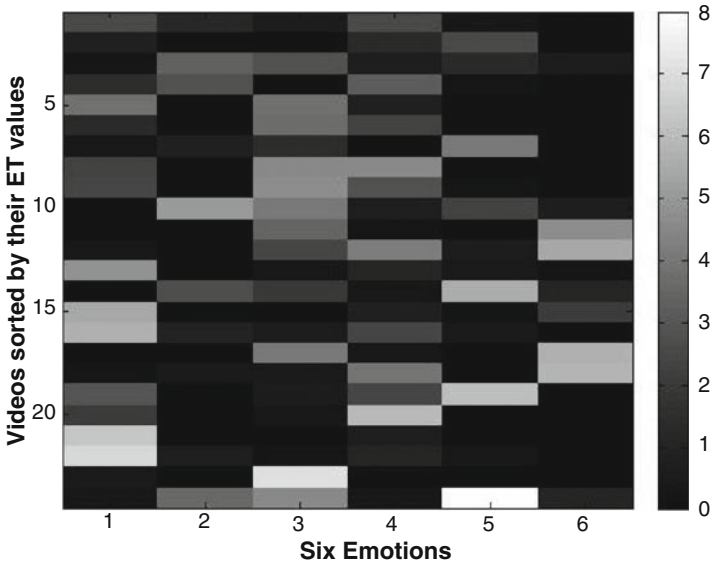


Fig. 2 The ratings of the video clips sorted by the ET value of videos (i.e., video 1 and video 24 have the lowest and highest ET values, respectively)

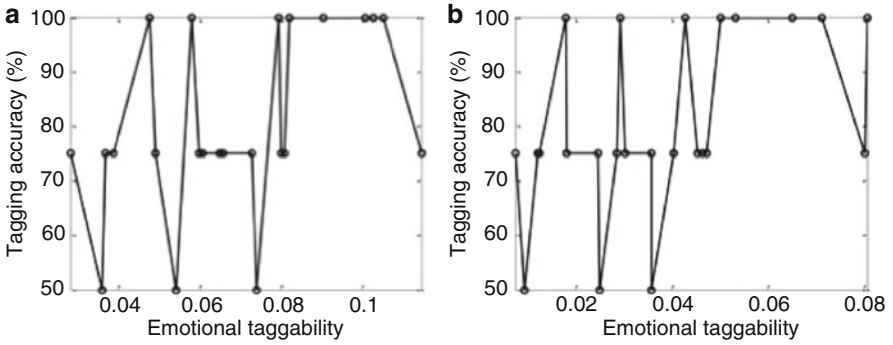


Fig. 3 Emotional taggability (ET) measure versus performance of the proposed emotional tagging system for two different definitions of the (**a**) ET1 (**b**) ET2

Further improvements to the work presented in this section might be to test the developed system with larger numbers of users in order to confirm the results obtained in the present work. Moreover, shorter interstimulus intervals can be investigated to assess the possibility of having systems with reduced time complexity for emotional tagging. Other emotion analysis approaches can be combined with the proposed system to provide additional information about the emotional dimensions of arousal and valence. To this end, EEG signals can be acquired during the consumption of multimedia contents, and the possibility of extracting arousal

and valence measures from the signals can be assessed. Moreover, multimodal emotion analysis can be performed by analyzing other physiological signals such as skin conductance, ECG and respiratory rate, as well as using visual and acoustic modalities.

4 MCA and EEG Signal Processing for Emotional Tagging

As mentioned in the previous sections, another approach to emotion assessment is to use the dimensional model (cf. Fig. 1). The estimated values of arousal and valence can be used directly as tags or they can be used to infer different emotional categories. In order to reliably estimate the levels of arousal and valence induced in users while consuming multimedia contents, we propose analyzing EEG/peripheral physiological signal acquired from the users and MCA of the audio and video channels of the content itself. In the following, we present two systems based on EEG/peripheral physiological signal analysis and MCA.

4.1 EEG/Peripheral Physiological Signal Analysis

In this study, music video clips are used as audiovisual stimuli in order to elicit different emotions. To this end, a relatively large set of music video clips (70 clips) was gathered. A subjective test was then performed to select the most appropriate test material. For each video, a 2-min highlight was extracted for the experiment. Six participants were asked to participate in the experiment, and their physiological signals (EEG and peripheral physiological signals) were recorded while they were watching the 20 selected music video clips. Participants were also asked to rate each video in terms of arousal, valence, and like/dislike. The experiments were performed in a laboratory environment with controlled temperature and illumination. Thirty-two active AgCl electrodes were placed according to the international 10–20 system, from which the EEG data were recorded at a rate of 512 Hz. At the same time, there are 13 peripheral physiological signals, namely, galvanic skin response (GSR), respiration, skin temperature, blood volume pulse by plethysmograph, EMGs of zygomaticus major and trapezius muscles (2 channels each), and 4-channel electrooculogram (EOG). GSR, also known as skin conductance, measures the electrical conductance of the skin. It varies with the moisture level controlled by the sympathetic nervous system and, thus, is used to capture the affective state, especially the arousal. It has been shown that a change in the magnitude of GSR and the intensity of the emotional experience are well associated in terms of arousal. In our experiments, GSR was measured by placing two electrodes on the distal phalanges of the middle and index fingers. The breathing activity of the users was recorded by using a stretch sensor around their abdomen. The amount of stretch in the rubber band of the sensor is measured as a voltage change.

In general, a decreased respiration rate is related to relaxation of a user, whereas negative emotions can cause irregular respiration patterns and momentary cessation of respiration may be due to surprising events and tense situations. The skin temperature is known to be related to the emotional state, i.e., arousing, negative emotions cause a decrease in temperature, whereas calm, positive emotions tend to increase the temperature [31]. A sensor was attached to the users' little fingers to record the skin temperature. A plethysmograph sensor was positioned on the thumb of a user in order to measure the blood volume pulse. It has been reported that the heart rate and its variability are subject to change according to the affective state of a user, e.g., anger, fear, and sadness cause increased heart rates [9] and pleasantness increases peak heart rate response [23]. Four sensors were used to record the EMG signals. Two of them were placed on the trapezius muscle of the neck to monitor head movements, and the other two were on the zygomaticus major muscle to detect the user's laugh or smile. Finally, the EOG signals were recorded by using four sensors around the eyes of each user, from which activities related to eye blinking were captured. It is known that the eye blinking rate is related to anxiety.

4.1.1 Data Processing

Before extracting features from EEG signals and learning a classification function, several preprocessing operations were applied to the recorded data in the following order. In order to remove the slow drifts and high-frequency noises from the acquired data, a sixth-order Butterworth band-pass filter with the cutoff frequencies of 0.6 Hz and 100 Hz was used. Filtering of the input sequence was performed in both forward and reverse time directions to remove all phase distortion, effectively doubling the filter order. The filtered EEG data was then downsampled from 512 to 256 Hz. The downsampling process filters the input data with a low-pass filter and then resamples the resulting smoothed signal at a lower rate. The recorded EEG signals are often contaminated with other non-cerebral artifactual signals such as eye blinking, eye movements, and muscle movements. These artifacts can potentially cause large-amplitude outliers in the EEG and subsequently result in a notable deterioration in classification accuracy. In this work, an orthogonal projection approach was used to remove any potential EOG and EMG artifacts from the recorded signal.

After preprocessing, the EEG signal was broken down into different segments such that each segment represented the EEG signal acquired during each run. More precisely, each segment began with the presentation onset of its corresponding video clip and lasted for the 120 s of video presentation. In order to extract descriptive and discriminating features from the EEG signal, the wavelet transform was used. This transform represents the signal in both time and frequency and provides precise information about transient events occurring in the signal. Furthermore, the wavelet time-frequency representation does not make any assumptions about signal stationarity and is capable of detecting dynamic changes due to its localization properties. The relative wavelet energies of each electrode together with relative

wavelet entropy of symmetrical electrode pairs are extracted as features. The latter feature can be considered as a novel asymmetry index for detecting any asymmetry in the distribution of energy over the right and left brain hemispheres due to a change of emotional state.

The peripheral physiological signals were also processed and the following features for recognition were extracted as follows. First, the signals were partitioned into multiple temporal segments. Physiological changes caused by affective states are usually observed on a longer term in comparison to the changes in EEG signals [34]. In order to capture such changes appropriately, relatively long moving windows were used for segmentation. We tested the recognition performance for various combinations of window parameters (i.e., length and moving rate): the length varied from 10 to 120 s and the moving rate ranged between 5 and 30 s. The best performance was observed by using a window having a length of 60 s and moving at every 15 s (so that neighboring segments have a 45 s overlap), which produced five segments for each music video clip. Then, for each segment, 30-dimensional features were extracted. We extracted the mean and standard deviation of each signal. In addition, the respiration rate, heart beat rate, and eye blinking rate were estimated from the breathing activity, blood volume pulse, and vertical EOG signals, respectively, and their mean and standard deviation values were used as features.

4.1.2 Results

Three different binary classification problems were posed: positive/negative valence, low/high arousal, and low/high liking. To this end, the participants' ratings by self-assessment during the experiment were used as the ground truth, and support vector machine (SVM) classifier was employed. The ratings for each of these scales were thresholded into two classes (low/high or negative/positive). On the nine-point rating scales, the threshold was simply placed in the middle. In order to explore the feasibility of developing a general purpose affect recognition system, which can be trained using data acquired from limited number of participants and can be applied for a naive user who never used the system before, a leave-one-participant-out cross-validation scheme was used as in [41]. In other words, a classifier was trained based on physiological signals acquired from five participants and was tested on the data acquired from the remaining one user. This was repeated until each participant was a test participant once. The results of single-trial classification using this approach are presented in Table 2.

As it can be seen in Table 2, the results obtained using EEG signals are relatively higher than those obtained using peripheral physiological signals. It can also be seen that the results vary across different users, implying the difficulty of developing a generic system which can be used by a new user without any training. However, if a classifier is trained using the data gathered from the same user, the results are relatively higher. In order to perform user-dependent classification, the leave-one-video-out cross-validation scheme was considered. In other words, the EEG signals

Table 2 Results of user-independent two-class classification of EEG and peripheral physiological signals for the valence, arousal, and like/dislike targets

	P1		P2		P3		P4		P5		P6		Avg.	
	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.
Valence	68.3	39	62.9	61	60.8	60	56.2	56	49.2	41	74.6	32	62	48.2
Arousal	58.3	52	57.5	51	63.7	52	43.7	32	49.6	52	58.3	53	55.2	48.7
Like/dislike	63.7	57	61.7	45	68.3	56	65.4	69	55.8	60	57.5	55	62.1	57

Table 3 Results of two-class classification of EEG and peripheral physiological signals for the valence, arousal, and like/dislike targets

	P1		P2		P3		P4		P5		P6		Avg.	
	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.
Valence	65	50	60	55	70	50	60	50	80	70	80	70	69.2	57.5
Arousal	–	–	60	45	80	65	85	55	95	35	90	65	82	53
Like/dislike	–	–	70	75	80	65	75	75	75	55	70	70	74	68

recorded from a given user corresponding to 19 video clips were used as training, and the remaining one video was used as a test set. This was repeated until all video clips were considered as a test set once. Table 3 presents the user-dependent classification results.

The results obtained in this study demonstrate the feasibility of using EEG signals as appropriate tools for emotion recognition and emotional annotation; however, since the EEG signal changes are very different across users, it is a challenging task to design a generic system which can lead to a good performance.

4.2 MCA

This section describes an approach to emotional characterization by analyzing the multimedia content itself. The aim of this section is to explain and show how affective multimedia content analysis is performed. To this end, first, the database of music video clips used in this study is described. Then, the features which are extracted from the music video clips (audio features and video features) are introduced, and finally, the classification results are presented.

The music video clips were taken from the DEAP (Database for Emotion Analysis using Physiological Signals) database [20]. Through a web-based subjective test, 40 music video clips were selected so that only the music video clips which induce strong emotions are used. More precisely, ten music video clips from each quadrant or arousal-valence space, which all had the strongest possible volunteer ratings with a small variation, were selected. More information about the selection procedure can be found in [20]. After selecting the test material, 32 participants (50% female), aged between 19 and 37 (mean age 26.9), participated in the experiment to create the DEAP. They were asked to watch the 40 selected music video clips. After watching

each music video clip, they were asked to perform the rating of their perceived emotion. The music video clips and the user's self-assessed emotions are used in this study.

4.2.1 Video Feature Extraction

Most previous work reported that there is a relationship between the low-level visual features of multimedia contents such as "lighting key," "shot length," "color," "motion" and emotional states perceived by people while watching these contents. Thus, these four different features were extracted from the music video clips. The following paragraphs explain these features in more details.

Filmmakers often use lighting as an important tool to evoke different emotions. They use multiple light sources and balance the direction and the intensity of light in order to create effects with the contrast of shadow and light or direct the attention of the viewer [37]. The algorithm used to compute the lighting key is based on the fact that the proportion of bright pixels in high-key shots and low-key shots are high and low, respectively. In [16] and [42], the authors showed that there is a direct relation between the average shot length of the content and the induced affective states. In [26] and [30], the authors proposed a method based on the color histogram to detect shot boundaries. They considered the differences between color histograms of frames belonging to a video sequence in order to detect hard cuts, fades, or dissolves. In this work, the proposed method in [1] is used. The algorithm is based on singular value decomposition (SVD) and extracts low-cost, multivariate color features to construct two-dimensional feature matrices. Many research studies on affective content analysis have found a relationship between colors and evoked emotions in spectators. For example, [16] shows that the colors "yellow," "orange," and "red" correspond to the emotions "fear" and "anger," while the colors "blue," "violet," and "green" are dominant colors when the spectator feels "high valence" and "low arousal." In this work, the color features are extracted in the following manner. From each frame of a video sequence, a color histogram is computed. For a given frame i, the hue component of the HSV space is used. Then, the maximum, minimum, mean, and median hue values of frame i are computed. The medians of these values over all frames of a video sequence are taken as features. Sun et al. [42] show the relation between the emotions of "joy," "anger," "sadness," and "fear" and the motion. A fast motion vector is computed for every fourth frame in a video sequence. The median and mean values of the motion vector absolute value of each four frames are computed. Finally, the mean value of all median and mean values are computed in order to construct two features.

4.2.2 Audio Feature Extraction

As reported in several research studies, sound can have a close relationship with the affective content of a music video clip [45]. In this work some characteristics

of sound were selected and used as features. In general, arousal is believed to be correlated with tempo (fast/slow) and pitch, while valence is associated with energy, harmony, and scale (major/minor). In order to capture some of these cues, the following features are computed for each music video clip using the Matlab music information retrieval MIR toolbox [24]. The zero-crossing rate is defined as the number of times the signal crosses the zero line (x-axis) per unit time. In other words, it is the number of times the signal changes its sign per time unit. Mel-frequency cepstral coefficients (MFCCs) can be seen as the description of the spectral shape of the sound. Most of the signal information can be found in low-frequency coefficients. Therefore, only the 13 first coefficients are used as features. Furthermore, the ΔMFCC is another feature which provides quantitative measures of the movement of the MFCC from every pair of subsequent frames.

4.2.3 Classification

After the feature extraction stage, the extracted features are fed into Gaussian mixture model (GMM) classifiers. For the training of the GMM classifiers, the expectation maximization (EM) algorithm is utilized, while eight Gaussian mixtures are created to describe each class. Experiments are conducted using the "leave-one-video-out" cross-validation scheme, which involves training the classifier with samples from all but one video, and testing its performance on the left-out video (test set). In other words, for performing this cross validation, all sequences of the test video clip are kept out and a classifier is learned based on the sequences of training video clips. This procedure is repeated until each video clip is considered as a test sample once. The test set is then compared to the GMMs trained for each class, resulting in a matching score. The benefits of the GMM classifier include superior classification performance and low computational complexity, as well as producing soft-output results which can be later used for the information fusion stage.

4.3 Results

The classification can be performed for each user separately, (as it is done in this work) or for the average ratings over all users. In the first case, we train the classifier for each user separately. The same features can be mapped to other emotions for another user, and each music video clip is classified into other emotions depending on the user. In the second case, the average "arousal" and "valence" values are considered to classify the video clips. Then the video clips are classified according to the average rating values into different emotions. The second classification protocol assumes that the subjective ratings for a video are all very close together with a small standard deviation. In other words, for this kind of classification, it is favorable that the video clips evoke the same or similar emotions in all users. Figure 4 presents the classification results per users. In order to perform this classification scheme, a

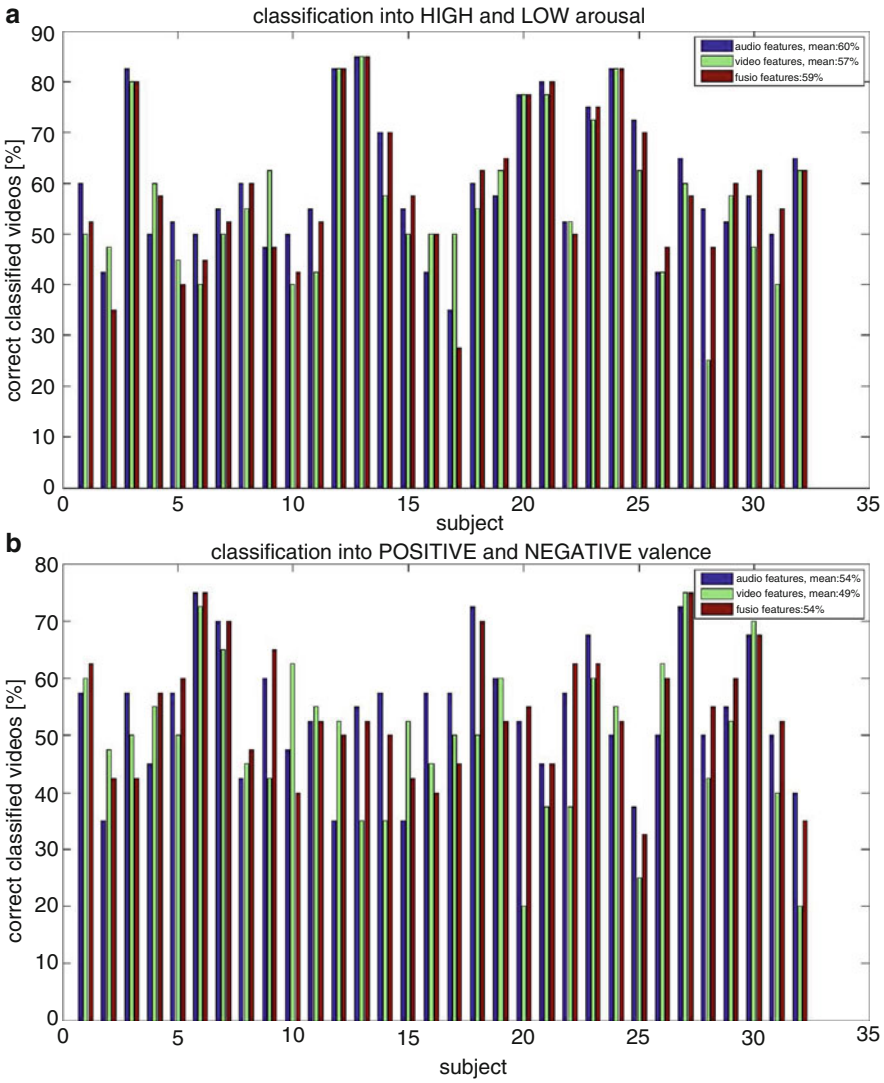


Fig. 4 Results of classification into (**a**) high and low arousal and (**b**) positive and negative valence

leave-one-video-out cross validation was considered. In other words, one video was used as a test video and the remaining video clips were used as the training set. This was repeated until each video is used as a test set once.

In order to perform the second classification scheme, namely, the social scheme, the ground truth for each video was constructed by computing the mean opinion score of ratings performed by all participants. Table 4 presents the classification results using audio, video, and audiovisual (concatenation of audio and video) features. Our results indicate that audio features carry more information on arousal,

Table 4 Comparison of the classification results using audio, video, and audiovisual features

	A	V	AV
Arousal	82.5%	77.5%	85%
Valence	57.25%	62.5%	70%

whereas for valence, visual information is of higher significance. The results also imply that the merged audiovisual modalities at feature level result in relatively higher classification accuracy.

Comparison of the results obtained using EEG signal processing and MCA leads to interesting fact. When analyzing the emotions induced in users during the watching of music video clips, emotion recognition using EEG signal processing outperforms that of MCA. However, since the EEG alterations are very user dependent, it is a challenging task to develop a generic classification system that can be used by new users without training. Furthermore, when analyzing a group of users, MCA is shown to be efficient in determination of the affective value of a multimedia content. This affective value will be precise for majority of the users (mean opinion score); however, different people might have different emotional experiences while watching them. An approach to overcome this shortcoming is to cluster the users in groups (profiles) and analyze the ratings of each group, separately.

5 Toward Implicit Emotional Annotation System for Social Media Retrieval

In Sect. 4.1, we presented an emotion recognition system based on analysis of EEG and other peripheral physiological signals. The recognition results of this system suggested that EEG signal processing can be used as an appropriate tool for determining the arousal, valence, and like/dislike dimensions of emotions, induced in users while consuming multimedia contents. In Sect. 4.2, MCA was shown to be an appropriate tool for estimating the information about the possible emotion that can be induced in majority of the users while consuming a multimedia content. Based on these results, we propose a practical implicit emotional annotation system as follows.

Figure 5 depicts a general overview of the implicit emotional annotation system and how it can be used in search and retrieval. Multimedia Item denotes the content under consideration, which could be from a Multimedia Database similar to those found on YouTube, Last.fm, Flickr, or professional content providers. Such content can be analyzed to retrieve low-, mid-, or high-level multimedia descriptors. Cues to both ongoing user mood and transient emotional reactions to media are analyzed by monitoring devices such as physiological signal acquisition devices, as well as multimodal input devices. Additional analysis and mining is performed by using user's existing social networks or by inferring preferences and relationships from

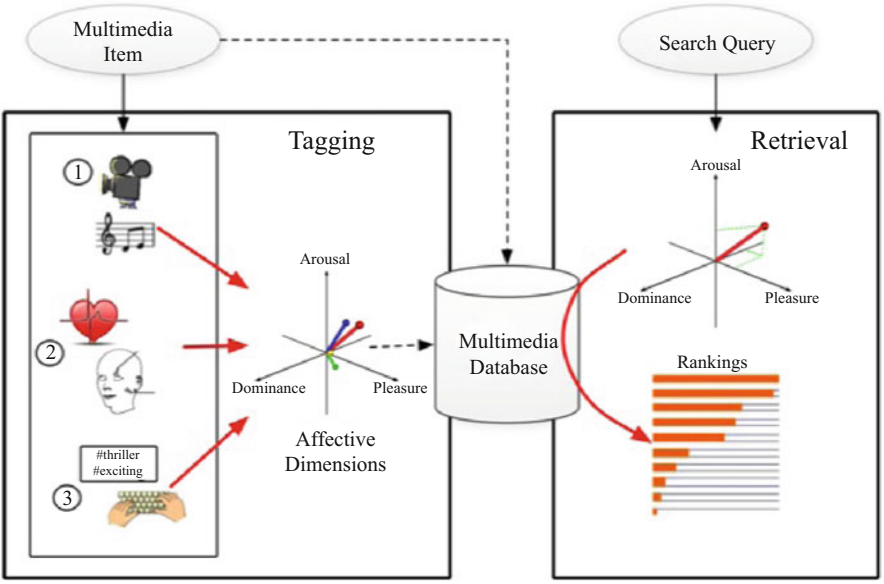


Fig. 5 Emotional indexing and retrieval of multimedia contents [17]

comparisons to other users with similar, accessible profiles. These correspond to various forms of user input, individual or through his/her social network, from which multimedia content can be accessed, recommended, or filtered.

References

1. Abd-Almageed, W.: Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing. In: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, pp. 3200–3203. IEEE, Piscataway (2008)
2. Adams, W., Iyengar, G., Lin, C., Naphade, M., Neti, C., Nock, H., Smith, J.: Semantic indexing of multimedia content using visual, audio, and text cues. EURASIP J. Appl. Signal Process. **2**, 170–185 (2003)
3. Aftanas, L., Reva, N., Varlamov, A., Pavlov, S., Makhnev, V.: Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. Neurosci. Behav. Physiol. **34**(8), 859–867 (2004)
4. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 971–980. ACM, New York (2007)
5. Bishop, C., en ligne), S.S.: Pattern Recognition and Machine Learning, vol. 4. springer, New York (2006)
6. Centeno, T., Lawrence, N.: Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis. J. Mach. Learn. Res. **7**, 455–491 (2006)
7. Chanel, G., Kronegg, J., Grandjean, D., Pun, T.: Emotion assessment: arousal evaluation using EEG's and peripheral physiological signals. Multimedia Content Representation, Classification and Security, pp. 530–537. Springer, Berlin/New York (2006)

8. Cowie, R.: Emotion-Oriented Systems: The Humaine Handbook. Springer, Heidelberg (2010)
9. Ekman, P., Levenson, R., Friesen, W.: Autonomic nervous system activity distinguishes among emotions. Science **221**(4616), 1208 (1983)
10. Fragopanagos, N., Taylor, J.: Emotion recognition in human-computer interaction. Neural Netw. **18**(4), 389–405 (2005)
11. Hanjalic, A., Xu, L.: Affective video content representation and modeling. IEEE Trans. Multimed. **7**(1), 143–154 (2005)
12. Healey, J.A.: Wearable and automotive systems for affect recognition from physiology. Ph.D. thesis, MIT (2000)
13. Hoffmann, U., Vesin, J., Ebrahimi, T., Diserens, K.: An efficient p300-based brain-computer interface for disabled subjects. J. Neurosci. methods **167**(1), 115–125 (2008)
14. Ishino, K., Hagiwara, M.: A feeling estimation system using a simple electroencephalograph. In: Proc. IEEE Int. Conf. Syst. Man Cybern. **5**, 4204–4209 (2003)
15. Joho, H., Jose, J., Valenti, R., Sebe, N.: Exploiting facial expressions for affective video summarisation. In: Proceedings of the ACM International Conference on Image and Video Retrieval, p. 31. ACM, New York (2009)
16. Kang, H.: Affective content detection using HMMs. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp. 259–262. ACM, New York (2003)
17. Kierkels, J., Soleymani, M., Pun, T.: Queries and tags in affect-based multimedia retrieval. In: IEEE International Conference on Multimedia and Expo, 2009. ICME 2009, pp. 1436–1439. IEEE New York, NY, USA (2009)
18. Kim, J., André, E.: Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2067–2083 (2008)
19. Kim, K., Bang, S., Kim, S.: Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. **42**(3), 419–427 (2004)
20. Koelstra, S., Muhl, C., Soleymani, M., Lee, J., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: a database for emotion analysis using physiological signals. IEEE Trans. Affect. Comput. **99**, 1–1 (2011)
21. Kostyunina, M., Kulikov, M.: Frequency characteristics of EEG spectra in the emotions. Neurosci. Behav. Physiol. **26**(4), 340–343 (1996)
22. Krause, C., Viemerö, V., Rosenqvist, A., Sillanmäki, L., Åström, T.: Relative electroencephalographic desynchronization and synchronization in humans to emotional film content: an analysis of the 4–6, 6–8, 8–10 and 10–12 Hz frequency bands. Neurosci. Lett. **286**(1), 9–12 (2000)
23. Lang, P., Greenwald, M., Bradeley, M., Hamm, A.: Looking at pictures- affective, facial, visceral, and behavioral reactions. Psychophysiology **30**(3), 261–273 (1993)
24. Lartillot, O., Toiviainen, P., Eerola, T.: A matlab toolbox for music information retrieval. Data Analysis, Machine Learning and Applications, pp. 261–268. Springer (2008)
25. Lee, J., Park, C.: Adaptive decision fusion for audio-visual speech recognition. Speech Recognition, Technologies and Applications, p. 550. InTech (2008)
26. Lienhart, R.: Comparison of automatic shot boundary detection algorithms. Proc. SPIE **3656**, 290–301 (1999)
27. Lin, Y., Wang, C., Jung, T., Wu, T., Jeng, S., Duann, J., Chen, J.: Eeg-based emotion recognition in music listening. IEEE Trans. Biomed. Eng. **57**(7), 1798–1806 (2010)
28. Lisetti, C.L., Nasoz, F.: Using noninvasive wearable computers to recognize human emotions from physiological signals. EURASIP J. Appl. Signal Process. **2004**(1), 1672–1687 (2004)
29. Lopatovska, I., Arapakis, I.: Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. Inf. Process. Manag. **47**(4), 575–592 (2011)
30. Mas, J., Fernandez, G.: Video shot boundary detection based on color histogram. In: Notebook Papers TRECVID2003, Gaithersburg, NIST (2003)
31. McFarland, R.: Relationship of skin temperature changes to the emotions accompanying music. Appl. Psychophysiol. Biofeedback **10**(3), 255–267 (1985)

32. Pantic, M., Vinciarelli, A.: Implicit human-centered tagging [social sciences]. IEEE Signal Process. Mag. **26**(6), 173–180 (2009)
33. Petrantonakis, P., Hadjileontiadis, L.: Emotion recognition from eeg using higher order crossings. IEEE Trans. Inf. Technol. Biomed. **14**(2), 186–197 (2010)
34. Picard, R., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**(10), 1175–1191 (2001)
35. Plutchik, R.: The nature of emotions. Am. Sci. **89**, 344 (2001)
36. Potamianos, G., Neti, C.: Stream confidence estimation for audio-visual speech recognition. In: Sixth International Conference on Spoken Language Processing. Beijing, China, October 16–20 (2000)
37. Rasheed, Z., Sheikh, Y., Shah, M.: On the use of computable features for film classification. IEEE Trans. Circuits Sys. Video Technol. **15**(1), 52–64 (2005)
38. Russell, J., Mehrabian, A.: Evidence for a three-factor theory of emotions. J. Res. Personal. **11**(3), 273–294 (1977)
39. Schaaff, K., Schultz, T.: Towards emotion recognition from electroencephalographic signals. In: Proceedings of International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, pp. 1–6 (2009)
40. Sebe, N., Cohen, I., Gevers, T., Huang, T.: Emotion recognition based on joint visual and audio cues. In: 18th International Conference on Pattern Recognition, 2006. ICPR 2006, vol. 1, pp. 1136–1139. IEEE, Washington, DC (2006)
41. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multi-modal affective database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **99**, 1–1 (2011)
42. Sun, K., Yu, J.: Video affective content representation and recognition using video affective tree and hidden markov models. Affective Computing and Intelligent Interaction, pp. 594–605. Springer, Berlin/New York (2007)
43. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: resources, features, and methods. Speech Commun. **48**(9), 1162–1181 (2006)
44. Wang, Y., Liu, Z., Huang, J.: Multimedia content analysis-using both audio and visual clues. Signal Process. Mag. IEEE **17**(6), 12–36 (2000)
45. Yang, Y., Chen, H.: Ranking-based emotion recognition for music organization and retrieval. IEEE Trans. Audio Speech Lang. Process. **19**(4), 762–774 (2011)
46. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009). doi:10.1109/TPAMI.2008.52

Part III
Privacy and Personalisation of Social Media

Privacy in Recommender Systems

Arjan J.P. Jeckmans, Michael Beye, Zekeriya Erkin, Pieter Hartel, Reginald L. Lagendijk, and Qiang Tang

Abstract In many online applications, the range of content that is offered to users is so wide that a need for automated recommender systems arises. Such systems can provide a personalized selection of relevant items to users. In practice, this can help people find entertaining movies, boost sales through targeted advertisements, or help social network users meet new friends.

To generate accurate personalized recommendations, recommender systems rely on detailed personal data on the preferences of users. Examples are ratings, consumption histories, and personal profiles. Recommender systems are useful; however, the privacy risks associated to gathering and processing personal data are often underestimated or ignored. Many users are not sufficiently aware if and how much of their data is collected, if such data is sold to third parties, or how securely it is stored and for how long.

This chapter aims to provide insight into privacy in recommender systems. First, we discuss different types of existing recommender systems. Second, we give an overview of the data that is used in recommender systems. Third, we examine the associated risks to data privacy. Fourth, relevant research areas for privacy-protection techniques and their applicability to recommender systems are discussed. Finally, we conclude with a discussion on applying and combining different privacy-protection techniques in real-world settings, making clear mappings to reflect typical relations between recommender system types, information types, particular privacy risks, and privacy-protection techniques.

A.J.P Jeckmans (✉) • P. Hartel • Q. Tang
Distributed and Embedded Security, Faculty of EEMCS, University of Twente, Enschede, The Netherlands
e-mail: a.j.p.jeckmans@utwente.nl; q.tang@utwente.nl; pieter.hartel@utwente.nl

M. Beye • Z. Erkin • R.L. Lagendijk
Information Security and Privacy Lab, Faculty of EEMCS, Delft University of Technology, Delft, The Netherlands
e-mail: m.r.t.beye@tudelft.nl; z.erkin@tudelft.nl; r.l.lagendijk@tudelft.nl

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_12,

1 Introduction

In recent years, online applications have become an important part of daily life for millions of users. People consume media (Youtube, Flickr, LastFM), do their shopping (Amazon, Ebay), and interact (Facebook, Gmail) online. Because the range and amount of content that is offered to users is often huge, automated recommender systems are employed. By providing personalized suggestions, these systems can help people find interesting media, boost sales through targeted advertisements, or help people meet new friends. Because of their automated nature, recommender systems can meet the demands of large online applications that operate on a global scale.

All recommender systems share a common trait: in order to generate personalized recommendations, they require information on the attributes, demands, or preferences of the user. Typically, the more detailed the information related to the user is, the more accurate the recommendations for the user are. Service providers running the recommender systems collect information where possible to ensure accurate recommendations. The information supplied can either be automatically collected or specifically provided by the user. Automatically collected information is the result of users interacting with the recommender systems and making choices based on recommendations. For example, page views on Ebay are used to automatically present a selection of recommended similar items (*recommendations for you*). Similarly, recommended videos on Youtube are influenced by recently viewed videos. Based on purchases by other users, items on Amazon are accompanied by package deals (*frequently bought together*) or related items (*customers who bought this item also bought*). Based on sites visited, Google serves personalized advertisements. Based on your friends and social interactions, Facebook suggests new friends to make. LinkedIn, based on a user's cv and connections, recommends interesting companies, job offers, and news. Vice versa, LinkedIn also recommends people to recruiters posting new job openings. Many dating sites, such as Match.com or PAIQ, recommend partner matches to its users. Many more examples of such systems exist and they will continue to exist, in the future. Users can also specifically provide information. In this way, users build their own profile specifying their likes and dislikes or containing general information (such as age and gender) about themselves. For example, LastFM and Youtube allow users to specify their favorites. Facebook allows listing profile information as well as interests.

However, potential threats to user privacy are often underestimated. The more detailed the information related to the user is, the larger the threat to the user's privacy is. In order to enhance their recommender systems, service providers are collecting and consolidating more and more information. For example, in recent privacy policy updates, Google stated that they consolidate information from all their services to a single profile. Facebook continues to expand its reach around the Internet, giving the ability to share more and *like* almost everything. Information might be abused by the service provider, sold to a third party, or leaked by a hacker. There is an inherent trade-off between utility (getting accurate personalized

recommendations) and privacy. Research into regulations, anonymization, and privacy-preserving algorithms aims to improve privacy, while maintaining utility. In this chapter, we will analyze the privacy risks associated with recommender systems and the research that helps to minimize these risks.

We first look at the state of the art of various types of recommender systems in use today (Sect. 2). Second, we give our categorization of the types of information generally involved in recommender systems, operation and how this is mapped to the various types of recommender systems. Third, we identify the privacy concerns in recommender systems and give our own classification of them. To see how the privacy concerns affect the recommender systems, the privacy concerns are mapped to the different types of information (Sect. 3). Fourth, we give an overview of existing research into state-of-the-art privacy-protection techniques (Sect. 4). The relationship between the research and privacy concerns is also given. Finally, we conclude and discuss (Sect. 5).

2 Recommender Systems

In this section, we give an overview of the different recommender system types. We then list the information present in recommender systems. Finally, we show what information is typically used in which recommender type. This relationship will serve as a basis, to describe the privacy concerns in the next section.

A recommender system provides a set of *items* (e.g., content, solutions, or other users) that is most relevant to a particular user of the system. Typically, recommender systems achieve this by predicting *relevance scores* for all items that the user has not seen yet. Items that receive the highest score get recommended (typically the top-N items or all items above a threshold t). The prediction is made by considering both the traits of the item and user. Typically, systems look at similarities between items, similarities between users, or relations between particular types of items and particular types of users. The performance of a recommender system is determined by looking at the recommendation accuracy, that is, the error between given and expected results.

2.1 Recommender System Types

Adomavicius and Tuzhilin [1] gave an overview of the state of the art in recommender systems and possible extensions. They list only the three popular types of that time: collaborative filtering, content-based, and hybrid. We make a distinction between basic recommender types and improved recommender types. Improved recommender types build upon basic recommender types by combining them or adding new information. Any improved recommender type can be combined with

any basic recommender type. The following basic recommender system types have been around for quite some time:

Collaborative Filtering. One of the first collaborative filtering recommender systems is Tapestry, by Goldberg et al. [16]. This system was designed to retrieve email messages from Usenet mailing lists, relevant to a user's particular interests. Goldberg et al. observed that conventional mailing lists are too static and rarely form a perfect match to a user's demands. Tapestry relies on what the authors termed *collaborative filtering techniques*, which are still widely used today. In collaborative filtering, each user rates content items. These ratings determine similarity between either users (similar users like similar items) or items (users like items similar to highly rated items). Different metrics exist to compute similarity. Recommended for the current user are those items that are rated highest by his most similar peers or contain those items that are rated most similar to his favorite items.

Content-based. Content-based recommender systems use item similarity to determine recommendations. Unlike the collaborative filtering method, item similarity is computed by item metadata. Examples of metadata are kitchen for restaurants, genre for movies, and artist for music. Recommended are those items that are most similar to the user's favorite items. An example of a content-based recommender system is Newsweeder, by Lang [24].

Demographic. When detailed information about the user's preferences is not available, demographic information can lead to somewhat personalized recommendations. Grundy, by Rich [32], is an example of this. Demographic information may include age, gender, country of residence, education level, etc. The demographic information is matched to a stereotype, and the items attached to this stereotype are recommended. Personalization for the user is limited due to the generalization to a stereotype.

Knowledge-based. When requiring a recommendation, the user enters his preferences in the recommender system. The system then outputs a (number of) potential recommendations based on (expert) knowledge contained in the system. Possibly, the user can give feedback, and the recommendation is refined. After a few iterations, the recommendation is tailored to the user. Entree [7] is an example of such a system, built to help diners find a suitable restaurant. In learning knowledge-based recommender systems, feedback from the user is fed back into the system to add to the knowledge [25].

The following improvements have been proposed to the basic recommender systems mentioned above:

Context-aware. In many application domains, contextual information is available, which can be used to improve recommendations. Common examples of contextual information are location, group dynamic, time, date, and purpose. While user preferences and domain knowledge are relatively static, context is highly dynamic in nature. Every recommendation, even for the same user, may have a completely new context. Adomavicius and Tuzhilin [2] provided a discussion

on contextual information in recommender systems. They showed three ways in which such contextual information can be added to existing recommender systems: (1) Use a prefilter to remove content items (and information associated with them) that do not fit the context from the system. (2) Use a postfilter to remove recommendations that do not fit the context. (3) Add the context to the model of the recommender system and use the contextual information during the recommendation process.

Ensemble. Ensembles of recommender systems combine several of the same type of recommender system to improve performance. The idea behind ensembles is to get multiple opinions before making a decision. Schclar et al. [34] detailed the use of ensembles on collaborative filtering.

Hybrid. Hybrid recommender systems, like ensembles, combine multiple recommender systems. However, in a hybrid system, multiple different types of recommender systems are combined. A comprehensive overview of different hybridization techniques is given by Burke [8]. As concluded by Burke, given a certain hybridization, not all basic recommender systems can be (straightforwardly) combined.

Social. The rise of online social networks increased the availability of a user's social information (e.g., the friendship network). Because friends typically share interests, the information they supply to the recommender system is more likely to fit with the user. Or alternatively, the social information can be used to infer communities of similar users. As an example, Konstas et al. [22] utilized the social information in LastFM to improve collaborative filtering.

2.2 Information in Recommender Systems

We now discuss the different types of information typically used in recommender systems. We do not aim to give a complete categorization of information used but instead to explore the diversity of information used in recommender systems.

Behavioral information is the implicit information that the recommender system can gather while the user interacts with the broader system. For example, product views in a webshop or not fully watching a movie on a video on demand site.

Contextual information describes to the context in which a recommendation query is made. Common examples of contextual information are location, social group, time, date, and purpose.

Domain knowledge specifies the relationship between a user stereotype and content items. Domain knowledge is usually static but can change over time.

Item metadata is descriptive information about content items. Examples of metadata are kitchen for restaurants, genre for movies, and artist for music.

Purchase or consumption history is the list of content that has previously been purchased or consumed by the user.

Table 1 Information that is present in different recommender system types

↓ Rec. sys./info. →	Behavioral	Contextual	Domain knowledge	Item meta-data	History	Recommendation	Feedback	Social	User attributes	User preferences
Collaborative filtering	•	·	·	·	•	●	·	·	·	●
Content-based	·	·	·	●	•	●	·	·	·	●
Demographic	·	·	●	·	·	●	·	·	●	·
Knowledge-based	•	·	●	•	·	●	•	·	•	●
Context-aware	•	●	•	•	·	·	·	·	•	·
Ensemble	·	·	·	·	·	●	·	·	·	·
Hybrid	·	·	·	·	·	●	·	·	·	·
Social	·	·	·	·	·	·	·	●	·	·

Recommendations are the output of a recommender system, typically a ranked list of items. In some systems, the relevance score for each content item is also given to the user.

Recommendation feedback is information about the recommendation provided by the user. Feedback can be expressed as positive, negative, or something more nuanced (stating a reason as well).

Social information describes the relationship between different users. Many sites allow users to specify a friendship relation (or similar) to other users, community membership, or both.

User attributes describe the user. Examples of user attributes are demographic information, income, and marital status.

User preferences are explicitly stated opinions about items or groups of items. Preferences are expressed by either a scalar measure (rating items on a scale of 1–5 stars), a binary indicator (keeping a list of favorites), or text (tags and comments).

2.3 Summary

Table 1 shows the type of information used by the type of recommender system. Information is ● almost always present, • sometimes present, or · almost never present in a recommender system. For improved recommender types, we mark what information is added to the basic recommender types. There is a clear distinction between the information used in the basic recommender types. The improved recommender types either add new types of information or combine multiple basic recommender systems.

3 Privacy Concerns in Recommender Systems

Because users need to reveal information in order to make use of the desired functionality of a recommender system, a trade-off exists between utility and user privacy. Obtaining accurate recommendations is one thing, but sharing personal information may also lead to privacy breaches. In this section, we will look into privacy in recommender systems and potential privacy concerns with a focus on user privacy.

3.1 Privacy and Confidentiality

The word privacy has many subtly different meanings, each with their own definition. Privacy on the Internet revolves mainly around *information privacy*. Kang [20] used the wording of the Information Infrastructure Task Force (IITF), as cited below:

> Information privacy is "an individual's claim to control the terms under which personal information – information identifiable to the individual – is acquired, disclosed or used."

This concept of information privacy is strongly related to the notion of *confidentiality*, from the field of information security, but not to be used interchangeably. Confidentiality is concerned with the secrecy of individual pieces of information. Information privacy focusses on the individual who is the subject of said information, the effects that disclosure have on this person, and his or her control and consent. In our overview of privacy-protection technologies, the focus will lie on preventing unwanted disclosure and usage of information, but not on the effects on the person.

When using online applications, users generally share a lot of (personal) information. Whether it is uploading ratings or comments, posting personal information on a profile, or making purchases, information is always shared within a particular *scope* [28]. Privacy involves keeping a piece of information in its intended scope. This scope is defined by the size of the audience (breadth), by extent of usage allowed (depth), and duration (lifetime). When a piece of information is moved beyond its intended scope in any of these dimensions (be it accidentally or maliciously), a privacy breach occurs. So, a breach may occur when information is disclosed to a party for whom it was not intended, when information is abused for a different purpose than was intended, or when information is stored beyond its intended lifetime.

Weiss [39] stated that on the traditional Web, privacy is maintained by limiting data collection, hiding users' identities, and restricting access to authorized parties only. Often, in practice, information and identity become closely linked and visible to large groups of people. Profiles may be publicly visible, comments can be seen by all viewers of a content item, and some sites list the last users to visit a particular page. It becomes harder for a user to monitor and control his personal information,

as more of it becomes available online. This problem mainly applies to systems where the user logs in to an account and where tools are available to express a user's preferences.

Often, users are not very aware of their (lack of) privacy. In a study on social network users in particular, Gross and Acquisti [18] showed that most users do not change the default privacy settings, while sharing a large amount of information on their profile. Tufekci [38] concluded in his case study that privacy-aware users are actually more reluctant to join social networks, but once they do join, they still disclose a lot of information. As opposed to social networks, in most recommender systems, privacy toward other users is probably not the largest issue. Users place a lot of implicit trust in service providers, expecting them to handle user information in a fair and conscientious way and continue to do so in the future. By using the system, users enter into a relationship with the service provider, who can generally view *all* information in the system, including private uploads, browsing and purchase behavior, and IP addresses. It is also the service provider who decides which information is stored, how long it is kept, and how it is used or distributed. Usually, privacy statements are offered to display the position the service provider takes and to acquire the user's consent. However, this leaves users little choice: they can either agree to the terms or will not benefit from using the system. The power balance is clearly in favor of the service provider.

3.2 Privacy Concerns

Privacy breaches can involve a variety of parties (fellow users, the service provider, or outsiders) and may be a deliberate act (snooping, hacking), or accidental (mismanagement, lingering data). Depending on the sensitivity of information involved, such incidents may have serious consequences. Lam et al. [23] already identified some threats to privacy in recommender systems. Their concern is the amount of (personal) information that is collected by the service provider and the potential leakage of this information. Independent of their work, we explicitly identify the privacy concerns in recommender systems and classify them as follows:

Data Collection. Many users are not aware of the amount and extent of information that a service provider is able to collect and what can be derived from this information. This may be due to the fact that privacy statements are seldomly read, and people have become used to pursuing online activities. Usually there is no way to opt-out of such data gathering, other than not using the system at all. As collection practices do not match with the users' expectations, this concern relates to the extent of information usage.

Data Retention. Online information is often difficult to remove, the service provider may even intentionally prevent or hinder removal of data. This is because there is commercial value in user information, for both competitive advantage through

analysis and/or data sales. Furthermore, information that is apparently erased from one place may still reside somewhere else in the system, for example, in backups, to be found by others. The data retention concern relates to the intended lifetime, as information can be available longer than intended.

Data Sales. The wealth of information that is stored in online systems is likely to be of value to third parties and may be sold in some cases. Users' ratings, preferences, and purchase histories are all potentially interesting for marketing purposes. Data sales usually conflicts with the privacy expectations of users. Even though data is often anonymized before being sold to protect user privacy, re-identification is a threat that is often overlooked or ignored. For example, the information published by Netflix as part of their recommender systems prize, though anonymized, allowed for re-identification [27]. Narayanan and Shmatikov linked the anonymized records to publicly available records (such as IMDb) based on rating similarity and time of rating. If two records give a similar rating to a movie around the same time, they are likely to be from the same person. A higher number of similar movie ratings (in rating and in time) increases the confidence of the link between the records. This concern relates mainly to the extent of information usage.

Employee Browsing Private Information. The service provider as an entity has full access to the information, and its employees might take advantage of this. This is in conflict with the intended breadth of the audience, and the privacy that the service provider has promised its users.

Recommendations Revealing Information. Recommendations inherently are based on the information contained in the recommender system. For example, in collaborative filtering that information is the ratings of all the users, or in knowledge-based recommender systems it is the expert knowledge. Each recommendation reveals a tiny piece of information about the private information. It is unclear how a large number of recommendations impact the disclosure of information. This could be used to reveal information about other users (compromising their privacy) or information about the recommender system itself (potentially leading to reverse engineering of the system). Here, we focus on the privacy of the user, not the security of the system. Ramakrishnan et al. [31] looked at the privacy of eccentric users (users with unusual ratings) from a graph perspective. When looking at recommendation results, these users are at a higher risk than average users. As eccentric users cannot hide in crowds of other users, when their data is used for making recommendation, other data is often not. The recommendations output by the system are then based on only a few users, with a strong correlation between the input of the eccentric users and the recommendation output. This is in conflict with the intended breadth of the audience.

Shared Device or Service. Privacy at home can be just as important as privacy online. When sharing a device like a set-top box or computer, or a login to an online service, controlling privacy toward family and friends may be difficult. For example, a wife who wants to hide from her husband the fact that she purchased

Table 2 Privacy concerns for user information in recommender systems

↓ Concern/info. →	Behavioral	Contextual	Domain knowledge	Item meta-data	History	Recommendation	Feedback	Social	User attributes	User preferences
Data collection	●	•	·	·	●	•	●	•	•	•
Data retention	•	•	·	·	•	•	•	•	•	•
Data sales	●	●	·	·	●	●	●	●	●	●
Employee	●	●	·	·	●	●	●	●	●	●
Recommendations	·	·	·	·	●	·	•	•	·	●
Shared service	•	·	·	·	●	●	•	·	·	●
Stranger views	·	•	·	·	●	•	•	●	●	●

a gift for him. Unless she has a private account, her husband might inadvertently see her purchase or receive recommendations based on it. Many would want to keep some purchases private from their kids or their viewing behavior from their housemates. While some services allow for separate accounts, this is not always possible. For example, targeted advertising works with cookies that are stored in the browser, which is implicitly shared on a computer. This is related to the intended breadth of the audience.

Stranger Views Private Information. Users can falsely assume some information to be kept restricted to the service provider or a limited audience, when in reality it is not. This can be due to design flaws on the part of the service provider or a lack of the user's own understanding or attention to his privacy. When a stranger views such private information, there is a conflict with regard to the intended breadth of the audience. Rosenblum [33] showed, for example, that information in social networks is far more accessible to a widespread audience than perceived by its owners.

3.3 Summary

Because recommender systems typically contain a large amount of information, often about its users, they form an interesting target for attack. Information could end up in the wrong hands or be misused by legitimate data holders. Given the amount and detail of information within recommender systems, the privacy concerns should be taken seriously. Table 2 gives an overview of how different concerns impact different information within recommender systems. In this table, impact is either high (●), medium (•), or low (·). We can see that the impact on domain knowledge and item metadata is low for all privacy concerns. This is due to

the fact that this data is not about the user but about the content items. The user's history and preferences have the highest privacy concerns. This is mainly due to their accurate representation of the user's opinions about items.

4 Research into Privacy-Protection Technologies

We have seen a wide variety of privacy issues associated with recommender systems. Research from many areas could be applied to alleviate some of the aforementioned concerns. We will provide an overview of research areas and briefly discuss their mechanisms, advantages, and limitations.

4.1 Awareness

Research in this mainly social field aims to enhance user awareness of the privacy issues that exist within online systems. It can aid users in specifying their privacy boundaries. The Platform for Privacy Preferences (P3P) [12] is an initiative that aims to provide websites with a standardized format in which they can define their privacy policy. Visitors of the website can then, through client-side *user agents* (e.g. plug-ins for their browser or applets), easily check the details of a privacy policy and see what will happen to information they submit. This system can help to increase user awareness but only for users that employ agents and if websites properly define their privacy policies and adhere to them.

Tsai et al. [37] showed that when privacy information is shown *more prominent*, and users are made more aware of the privacy consequences, privacy is taken into account when shopping online. Tsai et al.'s study also shows that some users are even willing to pay more for the product, if it means getting more privacy. Offering users the ability to opt-out for or opt-in to data collection would in many cases level the playing field between users and service providers.

4.2 Laws and Regulations

This legal field of research aims to find proper and broad laws and regulations that protect the users' privacy, while not greatly hindering businesses. It also focusses on compliance of both users and service providers to established laws and social conducts. Laws and regulations form an important and much needed tool. For example, the Article 29 Working Party has been working toward regulations for online behavioral advertising [21].

This legal approach runs after the technology, as specific laws dealing with personal information as related to the Internet often take long to be developed. Also, laws are generally used to solve matters *after* things go wrong, whereas most technical solutions attempt to *prevent* violations.

4.3 Anonymization

As pointed out in Sects. 2 and 3.2, sales of information can be a major source of revenue for service providers. If this were to be done without any further consideration for privacy, users might take offense and stop using the system (thus hurting revenue) or take justified legal action. Service providers may try to remove the privacy issues associated with data sales, by *obscuring the link between users and data sold* [36].

This can be done through anonymization, which involves removing any identifying (or identifiable) information from the data, while preserving other structures of interest in the data. As mentioned before, the information published by Netflix as part of their recommender systems' prize, though anonymized, allowed for re-identification [27]. This mainly stems from the fact that information can only be *partially* removed or obfuscated, while other parts *must be kept intact* for the dataset to remain useful. In the real world, it is difficult to predict which external sources of information may become available, allowing pieces of data to be combined into identifiable information.

When looking at anonymization during recommendation, Cissée and Albayrak [11] utilized trusted agents (essentially moving the trust around) to act as a relay and filter the information that is sent. This way, the user can interact (through the agent) with the recommender system in an anonymous way. The user hides his personal information from the service provider and is safe from the service provider linking his rating information to a person. However, the user still needs to trust that the agents (either hardware or software based) and the service provider do not collude.

4.4 Randomization and Differential Privacy

Similar to anonymization is randomization. In randomization (sometimes referred to as perturbation), the information fed into the system is altered to add a degree of uncertainty. Polat and Du [29] proposed a singular value decomposition predictor based on random perturbation of data. The user's data is perturbed by adding a random value (from a fixed distribution) to each of the ratings; unknown ratings are filled in with the mean rating. They go on to show the impact on privacy and accuracy and their inherent trade-off due to perturbation. In later work [30], their setting is different. A user wants two companies to collaboratively compute recommendations for him. This user acts as a relay for the two companies. The user's privacy is based on randomizing values. Berkovsky et al. [6] proposed to combine random perturbation with a peer-to-peer structure to create a form of dynamic random perturbation. For each request, the user can decide what data to reveal and how much protection is put on the data. Different perturbation strategies are compared based on accuracy and perceived privacy. Shokri et al. [35] added privacy by aggregating user information instead of perturbing. Aggregation occurs between users, without interaction with the recommender system. Thus, the

recommender system cannot identify which information is part of the original user information and what is added by aggregation. A degree of uncertainty is added to the user's information similar to randomization.

Recently the field of randomization is shifting toward differential privacy [13], which aims to obscure the link between single users' information in the input (the user's information) and output (the recommendation). This is accomplished by making users in released data computationally indistinguishable from most of the other users in that dataset. This is typically accomplished by adding noise to the inputs or output, to hide small changes that arise from a single user's contribution. The required level of noise depends on how and how often the data will be used and typically involves a balancing act between accuracy of the output and privacy of the input. Such indistinguishability also applies strongly to collaborative recommender systems, where a user should be unable to identify individual peers' ratings in the output he receives. As each recommendation leaks a little bit of information about the input (even with noise), with a larger number of recommendations, the added noise should be greater to provide the same level of privacy. McSherry and Mironov [26] proposed collaborative filtering algorithms in the differential privacy framework. Noise is added to the item covariance matrix (for item similarity). Since the item covariance matrix is smaller than the user covariance matrix, less noise needs to be added and more accuracy is preserved.

The drawback of these techniques is that the security of these methods is hard to be *formally proven*, as is done in classical cryptography. The noise levels in differential privacy techniques must not overwhelm the initial output data and thus remove utility of the results completely. At the same time, enough noise must be added in order to hide the contribution of a user. When combined with multiple computational results and external information, even more noise is needed to protect the privacy of a user.

4.5 Privacy-Preserving Cryptographic Protocols

We first give an overview of some of the tools used in privacy-preserving cryptographic protocols, before addressing the protocols themselves. Among the tools [17] are secure multiparty computations, secret sharing, homomorphic encryption, and zero-knowledge proofs.

Secure multiparty computations are a class of protocols that allow two or more parties to collaboratively compute a function based on input held by each of them. The output of this function can be given to one of the parties or all of them. Any function can be computed, but the complexity of the protocol depends on the function. For example, multiplication and integer comparison.

Secret sharing distributes a number of shares of a value among different parties. The shares of a fixed number of parties need to be combined in order to reconstruct the original value. With less than the fixed number of shares, no information about the value can be obtained. Some secret sharing schemes allow basic operations (such as addition) to be performed.

Homomorphic encryption allows one (or sometimes more) operation (e.g., addition or multiplication) on the encrypted values, by performing a corresponding operation on the ciphertexts. This allows anyone to compute a (basic) function on the encrypted values, without knowledge of the actual values. Decryption is then required to get the result of the function.

Zero-knowledge proofs allow a user to prove a property about a value, without revealing that value. For example, that a value is in a given range of possible values. To do this, the user first sends a commitment to the verifier. Then the verifier ask the user to open the commitment in a certain way. The commitment can only be opened correctly when the property of the value holds. With a certain probability, the user can correctly open the commitment even if the property does not hold. However, by running multiple zero-knowledge proofs, this percentage can be reduced.

Privacy-Preserving Cryptographic Protocols Without Server

Privacy-preserving cryptographic protocols without a central server aim to remove the trust that is placed in service providers by removing them from the picture. Secure multiparty computations protect the privacy of users against each other. Canny [9, 10] used a combination of secure multiparty computation, homomorphic encryption, and zero-knowledge proofs to create a privacy-preserving recommender protocols without a central server. The users collaborate to privately compute intermediate values of the collaborative filtering process. These intermediate values (based on all users) are then made public. In the next step, the users perform singular value decomposition and factor analysis, which leads to a model for recommendations. This model is made publicly available and can be used by each user independently to compute recommendations for themselves.

The system proposed by Hoens et al. [19] allowed trusted friends to collaboratively compute recommendations with each other. They rely on Facebook for retrieving friendship information and a server to facilitate asynchronous messaging. Homomorphic cryptography and secure multiparty protocols are used to compute the actual recommendations for a given item.

Because a decentralized structure works strongly toward taking power away from the service provider, it is contrary to existing business models. This means that existing companies are not likely to adopt such a structure or aid its development. Another drawback is the involvement of many users that is required to make (the model for) the recommendations. These users need to interact with each other, but not all users will be available at the same time. This can lead to considerable delays or a loss of accuracy.

Privacy-Preserving Cryptographic Protocols with Server

Privacy-preserving cryptographic protocols with a central server aim to make use of the centralization offered by the service provider, while using secure two-party

computation and encryption to ensure the privacy of the users. Good motivations for the service provider would be a reduced liability for the data collected, an increased perception of security among users (and thus, a competitive advantage), and adherence to possible stricter future laws.

Aïmeur et al. [3] provided a framework for collaborative filtering, where user information is separately stored over two parties. An agent has access to ratings, and the company has access to the items, so that they together can generate recommendations for the user. The centralized structure is preserved, but neither the agent nor the company can link the user's ratings to the items. Erkin et al. [14, 15] proposed a collaborative filtering algorithm based on homomorphic cryptosystems. In their framework, a central server acts as a mediator between the users and is in charge of combining the results given by different users. When desiring a recommendation, a user sends an encrypted request to the central server. The server distributes this request to other users that can work on the request by using the homomorphic properties of the cryptosystem. A secure two-party computation then determines for each user if their information should be included in the recommendation or not. The central server then combines the (still encrypted) results to generate the recommendation.

Basu et al. [4,5] proposed a privacy-preserving version of the slope one predictor for collaborative filtering. The assumption is that different parties hold different parts of the information, this essentially allows multiple companies to collaborate. They precompute the deviation and cardinality matrices under encryption and make the cardinality matrix public. Then the prediction for a single item can be computed under encryption and all parties collaborate to decrypt the result.

The drawback of these schemes (that add a layer of encryption) is efficiency. The homomorphic operations and secure two-party computations are always more expensive than their unprotected counterparts. In fact, the discrepancy is often huge. This results in poor efficiency and scalability for these protocols, an issue that the research tries to address.

4.6 Summary

Table 3 shows which research areas contribute to address which privacy concern: a ● indicates that the area is helpful to address a particular concern, a • indicates that the area is somewhat helpful, while a · indicates that the area does not seem applicable. The majority of the research areas focusses on protecting the user's information from the service provider. As can been seen in Sect. 3.3, the privacy concerns related to the service provider have a high privacy impact.

None of the research areas mentioned in this section can offer complete user privacy for all recommender systems. Privacy is multifaceted, as are the domains in which recommender systems are applied. Several areas will likely need to be combined to develop proper privacy-protection techniques for a given application. In addition, service providers should be encouraged or required to implement such solutions, and users need to be made aware of the benefits of using them.

Table 3 Privacy concerns and relevant research areas

↓ Concern/research →	Awareness	Law	Anonymization	Randomization	Protocols w/o server	Protocols w/ server
Data collection	•	●	·	•	●	●
Data retention	·	●	·	·	●	●
Data sales	·	●	●	•	●	●
Employee	·	●	·	•	●	●
Recommendations	·	·	·	●	·	·
Shared service	•	·	·	·	·	·
Stranger views	•	·	·	•	•	•

5 Conclusion

We have seen that recommender systems play an important role in the online experience of millions of people. While accuracy has been the focus in recommender system development, we argue that privacy should not be overlooked. We have seen that depending on the type of information utilized by a recommender system, various privacy concerns exist. The fact that trust in the service provider is not always justified further complicates matters. With increased information sharing, users must weigh the advantages of getting (more accurate) recommendations against the privacy risks and should more often be given the choice to opt-in or opt-out of data collection.

Many areas of research can help to protect user privacy, ranging from technical (e.g., system design and cryptography) to nontechnical (e.g., sociology and law). However, we must realize that one single research area cannot address all privacy concerns. Furthermore, we notice a trend in the different research areas. The areas of awareness and law do not focus on any single specific type of recommender system. However, the areas that provide technical solutions mainly focus on collaborative filtering recommender systems. The other types of recommender systems are barely (if at all) represented.

As commonly known, in the technical solutions, there is an inherent trade-off between privacy, accuracy, and efficiency. Randomization techniques increase privacy by lowering accuracy and leaving efficiency the same. Cryptographic and secure multiparty computation protocols increase privacy by lowering efficiency and leaving accuracy the same. However, when aiming for a specific trade-off in a certain scenario and goal, it is difficult to choose the right solution. Comparison is difficult because researchers use different datasets and different measures for accuracy. It is an open question how different privacy-protection techniques compare to each other when applied to the same dataset, with the same accuracy measure, and the same programming language and hardware.

Our conclusion is that in order to develop a full solution to protect user privacy, the strengths of several research areas will need to be brought together. Ideally, privacy-protection techniques are built into the system design. These privacy-protection techniques should not harm the operations of the recommender system. Therefore, the users and the service provider should not be overburdened, and the functionality and accuracy of the recommender system should not be hampered.

Acknowledgements The research for this work was carried out within the Kindred Spirits project, part of the STW Sentinels research program.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
2. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 217–253. Springer, New York (2011)
3. Aïmeur, E., Brassard, G., Fernandez, J., Mani Onana, F.: Alambic: a privacy-preserving recommender system for electronic commerce. Int. J. Inf. Secur. **7**(5), 307–334 (2008)
4. Basu, A., Kikuchi, H., Vaidya, J.: Privacy-preserving weighted slope one predictor for item-based collaborative filtering. In: Proceedings of the International Workshop on Trust and Privacy in Distributed Information Processing. Springer, Berlin/Heidelberg (2011)
5. Basu, A., Vaidya, J., Kikuchi, H.: Efficient privacy-preserving collaborative filtering based on the weighted slope one predictor. J. Internet Serv. Inf. Secur. (JISIS) **1**(4), 26–46 (2011)
6. Berkovsky, S., Eytani, Y., Kuflik, T., Ricci, F.: Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In: Proceedings of the 2007 ACM Conference on Recommender Systems, Minneapolis, pp. 9–16 (2007)
7. Burke, R.: Knowledge-based recommender systems. In: Encyclopedia of Library and Information Systems, vol. 69, pp. 180–200. Marcel Dekker, New York (2000)
8. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User Adapt. Interact. **12**, 331–370 (2002)
9. Canny, J.: Collaborative filtering with privacy. In: IEEE Symposium on Security and Privacy, Oakland, pp. 45–57 (2002)
10. Canny, J.: Collaborative filtering with privacy via factor analysis. In: Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval, Tampere, pp. 238–245 (2002)
11. Cissée, R., Albayrak, S.: An agent-based approach for privacy-preserving recommender systems. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '07, pp. 182:1–182:8. ACM, New York (2007)
12. Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J.: The platform for privacy preferences 1.0 (p3p1.0) specification. http://www.w3.org/TR/P3P/
13. Dwork, C.: Differential privacy. In: Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, 10–14 July 2006, Proceedings, Part II, pp. 1–12 (2006)
14. Erkin, Z., Beye, M., Veugen, T., Lagendijk, R.L.: Privacy enhanced recommender system. In: Thirty-First Symposium on Information Theory in the Benelux, Rotterdam, pp. 35–42 (2010)

15. Erkin, Z., Beye, M., Veugen, T., Lagendijk, R.L.: Efficiently computing private recommendations. In: International Conference on Acoustic, Speech and Signal Processing-ICASSP, Prague, pp. 5864–5867 (2011)
16. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Commun. ACM **35**, 61–70 (1992)
17. Goldreich, O.: Foundations of cryptography: a primer. Found. Trend Theor. Comput. Sci. **1**, 1–116 (2005)
18. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: WPES '05: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, pp. 71–80. ACM, New York (2005)
19. Hoens, T.R., Blanton, M., Chawla, N.V.: A private and reliable recommendation system for social networks. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), Minneapolis, pp. 816–825 (2010)
20. Kang. J.: Information privacy in cyberspace transactions. Stanf. Law Rev. **50**(4), 1193–1294 (1998)
21. Kohnstamm, J.: Opinion 2/2010 on online behavioural advertising. Technical report 00909/10/EN WP 171, article 29 data protection working party, 6 2010. http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp171_en.pdf
22. Konstas, I., Stathopoulos, V., Jose, J.M.: On social networks and collaborative recommendation. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, pp. 195–202. ACM, New York (2009)
23. Lam, S., Frankowski, D., Riedl, J.: Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In: Müller, G. (ed.) Emerging Trends in Information and Communication Security. Lecture Notes in Computer Science, vol. 3995, pp. 14–29. Springer, Berlin/Heidelberg (2006)
24. Lang, K., Newsweeder: learning to filter netnews. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 331–339. Morgan Kaufmann, San Francisco (1995)
25. Lorenzi, F., Ricci, F.: Case-based recommender systems: a unifying view. In: Intelligent Techniques for Web Personalization. Lecture Notes in Computer Science, vol. 3169, pp. 89–113. Springer, Berlin/Heidelberg (2005)
26. McSherry, F., Mironov, I.: Differentially private recommender systems: building privacy into the netflix prize contenders. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, pp. 627–636 (2009)
27. Narayanan, A., Shmatikov, V.: How to break anonymity of the netflix prize dataset. In: CoRR: Computing Research Repository, pp. 1–24. Cornell University, Ithaca (2006)
28. Palen, L., Dourish, P.: Unpacking "privacy" for a networked world. In: CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 129–136. ACM, New York (2003)
29. Polat, H., Du, W.: Svd-based collaborative filtering with privacy. In: Proceedings of the 2005 ACM Symposium on Applied Computing, Santa Fe, pp. 791–795 (2005)
30. Polat, H., Du, W.: Privacy-preserving top-n recommendation on distributed data. J. Am. Soc. Inf. Sci. Technol. **59**, 1093–1108 (2008)
31. Ramakrishnan, N., Keller, B.J., Mirza, B.J., Grama, A.Y., Karypis, G.: Privacy risks in recommender systems. IEEE Internet Comput. **5**(6), 54–62 (2001)
32. Rich, E.: User modeling via stereotypes. Cognit. Sci. **3**(4), 329–354 (1979)
33. Rosenblum, D.: What anyone can know: the privacy risks of social networking sites. IEEE Secur. Priv. **5**(3), 40–49 (2007)
34. Schclar, A., Tsikinovsky, A., Rokach, L., Meisels, A., Antwarg, L.: Ensemble methods for improving the performance of neighborhood-based collaborative filtering. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pp. 261–264. ACM, New York (2009)

35. Shokri, R., Pedarsani, P., Theodorakopoulos, G., Hubaux, J.-P.: Preserving privacy in collaborative filtering through distributed aggregation of offline profiles. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pp. 157–164. ACM, New York (2009)
36. Sweeney, L.: K-anonymity: a model for protecting privacy. IEEE Secur. Priv. **10**(5), 557–570 (2002)
37. Tsai, J.Y., Egelman, S., Cranor, L., Acquisti, A.: The effect of online privacy information on purchasing behavior: an experimental study. Inf. Syst. Res. **22**, 254–268 (2011)
38. Tufekci, Z.: Can you see me now? audience and disclosure regulation in online social network sites. Bull. Sci. Technol. Soc. **28**(1), 20–36 (2008)
39. Weiss, S.: The need for a paradigm shift in addressing privacy risks in social networking applications. In: The Future of Identity in the Information Society, Brno, vol. 262, pp. 161–171. IFIP International Federation for Information Processing. Karlstad (2008)

Geotag Propagation with User Trust Modeling

Ivan Ivanov, Peter Vajda, Jong-Seok Lee, Pavel Korshunov, and Touradj Ebrahimi

Abstract The amount of information that people share on social networks is constantly increasing. People also comment, annotate, and tag their own content (videos, photos, notes, etc.), as well as the content of others. In many cases, the content is tagged manually. One way to make this time-consuming manual tagging process more efficient is to propagate tags from a small set of tagged images to the larger set of untagged images automatically. In such a scenario, however, a wrong or a spam tag can damage the integrity and reliability of the automated propagation system. Users may make mistakes in tagging, or irrelevant tags and content may be added maliciously for advertisement or self-promotion. Therefore, a certain mechanism insuring the trustworthiness of users or published content is needed. In this chapter, we discuss several image retrieval methods based on tags, various approaches to trust modeling and spam protection in social networks, and trust modeling in geotagging systems. We then consider a specific example of automated geotag propagation system that adopts a user trust model. The tag propagation in images relies on the similarity between image content (famous landmarks) and its context (associated geotags). For each tagged image, similar untagged images are found by the robust graph-based object duplicate detection, and the known tags are propagated accordingly. The user trust value is estimated based on a social feedback from the users of the photo-sharing system, and only tags from trusted users are propagated. This approach demonstrates that a practical tagging system significantly benefits from the intelligent combination of efficient propagation algorithm and a user-centered trust model.

I. Ivanov (✉) • P. Vajda • P. Korshunov • T. Ebrahimi
Multimedia Signal Processing Group (MMSPG), École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Vaud, Switzerland
e-mail: ivan.ivanov@epfl.ch; peter.vajda@epfl.ch; pavel.korshunov@epfl.ch; touradj.ebrahimi@epfl.ch

J.-S. Lee
School of Integrated Technology, Yonsei University, Incheon 406-840, South Korea
e-mail: jong-seok.lee@yonsei.ac.kr

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_13, © Springer-Verlag London 2013

1 Introduction

Social networks and photo-sharing websites have become increasingly popular in recent years. Their services typically focus on building online communities of people who interact with each other by sharing their own interests or activities and exploring shared content of others. Such social networks have become a popular way to disseminate different types of information, such as photo, video, text, and audio. For example, a user uploads a wedding album to let other people from the online community comment or rate the photos. This sharing trend has resulted in a continuously growing volume of publicly available photos on such websites like Flickr,[1] Picasa,[2] or Photobucket,[3] as well as social networks like Facebook[4] and Google+.[5] For instance, Photobucket hosts more than eight billion photos [17] seven billion photos are hosted on Picasa [17], and six billion photos on Flickr [42]. Facebook has more than 250 million photos posted to its network every day [32] and approximately 100 billion photos stored on its servers [2], while 3.4 billion photos have been uploaded to Google+ [2], in the first 100 days of it being open to the public. This large volume of multimedia content poses significant challenges for efficient search, retrieval, and processing of the shared content.

Tagging is one of the popular methods to categorize large volume of photos. It is a process by which users assign short textual annotations to photos (in the form of keywords) to describe them and to provide additional information for search engines, online photo albums, and for people browsing the photo collections. Tags, when combined with search technologies, are essential in resolving user queries targeting shared photos. The success of social networks such as Flickr, Google+, and Facebook proves that users are willing to provide tags through manual annotations. Different users annotating the same photo can enrich the information about that photo. However, tagging a lot of photos by hand is a time-consuming task. Users typically tag a small number of the shared photos only, leaving most of the other photos with incomplete metadata. This lack of metadata decreases the precision of search because photos without proper annotations are typically much harder to retrieve than correctly annotated photos. Therefore, to help people organize and browse large collections of personal photos in an effective way, it is important to develop robust and efficient algorithms for automatic tagging or tag propagation.

Another important challenge in tagging is to identify most appropriate tags for given content and, at the same time, to eliminate noisy or spam tags. The shared photos are sometimes assigned with inappropriate tags for several reasons. First of all, users are human beings and make mistakes. It is also possible that misleading

[1] http://www.flickr.com

[2] http://picasa.google.com

[3] http://www.photobucket.com

[4] http://www.facebook.com

[5] http://plus.google.com

tags are assigned for advertisement purposes, self-promotion, or to increase the rank of a particular tag in the search engines. Consequently, free-form keywords (tags) assigned to photos carry a significant risk that wrong or irrelevant tags eventually prevent users from the intended benefits of annotated photos. Finally, wrong machine tags, such as longitude and latitude, can be automatically assigned to images captured with cameras equipped with GPS devices due to bad or noisy communication channels with GPS satellites or wireless access points. Kennedy et al. [15] analyzed the Flickr website and revealed that the tags provided by users are often imprecise and only around 50% of tags are truly related to an image. Beside the tag-photo association, spam objects can take other forms, that is, possibly manifesting as a spam photo or a spam user (spammer). Therefore, for the practical tag propagation system, it is important to consider user trust information derived from users' tagging behavior.

Trust provides a natural security policy stipulating that users or photos with low trust values should be investigated or eliminated. Trust can predict the future behavior of users in order to avoid undesirable influences of untrustworthy users. Trust-based schemes can be used to motivate users to positively contribute to social networks and/or penalize adversaries that are trying to disrupt the network. Therefore, the distribution of the trust values associated with either users or photos in a social network can represent the health of the network and used in a spam-free tag propagation algorithm.

In this chapter, we focus on trust aspects and trust models used in social networks and applicability of these models in automatic tag propagation systems. In Sect. 2, we first discuss geotagging and how it is used in various social networks and media retrieval systems. In Sect. 3, we introduce several techniques used for combatting noise and spam through trust modeling in social tagging systems. In Sect. 4, we present detailed overview of several trust modeling approaches, which are specific to geotagging systems. And, in Sect. 5, we demonstrate the advantages of using trust modeling on an example of automatic geotag propagation system in travel-related photos. We conclude this chapter with Sect. 6.

2 Geotagging in Social Networks and Sharing Websites

In the last few years, an important trend in multimedia understanding is modeling and extracting value from geographical context, such as GPS coordinates, and visual content, such as a digital representation (description) of a photo. Different research problems and significant approaches in this field are summarized by Luo et al. [24]. In this section, we focus on some of the representative image retrieval approaches that rely on a variety of image or landmark descriptors combined with geographic information. These approaches are summarized in Table 1.

A pioneering paper in this area by Hays and Efros [8] proposed an algorithm called IM2GPS to estimate the locations of a single image using a purely data-driven scene matching approach. Given a test image, the algorithm finds the visual

Table 1 Summary of representative recent techniques that combine geographical context and visual content for automatic geotagging of images

Reference	Descriptor	Method	Application
Hays and Efros [8]	Visual features	The probability distribution for the location of an unknown image is found on the globe using a purely data-driven scene matching	Non-landmark (scene) location recognition
Kennedy and Naaman [16]	Visual and textual features	For a given location, diverse and representative images are generated based on geotagged community images	Visual summary of landmarks
Zheng et al. [47]	Visual features and GPS coordinates	Travel blogs and geotagged images are analyzed, and a list of tourist landmarks is established based on the information from nearest neighbors	Landmark recognition
Quack et al. [33]	Visual and textual features and GPS coordinates	Objects and events are retrieved from a large-scale collection of geotagged images using pair-wise similarity	Event/scene understanding

nearest neighbors in the database and estimates a geolocation of the image from the GPS coordinates of the tagged nearest neighbors. The estimated image location is represented as a probability distribution over the Earth's surface. However, the IM2GPS approach showed low recognition accuracy due to low-level features. While IM2GPS uses a set of more than six million training images, its general applicability is inconclusive because the performance was verified only on 237 hand-selected test images.

Kennedy and Naaman [16] presented a method to search representative landmark images from a large collection of geotagged images. This method uses tags and the geographical location representing a landmark. The visual features (global color and texture features and scale invariant feature transform (SIFT)) are analyzed to cluster landmark images into visually similar groups. The method has been proven to be effective for extraction of the representative image sets for a given landmark. But since it cannot be applied to untagged images, its applicability is limited.

The recent work of Zheng et al. [47] automatically finds frequently photographed landmarks from a large collection of geotagged photos. The authors perform clustering on GPS coordinates and visual texture features from the image pool and extract landmark names as the most frequent tags associated with the particular visual cluster. Additionally, they extract landmark names from the travel guide articles, such as Wikitravel,[6] and visually cluster photos gathered

[6] http://www.wikitravel.com

by querying Google Images.[7] However, the test set they use is quite limited – 728 images in total for a 124-category problem or less than six test images per landmark.

Another application that combines textual and visual techniques has been proposed by Quack et al. [33]. The authors developed a system that crawls photos on the Internet and identifies clusters of images referring to a common object (physical items on fixed locations) and events (special social occasions taking place at certain times). The clusters are created based on the pair-wise visual similarities between the images, and the metadata of the clustered photos are used to derive labels for the clusters. Finally, Wikipedia[8] articles are attached to the images, and the validity of these associations is checked. Gammeter et al. [6] extend this idea toward object-based auto-annotation of holiday photos in a large database that includes landmark buildings, statues, scenes, and pieces of art, with the help of external resources such as Wikipedia. In both [33] and [6], GPS coordinates are used to pre-cluster objects which may not be always available.

Most of the photo-sharing websites (e.g., Flickr, Picasa, Panoramio,[9] Zooomr,[10]) provide information about where images were taken in form of maps or groups. This information is either provided by an external GPS sensor and stored as image metadata (exchangeable image file format (EXIF) [35], International Press Telecommunications Council (IPTC) [11]) or manually annotated via geocoding.

The main disadvantage of the above systems is that they rely on GPS coordinates to derive geographical annotation, which is not available for the majority of web images and photos, since only a few camera models are equipped with GPS devices. Furthermore, a GPS sensor in a camera provides only the location of the photographer instead of that of the captured landmark, which may be up to several kilometers away. Therefore, the GPS coordinates alone may not be enough to distinguish between two landmarks within a city. Describing landmarks through location names rather than GPS coordinates is not only more reliable but also more expressive. A recent study by Hollenstein and Purves [10] indicated that geotagging should follow the way people actually describe locations, that is, it is more convenient to use Church of Saint Sava in Belgrade rather than latitude 44.798083 and longitude 20.46855. Therefore, there is a growing interest in the research community to derive geographic locations of the scenes in photos based on visual and text features.

[7] http://images.google.com

[8] http://www.wikipedia.org

[9] http://www.panoramio.com

[10] http://www.zooomr.com

3 Trust Modeling in Social Media

When information is exchanged on the Internet, malicious individuals are everywhere trying to take advantage of the information exchange structure for their own benefit, while bothering and spamming others. Before social tagging became popular, spam content was observed in various domains: first in e-mail (e.g., [34]) and then in web search (e.g., [5]). Peer-to-peer (P2P) networks have been also influenced by malicious peers, and thus various solutions based on trust and reputation have been proposed, which dealt with collecting information on peer behavior, scoring and ranking peers, and responding based on the scores [27]. Nowadays, even blogs are spammed [36]. Ratings in online reputation systems, such as eBay,[11] Amazon[12] and Epinions,[13] are very similar to tagging systems, and they may face the problem of unfair ratings by artificially inflating or deflating reputations [14]. Several filtering techniques for excluding unfair ratings are proposed in the literature (e.g., [41, 46]). Unfortunately, the countermeasures developed for the e-mail and web spam do not directly apply to social networks and photo-sharing websites [9].

In order to reduce or eliminate spams in social networks, various antispam methods have been proposed in the state-of-the-art research. Heymann et al. [9] classified antispam strategies into three categories: prevention, detection, and demotion. *Prevention-based approaches* aim at making it difficult for spam content to contribute to social networks by restricting certain access types through interfaces (such as CAPTCHA [39] or reCAPTCHA [40]) or through usage limits (such as tagging quota, e.g., Flickr introduced a limit of 75 tags per photo [45]). *Detection approaches* identify likely spams either manually or automatically by making use of, for example, machine learning (such as text classification) or statistical analysis (such as link analysis), and then deleting the spam content or visibly marking it as hidden to users. Finally, *demotion-based approaches* reduce the prominence of content likely to be spam. For instance, rank-based methods produce ordering of a network's content, tags, or users based on their trust scores. The prevention-based approaches can be considered as a type of precaution to prevent spammers. However, they cannot completely secure a social network. Some studies, for example, [29], showed that CAPTCHA systems can be defeated by computers with around 90% of accuracy, using, for example, optical character recognition or shape context matching. Even if prevention methods were perfect, there would be still possibility that the social networks get polluted with spam (malicious) or irrelevant tags. Therefore, detection and demotion via trust modeling are required to keep a network free of noise and spam.

In a social network with tagging capability, spam or noise can be injected at three different levels: spam content (in our case photos, but might be any piece

[11] http://www.ebay.com
[12] http://www.amazon.com
[13] http://www.epinions.com

of information – videos, textual documents, or web pages), spam tag-content association, and spammer [25]. Trust modeling can be performed at each level separately (e.g., [25]), or different levels can be considered jointly to produce trust models, for example, to assess a user's reliability, one can consider not only the user profile but also the content that the user uploaded to a social network (e.g., [20]). Trust modeling approaches can be categorized into two classes according to the target of trust, that is, content and user trust modeling [12].

Content trust modeling is to classify content (e.g., web pages, images, videos) as spam or legitimate. In this case, the target of trust is a content, and thus, a trust score is given to each content. Approaches for content trust modeling utilize features extracted from content information, users' profiles, and/or associated tags to detect specific spam content.

Gyongyi et al. [7] proposed an algorithm called TrustRank to semiautomatically separate reputable from spam web pages. TrustRank relies on an important empirical observation called approximate isolation of the good set: good pages seldom point to bad ones. It starts from a set of seeds selected as high-qualified, credible, and popular web pages in the web graph and then iteratively propagate trust scores to all nodes in the graph by splitting the trust score of a node among its neighbors according to a weighting scheme. TrustRank effectively removes most of the spam from the top-scored web pages; however, it is unable to effectively separate low-scored good sites from bad ones, due to the lack of distinguishing features. In search engines, TrustRank can be used either solely to filter search results or in combination with PageRank and other metrics to rank content in search results.

Wu et al. [43] proposed a computer vision-based technique that discriminates spam images from legitimate ones. By assuming that images containing text are likely to be spam (e.g., banners), they identified a number of useful low-level image features detecting embedded text and computer-generated graphics. Then, pattern classification using support vector machines (SVMs) was performed to classify spam and nonspam images. Although they reported a high detection rate with a low false-positive rate, this approach has limitations in that the discriminant capability of the used features may be limited, and, moreover, the assumption that images containing text or computer-generated images are likely to be spam may not be true in some cases.

In *user trust modeling*, trust is given to each user based on the information extracted from a user's profile, his/her interaction with other participants within the social network and/or the relationship between the content and tags that the user contributed to the social network. Given a user trust score, the user might be flagged as a legitimate user or spammer. Most of user trust modeling techniques use machine learning approaches applied to features specific to considered social network domains.

Krause et al. [20] employed a machine learning approach to identify spammers in BibSonomy.[14] They investigated features considering information about a user's

[14] http://www.bibsonomy.org

profile (e.g., number of digits in the username and the e-mail address), location (e.g., number of spam users with the same IP), bookmarking activity (e.g., number of tags per post), and context of tags (e.g., user co-occurrences with spammers related to tags, content, and tag-content pairs). By making use of these features and SVM or naive Bayes classifier, they were able to distinguish legitimate users from malicious ones. It was found that the cooccurrence features describing the usage of a similar vocabulary and content usage are the most promising.

Markines et al. [25] proposed six different tag-, content-, and user-based features for automatic detection of spammers in BibSonomy. First, tag- and content-based features are averaged across each user's posts, then combined with user-based features, and finally fed into a supervised learning algorithm (such as LogitBoost or AdaBoost) to discriminate spammers from legitimate users. It was shown that TagSpam feature (probability that a particular tag is used to spam, aggregated across all tags assigned to a content) is the best predictor of spammers among all other features because spammers tend to use certain "suspect" tags more than legitimate users. DomFp feature (likelihood that a content is spam based on its structure) also appeared important but may not be available since it relies on an infrastructure to enable access to the content, and therefore, its feasibility depends on the circumstances of a particular social tagging system.

Noll et al. [31] introduced the time of tagging as an additional dimension for assessing the trust of a user in Delicious.[15] They proposed a graph-based algorithm, called SPEAR (SPamming-resistant Expertise Analysis and Ranking). It computes the expertise score of a user and the quality score of a content which are dependent on each other. The time of tagging is considered so that the earlier a user tags a content, the more expertise score he/she receives. These two scores are calculated iteratively in a similar way to that of the Hyperlink-Induced Topic Search (HITS) algorithm. It was shown that SPEAR produces better ranking of users than the HITS method. SPEAR was able to demote different types of spammers (flooders, promoters, and trojans [31]) and remove them from the top of the ranking.

It can be noted that approaches based on user trust modeling are more common than content trust modeling. One reason is that the user-centered model is simpler to describe than content-centered. Also, user trust models can quickly adapt to the constantly evolving and changing environment in social systems due to the type of features used for modeling and thus be applicable longer than content trust models, without need for creation of new models. On the other hand, user trust modeling has a disadvantage of "broad brush," that is, it may be excessively strict if a user happens to post one bit of questionable content on otherwise legitimate content. Trustworthiness of a user is often judged based on the content that the user uploaded to a social system, and thus, "subjectivity" in discriminating spammers from legitimate users remains an issue for user trust modeling as in content trust modeling.

[15] http://www.delicious.com

4 Trust Modeling in Geotagging Applications

From the general trust modeling described in the previous section, we now shift the discussion to a more specific problem of geotagging the shared content and efficient propagation of such tags throughout the untagged content. In this section, we present and discuss several techniques for combatting noise and spam through trust modeling in social tagging systems. First, we introduce the model of a social tagging system. Then we present in details the five recent techniques for trust modeling that are suitable for geotagging and can be used in geotag propagation systems.

The model of a social tagging system [26] consists of *users* who interact with the system, *content* (resources or documents) which might be any piece of information (e.g., photos, videos, textual documents, or web pages), and *tags* which are descriptions assigned to the piece of the content by users. The action of associating a tag to a content by a user is usually referred to as *tag assignment* [22]. Depending on the system under consideration, a user can assign one or several tags to each type of content. Following notations are used in formal description of the trust models: U is a set of users u, D denotes a set of documents (content) d, T is a set of tags t, and a set of tag assignments p is denoted as $P \in U \times D \times T$.

Table 2 summarizes five trust modeling approaches, which we then describe in more details (in the same order as they are presented in the table). These methods

Table 2 Summary of five trust modeling techniques used for combatting noise and spam in social tagging systems

Reference	Content	Method	Dataset
Koutrika et al. [19]	Bookmarks	A coincidence-based model for query-by-tag search, which estimates the level of agreement among different users in the system for a given tag	Delicious, real, and simulated
Liu et al. [22]	Bookmarks	An iterative approach to identify spam content by its information value extracted from the collaborative knowledge	Delicious, real
Xu et al. [44]	Bookmarks	An iterative approach to compute the goodness of each tag with respect to a content and the authority scores of the users	MyWeb 2.0, real
Krestel and Chen [21]	Bookmarks	A TrustRank-based approach using features which model tag cooccurrence, content cooccurrence, and cooccurrence of tag-content	BibSonomy, real
Ivanov et al. [13]	Images	An approach based on the feedback from other users who agree or disagree with a tag associated with an image	Panoramio, real

are different in the targeted media content, for which the geotagging is intended, the application they are used in, and the required level of participation from the users of the geotagging system.

4.1 A Coincidence-Based Model

Koutrika et al. [19] were the first to explicitly discuss methods of tackling spamming activities in social tagging systems. The authors studied the impact of spamming through a framework for modeling social tagging systems and user tagging behavior. They proposed a method for ranking content matching a tag based on taggers' reliability in social bookmarking service Delicious. Their coincidence-based model for query-by-tag search estimates the level of agreement among different users in the system for a given tag. A bookmark is ranked high if it is tagged correctly by many reliable users. A user is more reliable if his/her tags more often coincide with other users' tags.

In more formal way, the following calculations are performed:

$$c(u) = \sum_{d,t:\exists P(u,d,t)} \sum_{u_i \in U : u_i \neq u} |p : \exists P(u_i, d, t)| \tag{1}$$

$$\text{score}(d,t) = \frac{\sum_{u:\exists P(u,d,t)} c(u)}{\sum_{u \in U} c(u)} \tag{2}$$

$$\text{trust}^{\text{Koutrika}}(u) = \sum_{d,t:\exists P(u,d,t)} \text{score}(d,t) \tag{3}$$

where $c(u)$, coincidence factor of the user u, is the number of other users u_i who assigned the same tag t to the same document d as the user u did. Score of the document d with respect to the tag t, denoted as $\text{score}(d,t)$, is calculated as a normalized value of c over all users who assigned t to d. Finally, a trust value of the user u, $\text{trust}^{\text{Koutrika}}(u)$, is the sum of $\text{score}(d,t)$ over all tag assignments by u.

Koutrika et al. performed a variety of evaluations of their trust model on controlled (simulated) dataset by populating a tagging system with different user tagging behavior models, including a good user, bad user, targeted attack model, and several other models. Using controlled data, interesting scenarios that are not covered by real-world data could be explored. It was shown that spam in tag search results using the coincidence-based model is ranked lower than in results generated by, for example, a traditional occurence-based model, where content is ranked based on the number of posts that associate the content to the query tag.

4.2 A Wisdom of Crowds Model

Liu et al. [22] proposed a simple but effective approach for detecting spam content in Delicious, by harvesting the wisdom of crowds. An information value of a bookmark is defined as the average number of times that each tag of the content is assigned by different users. A low information value of a bookmark indicates a divergence from crowds, which can be considered as a spam content. Furthermore, this method was extended to user trust modeling by aggregating the information values for each user.

All measures are defined as follows:

$$it(d,t) = \frac{|u : \exists P(u,d,t)|}{\sum_{t' \in T} |u : \exists P(u,d,t')|} \tag{4}$$

$$ic(u,d) = \frac{\sum_{t:\exists P(u,d,t)} it(d,t)}{|t : \exists P(u,d,t)|} \tag{5}$$

$$I(d) = \frac{|u : \exists P(u,d,.)|}{\sum_{d' \in D} |u : \exists P(u,d',.)|} \tag{6}$$

$$\text{trust}^{\text{Liu}}(u) = \sum_{d:\exists P(u,d,.)} I(d) \cdot ic(u,d) \tag{7}$$

where $it(d,t)$ represents the tag's t tagging information value with respect to document d and $ic(u,d)$ is the information value of the content (document) d with respect to user u. The importance of the document d is defined with $I(d)$. Finally, a trust value of the user u, $\text{trust}^{\text{Liu}}(u)$, is calculated as the weighted average of the information value of the content tagged by user u, with the importance of the document as weight.

An interesting point is that, for the time being, Liu et al. collected the largest dataset for trust modeling by crawling Delicious [12]. This dataset had around 82,000 users, 1.1 million tags, 9.3 million bookmarks, and 17.4 million tag-bookmark associations.

4.3 An "Authority" Model Based on Goodness of Tags

Xu et al. [44] introduced the concept of "authority" in social bookmarking systems, where they measured the goodness of each tag with respect to a content by the sum of the authority scores of the users who have assigned the tag to the content. Authority scores and goodness are iteratively updated by using HITS algorithm, which was initially used to rank web pages based on their linkage on the web [18].

Following measures are defined and iteratively calculated:

$$s_{i+1}(d,t) = \sum_{u:\exists P(u,d,t)} \text{trust}_i^{\text{Xu}}(u) \tag{8}$$

$$\text{trust}_i^{\text{Xu}}(u) = \frac{\sum_{d,t:\exists P(u,d,t)} s_i(d,t)}{|t : \exists P(u,.,t)|} \tag{9}$$

where $i \in [1 \ldots Q]$, $s_i(d,t)$ is the goodness of each tag t with respect to a content d, and $\text{trust}_i^{\text{Xu}}(u)$ represents a trust value (authority score) of the user u. Initial settings in this iterative approach are $s_0(d,t) = 0, \forall t, d$ and $\text{trust}_0(u) = 1, \forall u$. The number of iterations is set to $Q = 100$.

4.4 A Cooccurrence Model

In contrast to the approach of Xu et al. [44], Krestel and Chen [21] iteratively updated scores for users only. The authors proposed to use a spam score propagation technique to propagate trust scores through a social graph in BibSonomy, where edges between nodes (in this case, users) indicate the number of common tags supplied by users, common content annotated by users, and/or common tag-content pairs used by users. Starting from a manually assessed set of nodes labeled as spammers or legitimate users with the initial spam scores, a TrustRank metric is used to calculate and iteratively update spam scores for all users. TrustRank metric is previously introduced in Sect. 3.

All measures are calculated as follows:

$$W(u_1,u_2) = |t : \exists P(u_1,.,t), P(u_2,.,t)| + |d : \exists P(u_1,d,.), P(u_2,d,.)| + |d,t : \exists P(u_1,d,t), P(u_2,d,t)| \tag{10}$$

$$Tr(u_1,u_2) = \frac{W(u_1,u_2)}{\sum_{v \in U} W(u_1,v)} \tag{11}$$

$$\text{trust}_i^{\text{Krestel}}(u) = \alpha \cdot \sum_{v \in U} Tr(u,v) \cdot \text{trust}_{i-1}^{\text{Krestel}}(v) - (1-\alpha)d(u) \tag{12}$$

where $i \in [1 \ldots Q]$, $W(u_1,u_2)$ is the weight of the edge between users u_1 and u_2 in the social graph and $Tr(u_1,u_2)$ is the corresponding transition matrix. A trust value of the user u, $\text{trust}_i^{\text{Krestel}}(u)$, is iteratively calculated. Initial setting in this iterative approach is $\text{trust}_0(u) = d(u), \forall u$, where $d(u)$ represents the trust values of the seed users. The number of iterations is set to $Q = 100$.

The approach of Krestel and Chen is more sophisticated than the approach of Xu et al. [44] in that multiple relationships, such as tag cooccurrence, content cooccurrence and tag-content cooccurrence, can be taken into account rather than considering only the tag-content pairs shared by users.

4.5 User Reliability-Based Model

In this section, we describe our own approach for user trust modeling in image tagging, which was proposed in [13]. First, we evaluate the trust or reliability of users by making use of their past behavior in tagging. We want to distinguish between users who provide reliable geotags and those who do not. After user evaluation and trust model creation, tags will be propagated to other photos in the database only if the user is trusted. Assuming that there are L users who tag M training images, a matrix $R_{i,u}$, $i \in [1 \ldots M]$ and $u \in [1 \ldots L]$, is defined as:

$$R_{i,u} = \begin{cases} 1, \text{ if user } u \text{ tags image } i \text{ correctly} \\ 0, \text{ otherwise} \end{cases} \tag{13}$$

The process of comparing the propagated tags to ground truth tags can be done automatically using tag similarity measures, for example WordNet [3] or Google distance [4] measures. Nevertheless, we considered only manually defined ground truth for our experiments.

A trust value for user u, $\text{trust}^{\text{Ivanov}}(u)$, is computed as the percentage of the correctly tagged images among all images tagged by user u:

$$\text{trust}^{\text{Ivanov}}(u) = \frac{\sum_{i=1}^{M} R_{i,u}}{M} \tag{14}$$

Only tags from users who are trusted are propagated to other photos in the dataset. In other words, if the user trust value, $\text{trust}^{\text{Ivanov}}(u)$, exceeds a predefined threshold $\hat{T}$, then all his/her tags are propagated. Otherwise, none of his/her tags are propagated.

In this approach, ground truth data are used for the estimation of the user trust value. However, for a practical photo-sharing system, such as Panoramio, it is not necessary to collect ground truth data since user feedback can replace them. The main idea is that users evaluate tagged images by assigning a true or a false flag to the tag associated with an image. If the user assigns a false flag, then he/she needs to suggest a correct tag for the image. The more misplacements a user has, the more untrusted he/she is. By applying this method, spammers and unreliable users can be efficiently detected and eliminated. Therefore, the user trust value is calculated as the ratio between the number of true flags and all associated flags over all images tagged by that user. The number of misplacements in Panoramio is analogous to the number of wrongly tagged images in our approach.

In case that a spammer attacks the system, other users can collaboratively eliminate the spammer. First, the spammer wants to make other users untrusted, so he/she assigns many false flags to the tags given by the trusted users and sets new wrong tags to these images. In this way, the spammer becomes trusted. Then, other users correct the tags given by the spammer, so that the spammer becomes untrusted and all of his/her feedbacks in the form of flags are not considered in the whole system. Finally, previously trusted users, who were untrusted due to spammer attack,

recover their status. Following this scenario, the user trust value can be constructed by making use of the feedbacks from other users who agree or disagree with the tagged location. However, due to the lack of a suitable dataset which provides user feedback, the evaluation of the user trust scenario is based on the simulation of the social network environment as described in details in [13].

5 An Automated Geotag Propagation System

Based on the user reliability trust modeling described in Sect. 4.5, we built the solution for geotag propagation between images. The main innovation of such system is the combination of object duplicate detection and user trust modeling for accurate and reliable geotag propagation. The system architecture has been proposed previously in [13] and is illustrated here in Fig. 1. It contains three functional modules, each of which has a specific task: object duplicate detection, tag propagation, and user trust modeling. As the focus of this chapter is on trust modeling, the object duplicate detection [38] and tag propagation [13] modules are only summarized briefly below.

The system takes a small set of training images with associated geotags to create the corresponding object (landmark) models. These object models are used to detect objects duplicated in a set of untagged images. As a result, matching scores

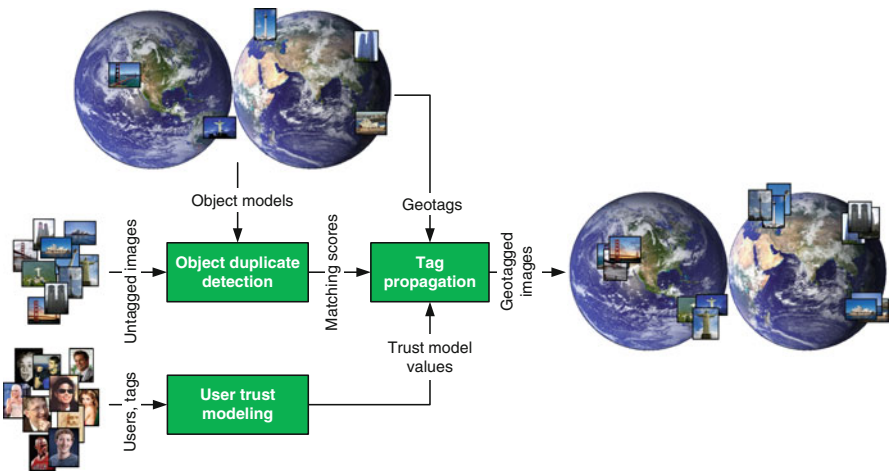


Fig. 1 Overview of the system for geotag propagation in images. The object duplicate detection is trained with a small set of images with associated geotags. The created object (landmark) models are matched against untagged images. The resulting matching scores serve as an input to the tag propagation module, which propagates the corresponding tags to the untagged images. Given a user trust model, only the tags from reliable users are propagated

between the models and the images are obtained. According to the scores, the tag propagation module makes decisions about which geotags should be propagated to the individual images. Given a user trust model which describes the tagging reliability of each user, only the tags from the users who are trusted are propagated to the photos in the dataset.

5.1 Object Duplicate Detection

The goal of the object duplicate detection module is to detect the presence of a target object in an image based on an object model created from training images. Duplicate objects may vary from their perspective, have different size, or be modified versions of the original objects after minor manipulations, as long as such manipulations do not change their identity. This is especially true for images related to travel, where tourists tend to take a lot of photos from different distances and viewpoints around a famous landmark. The basic idea of applying object duplicate detection for geotag propagation is that travel images typically depict distinctive landmarks (buildings, mountains, bridges, etc.), which can be considered as object duplicates.

Training is performed as follows: given a set of images, features are extracted, and a spatial graph model describing the object, that is, landmark, is created for each of the landmarks. In our case, one training image per landmark is used to create a graph model. First, regions of interest (ROIs) in an image are extracted using the Hessian affine detector [28], and each of these regions is described using SIFT features [23]. These features are robust to arbitrary changes in viewpoints. Then, hierarchical k-means clustering [30] is applied to the features, to group them based on their similarity. The result of the hierarchical clustering is used for the fast approximation of the nearest neighbor search, to efficiently resolve feature matching in the test phase. Finally, a spatial graph model is constructed to improve the accuracy of the feature matching with a test image. The graph model considers the scale, orientation, position, and neighborhood of features. The nodes of the graph are the features of the training images. The edges of the graph connect features with their spatial nearest neighbors. The attributes of edges are the distance and orientation of the neighbors. These attributes are important for the matching step in the test phase.

To detect the presence of the landmark within a test image, the features are extracted from the image in the same way described above. These features are matched to those in the graph model derived from the training images. Feature matching is performed using a one-to-one nearest neighbor matching, where the hierarchical clustering is used to efficiently resolve the nearest neighbor search. Considering only matched features and their positions, a spatial graph model of the query image is constructed in the same way described in the training phase. Then, graph matching is applied between two graph models to identify the local correspondences between regions in the training and the test image. Finally, for

the global object matching and matching score computation, the general Hough transform [1] is applied on the nodes of the matched graph. The matching scores represent the pair-wise comparison of training and test images.

More details about the proposed object duplicate detection approach are presented in [37, 38].

5.2 Tag Propagation

The goal of the tag propagation module is to propagate the geotags from the tagged to the untagged images according to the matching scores, provided by the object duplicate detection module. As a result, labels from the training set are propagated to the same object found in the test set.

The geographical metadata (geotags) embedded in the image file usually consist of location names and/or GPS coordinates but may also include altitude, viewpoint, etc. Two of the most commonly used metadata formats for image files are EXIF and IPTC. In this chapter, we consider the existing IPTC schema and introduce a hierarchical order for a subset of the available geotags, namely, city (name of the city where image was taken) and sublocation (area or name of the landmark), for example, Paris (Eiffel Tower) and Budapest (Parliament).

It was shown in [13] that tag propagation module supports two application scenarios: closed and open set problem. In the *closed set problem*, each test image is assumed to correspond to exactly one of the known (trained) landmarks. Therefore, the image gets assigned to the most probable trained landmark, based on the matching scores provided by the object duplicate detection module, and the corresponding tag is propagated to the test image. However, in the *open set problem*, the test picture may correspond to an unknown landmark, and then either one geotag or none will be propagated to the test image.

5.3 Experiments and Results

In Ivanov et al. [13], we argued that our approach to user trust modeling requires a small number of images to learn models for geotag propagation. We evaluated the approach on a dataset of 1,320 images of famous landmarks (such as Bird's Nest Stadium, Sagrada Familia, Reichstag, Golden Gate Bridge, and Eiffel Tower) downloaded from Google Images, Flickr, and Wikipedia. All landmarks were split into different groups, such as castles, churches, bridges, towers/statues, stadiums, and ground structure. More details on the dataset are available in [13].

At first, we evaluated the automatic geotag propagation algorithm without including users and their mistakes in the annotation process. We showed that the object duplicated detection approach performs the best for the landmarks like castles or other buildings which have more salient regions, while landmarks that

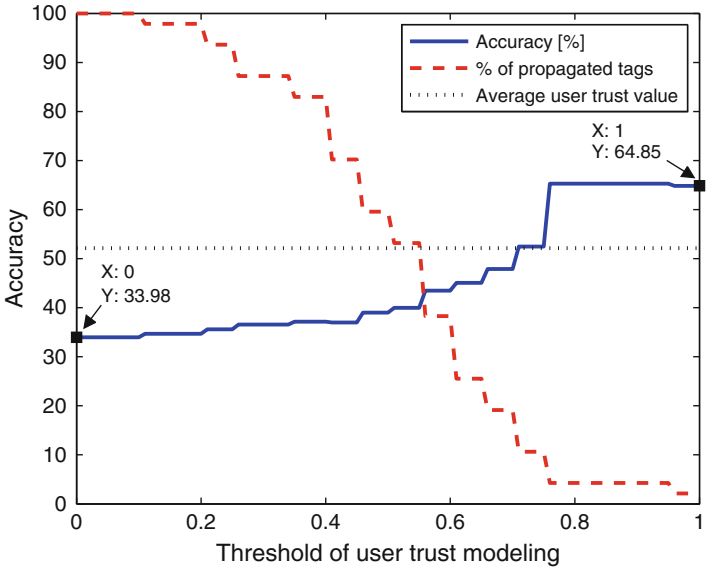


Fig. 2 The recognition rate of the geotag propagation system and the percentage of the propagated tags versus the threshold $\hat{T}$ for the user trust modeling

belong to tower and stadium groups perform worse because these landmarks do not have enough discriminative features or due to large variety of viewpoints. The accuracy measured as an average recognition rate across all landmarks is 71%. The recognition errors are solely caused by the object duplicate detection.

Then, the users are introduced in the system in order to simulate a real social network and evaluate the algorithm, which combines object duplicated detection with user trust modeling. The methodology used in this experiment is to extract a subnetwork from a large social network, in a way that every user in this subnetwork annotates every landmark in the subset of the dataset. In our experiments, each of 47 users is asked to annotate 66 images. Upon this sub-network, we build up an automatic propagation system in order to decrease the annotation time and increase the accuracy of the system. In this case, our system relies on user-provided tags, which may sometimes be spam annotations given on purpose or wrong tags given by mistake. The users are evaluated, and only tags from users whose trust model exceeds a predefined threshold are propagated to other images of the database.

Figure 2 shows the accuracy (recognition rate) of the system and the percentage of the number of propagated tags versus the threshold set for the user trust modeling. The optimal accuracy using object duplicate detection for geotag propagation is 71%. However, in this scenario, the error of the user tagging step leads to a decrease of the performance. This error is caused by wrong tags given by the users. The optimal results can be reached if we set the threshold $\hat{T}$ to a high value, but then the number of propagated tags becomes very low. On the other hand, when the

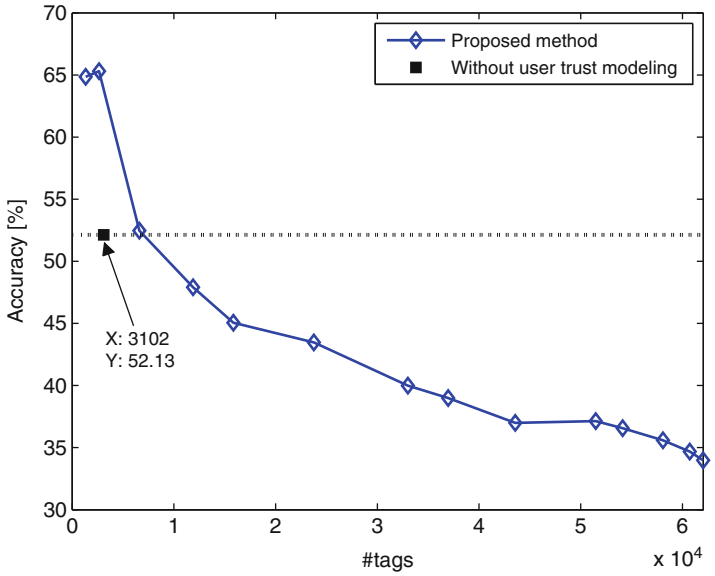


Fig. 3 The recognition rate of the geotag propagation system versus the number of the propagated tags

threshold is low, more tags are propagated. These curves could be used to determine an appropriate threshold for the proposed user trust model. The higher the threshold for the user trust model is, the more reliable the geotag propagation system is. At a threshold of 0, the accuracy of the system is equal to that without a user trust model, since all the user tags are propagated. In this case, the accuracy of the system is 34%. The figure also shows the average user trust value of 52%, which is the same as the accuracy when the users tag all the images in the dataset (1,320 images) and not only 66 images. Therefore, if we consider a large social network system where landmarks and users are selected in a way that each landmark is annotated by each user, our system shows that the best performance is achieved by choosing the most trusted user and propagating his/her annotations through the whole database of images. More precisely, in our dataset, the user annotates $1{,}320/66 = 20$ times less images, and the performance of the system (recognition rate) increases from value of 52 to 65%. As a conclusion, by using the proposed model, less manual tagging is needed, while the performance of the system increases significantly.

Figure 3 illustrates the relationship between the accuracy of the tag propagation system and the number of propagated tags by plotting them against each other. The maximum number of propagated tags can be much higher than the number of images, since several tags can be assigned to an image by different users. The black marker indicates the average tagging accuracy of the system without the user trust model and tag propagation presented in this chapter. In this case, if users tag 47×66 =3,102 photos (47 users in our experiments and each of them tags 66

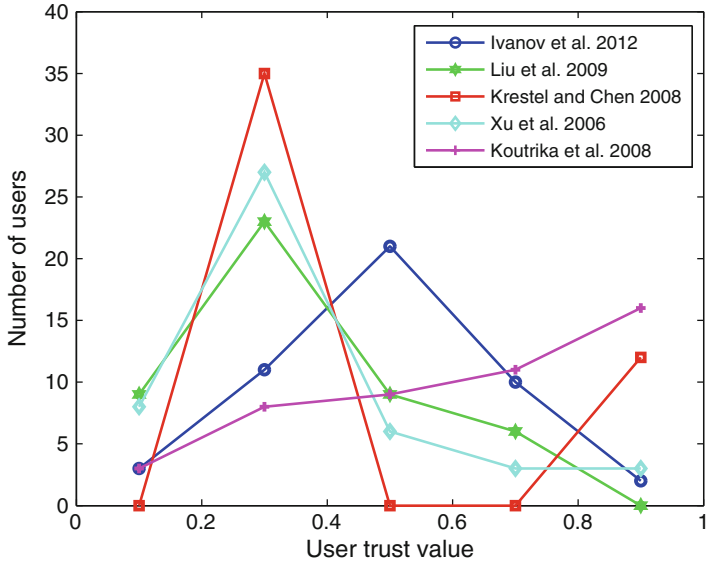


Fig. 4 The distribution of the normalized trust values for different user trust models. Different user trust models are depicted with different line colors and different markers. The results show wide variety of distributions, mainly not uniform, which leads to a conclusion that users possess different knowledge in landmarks recognition, and thus, people are more or less reliable in geotagging

images), the average accuracy of 52% can be achieved. This is equivalent to what we currently have in Flickr or Panoramio, where users simply tag photos independently, and these tags are not being propagated. However, by introducing a user trust model and tag propagation into the system, we can improve the accuracy of the system and propagate more correct tags to untagged images in the dataset. This is depicted with the left part of the blue curve, which is above the dashed line; we can still propagate more than 6,000 tags, twice more than without a trust model, from trusted users, while keeping accuracy higher than 52%.

To compare different user trust models, we analyze the distribution of their trust values given the manually assigned tags by the human participants. The values for each trust model were computed as described in Sect. 4. Obtained user trust values were normalized to 1. Then, the trust values were split into five equally distributed histogram bins with the following ranges: 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1. Figure 4 shows the distribution of the total number of users with trust values in different bins for each of the trust model. From the results, it can be noted that the distributions for most of the user trust models are not uniformed. However, the tags to our dataset assigned by the human participants can be regarded following a uniform distribution, assuming, participants unbiasedly tagged the depicted generally well-known landmarks. Therefore, useful, adequate, and practical user trust model should also reflect this uniformity in the gathered tags from participants. From Fig. 4, we can notice that only two out of five compared user

trust models, Koutrika et al. [19] and Ivanov et al. [13], demonstrate the uniformity in their assignment of the trust values to the participated users, while the rest of the models mark majority of the users as untrusted.

6 Conclusion

In this chapter, we have presented different approaches for automatic geotagging and trust modeling in social tagging systems. The problem of having trustworthy geotags of the content is important in social networks because of their increasing popularity as means of sharing interests and information. Especially photo sharing and tagging is becoming more and more popular. Among other tags, geotags in form of geographical locations provide efficient information for grouping or retrieving images. Since manual annotation of these tags is time consuming, automatic tag propagation based on visual similarity offers a very interestingly good solution.

The particular focus of this chapter is on the system for automatic geotag propagation by associating locations with distinctive landmarks and using object duplicate detection for tag propagation. The adopted graph-based approach reliably establishes the correspondence between a small set of tagged images and a large set of untagged images. Based on these correspondences and a trust value of the model derived for each user, only reliable geotags are propagated, which leads to a decrease of tagging efforts. We have analyzed the influence of wrongly annotated tags, which causes even more wrongly propagated tags in the database. By considering user trust models, the accuracy of the system could be considerably improved. In this way, the proposed user trust model can be generalized to photo-sharing platforms such as Panoramio or Flickr.

Most of the current techniques for noise and spam reduction focus only on textual tag processing and user profile analysis, while visual features of multimedia content can also provide useful information about the relevance of the content and content-tag relationship. In the future, a promising research direction would be to combine multimedia content analysis with conventional tag processing and user profile analysis.

Acknowledgements This work was supported by the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management (IM2), the Swiss National Science Foundation Grant "Multimedia Security" (number 200020-113709), and partially supported by the European Network of Excellence PetaMedia (FP7/2007-2011).

References

1. Ballard, D.H.: Generalizing the Hough transform to detect arbitrary shapes. Pattern Recogn. **13**(2), 111–122 (1981)
2. Barnett, E.: 3.4 billion photographs on Google+ in 100 days. http://www.telegraph.co.uk/technology/google/8838196/3.4-billion-photographs-on-Google-in-100-days.html (2011)

3. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)
4. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google similarity distance. IEEE Trans. Knowl. Data Eng. **19**(3), 370–383 (2007)
5. Fetterly, D., Manasse, M., Najork, M.: Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In: Proceedings of the ACM WebDB, Paris, pp. 1–6 (2004)
6. Gammeter, S., Bossard, L., Quack, T., van Gool, L.: I know what you did last summer: object level auto-annotation of holiday snaps. In: Proceedings of the ICCV, Kyoto, pp. 614–621 (2009)
7. Gyongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with TrustRank. In: Proceedings of the VLDB, Toronto, pp. 576–587 (2004)
8. Hays, J., Efros, A.A.: IM2GPS: Estimating geographic information from a single image. In: Proceedings of the IEEE CVPR, Anchorage, pp. 1–8 (2008)
9. Heymann, P., Koutrika, G., Garcia-Molina, H.: Fighting spam on social web sites: a survey of approaches and future challenges. IEEE Internet Comput. **11**(6), 36–45 (2007)
10. Hollenstein, L., Purves, R.: Exploring place through user-generated content: using Flickr to describe city cores. J. Spat. Inf. Sci. **1**(1), 21–48 (2010)
11. International Press Telecommunications Council: IPTC Photo Metadata Standard, IPTC Core 1.1 and IPTC Extension 1.1. Technical report, London (2009)
12. Ivanov, I., Vajda, P., Lee, J.S., Ebrahimi, T.: In tags we trust: trust modeling in social tagging of multimedia content. IEEE Signal Proc. Mag. **29**(2), 98–107 (2012)
13. Ivanov, I., Vajda, P., Lee, J.S., Goldmann, L., Ebrahimi, T.: Geotag propagation in social networks based on user trust model. MTAP **56**(1), 155–177 (2012)
14. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Syst. **43**(2), 618–644 (2007)
15. Kennedy, L.S., Chang, S.F., Kozintsev, I.V.: To search or to label?: predicting the performance of search-based automatic image classifiers. In: Proceedings of the ACM MIR, Santa Barbara, pp. 249–258 (2006)
16. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: Proceedings of the WWW, Beijing, pp. 297–306 (2008)
17. Kessler, S.: Mashable Infographics – Facebook Photos by the Numbers. http://www.mashable.com/2011/02/14/facebook-photo-infographic (2011)
18. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. JACM **46**(5), 604–632 (1999)
19. Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P., Garcia-Molina, H.: Combating spam in tagging systems: an evaluation. ACM TWEB **2**(4), 22:1–22:34 (2008)
20. Krause, B., Schmitz, C., Hotho, A., G., S.: The anti-social tagger: detecting spam in social bookmarking systems. In: Proceedings of the ACM AIRWeb, Beijing, pp. 61–68 (2008)
21. Krestel, R., Chen, L.: Using cooccurrence of tags and resources to identify spammers. In: Proceedings of the ECML PKDD, Antwerp, pp. 38–46 (2008)
22. Liu, K., Fang, B., Zhang, Y.: Detecting tag spam in social tagging systems with collaborative knowledge. In: Proceedings of the IEEE FSKD, Tianjin, pp. 427–431 (2009)
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
24. Luo, J., Joshi, D., Yu, J., Gallagher, A.: Geotagging in multimedia and computer vision – a survey. MTAP **51**(1), 187–211 (2011)
25. Markines, B., Cattuto, C., Menczer, F.: Social spam detection. In: Proceedings of the ACM AIRWeb, Madrid, pp. 41–48 (2009)
26. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: Proceedings of the ACM HT, Odense, pp. 31–40 (2006)
27. Marti, S., Garcia-Molina, H.: Taxonomy of trust: Categorizing P2P reputation systems. Comput. Netw. **50**(4), 472–484 (2006)
28. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proceedings of the ECCV, Copenhagen, pp. 128–142 (2002)

29. Mori, G., Malik, J.: Recognizing objects in adversarial clutter: breaking a visual CAPTCHA. In: Proceedings of the IEEE CVPR, Madison, pp. I–134–I–141 (2003)
30. Nister, D., Stewenius, H.: Robust scalable recognition with a vocabulary tree. In: Proceedings of the IEEE CVPR, New York, pp. 2161–2168 (2006)
31. Noll, M.G., Yeung, C.A., Gibbins, N., Meinel, C., Shadbolt, N.: Telling experts from spammers: expertise ranking in folksonomies. In: Proceedings of the ACM SIGIR, Boston, pp. 612–619 (2009)
32. Parr, B.: Mashable Infographics – Facebook by the Numbers. http://www.mashable.com/2011/10/21/facebook-infographic (2011)
33. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: Proceedings of the IEEE CIVR, Niagara Falls, pp. 47–56 (2008)
34. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A bayesian approach to filtering junk e-mail. Technical report, WS-98-05, Madison (1998)
35. Technical Standardization Committee on AV & IT Storage Systems and Equipment: Exchangeable image file format for digital still cameras: Exif Version 2.2. Technical report, JEITA CP-3451, Tokyo (2002)
36. Thomason, A.: Blog spam: A review. In: Proceedings of the CEAS, Mountain View (2007)
37. Vajda, P., Goldmann, L., Ebrahimi, T.: Analysis of the limits of graph-based object duplicate detection. In: Proceedings of the Symposium on Multimedia, San Diego, pp. 600–605 (2009)
38. Vajda, P., Ivanov, I., Goldmann, L., Lee, J.S., Ebrahimi, T.: Robust duplicate detection of 2D and 3D objects. IJMDEM **1**(3), 19–40 (2010)
39. von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: CAPTCHA: using hard AI problems for security. In: Proceedings of the Eurocrypt, Warsaw, pp. 294–311 (2003)
40. von Ahn, L., Maurer, B., Mcmillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-based character recognition via web security measures. Science **321**(5895), 1465–1468 (2008)
41. Whitby, A., Jøsang, A., Indulska, J.: Filtering out unfair ratings in bayesian reputation systems. In: Proceedings of the IEEE AAMAS, New York, pp. 106–117 (2004)
42. Wikimedia Foundation Inc.: Wikipedia–Flickr. http://en.wikipedia.org/wiki/Flickr (2012)
43. Wu, C.T., Cheng, K.T., Zhu, Q., Wu, Y.L.: Using visual features for anti-spam filtering. In: Proceedings of the IEEE ICIP, Genoa, vol. 3, pp. III – 509–512 (2005)
44. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: collaborative tag suggestions. In: Proceedings of the ACM WWW, Edinburgh (2006)
45. Yahoo! Inc.: Flickr – Tags. http://www.flickr.com/help/tags (2011)
46. Yang, Y., Sun, Y.L., Kay, S., Yang, Q.: Defending online reputation systems against collaborative unfair raters through signal modeling and trust. In: Proceedings of the ACM SAC, Honolulu, pp. 1308–1315 (2009)
47. Zheng, Y.T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T., Neven, H.: Tour the World: building a web-scale landmark recognition engine. In: Proceedings of the IEEE CVPR, Miami, pp. 1085–1092 (2009)

Context-Aware Content Adaptation for Personalised Social Media Access

Hemantha Kodikara Arachchi and Safak Dogan

Abstract Accessing, sharing, and delivering social media involve online user-to-user social interactions. These interactions can be characterised by one-to-many or many-to-many communications established amongst numerous users. This in turn makes the whole environment a very heterogeneous one, comprising diverse user devices, fixed and mobile access network technologies, content representation formats, user needs and preferences, and media usage and consumption environment characteristics. Owing to this very heterogeneous nature of social media access, personalised access to the social media is a significant challenge for maximising the user experience and satisfaction. This chapter presents context-aware social media content adaptation as the key technology to address this challenge. It introduces the importance of context awareness in personalised social media access, and how it can be coupled with content adaptation and adaptation decision-taking mechanisms to provide a complete solution for both the technical and non-technical (i.e. social) challenges faced. Detailed discussions focus on describing various adaptation and decision-taking types and operations while also pointing at a number of open issues for future research, so as to address those highlighted challenges for realising true personalised social media access environments.

1 Introduction

In the context of networked digital media delivery, personalised access to multimedia content embodies all of the necessary strategies that are needed for effectively responding to the ever-changing and increasing preferences of users of media

H. Kodikara Arachchi • S. Dogan (✉)
I – Lab Multimedia Communications Research, University of Surrey, Guildford, UK
e-mail: S.Dogan@surrey.ac.uk

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_14,

services. These strategies encompass the development of several key technologies, most notably content adaptation, for providing the users with what they would like to receive, watch, see, hear, and read, in the way that they prefer.

In an era, in which content has fast become ubiquitous for consumption, personalised access to content and content services has also become extremely demanding with the availability of many media networking and access technologies as well as fancy and affordable user devices. Challenges arise due to both technical and social reasons. Technically, the vast variety of access network technologies, content types and formats, and terminals have led to tremendously heterogeneous digital media distribution and consumption environments. Social challenges have been created as a result of users' various demands, preferences, and expectations from the services that are provided. In turn, both render access to content very challenging. Particularly, with the proliferation of the recent social networking and content sharing facilities, these challenges have been exacerbated, as both the number and demands of users have significantly increased and varied.

Nowadays, there seem to be a large number of social networks and thus networked communities of users active over the Internet for every need or reason, which promote communication of people and sharing of content with each other, regardless of their diverse age groups, backgrounds, locations, special requirements, preferences, etc. The growth in the amount of user interest in social networks has made content sharing a part of the daily life. Consequently, there is an increasing tendency of sharing more and more media content without paying much attention to the diverse content types produced and shared, the user devices used to capture, generate, and consume those various content types, and the communication links that are used as the access networking backbone with their respective limitations.

In light of the above discussions, this chapter proposes to provide in-depth knowledge on how context-aware content adaptation, as a key technology, is able to constitute an efficient strategy to address those aforementioned challenges for personalised social media access. Therefore, this chapter first provides a close look at the context, which drives the content adaptation. Often, context imposes constraints and requirements for the information to be delivered to a given user or a set of users. For instance, available network bandwidth limits the data rate of a video delivered to a user. Since maintaining the optimum user perception towards the consumed media is a key requirement, accurate sensing and understanding of the contextual information plays a significant role in efficient social media sharing and delivery applications. Hence, context sensing and classification technologies are introduced, and context in social media consumption environments is analysed in detail.

The heart of a content adaptation system is the signal processing techniques, which perform the actual adaptation. The signal processing techniques used to make the input media suitable for a content consumption environment, which is described by the contextual information, are investigated in this chapter. Typically, adaptation mechanisms can be located anywhere in the end-to-end media delivery chain. In terms of their functionality, these mechanisms can perform a variety of content adaptation operations and can be classified in terms of level of adaptation

applied on the content. Some of these adaptation mechanisms perform signal level transforms, such as transcoding, transrating, and transmoding. Others may apply semantic level transforms, such as cropping and highlighting of audiovisual attention areas, while another group of adaptation mechanisms transforms the input signal at structural level, such as summarising and mosaicing a lengthy video stream. Besides the abovementioned classification, the adaptation mechanisms can further be categorised in terms of whether they are executed offline or online, and also depending on the purpose of adaptation. These classification strategies are discussed in detail in this chapter. Subsequently, in-depth discussion of the signal processing techniques and tools, which are most applicable for social media applications, are also presented.

Once the contextual information is obtained, choosing the most efficient and effective signal processing technique from a large number of available techniques for adaptation is a challenging issue. This is the exact role of an adaptation decision-taking algorithm. Due to the cognitive nature of its functionality, adaptation decision taking is considered as the brain of a content adaptation system. This chapter presents a comprehensive discussion on the adaptation decision-taking algorithms. First, context reasoning, which involves processing of raw contextual information to unearth constraints and requirements that solicit the need for adapting the media, is explored. Subsequently, two adaptation decision-taking approaches, namely knowledge-based and utility-based adaptation decision taking, are elaborated.

Social media adaptation is not always as straightforward as it is anticipated in a user-to-user communication scenario. There are many technological and non-technological (i.e. social) issues that may inhibit or discourage performing content adaptation. Thus, last but not least, this chapter elaborates on those challenges with necessary pointers to open issues and potential directions for addressing them efficiently. Finally, this chapter is summarised and concluded.

1.1 Ambient Intelligence for Personalised Social Media Access

In the beginning of the pervasive social networking age, like many earlier multimedia services, social media access, sharing, and distribution were also limited to mostly home and office environments with devices connected to fixed networks supporting broadband links. This made the provision of such services rather easy and manageable, as dedicated links and terminals were targeted while providing social networking and content sharing with various online communities. This also meant that the intelligence for providing device and/or network centric media content customisation capabilities during online social activities mainly lay centrally within the networking infrastructures for supporting those services with acceptable quality of service (QoS) levels. Nevertheless, with changing times and fast proliferation of ubiquitous computing and communication systems, a great number (if not all) of the media services that we used to receive only in our homes have migrated rapidly

to the mobile environments. Without doubt, this has led to a wider uptake of social media access and content sharing with even wider communities, who were not fully foreseen prior to such migration.

Thus, not only has this resulted in the need to cater for a wide range of terminals (both fixed and mobile) and networks (both wired and wireless), as was the main focus of the past media services, but also accelerated the research and technology development activities for placing users themselves in the spotlight as the real consumers of media content at last. In turn, personalisation aspects have gained significant weight for accessing and sharing the social media while paving the path towards enhancing the quality of experience (QoE) of users of those services. Evidently, this has also presented an alternative to the centralised approach through realising distributed intelligence within the networked digital media delivery chain to provide personalised social media access.

Distributed intelligence starts from the surroundings of the user, which can influence the access and delivery of social media content and its successful sharing, consumption, and wide-scale acceptance by all parties that take part in an online social interaction. This concept is widely referred to as ambient intelligence in the computing literature [1, 2], where electronic environments that are sensitive and responsive to the presence of people (in our case, social media users) are addressed. This is a natural evolution of ubiquitous computing and communication systems that are enabled by the smart equipment and devices with sensory elements as well as intelligent networks and networking peripherals. The intelligence is embedded within the systems that are context aware, where personalised and adaptive services can be supported with an anticipated level of user satisfaction.

In personalised access to social media, ambient intelligence thus plays a very significant role on the successful user centric delivery and consumption of content towards enhancing the user QoE, which is easily affected by the user preferences. The core of ambient intelligence relies on the sensed context in the environment. In fixed environments, the ambient effects can be predicted (if not measured) accurately, as these effects generally present measurable and predetermined variations. On the contrary, mobile environments have an ever-changing nature, which demands continuous monitoring and sensing of the changing context. To exemplify this argument simply, one can consider the differences in access to a social networking and content sharing platform using public and private fixed user devices. As two different devices are fixed to either public or private context, based on where they are located, a dedicated protocol can easily be followed for access granting and content sharing in both cases. However, while using a mobile smart device, the boundaries for public and private access may become blurry when the user location frequently changes. Therefore, this situation calls for regular context updates to sense and follow the variations, so as to be responsive to the new conditions accordingly.

Interaction with the ambient intelligence within the environment can be achieved both automatically, where devices and computing elements react autonomously, and semiautomatically, where user interacts with the execution of the protocols and

algorithms through various user interfaces. This is particularly important, where user preferences can be provided as a live and accurate feedback to a content adaptation system that serves the user for personalised social media access and consumption.

Further intelligence that supports personalised access to social media content lies in the architectures of servers, access networks, and terminals in an end-to-end media access, sharing, and delivery chain. Such intelligence is provided by employing necessary media content adaptation capabilities at those different points within the delivery chain, which will be further introduced and discussed in the subsequent sections.

1.2 Need for Content Adaptation in Social Media Access

In a nutshell in the context of this chapter, the term social media refers to the rich multimedia content that is exchanged, shared, consumed, and enjoyed over the Internet socially. Since access, sharing, and delivery of social media involve an online social interaction, they also result in connecting a very wide range of users, who may or may not know each other physically, in one-to-many or many-to-many communication modes. Such diversity means a heterogeneous media access, sharing, and delivery environment, which comprises different user devices, and access network technologies, as well as various content representation formats, user preferences and 'intentions', and media usage and consumption environment characteristics.

The success and wide proliferation of the Internet and mobile communication systems together have motivated the development of various enhanced-capacity fixed and wireless networking technologies (e.g. 3G, LTE, WLAN, WiMAX, broadband Internet) [3]. This has encouraged growth in the variety of formats with which media content can be represented for various media applications and services, including social media access and sharing. These services supported by such networks helped to foster the vision of being connected at anywhere, anytime, and with any device particularly for pervasive social media applications. However, the coexistence of the different networking infrastructures and services has also led to an increased heterogeneity of compressed media communication/delivery systems and application scenarios, in which a wide range of user terminals with various capabilities access rich media content over a multitude of access networks with different characteristics.

In a typical social media access scenario, content may be captured, preprocessed, and prepared for exchanging and sharing using a particular format, which may be specific to a capturing device (e.g. a high-quality handy-cam). In addition, the user, who shares this content with his/her peers in a social networking platform, may wish to draw the attention to specific parts of the content through either manual or automatic editing and metadata tagging. The device is connected to an

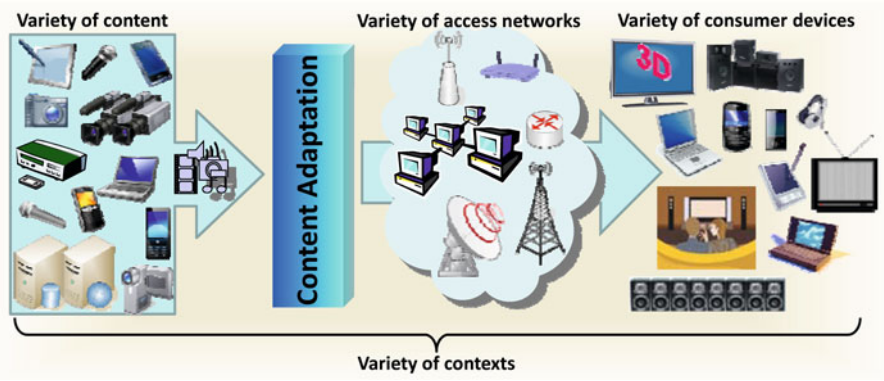


Fig. 1 The big picture of the UMA concept and need for content adaptation

access network or a bundle of networks directly or through another networked device (e.g. a multimedia tablet or desktop PC) for transmitting the information with their respective requirements, such as bandwidth limitations, delay, jitter, and error performances, etc. All of these pose non-trivial challenges for sharing this social media content with the intended level of QoE, as the other users may or may not be able to match their terminal capabilities, access network characteristics, or content formats for accurate processing, rendering, displaying, and presentation. Furthermore, their preferences may not adequately mirror those of the content creator's/producer's, as for instance they may wish to focus their attention to other portions of the content where it is more attractive to their taste.

As social media is an intensely user centric concept, it is also important to cater for a mechanism to provide user feedback into the content access and sharing platform. This necessitates the consideration of various modes of user interaction, again which may not have similarities with the modes used for preparing the content in the first instance, may not comply with the specifications set by the content producer or may have specific requirements due to content usage environment. The environment, in which the content is acquired and consumed, may also impose changes in the way the content is rendered and presented in contrast to the originally captured and shared content, as previously discussed. All of the aforementioned mismatches may thus hinder a guaranteed or acceptable level of QoS as well as QoE of the social media users.

The mismatches between the media content properties and several network and/or device-centric features, usage environment characteristics, as well as diverse user preferences call for efficient content delivery and sharing systems featuring effective context-aware content adaptation mechanisms for personalised social media access. In general, this concept has been widely addressed in literature with the theme of universal multimedia access (UMA) [4, 5]. The big picture addressing the need for content adaptation in the UMA concept is depicted in Fig. 1.

1.3 Core Essentials of Content Adaptation

Content adaptation refers to the set of mechanisms and technologies to provide effective solutions for meeting various mismatches between media content properties, diverse access network and device centric features, user preferences, and usage environment characteristics in a social media access and content sharing scenario. The overall aim here is to satisfy a set of technology imposed constraints for seamless and personalised social media content access, sharing, and distribution while improving user experience in terms of perceived media quality and satisfaction from the delivered social media services. Thus, content adaptation is the process of transforming an input media stream to an output media or augmented media representation format by manipulating the input media signal at different levels (signal, structural, or semantic) in order to meet diverse resource constraints and user preferences while optimising the overall utility and usability of the media content.

Three core components determine the operational success of content adaptation: context awareness, actual media adaptation, and adaptation decision taking. An adaptation system should be aware of the contextual information that characterises the social media access, sharing, delivery, and consumption environment. The features of the user device that is used to access the social media content, the characteristics of the access network that the device is connected to, user's location, content access time, the ambient conditions of the environment where content is accessed or shared, etc. are a few of the key contextual drivers that influence the way the content can be used. Through accurate sensing and monitoring these contextual elements, the selection and working of content adaptation methods can be made responsive to the operational specifications of the content access and sharing platform adequately. Actual media adaptation involves applying a suite of signal processing techniques on the input media signals, so as to minimise the effects of the aforementioned mismatches between content properties and usage environment conditions with the help of contextual information that is collected and inferred. All of this process is overseen by an adaptation decision-taking mechanism, which constitutes the intelligent component of a content adaptation system. Given a context, there can be a number of media adaptation methods that can be applied, which achieve similar results. Nevertheless, the decision to select and apply the most effective method is the main responsibility of the adaptation decision-taking component. The core essentials for performing the adaptation of social media for personalised access are discussed in the following sections of this chapter in detail.

1.4 Middleware Architecture for Social Media Adaptation

Distributed intelligence for handling content processing to provide personalised social media access within the networked digital media delivery chain requires a distributed approach to social media adaptation. Social networks are the prime

examples of person-to-person or in other words user-to-user communication and media sharing platforms. In such a networking platform, users can act both as content producers and consumers, where the focus is on assistive technologies that are aimed to adapt to the heterogeneous usage environments rather than users trying to adapt to technology specifications. Thus, once content is shared with peers in online social communities, it is not the main concern of the content producer to ensure the shared social media is accessed and delivered to all parties at their preferred levels of user experience or in compliance with their service level requirements adequately.

Indeed, in most cases, content (captured and prepared by professionals or lay users alike) is uploaded to a content sharing platform and expected that it will be accessed and downloaded by every user with similar experience successfully. Nevertheless, it is not always the case in reality, as a high-definition (HD) resolution media content uploaded using a high-broadband Internet connection for sharing with others cannot be easily accessed and downloaded by users in full resolution or fidelity over a mobile wireless network (e.g. 3G) with limited bandwidth using a smart phone with a display size at a fraction of the HD resolution and low-capacity processing power.

Thus, social media adaptation is needed, so that all users in an online social circle can access the same content at varying fidelities that suit their needs and preferences as well as their technology requirements. Adaptation is an application layer middleware operation, which sits between content coding and decoding acting as a processing tool. The strategic positioning of the adaptation and decision mechanisms in a social media delivery platform is the determining factor for the success and effectiveness of content adaptation for personalised social media access. Three middleware architectures have been widely reported in literature based on this strategic positioning: server, network, and terminal-side architectures, as illustrated in Fig. 2. These architectures have their respective advantages and disadvantages associated to the factors (e.g. content processing load, transmission bandwidth need) that affect adaptation operations that can be performed [6–8].

In server-side architecture, content is adapted at the main content server level, where a number of copies of the content at different resolutions and/or qualities can be prepared for distribution when needed [8, 9]. Similarly, content specific preprocessing, scalable layering, error resilience addition, or multiple description-based encoding can be applied at the server for targeting heterogeneous communication and content delivery networks. The server is the sole responsible for gathering the user preferences and requests, device capabilities, and the available bandwidth on the network, performing adaptation, and sending the adapted content to the user terminal in this architecture. The main advantage is that it allows both offline and online media adaptation for social media access, while also providing the author of the social media content with tools for better controlling the operation and presentation of the shared media. The main disadvantage is the computational load on the server, particularly for offline adaptations.

Content can also be adapted using a network-side middleware architecture, which can be located within the social media access, sharing, and delivery platform. This

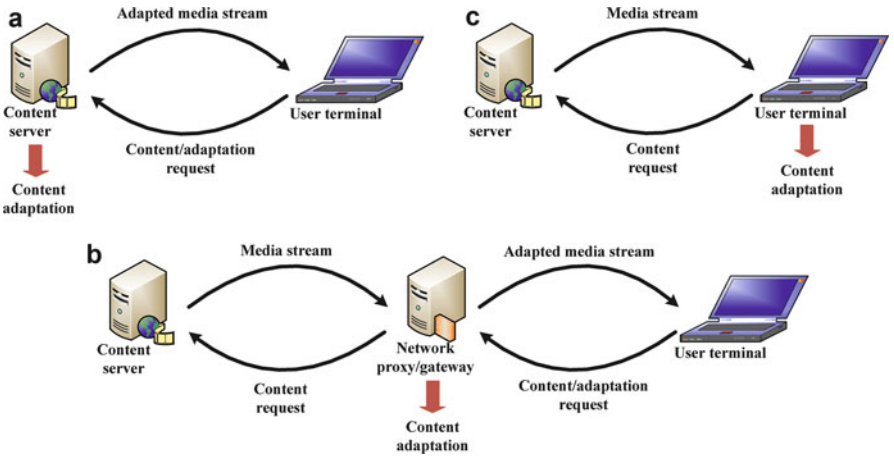


Fig. 2 Location/architecture of adaptation. (**a**) Server side middleware architecture, (**b**) Network side middleware architecture, (**c**) Terminal side middleware architecture

point can be referred to as a gateway or proxy node on the edges of heterogeneous networks, so as to match mainly network centric diversities that significantly influence both the QoS and QoE of the received media at the user terminal end. Device capabilities can also be addressed in this kind of an adaptation mechanism. The adaptation requests are sent to the relevant proxy or gateway rather than all the way to the content server. When the proxy/gateway receives the request, it acquires the adequate context from the user and his/her terminal, and it connects to the server on behalf of the user/terminal to gather the adequate content information. Then, the proxy/gateway decides on the adaptation operation, executes it, and sends the adapted media to the user. Thus, at such network points, the specifics of the social media content are mainly adapted based on the characteristics of the next host transmission medium for service level requirements, such as access bandwidth matching and error robustness insertion, as well as spatio-temporal scalability of the content can be provided for addressing various terminals in the usage environment. This type of architecture is advantageous in terms of bandwidth, as it takes the advantage of the bandwidth between the server and proxy or gateway, which is relatively high in most cases [8, 10]. The network-side architecture alleviates the disadvantages of the server and terminal-side middleware architectures [7].

Media adaptation can also be performed at the terminal end depending on the device capabilities in the terminal-side architecture. This architecture has a disadvantage that it can be sometimes impossible to perform adaptation due to the limited processing power of user devices, such as PDAs and mobile phones [8, 11]. Moreover, sending high-quality media content to the user terminal and leaving the adaptation to the device may cause high network traffic all the way throughout the social media access, sharing, and delivery chain [7]. However, this kind of architecture also has its advantages, such that adaptation on the user terminal

targets accurate rendering and presentation of the content, while also providing a valuable means for better personalised content access taking into account the user preferences. Preferences can be easily input to the content adaptation system through simple user interaction modes supported on the user terminal.

In a typical media access, sharing, and delivery scenario, it is not necessary having to choose and abide by one point of adaptation, as performing distributed adaptation at a combination of various adaptation locations may result in enhanced quality media content at the receiving end. Media adaptation can be applied in a co-operative fashion, such that at all three strategic adaptation points, a set of different requirements and constraints can be addressed by employing accurately selected media adaptation method and middleware architecture. This can be achieved by applying the adaptation decision-taking mechanism also in a distributed fashion to closely monitor, supervise, and drive the adaptation operations to be performed at different positions in line with the contextual information collected at those respective locations. In this way, distributed media adaptation architecture can be realised, which aids in distributing and balancing the adaptation load across servers, network proxies/gateways, and terminals adequately [7, 12].

2 Context Awareness in Media Networking

According to [13], a system is context aware if it uses context to provide relevant information and/or services to the user. Context awareness refers to the state of being aware of the context in a sensed environment [14]. As previously introduced, it is one of the core essentials for performing media adaptation to address context-related constraints and requirements in a digital media access, sharing, and delivery environment successfully and effectively. Context-aware systems have thus ability to sense, infer, and react to contextual information, such as network conditions, terminal capabilities, natural surrounding environment characteristics, and user preferences, by adapting their behaviour dynamically [15]. The use of contextual information to facilitate decision taking on how to adapt media content based on that information is key for implementing meaningful media adaptation operations that meet users' expectations while also satisfying their usage environment constraints.

2.1 Definition of Context

If a user and application are in interaction, any information sensed at the time of this interaction can be identified as context [16, 17]. Context has numerous definitions in literature depending on the context-aware application it is used in [18]. The most generic definition is given in [19], where context has been defined as the information to characterise the situation of person, place, and object, which are in interaction with each other.

Thus, contextual information can be any kind of information that characterises or provides additional information regarding any feature or condition of a complete content delivery and consumption environment. Use of context is particularly important for deciding adequate adaptation parameters that are needed in adaptation operations. Contextual information can be exploited in content adaptation systems as metadata, which is required for driving the adaptation decision-taking algorithms to determine adequate content adaptation methods in response to the constraints imposed by the context of usage.

Contextual metadata used in the decision algorithms describes the characteristics and conditions of the context of usage for networks (e.g. bandwidth availability, error rates, jitter), user terminals (e.g. screen size, CPU, codec capability), natural surrounding environment (e.g. ambient illumination conditions, noise level), and users (e.g. disabilities such as hard of hearing or colour perception defects, preferences such as language, summary of images versus video, a specific view in multi-view content).

2.2 Context Gathering and Classification

Context-aware systems need to acquire the contextual information, so as to process and reason about this information to further formulate concepts and take decisions when and how best to react to it. Gathering contextual information takes place in three steps: sensing low-level context, inferring high-level context, and sensing changes in the context. The first step relates to the generation and representation of basic contextual information, as it can be directly generated by a software or hardware application. In most cases, low-level context can thus be acquired through the use of physical tools, such as sensors placed on user devices or in the natural environment surrounding the user. Automatically generated data from software modules also allows collecting necessary low-level contextual information concerning terminal capabilities and network characteristics. On the other hand, information addressing the user preferences and his/her natural surrounding environment requires the use of dedicated hardware as well as occasional (or sometimes even regular) manual intervention from the user. Dedicated hardware includes visual and aural sensors, such as video cameras and microphones, as well as other user terminal interfaces. User preferences can be implicitly created based on usage history, but most often they require the explicit inputting by the user, at least during system start-up.

Based on the basic (or low-level) contextual information, applications may infer higher-level concepts in the second step. For instance, emotions of a user or his/her physical state can clearly affect his/her preferences about a media content or service, particularly in social media applications. This information can be inferred by analysing low-level contextual information obtained by imaging, sound, or medical sensors. Similarly, location data may be collected as low-level information using geographical coordinates during the context gathering stage. When this data is

analysed for inferring the associated high-level context, the type of physical space the user is in can be determined, such as indoors, outdoors, train station, bus stop, sports stadium, etc.

Once low-level context is acquired, and high-level context has been inferred from it, the third step is to monitor and sense the changes that occur in the contextual information. Therefore, the acquisition of context, at least of some types of contextual information, should be a continuous process, regardless of the fact that it is a periodic process or not. In this way, changes in the context can be perceived by the context-aware application. Evidently, this necessitates the reasoning about the basic contextual information to be a continuous process, which is conducted whenever changes in the basic contextual information are detected.

Typically, context can be classified into four classes in a context-aware system [18]:

- Resources context
  - Description of terminals in terms of hardware and software (available processor, battery, screen size, operating system, codecs, etc.)
  - Description of networks (available bandwidth, maximum capacity, bit errors, packet losses, etc.)
  - Description of multimedia servers (maximum number of users, maximum throughput, etc.)
  - Description of transcoding engines in terms of their hardware and software (input/output formats allowed, bit rate range supported, etc.)
- User context
  - Description of user's general characteristics (gender, age, nationality, etc.)
  - Description of user's preferences in terms of content consumed and his/her interests (type or language of the media preferred, action movies versus comedy, etc.)
  - Description of user's emotions (anxious, happy, sad, etc.)
  - Description of user's status (walking, talking, etc.)
  - Description of history/log of actions performed by the user
- Physical context
  - Description of environmental information surrounding the user (location, temperature, ambient illumination conditions, sound/noise levels, etc.)
- Time context
  - Description of time at which the context is measured, variations in the context have occurred or scheduling of future events (time of the day, frequency, etc.)

Besides, there are a few other factors that affect the contextual information that is gathered, which can be summarised as follows:

- Accuracy or level of confidence of the contextual information
- Validity period

  - Contextual information may be static, thus having an unlimited validity period (e.g. general characteristics of a user are static and do not require any additional information to be used), or it may be dynamic and valid for only a given period of time (for instance, user's emotions and conditions of the natural environment, such as illumination or background noise, are dynamic).

- Dependencies on other types of contextual information
  - Reasoning about one type of contextual information may depend on other types of contextual information.

2.3 Use of Context in Social Media Adaptation

In a typical end-to-end social media access, sharing, and delivery platform, contextual information can be obtained by the following available context providers:

- Users: via their terminal devices
- Network operators: via network equipment, including network probes
- Content providers: via streaming servers, encoders, databases, etc.
- Terminal equipment vendors: via a variety of supported device specific sensors
- Natural surrounding environment: via user terminals and other environmental sensors

Contextual information, once acquired with the help of the aforementioned context providers, is used to exert constraints and requirements on the social media that is accessed, shared, and delivered. Context use in media adaptation is illustrated in Fig. 3. Context is usually fed into the media adaptation mechanisms in the decision-taking phase in the form of Universal Environment Description (UED) and Universal Constraints Description (UCD) tools, as described by Part-7 of the MPEG-21 Standard on Digital Item Adaptation [20–22]. UEDs are used for describing the environment, in which the media content is transmitted, stored, used, and consumed. The necessary information for providing the UEDs comes from contextual data. They are divided into four main groups: user characteristics, terminal capabilities, network characteristics, and natural (usage) environment conditions. Together with UEDs, UCDs are also used in adaptation decision taking. They provide the description of limitations and optimisation constraints on possible adaptation operations. This tool enables users to further constrain the usage of media content; for instance, the resolution of the rendering device can be constrained to satisfy needs of the application or device display size using the UCDs. Social media providers may also restrict the way their content is adapted and used, such as imposing a minimum level of quality or user profile during consumption by help of this tool, which provides a control over the adaptation operations performed upon the content. Constraints that are described by UCDs can be derived from UEDs and can also be used to further constrain them.

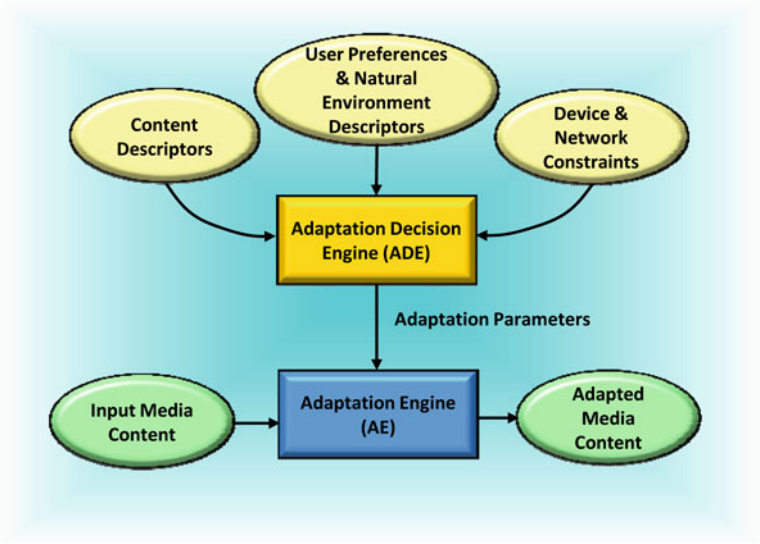


Fig. 3 Context use in media content adaptation

Processing and analysing a set of low-level context elements gathered with the help of context providers allow inferring high-level contextual information in a social media access, sharing, and delivery scenario. For example, natural environment-related contextual data may lead to the understanding of whether the social media is accessed and consumed by the user in a private or public location (e.g. at home or work place). In turn, this knowledge helps the social media adaptation system make necessary personalised adaptations according to the situation either by allowing the required adaptation or offering an alternative solution. A mobile media consumption environment inferred through relevant context providers results in adaptation of social media content shared by the content provider to the mobile and wireless context. This may include network bandwidth and device display size matching, user preferences-based content summarisation where full version of content cannot be processed or viewed on a smaller user terminal, etc.

Together with all of the context elements, content-related metadata also enables taking better decisions for personalised social media access. This is particularly desirable in social networking applications, as content is usually shared with all online communities without explicit knowledge of who is actually receiving the content. It is very common that most of the social media content shared in social networks is not appropriate for all users. Therefore, based on the content metadata, either added by the content provider manually or extracted and embedded within the media stream automatically, social media content adaptation can be performed to suit the preferences and needs of a diverse range of users, who may have different backgrounds, requirements, age ranges, understandings of the available content, etc.

3 Signal Processing Techniques for Content Adaptation

Any adaptation operation that transforms input media into a target form involves signal processing. Often, this is the most computationally expensive stage of the entire media adaptation operation. Signal processing tools can be deployed at any point in an end-to-end social media access, sharing, and delivery chain. Some of the tools intercept media even before they leave the server or producer while other operations are performed at the user terminal or in the communication network, as was illustrated in Fig. 2. Figure 4 shows a few of the example signal processing technologies that can be deployed to perform social media adaptation in the end-to-end communication chain.

Signal processing mechanisms available for social media personalisation can be classified in a number of ways. Some of these mechanisms perform signal level transforms. This type of transforms preserves key information enclosed in the source media as much as possible while transforming the media into the target format. However, the quality of the media content may degrade during the transform. Transcoding, transrating, and transmoding are a few examples of signal level adaptations [23, 24]. This group of signal processing mechanisms is the most common adaptation operations that can be used for personalising social media. Another type of signal processing mechanisms performs semantic level transforms, such as cropping and highlighting of audiovisual attention areas [25]. The aim of this class of adaptations is to improve the users' ability to grasp the most important information. For example, when a large photograph is viewed on a small mobile display, it is difficult to recognise people posing in the picture. However, when each person is zoomed in across the entire display, it is easier to recognise

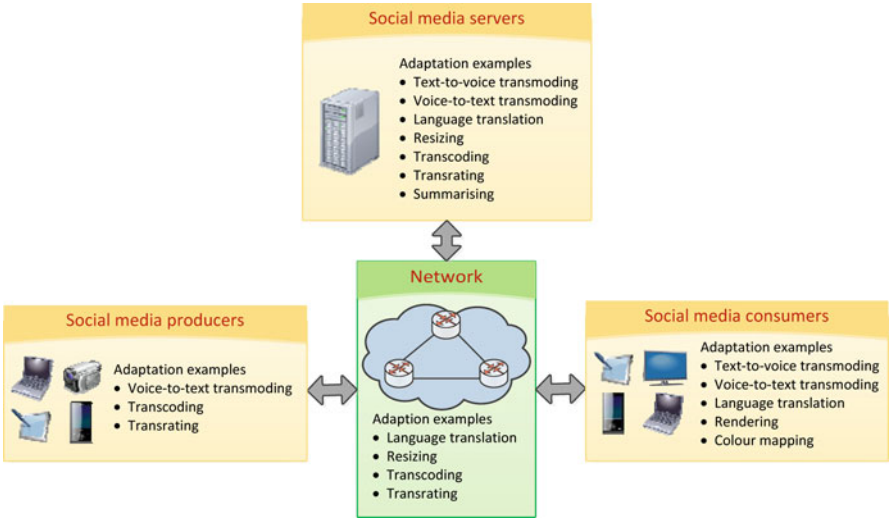


Fig. 4 Typical social media adaptation examples

them [26]. Cropping operation can be used for achieving this. The last group of signal processing mechanisms transforms the input signal at structural level, such as summarising and mosaicing a lengthy video stream [27, 28]. These operations attempt to provide a concise representation of large scale content in the forms of still images, shorter video segments, graphical representations, or textual descriptors that are more appropriate for a busy social media consumer. Moreover, this group of processing mechanisms also provides an easy means for deciding whether to consume the lengthy version of the social media content.

Besides the abovementioned classification, the adaptation mechanisms can also be categorised in terms of whether they are executed offline or online, and also depending on the purpose of adaptation. Offline adaptation is performed well before the content is consumed, often when the media is created. The objective is to keep as many adapted versions as necessary, so that users are given the option to choose the most adequate version for their preferences and needs. This approach is appropriate when online adaption is not feasible. However, the storage requirement increases with the proliferation of potential usage scenarios, each of which may have unique requirements. In contrast, the online adaptation is performed when social media is requested for sharing or consuming. The adaptation may be performed on real-time or non-real-time basis depending on the application. For instance, for streaming, real-time adaptation is necessary to enable pause-free continuous playback of the media. The disadvantage of online adaptation is the massive demand for computational resources for performing complex operations that need to be accomplished within a very limited time interval to guarantee the uninterrupted media playback. However, for on-demand applications that operate on the basis that users wait until the adapted media is ready before consuming it, the processing time delay pressure is less intensive.

Adaptation can also be classified in terms of the purpose of adaptation, such as content selection, content processing, and presentation adaptations are a few purpose-driven adaptation classes. Some of the adaptations are performed to assist users to select certain content. Summarising and mosaicing are classic examples. In this way, users can be given more informative ways of choosing the right content that they want to share or consume. Content processing may also trigger adaptations. For instance, a spatial resolution of a video needs to be reduced to add a decorative border around it. Similarly, creating a picture-in-picture effect not only requires spatial subsampling of one of the two videos (i.e. the secondary video) but also needs temporal resolution matching of the main and secondary videos if they are of different frame rates [29]. Adaptation for media presentation is required to match media characteristics to presentation devices or medium. For example, the layouts of web pages are often modified to better suit small displays such as those fitted on smart phones. Similarly, 5.1 multichannel audio needs to be re-rendered to match its features to a stereo sound reproduction system.

In light of the above discussion, the following subsections briefly describe the most frequently used signal processing mechanisms for realising personalised social media access.

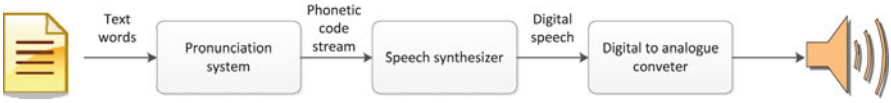


Fig. 5 Text-to-speech synthesiser

3.1 Text-to-Speech and Speech-to-Text Transmoding

Traditionally, text-to-speech and speech-to-text transmoding have been used as human-machine interaction (HCI) tools for people with audiovisual impairments. Even though this tool was a life saver for those who were unable to read or hear, until recently they were not welcomed by the general public due to poor quality of experience. The robotic-sounding synthesised speech generated by text-to-speech transmoding tools were dull and often hard to understand. Similarly, speech-to-text conversion results were very unreliable in most practical usage environments. However, this attitude has changed in time due to rapid improvements in relevant algorithms. At the same time, powerful hardware platforms that can offer sufficient computational resources for these sophisticated algorithms to operate effortlessly are becoming ubiquitous. Therefore, deploying such algorithms on mobile devices, on which these technologies are most frequently used, is commercially viable nowadays. As a result, both text-to-speech and speech-to-text tools are becoming 'must have' tools for the social media users on the go. For example, an application that reads out incoming text messages is a standard tool in recent popular smart phones. Likewise, there is an increasing trend to listen to, rather than reading, the text posted on social networking sites. At the same time, those who are not fond of fiddly touch controls and keyboards on smart phones will soon be able to enjoy all-voice-based interactivity.

Even though both speech-to-text and text-to-speech transmoding are straightforward tasks for a human being, for machines, they are unimaginably hard computational operations that involve a number of sophisticated signal processing techniques. The text-to-speech transmoding is generally achieved using a synthesiser that mimics the human vocal system [30]. A simplified block diagram of a typical text-to-speech synthesiser is shown in Fig. 5. The input text stream is analysed by the pronunciation system, which produces a series of phonetic codes. These phonetic codes consist of phoneme codes and prosody indicia. Speech synthesiser produces the digital version of the speech signal using these codes. Finally, the digital to analogue converter converts the digital speech signal to its analogue form, which can be played by the speaker.

In contrast, speech recognition maps a continuous speech signal into a sequence of recognised words [31]. Modern speech recognition systems often use hidden Markov models (HMM) and the Viterbi algorithm to recognise words from speech signals. A simplified block diagram of an automatic speech recognition system is shown in Fig. 6. The feature extraction module, first, preprocesses the input audio

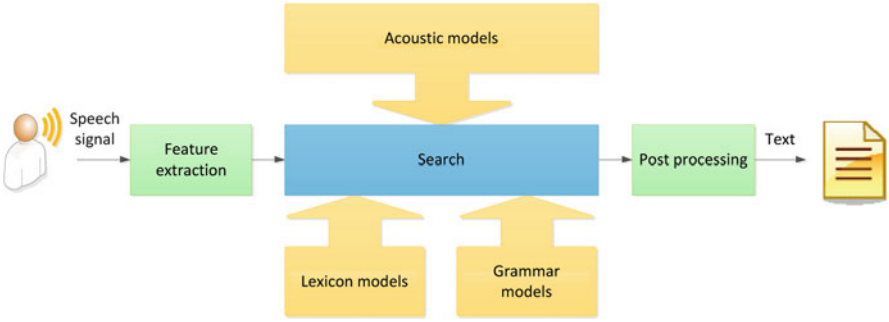


Fig. 6 Automatic speech recognition system

signal. It performs operations such as noise cancellation, normalisation, and pitch correction. Subsequently, it extracts feature characteristics of the speech signal. The resulting stream of feature vectors is used by the search module. The search module performs the core acoustic pattern matching operations of speech recognition. Three types of models are employed by the search module to decode the speech signals as shown in the figure [32, 33]. The first type, the acoustic models, assigns probabilities to speech sound (phone). The second set of models, the lexicon models, specifies phone sequences for words, and the last type is the language models, which specifies the probability of a sequence of words for a given language. The post-processing module, subsequently, performs application-specific tasks including normalisation of output formats such as formatting date in de facto standard styles (e.g. 'second of January twenty twelve' into 'January 02, 2012').

3.2 Language Translation

Language often limits the audience for social media. Not everybody can enjoy a video clip with a soundtrack of an unfamiliar language. Therefore, automatic language translation can enhance not only the reach of social media but also the interactivity between different ethnic groups beyond linguistic boundaries. This tool can be used in many ways in social media domain, as shown in Fig. 7. The most straightforward adaptation is to translate text-based material into a text in a different language (i.e. the target language). Together with speech recognition and speech synthesis, more advanced uses are also possible. Automatic insertion of subtitles in target language is one of the examples of combined use of automatic language translation and speech recognition. Addition of speech synthesis to the abovementioned example also enables re-dubbing of video in target language.

Automatic language translation is technically identified as machine translation, which is defined as the use of software to translate text or speech from one natural

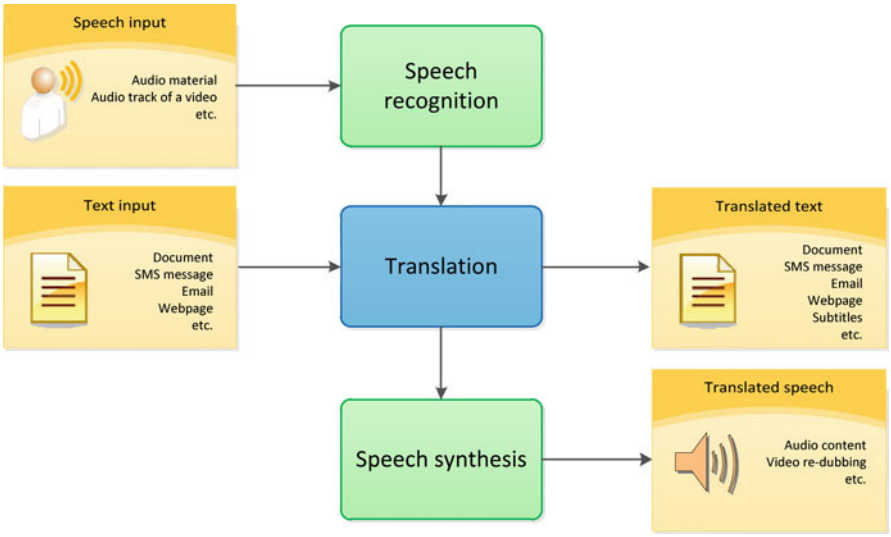


Fig. 7 Use of language translation in social media

language to another [34]. The key challenge in machine translation is to produce output text or speech maintaining the semantic, pragmatic, structural, lexical, and spatial in variances [35]. Maintaining the meaning of the source text in the texting target language is known as semantic invariance. Pragmatic invariance refers to preserving the implicit intent of the text or speech. This includes conveying the politeness, intent, and urgency of the message. Maintaining the syntactic structure of the source text is identified as structural invariance. Lexical invariance defines the ability of preserving one-to-one mapping of words or phrases from source to target language. Preserving external characteristics of the source material, such as its length, location on the page, and (in case of speech translation) synchronicity with the associated video, in the target language, is known as spatial invariance. Success of maintaining the abovementioned characteristics governs the quality of translation. The quality of translation has not been high enough to gain complete satisfaction from the users yet [36]. A tremendous effort is still needed from the research community to perfect the technology for widespread use in social media applications. Nevertheless, machine translation tools are commonly available for the professional and non-professional use. Amongst them, Google and Microsoft translation tools, which are integrated into a number of online as well as offline service and tools, are perhaps the most commonly used translation tools. However, they are mostly text-to-text translators. Meanwhile, dedicated portable speech-to-speech translators, such as ECTACO Travel SpeechGuard TL-4, are also available in the consumer market. Smart phone applications such as SpeechTrans, which supports most European and Far Eastern languages, are also becoming available in the consumer market.

A number of machine translation technologies have been discussed in literature [37]. The rule-based approach defines a set of rules to map original text into the target. In contrast, statistical translation technology uses statistics based on bilingual text corpora. Another approach is the example-based approach, which uses bilingual corpus as its main knowledge base, and translation is achieved by analogy. Hence, it can also be considered as a case-based reasoning approach.

3.3 Transcoding

Transcoding can broadly be defined as an operation that converts one encoded signal to another. This is a vital signal processing operation to ensure any media content is available to its target audience, who may not be able to consume the media in its original form due to various reasons (usage environment constraints), including unavailability of the correct software, network bandwidth limitations, and processing power constraints. Transcoding can thus be used to address a number of the usage environment constraints. Hence, this operation has a number of target objectives such as format conversion, bit rate reduction (transrating), variable-bit-rate to constant-bit-rate conversion, spatial resolution reduction (resizing), and temporal resolution reduction. A classic use case for format transcoding has emerged from the growing number of smart phones that do not support popular Adobe Flash format. Users of the phones, which are shipped without any Flash support, are thus unable to view content in social media portals such as YouTube, unless the media are transcoded to a format that those phones support, such as H.264/AVC and MPEG-4.

Transcoding is commonly achieved in two stages: decoding the source media into an intermediate format and re-encoding into the target format [23]. Extra stage may be necessary in some applications including image and video resizing and cropping, audio re-sampling, etc. Most straightforward transcoding approach would be to fully decode and re-encode. This is the most feasible solution for achieving objectives like arbitrary resizing of video and transcoding between encoding standards with no common syntax. However, it is usually very costly due to the computationally intensive re-encoding operation. Thus, the computational complexity reduction in transcoding is one of the hottest research areas in transcoding research [23, 38]. Approaches such as reusing as much information as possible from the source content have been discussed in literature to address this issue [25]. On the other hand, transcoding objectives such as transrating while maintaining the same encoding format can be achieved without fully decoding (i.e. through partial decoding) the content [39]. Thus, two transcoding approaches can be identified depending on the intermediate format, namely, full decoding- and partial decoding-based approaches. These approaches are summarised in Fig. 8.

Some of the state-of-the-art encoding formats have been developed with support for low-complexity transcoding. A number of transcoding operations can be

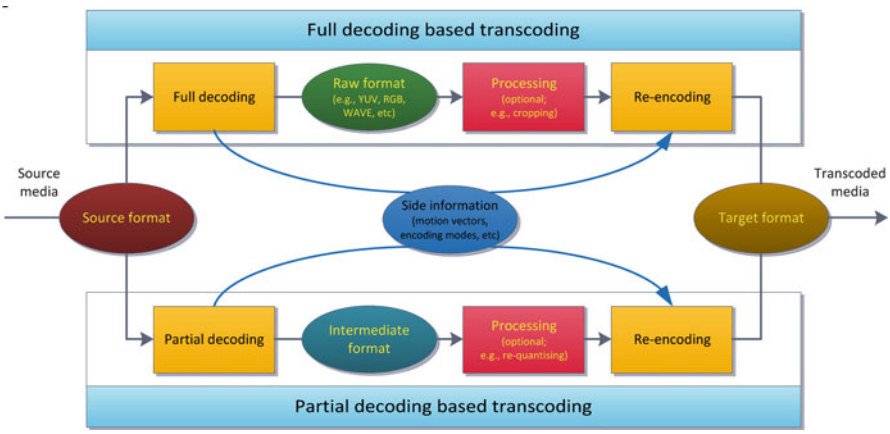


Fig. 8 Transcoding approaches

achieved simply by dropping unnecessary information in scalability extension of H.264/AVC [40] compatible video. These operations include common transcoding operations such as spatial and temporal resolution reduction and bit rate reduction. This is achieved utilising incremental encoding and smart packetisation approaches.

In general, transcoding is a lossy process unless the target encoding format is a lossless one. Moreover, if the transcoding is performed on-demand basis, computational complexity becomes a critical factor to consider before commercially deploying of such a technology. Considering, the above two factors, a measure to assess the efficiency of the transcoding operation, E_T, can be defined as follows:

$$E_T = \alpha \cdot Q_D + \beta \cdot C \tag{1}$$

where Q_D is the quality degradation, C is the complexity, and α and β are normalisation constants that assign the relative importance of the quality degradation and complexity, respectively. Based on the above definition, the lower the E_T, the better the transcoding technique. It should also be noted that quality degradation is linked to the complexity. For example, it is possible to reduce the transrating complexity of H.264/AVC-coded media by reusing motion vectors from the source media. However, this may incur a significant rate-distortion performance loss. The quality degradation can be minimised, however, by refining the motion vectors using rate-distortion optimisation, which in turn increases the computational complexity [41]. The state-of-the-art transcoding technologies make sure the loss of quality is marginal compared to the complexity reduction.

Even though transcoding is mostly associated with audiovisual media, it can also be easily applied to text-based media. For instance, a Microsoft Word document can easily be 'transcoded' to a HTML format that can be viewed on any device equipped with a web browser.

4 Adaptation Decision Taking for Heterogeneous Media Access

Adaptation decision taking is the brain of any context-aware content adaptation system. The role of an adaptation decision taking is multifold:

1. To select the most efficient and effective signal processing mechanisms from a large number of available mechanisms that effectively address the constraints described by the contextual information received from various context sensors
2. To determine the configuration parameter settings for the selected signal processing mechanisms (such as the output bit rate and encoding format for video transcoding)
3. To determine the most feasible location (i.e. at the server, a network node/proxy, or user terminal) for performing the adaptation operation

The ultimate objective of social media adaptation is to maximise the user experience under constraints described by the context. For instance, assume that a mobile user moves to an area where the data rate is not sufficient, due to poor signal strength, for the video stream he/she is consuming. This scenario can be addressed by reducing the bit rate of the content. From the signal processing point of view, though, there are a number of different ways to achieve lower bit rates including transrating and reducing spatial and/or temporal resolutions. Deciding the operation that produces the best quality video stream is the challenging task for the adaptation decision-taking stage.

To achieve the abovementioned objectives, first of all context has to be interpreted to uncover constraints, which calls for media adaptation. Technically, this is identified as context reasoning. Context reasoning is defined as inferring new information relevant to the application from a vast amount of data received from various context sensors [42]. Data received from various context sensors is often either incomplete or inaccurate. Thus, the first task of context reasoning is to perform corrective measures to overcome potential issues to improve the validity of the context. Subsequently, raw data has to be transformed into meaningful information. For the adaptation decision-taking application, constraints imposed by the detected context have to be inferred. Literature discusses a number of techniques that can be used for context reasoning such as ontology-, rule-, and case-based reasoning [43].

Once the essential contextual information is deduced, the next stage of adaptation decision taking is to determine the output format and associated configuration settings (e.g. for video: encoding format, encapsulation/container format, bit rate, frame rate, spatial resolution). In some cases, finding a suitable signal processing technology for performing certain types of adaptations in a single step may not be possible. In such scenarios, it is necessary to apply a number of signal processing operations on the source media in a sequence to generate the target format. As a result, the adaptation decision taking should consider the available

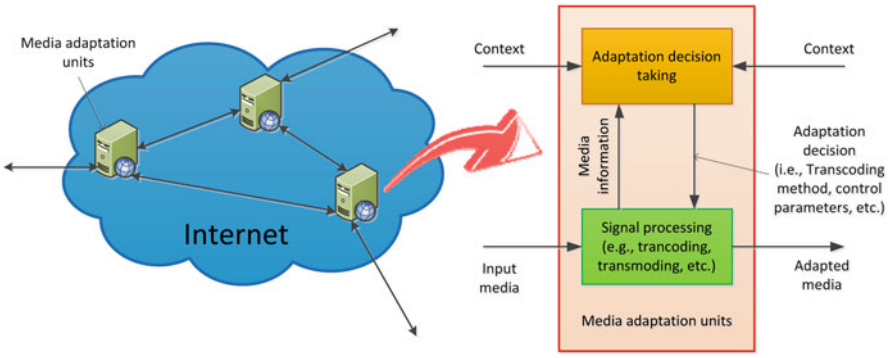


Fig. 9 Adaptation decision taking and signal processing as an integrated unit

set of technologies and their limitations, as additional set of constraints. Hence, the following extra information must also be derived during adaptation decision taking:

- Sequence of signal processing operations that are required to perform in order to generate the required output adapted bit stream
- Sequential order, in which the signal processing operations are applied

Adaptation decision taking can be implemented as an integral part of the signal processing mechanism, as shown Fig. 9. This integrated entity is called the media adaptation unit (MAU) [44]. For example, a mobile service provider can commission transcoders at base stations to dynamically control the spatial and temporal resolution, and the bit rate of the media before forwarding them to mobile phones to better utilise the limited base station capacity. Adaptation decisions are taken locally, and therefore sporadic context changes can be promptly served by dynamically changing the configuration settings.

An alternative to the aforementioned integrated architecture is the decoupled architecture, in which adaptation decision taking is implemented as a stand-alone service, often by a dedicated module known as adaptation decision engine (ADE) or adaptation decision-taking engine (ADTE) [45]. The advantage of this architecture is that the signal processing elements (often identified as adaptation engines – AEs) scattered throughout the communication chain can be coordinated to optimise the overall performance of the network. This can be explained with a scenario, where there is a bandwidth limitation in a downstream edge (e.g. user's access link). This scenario can be addressed by transrating media content at the network edge. However, the media servers still need to deliver high-bandwidth media all the way to the edge network. This is a waste of precious communication resources in the core network as well as the social media service provider, given the fact that the high bit rate version cannot be delivered to the user as it is. Alternatively, the adaptation can be performed at the server-side for improving network resource utilisation. Now,

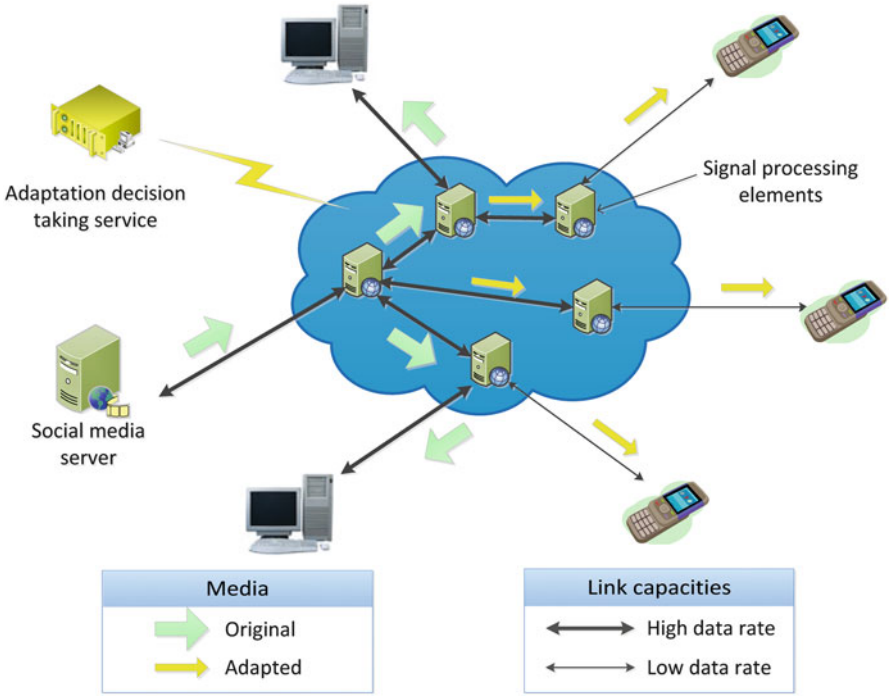


Fig. 10 Adaptation decision taking as a stand-alone service

assume that more than one user of the aforementioned edge network consumes the same media, and a number of them do not have any bandwidth issues. In such a scenario, it is more appropriate to perform the adaptation back in the edge network. Therefore, the place of adaptation is context dependent and is also an important decision that adaptation decision taking should consider. A stand-alone adaptation decision-taking architecture can provide necessary flexibility to achieve this task. Figure 10 illustrates this architecture in detail while also showing the most appropriate location of adaptation.

Nevertheless, the advantages of both integrated and stand-alone adaptation decision architectures can be obtained by commissioning a two-tier distributed adaptation decision architecture, in which a set of MAUs are overseen by a top-level adaptation decision-taking service. High-level decisions such as choosing the signal processing mechanism (e.g. transcoding) and their locations can be top-level decisions. Low-level parameters such as the output data rate can be determined in the lower layer by the MAU.

Out of the abovementioned architectures, the most feasible choice of social media adaptation decision taking is the integrated architecture since necessary infrastructure support is not yet available to realise the stand-alone architecture. The adaptation is currently possible at the content adaptation modules available

at social media servers, edge networks, and user terminals. Coordinating those adaptation modules is also virtually impossible currently. Unavailability of suitable protocols to communicate amongst the adaptation decision-taking servers and signal processing modules is deemed to be the major problem at the moment. Besides, security implications of intercepting content have to be addressed in order to gain mainstream recognition for both stand-alone and distributed architectures [46].

4.1 Adaptation Decision-Taking Approaches

The adaptation decision-taking problem is a constrained optimisation problem described as follows:

> Given any input media, define the output media that intended users can consume and enjoy under the constraints defined by the contextual information in such a way that their experience is maximised.

According to the above definition, the prime objective of adaptation is to maximise the user experience. Two different approaches have been suggested in literature:

1. Knowledge-based adaptation decision taking
2. Utility-based adaptation decision taking

The former technique uses a set of predefined rules [6] to find a suitable set of adaptation parameters to satisfy the constraints, while the latter attempts to define the bit stream format that maximises the utility [47, 48].

The knowledge-based approach considers adaptation decision taking as determining adequate signal processing sequences as a classical state-space planning problem [49]. A sequence of operations is determined to transform any given media to a format that maximises the user experience. Estimating user experience is a non-trivial problem by itself. Considering the practical difficulties, simple approximation models are often used for quantifying the user experience even though they do not accurately predict every individual's perception. Some examples are peak signal-to-noise ratio (PSNR), peak signal-to-perceptible noise ratio (PSPNR), and video quality metric (VQM) for video and signal-to-noise ratio (SNR) for audio [50].

The utility-based adaptation defines three important spaces – adaptation, resources, and utility, as shown in Fig. 11 [47, 48]. The adaptation space defines all the possible signal processing options that can be performed on any given media. Spatial and temporal resolution reduction and transrating are a few adaptation examples (a1, a2, and a3 axis, respectively, as in Fig. 11) for encoded video. Each point in adaptation space has corresponding points in resource and utility spaces. The resource space represents the resource requirements, including communication bandwidth requirement and minimum specifications of the user terminal(e.g. processing power and display resolution). Utility is the user satisfaction after the adaptation. It can be measured either subjectively or objectively. However, subjective method is

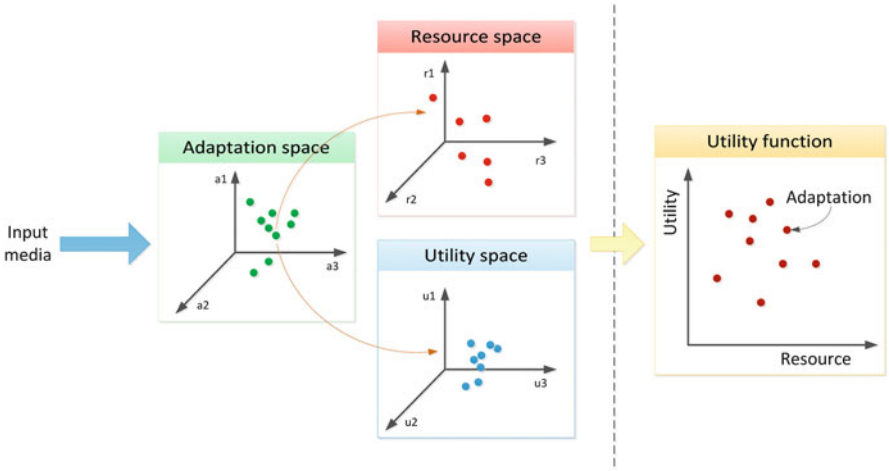


Fig. 11 Adaptation-resource-utility space concept and the utility function

not practical for serving for social media adaptation. Therefore, using objective measures, such as PSNR, PSPNR, and VQM, is the most feasible alternative. The utility function maps resource requirements into utility for each of the adaptation operations. The principle of the utility-based adaptation decision taking is simply to find the adaptation operation that produces the best utility while not exceeding the resource budget.

Even though the adaptation decision taking is classified as knowledge-based and utility-based, in a broader sense, the utility could well be a part of a knowledge-based approach [51].

5 Challenges in Social Media Adaptation

Subsequent to identifying and introducing context-aware content adaptation as one of the key technologies that enables personalised access to social media in the previous sections, it is now appropriate to examine a suite of prime issues that may set significant challenges for performing social media adaptation in the broader sense. As clearly stated earlier, social media is strictly a user centric concept, and thus it bears wide-scale one-to-many or many-to-many user-to-user communication and media sharing properties. This leads to the need for developing personalisation techniques to cater for an immense variety of user preferences, needs, characteristics, profiles, etc. Providing one solution to fit all of those is not possible due to both technological and non-technological (in other words, social) reasons, which constitute the prime challenges against social media adaptation, unless they are clearly identified and addressed.

5.1 Technological Challenges

Technological challenges are the challenges that may affect or inhibit the operation of the social media adaptation tools due to technological reasons. There may be several reasons causing such challenges; however, the following four have been identified in this subsection as the major issues to highlight.

Digital rights management (DRM) provides a set of access control technologies, which are used by the content providers to control the use of their digital content by unintended and unauthorised users in a digital value chain [52–56]. From this perspective, DRM supports intellectual property protection, which allows only authorised users' access and use of the protected digital content based on the rights expressions available, which govern the DRM-protected digital content [57]. Although intellectual property protection is important and necessary, particularly in today's social media world where even ordinary consumers have become content producers, it also has an inhibiting effect on the widespread access, sharing, and use of social media by users, sometimes even within the same social circle as the content provider. Thus, DRM can be a major limiting factor for a good personalisation practice on the social media content during adaptation, which requires particular attention. In literature, the importance of this challenge has been noted through specific research publications that propose methods to perform DRM-enabled media adaptation [45, 46, 58]. In general, those methods look at innovative ways to adapt media content while respecting the authorisation requirements for some types of adaptation operations, which are imposed by the DRM regulations that are set by the content providers.

Secure media applications via encryption of the media content are another limiting factor for the widespread access, sharing, and use of social media. Encryption also influences the required content adaptation operations, so as to provide personalised social media access to wider ranges of users. An encrypted content cannot be decrypted without sharing the encryption key with the user terminal. Without having such a key, it is very difficult to adapt the content if the media properties and media streams cannot be accessed by the adaptation middleware or tool. In literature, several research publications have been provided to address this significant issue, which investigate methods for adapting secure (i.e. encrypted) media content without the need for decryption or through partial decryption [59–63]. Nevertheless, true personalisation demanded by the social media users needs access to the media content itself, so that necessary adaptation can take place as requested. It is possible to achieve this by sharing the adaptation load with the content encoding and encryption stages at the content provider's side to some extent. Alternatively, the encryption key can be shared with the third-party adaptation middleware, which in turn puts sensitive content at risk of potential attacks by those who should not have the access. For a more generic solution to the problem, a fully automated blind secure media adaptation solution is yet to be realised, so that both proxy/gateway-based and user terminal supported personalised adaptations can be performed effortlessly and effectively.

Copyright issues also pose notable challenges for adapting social media content. As with DRM-based intellectual property protection, copyright protection of digital content limits the possible media adaptation operations that can be performed on social media that is shared amongst online social communities. Legitimate adaptation operations can still be applicable, as the copyright terms permit; however, it cannot be guaranteed that such operations will still have a wide enough scope to be able to address all of the diverse preferences or requests of a wide range of users.

Last but not least, new challenges arise for indexing, search, and retrieval of tremendous amounts of social media content uploaded everyday by many users for sharing with diverse online communities. Social media content adaptation technologies developed to cater for personalised access to social media hence should also operate hand-in-hand with personalised and fast social media search and content retrieval solutions, which are further elaborated in the other chapters of this book.

5.2 Non-technological (i.e. Social) Challenges

Non-technical challenges are those related to social aspects that may inhibit the operation of content adaptation tools for achieving efficient personalised access to social media. Below, a representative set of the most prominent challenges has been examined in detail to highlight the types of issues that can be often encountered during social media adaptation. However, it should be noted that certainly this does not provide a comprehensive list, which can be extended by the readers and upon their experience gained through various user activities in social media access, sharing, and distribution platforms.

Under no circumstance, the privacy of users should be jeopardised during social media adaptation for personalisation or any other purposes. As described in the earlier parts of this chapter, context-aware content adaptation relies on an adaptation decision-taking mechanism as an integral part of the overall adaptation process, which takes inputs from various external context providers for deciding on the best suitable adaptation method for a given situation. In addition to terminal capabilities, network characteristics, media content specifics, and natural environment properties, the diverse range of contextual information also comprises sensitive data on users' personal profiles, and thus their special needs, preferences, and personalisation requests to assist with adaptation. The contextual information is usually stored at a suitable location (e.g. at a dedicated middleware, a third-party context provider's database, partially on user terminal) in the social media access, sharing, and delivery chain, and hence obtained and processed when required by the adaptation decision mechanisms.

It is very important to collect and store such personal contextual data with necessary protection due to today's growing concern on user privacy, where exchange of sensitive personal information amongst different systems has become a common practice, particularly in heterogeneous social networking scenarios

[64, 65]. Contextual information requires similar treatment in terms of protection and privacy issues, and there are examples of different generic privacy protection system implementations reported in literature [66–69]. The good practice for protecting privacy is to define a privacy model, which identifies the information to be protected, so as to set strict privacy rules for protecting certain context depending on who it is from, who it is addressed to, and what its intended use is, while also respecting the necessary issues related to the ethics.

In social media access, sharing, and distribution scenarios such as social networks, the common trend is that user is expected to be responsible for defining his/her own privacy preferences. However, this may not always be practical to implement, as can be noted from the growing concern amongst social media and social networking users today. Instead, such sensitive information should be defined beforehand, and necessary privacy protection systems should be put in place and made transparent to the users while performing social media content adaptation. Here, the aim should be to protect the privacy of social media users while adapting their content for providing personalised access by protecting and not revealing their personal contextual data.

Privacy protection can thus pose significant challenges for supporting personalised access to social media through context-aware content adaptation. Large numbers of social media users and their very diverse personalisation requests due to various personal preferences may equally set strict challenges for adapting the content as required. To further elaborate on this with an example, one can assume a scenario where some media content is uploaded for sharing through social networking. It is likely that some of the users in this online social network would prefer to receive a part or parts of the shared content that concerns them while the others would like to receive the other parts or all of it. Within those groups of users, some prefer to receive only the audio content due to their prevalent contexts (e.g. mobile, driving or busy doing something else hence cannot watch, eyesight problems, disabilities), and some others are happy with the visual content at the expense of text or audio components due to other yet similar reasons. Indeed, this example can be taken a few steps further down to address each user's personalisation request individually, who accesses the one piece of content shared in such a heterogeneous online social community. In this way, the complex dynamics and multidimensional nature of media content adaptation becomes much clearer, where while responding to each and every personalisation request is very important, satisfying both the content provider's set rules and other users' comfort and their personalisation aspects is equally needed. A similar example can be formulated around another social media content that is shared, which contains scenes of an obscene nature for a particular group of users while it is acceptable to others. Thus, it is the sole responsibility of the adaptation decision mechanism to make use of the personal profiles provided in the form of contextual data accordingly, so as to decide on the most adequate adaptation solution to respond to every type of personalisation request in the most effective way possible. Evidently, the greater the diversity of requests for such personalised adaptations, the harder the challenge becomes that is faced by the social media content adaptation technologies.

Inaccurate adaptation of social media may result in adapted content that is completely 'out of context', which in turn may be disturbing for some individuals with potential psychological or social impacts on their online social behaviour. This may particularly happen if content adaptation is performed where semantics of the content are not carefully managed due to lack of availability of necessary contextual data inputs for inferring high-level context from low-level contextual information. It is of paramount importance that the content adaptation solutions that are selected and applied should provide acceptable outcomes to the precise needs of the users and their personalised social media access requests. For instance, a chosen adaptation operation leading to cropped images, video material, or random extracts from audio data, as requested by a mobile user with a small resource-restricted device, may end up with changing the entire meaning of the content compared to its complete version. Similarly, another adaptation operation performed to provide a selection of a view or a subset of views from multi-view/free-viewpoint video content without considering the overall message conveyed in the shared media may result in providing the user with the wrong angle to observe some events in the delivered visual scene, which may then turn out to be open for misinterpretations. This may cause unintentional stir in user's expectations as well as his/her mood and attitude towards the content that is shared through social media access technologies. To avoid this from occurring, social media content adaptation systems should also pay utmost attention to the semantics of the content that is being adapted, so as to retain its meaning as a whole, which may not always be an easy task to achieve, particularly under stringent low operational delay and computational complexity requirements.

6 Conclusion

In this chapter, context-aware content adaptation has been presented as an enabling technology for providing personalised access to social media. For this purpose, the rationale behind the need for adapting content and possible middleware structures that perform such operation in social media access, sharing, and delivery scenarios have been introduced firstly. Following, the importance of the use of contextual information and the different context types that are used in performing content adaptation in such scenarios have been emphasised. Next, discussions have been focused on describing the techniques for actual social media adaptation operations in detail, where the content adaptation tools used and the decision-taking mechanisms for selecting the most suitable adaptation operations for a given situation have been revealed. Last, both technical and non-technical (i.e. social) challenges have been highlighted, which are thought to influence the accuracy and performance of those possible social media content adaptation options. Particularly, in this final section, a number of key pointers are provided to indicate open issues for future research, so as to address those highlighted challenges for realising true personalised social media access environments. It is worthwhile mentioning

here that due to the multidimensional and multidisciplinary nature of the open issues, providing personalised access to available social media through context-aware content adaptation calls for inter-disciplinary and multinational/multicultural research and technology development efforts for finding effective solutions to the diverse technological and social challenges identified in this chapter.

Acknowledgments The authors would like to thank their past and present I-Lab colleagues as well as the partners of the EU-sponsored collaborative research projects, particularly those participated in VISNET II NoE, who provided the inspirations for composing some of the discussions presented in this chapter.

References

1. Aarts, E., Harwick, R., Schuurmans, M.: Ambient intelligence. In: Denning, P.J. (ed.) The Invisible Future: The Seamless Integration of Technology into Everyday Life, pp. 235–250. McGraw-Hill, New York (2002)
2. Zelkha, E., Epstein, B.: From devices to ambient intelligence. In: Proceedings of Digital Living Room Conference, Laguna Niguel, USA, June 1998
3. Sauter, M.: Beyond 3G – Bringing Networks, Terminals and the Web Together: LTE, WiMAX, IMS, 4G Devices and the Mobile Web 2.0. Wiley, Chichester (2009)
4. Vetro, A., Christopoulos, C., Ebrahimi, T. (eds.): Special issue on universal multimedia access. IEEE Signal Process. Mag. **20**(2), 16–73 (2003)
5. Pereira, F., Burnett, I.S., Chang, S.-F. (eds.): Special issue on multimedia adaptation. Signal Process. Image Commun. **18**(8), 597–768 (2003)
6. Martinez, J.M., Valdes, V., Bescos, J., Herranz, L.: Introducing CAIN: a metadata-driven content adaptation manager integrating heterogeneous content adaptation tools. In: Proceedings of the 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005), Montreux, 13–15 April 2005
7. Sofokleous, A.A., Angelides, M.C.: DCAF: an MPEG-21 dynamic content adaptation framework. Multimed. Tool. Appl. **40**(2), 151–182 (2008)
8. Ardon, S., Gunningberg, P., Landfeldt, B., Ismailov, Y., Portmann, M., Seneviratne, A.: MARCH: a distributed content adaptation architecture. Int. J. Commun. Syst. **16**(1), 97–115 (2003)
9. Fawaz, Y., Berhe, G., Brunie, L., Scuturici, V.-M., Coquil, D.: Efficient execution of service composition for content adaptation in pervasive computing. Int. J. Digit. Multimed. Broadcasting **2008**, 1–10 (2008). article ID 851628
10. Kaced, A.R., Moissinac, J.-C.: Secure intermediary caching in mobile wireless networks using asymmetric cipher sequences based encryption. Lect. Note. Comput. Sci. Mobile Ad-Hoc Sens. Netw. **4864**, 725–736 (2007)
11. Lei, Z., Georganas, N.D.: Context-based media adaptation in pervasive computing. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE 2001), vol. 2, pp. 913–918, Toronto, Ontario, 13–16 May 2001
12. Hutter, A., Amon, P., Panis, G., Delfosse, E., Ransburg, M., Hellwagner, H.: Automatic adaptation of streaming multimedia content in a dynamic and distributed environment. In Proceedings of the IEEE International Conference on Image Processing (ICIP 2005), pp. 716–719, Genoa, 11–14 September 2005
13. Pokraev, S., Costa, P.D., Pereira Filho, J.G., Zuidweg, M., Koolwaaij, J.W., van Setten, M.: Context-aware services: state-of-the-art. TelematicaInstituut, Technical Report TI/RS/2003/137, November 2003

14. Carreras, A., Andrade, M.T., Masterton, T., Kodikara Arachchi, H., Barbosa, V., Dogan, S., Delgado, J., Kondoz, A.M.: Contextual information in virtual collaboration systems beyond current standards. In: Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2009), pp. 209–213, London, 6–8 May 2009
15. Naguib, H., Coulouris, G., Mitchell, S.: Middleware support for context-aware multimedia applications. In: Proceedings of the IFIP TC6/WG6.1 3rd International Working Conference on New Developments in Distributed Applications and Interoperable Systems, pp. 9–22, Krakow, Poland, 17–19 September 2001
16. Andrade, M.T., Bretillon, P., Castro, H., Carvalho, P., Feiten, B.: Context-aware content adaptation: a systems approach. In: Proceedings of the European Symposium Mobile Media Delivery (EUMOB 2006), Sardinia, 20 September 2006
17. Lum, W.Y., Lau, F.C.M.: A context-aware decision engine for content adaptation. IEEE Pervasive Comput. **1**(3), 41–49 (2002)
18. Andrade, M.T., Delgado, J., Carreras, A., Nasir, S., Kodikara Arachchi, H., Dogan, S., Uzuner, H., Nur, G.: First developments on context-based adaptation. Networked Audiovisual Media Technologies (VISNET II NoE), Technical Project Deliverable D2.1.1, August 2007
19. Dey, A.K.: Providing architectural support for building context-aware applications. Ph.D. thesis, College of Computing, Georgia Institute of Technology, Atlanta (2000)
20. Information Technology-Multimedia Framework (MPEG-21)-Part 7: Digital Item Adaptation. ISO/IEC Standard ISO-IEC 21000–7:2007, December 2007
21. Vetro, A., Timmerer, C., Devillers, S.: Digital item adaptation – tools for universal multimedia access. In: Burnett, I.S., Pereira, F., Van de Walle, R., Koenen, R. (eds.) The MPEG-21 Book, pp. 243–281. Wiley, Chichester (2006)
22. Timmerer, C., Devillers, S., Vetro, A.: Digital item adaptation – coding format independence. In: Burnett, I.S., Pereira, F., Van de Walle, R., Koenen, R. (eds.) The MPEG-21 Book, pp. 283–331. Wiley, Chichester (2006)
23. Vetro, A., Christopoulos, C., Sun, H.: Video transcoding architectures and technique: an overview. IEEE Signal Process. Mag. **20**(2), 18–29 (2003)
24. Demircin, M.U., van Beek, P., Altunbasak, Y.: Delay-constrained and R-D optimized transrating for high-definition video streaming over WLANs. IEEE Trans. Multimed. **10**(6), 1155–1168 (2008)
25. Kodikara Arachchi, H., Dogan, S., Uzuner, H., Kondoz, A.M.: Utilisingmacroblock SKIP mode information to accelerate cropping of an H.264/AVC encoded video sequence for user centric content adaptation. In: Proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS 2007), Barcelona, 28–30 November 2007
26. Liu, H., Xie, X., Ma, W.Y., Zhang, H.J.: Automatic browsing of large pictures on mobile devices. In: Proceedings of the 11th ACM International Conference on Multimedia (Multimedia 2003), Berkeley, 2–8 November 2003
27. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. J. Vis. Commun. Image Represent. **19**(2), 121–143 (2008)
28. Szeliski, R.: Image mosaicing for tele-reality applications. In: Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision, pp. 44–53, Sarasota, 5–7 December 1994
29. Dawson, T.P., Read, C.J.: Multimedia network picture-in-picture. US 20,040,168,185, 26 Aug 2004
30. Schroeter, J.: Text to-speech (TTS) synthesis. In: Dorf, R.C. (ed.) Circuits, Signals, Speech and Image Processing: The Electrical Engineering Handbook, 3rd edn. CRC Press, Boca Raton (2006)
31. Brill, E., Mooney, R.J.: An overview of empirical natural language processing. AI Mag. **18**(4), 13–24 (1997)
32. Selfridge, E., Arizmendi, I., Heeman, P., Williams, J.: Stability and accuracy in incremental speech recognition. In: Proceedings of the 12th Annual SIGdial Meeting on Discourse and Dialogue, Portland, 17–18 June 2011

33. Hosom, J.P.: Automatic speech recognition. In: Bidgoli, H. (ed.) Encyclopaedia of Information Systems, vol. 4, pp. 155–169. Academic, San Francisco (2003)
34. Hutchins, W.J., Somers, H.L.: An Introduction to Machine Translation. Academic Press, London (1992)
35. Carbonell, J.G., Tomit, M.: New approaches to machine translation. In: Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Colgate University, Hamilton, New York, 14–16 August 1985
36. Hutchins, J.: Commercial systems. In: Somers, H. (ed.) Computers and Translation: A Translator's Guide. John Benjamins Publishing Company, Philadelphia (2003)
37. Hutchins, J.: Current commercial machine translation systems and computer-based translation tools: system types and their uses. Int. J. Transl. **17**(1–2), 5–38 (2005)
38. Ahmad, I., Wei, X., Sun, Y., Zhang, Y.-Q.: Video transcoding: an overview of various techniques and research issues. IEEE Trans. Multimed. **7**(5), 793–804 (2005)
39. Thomas, N., Bull, D., Redmill, D.: A novel H.264 SVC encryption scheme for secure bit-rate transcoding. In: Proceedings of the 27th Picture Coding Symposium (PCS 2009), Chicago, 6–8 May 2009
40. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans. Circuit. Syst. Video Technol. **17**(9), 1103–1120 (2007)
41. Xin, J., Lin, C.-W., Sun, M.-T.: Digital video transcoding. Proc. IEEE **93**(1), 84–97 (2005)
42. Nurmi, P., Floreen, P.: Reasoning in context-aware systems. PhD Thesis. University of Helsinki, Department of Computer Science (2004)
43. Bikakis, A., Patkos, T., Antoniou, G., Plexousakis, D.: A survey of semantics-based approaches for context reasoning in ambient intelligence. In Proceedings of the European Conference on Ambient Intelligence (AmI 2007), Darmstadt, 7–10 November 2007
44. Kassler, A., Schorr, A.: Generic QoS aware media stream transcoding and adaptation. In: Proceedings of the Packet Video Workshop (PV 2003), Nantes, 28–29 April 2003
45. Andrade, M.T., Dogan, S., Carreras, A., Barbosa, V., Kodikara Arachchi, H., Delgado, J., Kondoz, A.M.: Advanced delivery of sensitive multimedia content for better serving user expectations in virtual collaboration applications. Multimed. Tool. Appl. **58**(3), 633–661 (2012)
46. Carreras, A., Delgado, J., Rodriguez, E., Barbosa, V., Andrade, M.T., Kodikara Arachchi, H., Dogan, S., Kondoz, A.M.: A platform for context-aware and digital rights management-enabled content adaptation. IEEE Multimed. **17**(2), 74–89 (2010)
47. Kim, J.-G., Wang, Y., Chang, S.-F.: Content-adaptive utility-based video adaptation. In: Proceedings of the IEEE International Conference on Multimedia Computing and Expo (ICME 2003), Baltimore, 6–9 July 2003
48. Wang, Y., Kim, J., Chang, S.-F.: Content-based utility function prediction for real-time mpeg-4 video transcoding. In: Proceedings of the IEEE International Conference on Image Processing (ICIP 2003), pp. 189–192, Barcelona, 14–18 September 2003
49. Jannach, D., Leopold, K., Timmerer, C., Hellwagner, H.: A knowledge-based framework for multimedia adaptation. Appl. Intell. **24**(2), 109–125 (2006)
50. Chikkerur, S., Sundaram, V., Reisslein, M., Karam, L.J.: Objective video quality assessment methods: a classification, review, and performance comparison. IEEE Trans. Broadcast. **57**(2), 165–182 (2011)
51. Lopez, F., Nur, G., Dogan, S., Kodikara Arachchi, H., Mrak, M., Martinez, J.M., Garcia, N., Kondoz, A.: Improving scalable video adaptation in a knowledge-based framework. In: Proceedings of the 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010), Desenzano Del Garda, 12–14 April 2010
52. Lauf, S., Rodriguez, E.: IPMP components. In: Burnett, I.S., Pereira, F., Van de Walle, R., Koenen, R. (eds.) The MPEG-21 Book, pp. 117–138. Wiley, Chichester (2006)
53. DeMartini, T., Kalter, J., Nguyen, M., Valenzuela, E., Wang, X.: Rights expression language. In: Burnett, I.S., Pereira, F., Van de Walle, R., Koenen, R. (eds.) The MPEG-21 Book, pp. 139–212. Wiley, Chichester (2006)

54. Barlas, C., Dow, M., Rust, G.: The MPEG-21 rights data dictionary and new approaches to semantics. In: Burnett, I.S., Pereira, F., Van de Walle, R., Koenen, R. (eds.) The MPEG-21 Book, pp. 213–242. Wiley, Chichester (2006)
55. Lin, E.I., Eskicioglu, A.M., Lagendijk, R.L., Delp, E.J.: Advances in digital video content protection. Proc. IEEE **93**(1), 171–183 (2005)
56. Bormans, J., Gelissen, J., Perkis, A.: MPEG-21: the 21st century multimedia framework. IEEE Signal Process. Mag. **20**(2), 53–62 (2003)
57. Wang, X., DeMartini, T., Wragg, B., Paramasivam, M., Barlas, C.: The MPEG-21 rights expression language and rights data dictionary. IEEE Trans. Multimed. **7**(3), 408–417 (2005)
58. Carreras, A., Rodriguez, E., Dogan, S., Kodikara Arachchi, H., Perramon, X., Delgado, J., Kondoz, A.M.: Architectures and technologies for adapting secured content in governed multimedia applications. IEEE Multimed. **18**(4), 48–61 (2011)
59. Apostolopoulos, J.G., Wee, S.J.: Secure scalable streaming enabling transcoding without decryption. In: Proceedings of the IEEE International Conference on Image Processing (ICIP 2001), vol. 1, pp. 437–440, Thessaloniki, 7–10 October 2001
60. Apostolopoulos, J.G.: Secure media streaming and secure adaptation for non-scalable video. In: Proceedings of the IEEE International Conference on Image Processing (ICIP 2004), vol. 3, pp. 1763–1766, Singapore, 24–27 October 2004
61. Zeng, W., Lan, J., Zhuang, X.: Security for multimedia adaptation: architectures and solutions. IEEE Multimed. **13**(2), 68–76 (2006)
62. Kodikara Arachchi, H., Perramon, X., Dogan, S., Kondoz, A.M.: Adaptation-aware encryption of scalable H.264/AVC video for content security. Signal Process. Image Commun. **24**(6), 468–483 (2009)
63. Hellwagner, H., Kuschnig, R., Stutz, T., Uhl, A.: Efficient in-network adaptation of encrypted H.264/SVC content. Signal Process. Image Commun. **24**(9), 740–758 (2009)
64. Carreras, A., Delgado, J., Rodriguez, E., Tous, R.: The impact of contextual information on user privacy in social networks. In: Proceedings of the 1st Workshop on Privacy and Protection in Web-Based Social Networks, pp. 35–44, Barcelona, 8–12 June 2009
65. Zhu, Y., Hu, Z., Wang, H., Hu, H., Ahn, G.-J.: A collaborative framework for privacy protection in online social networks. In: Proceedings of the 6th International Conference on Collaborative Computing (CollaborateCom 2010), pp. 40–45, Chicago, 9–12 October 2010
66. Tumer, A., Dogac, A., Toroslu, I.H.: A semantic based privacy framework for web services. Computer Science: Intelligent Techniques for Web Personalisation, vol. 3169, pp. 289–305, Springer-Verlag GmbH, Berlin, November 2005
67. Sheppard, N.P., Safavi-Naini, R.: Protecting privacy with the MPEG-21 IPMP framework. Computer Science: Privacy Enhancing Technologies, vol. 4258, pp. 152–171, Springer-Verlag GmbH, Berlin, December 2006
68. Kenny, S., Korba, L.: Applying digital rights management systems to privacy rights. Comput. Secur. **21**(7), 648–664 (2002)
69. Rodriguez, E., Rodriguez, V., Carreras, A., Delgado, J.: A digital rights management approach to privacy in online social networks. In: Proceedings of the 1st Workshop on Privacy and Protection in Web-based Social Networks, pp. 45–53, Barcelona, 8–12 June 2009

Part IV
Applications and Services

Research in Social Media: How the EC Facilitates R&D Innovation

Loretta Anania

PetaMedia might be remembered as one of the 'last of the Mohicans' in that 'Networks of Excellence' of its kind are an endangered species, as far as EU contracts go.[1] Collaborative R&D projects in the EU framework programmes are a proven and effective instrument to address scientific and societal problems, reinforcing global competitiveness as part of a single 'European Research Area'. PetaMedia was given a grant to 'spread excellence' and to achieve a 'lasting integration' both fruitful and collaborative in nature. On average NoE grants had 11 separate organizations signing the contract, but PetaMedia had just four 'core' partners and other loosely tethered 'affiliated partners'. Their goal was to co-operate as part of a 'Virtual Centre of Excellence' that would attract the most ingenious researchers in the area of peer-to-peer networking and tagged media (hence the acronym, which French speakers found peculiar).

As far as scalability is concerned, PetaMedia was suitably ambitious. Three billion users are now connected via the global Internet. One billion of them use social networks, and the estimated size of today's digitized content is in the terabyte to hexabyte range.[2] The size of the Internet and the 'Internet economy' affected by it grows exponentially. No wonder that social media retrieval is a dynamic and

---

The views expressed in the article are the sole responsibility of the author and in no way represent the view of the European Commission and its services.

[1] At the time of the PetaMedia grant, in DG Information Society's Directorate D Convergent Networks and Services, the networks of excellence were only 5 out of a total 104 grants (another 20 were IPs, 67 STREPS, and 12 CSAs). With more calls, the statistics have changed, but NoEs are excluded from many objectives, after a critical review of NoEs by the Court of Auditors.

[2] Others forecast wilder estimates: The May 2012 SMART study by Jonathan Cave of RAND Europe, Gabriella Cattaneo of IDC EMEA, and other claims that the 'digital universe' has grown

L. Anania (✉)
Scientific Officer at DG Connect, European Commission, Brussels, Belgium
e-mail: loretta.anania@ec.europa.eu

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_15, © Springer-Verlag London 2013

protean field of activity. As personal electronic devices get fancier, cheaper, smarter, and always connected, new forms of social communication and data sharing across people, across computers, and across organizations take shape and take power (in the figurative and literal sense, electrical power consumption).

PetaMedia was part of a movement to develop a new decentralized infrastructure for the 'future Internet'. Today's Internet is more centralized than it once was: take Facebook, a centralized social network; Google docs, a centralized group and document management system; and YouTube, a centralized media hosting facility. To counter the big players, innovation activities in research projects targeted the expansion of user-centred media, building on personalized, shared content search and the expansion of new user capabilities. This could be called a bottom-up approach as opposed to industrial policy based on the competitiveness of the fittest (where rent and profits are based primarily on maintenance or acquisition of dominant position in established markets). The user-centred digital end-to-end media were challenging the position of traditional media and circumventing paid media access regimes.

Social change is shaped by information and communication technologies. Consider, for example, the revolution in citizen media journalism. Communication channels like YouTube are having bigger and faster impact on the on-line audience of Internet users. The media content produced 'non-professionally' by the 'me broadcast generation' of bloggers and tweeters, and their flow of 'user-generated content' fills the net with a variety of audiovisual contents.

Traditional and new forms compete for attention. A recent Internet book author complains that user-generated content is an excuse for consumerization of the centralized net and its viral advertisement (viral as in contagion and social spam that spreads across social media, by friends of friends, clogging the electronic highways and creating on-line addictions or attention disorder illness). Others trumpet enthusiasm for an Internet where curiosity is king: consider that TED Talks viewing grew a hundredfold between 2009 and 2011.

During this very same period, PetaMedia (whose contract was from March 2009 to October 2011) was well positioned to address these issues and to measure or experiment with media 'sharing' and 'searching' in a peer-to-peer (as opposed to traditional client/server access architectures). It pioneered the convergence of Internet search and social media, particularly in the sphere of entertainment and for video-streaming forms of narrowcasting. The methods applied in research consist of applying content analysis techniques, statistical machine learning, and information retrieval to any type of 'multimedia' content using fancy algorithms and clever hardware and software implementation. An advisory group of wise men from industry was created to turn the project ideas into 'potential business ideas' and 'exploitation plans'. Two initial applications were 'Spud TV' (for

48% since 2011 to a difficult-to-imagine 2.7 zettabytes of data. The sum of all human speech (if it were to be recorded) would take up 42 zettabytes of storage, and by 2020, the 'Internet economy' is expected to reach five trillion euro.

couch potato ladies and Joe six packs) and 'near 2 me' (mobile geolocation): in both use cases, the user quality of experience feedback was key to continue or discontinue or reformulate the application for possible media product or web service developments.

When business models scale big, interoperability or 'digital convergence' of previously separate and inaccessible traditional media systems also gives power (not necessarily the licence) to create new value models. Mining, fishing, and exploiting data across borders can add regulatory complexity to networked media practices across the public and private sphere. Even 'tagging' poses the question of who owns the metadata and what IPR are created by a process or collective intelligence or by a virtual entity. Much to the surprise of the computer scientists who developed them, the fist exploitation of OS search engine project results were used by lawyers working on very proprietary 'brand infringement' cases. Other uses were by the security agencies looking for large-scale 'face recognition' and 'event detection' to identify criminals. One news agency project participant used search engine metadata tools to detect that the so-called final photo of Osama bin Laden was faked, a detection which proved to be correct, and he was thankful for not publishing it, whereas news competitors who were less technology-savvy had printed it as real.

Licit or illicit uses of data processing machinery are an ethical and legislative concern to the whole Internet economy and need to be solved at political level. As the electoral success of Pirate Party representatives (first in Sweden and more recently in Germany) indicates, the future Internet will be a political battleground as well as the engineers' sandbox we know and love to play with. There is clear citizen concern for the 'openness' and free speech policy of the Internet and also for accountability and for the protection of critical infrastructure. Governance of the information societies raises the dilemma: do we want 'more markets or more hierarchies'? Current solutions to this problem pose substantial political challenges, to which R&D can contribute possible answers, test the efficiency or economy of the choices, and offer alternative visions.

As I recall, during its last 5 years the network media unit (which no longer exists due to internal reorganization)[3] had an annual R&D budget in the range of 40–50 million, spread over 50 projects lasting on average 3 years each. The target outcome was to widen access to a panoply of social media based on Internet connectivity interoperability and possibly common standards. Many applications were tested, including multimedia broadcasting or web TV, multimedia and multimodal search, serious gaming, immersive interactive experiences, distributed cinema, music web, augmented and mixed reality, and personalization of media. User-centred research emphasized interaction design, use case analysis, and objective quality of experience metrics.[4] Benchmarking and evaluation of 'multimedia' or cross media was noteworthy: the PetaMedia contest called 'MediaEval' has gained international

[3]The projects will continue their work in other units.

[4]Example are the QoEMEX international conference and the COST action project that supports QoE metrics.

recognition and a dedicated group of followers. Rich media, content enhancement, and interactive 'social computing' were common features across networked media R&D projects.

As digital media content proliferates, search engine technologies or programme guides will get embedded in every large-scale web system that offers meaningful and fast content retrieval to the end user. What disappears from the front end is probably getting far more complex at the back end and network side. Developers need to figure out how new technologies and plug-in bric-a-brac are assembled and designed to work coherently. On the other hand most end users want to be pampered into using social media apps without needing to read a manual or sign a privacy and consent decree before every click. Social media applications or gaming must make us feel clever and curious. Social network and web search or gaming providers know this. Computer literacy is highest among the young digital natives, who hop across browsers and search bars to tap vast and fresh information sources, exploiting the available scale and scope of broadband connectivity anywhere, anytime, and at minimal cost or effort. Despite all the freeware and entitlement culture of the digital natives, we all want new technologies that respect good 'old' design principles (to be reliable, robust, secure, trustworthy, ergonomic). Some argue that future development of user-driven applications in shared networked environments will depend on the availability of open-source libraries and search computing platforms that implement cutting-edge algorithms. All links were created equal, but thanks to social media and the 'human-in-the-loop', some links are more equal than others. Not everyone can guarantee to offer trustworthy spam-protected and reliable sub-100 ms latency social media retrieval with real-time or near real-time performance!

The networked media projects formed three clusters:

1. Media platforms and content delivery cluster. Here part of the effort went to developing common standards, such as DVD-H or 3D RUCOD or MPEG and JPEG, or interoperability. Projects like P2PNEXT developed set-top boxes and IPTV applications. Interoperability across formats and devices was tested and validated, and usability requirements were gathered and contributed to the development of common standards (which is a long-term process). A subset of projects worked on content-centric media and the future of Internet traffic. Eight projects banded together to contribute to the IETF forum, by means of a task force devoted to content-centric networking architectures under the banner of a Future Internet Assembly.[5] Some work examined traffic and user-generated content distribution or investigated 'context-aware' networks with sensors and actuators.
2. Immersive interactive 3D media cluster. Projects tested Internet-distributed entertainment and gaming, music creation and gesture interaction, as well as

[5]Future Internet Design Principles, FIA Arch Group, January 2012 edited by Dimitri Papadimitriou and Theodore Zahariades.

production of HDTV and 3D cinema distribution to the home. One worked on augmented reality and 3D CAVE (avatars and robots) to be present at a distant location and digital beaming of our three senses (hearing, touching, seeing) for embodied telepresence or teleportation. One developed serious gaming for physical or neuro-rehabilitation and tested the technologies in hospitals to find out if the cure worked. Interactive media applications used high-fidelity motion, sound, and gesture capture. There were some interdisciplinary projects, which investigated not just the engineering but the human perceptual mechanisms as well as the ethical aspects and media effects.

3. Multimedia search cluster. As digital content explodes search engines need to be embedded in every large-scale web system that offers meaningful content retrieval services to the end user. PetaMedia contributed to further development of Tribler, an OS peer-to-peer search engine, and to the design of a future Internet architecture. Among the applications that were developed were a '3D search engine' and a 'multimodal search engine' that allowed musical tune retrieval by tapping of fingers or feet (as opposed to traditional keyword search) to get a mazurka or a waltzing Matilda tune that was congenial.[6] PetaMedia once investigated picture search based on 'brain interfaces' (search of personal photo collections and 'search by thinking' wearing an alpha-wave cap as interface as opposed to typing keywords; here it is emotional feedback and statistical methods to help the machine predict what its individual user wants to tag). PetaMedia addressed both user 'tagging' behaviour and real-time personalized multimedia retrieval. Service prototyping examples include the near-2-me social media application to examine push versus pull services, where 'push' refers to paid or free content.

A common dispute at the cluster meetings was the effort to disambiguate the questions:

* Search for 'where' in the network (location, storage, centralized or decentralized IP architecture, cloud computing and reduced power consumption)
* Search for 'what' (social content indexing, web content crawling, automatic summarization, indexing, refreshing content, ranking, optimal presentation of retrieved results, user intentions)
* Search 'by whom' (deep packet inspection, query (re) formulation, profiling, curtailing)

Underlying the scientific developments related to social media and search systems is the need to establish a suitable basis for personalization, semantic interpretation, and a deeper understanding of user intentions in query formulation. Profiling as practised today is far from achieving personalization, and it is reaching its limits both technical and legal enforcement (data protection, privacy, trust).

[6]Reference is to VICTORY and iSEARCH project results.

EU collaborative research exploring social media and user-centred applications will continue to investigate these concerns in the European Commission's Work Programme for 2013.[7]

User behaviour evolves and changes technology in ways that were not anticipated by its original inventors or developers. If until last year 80% of users entered their Internet via a typed keyword search, by April 2011, social networks had taken over as the favourite end-user entry point into the picture world of Internet. Catching the essence of what different users want while offering them products they actually need at a cost they can afford can determine market success. Monitoring what the dominant Internet providers rank as 'premium' and providing cheaper alternatives are parts of a competitive web entrepreneur's edge, a possible new business or niche market. Web crawling, refreshing content, linking and indexing all html pages, summarizing voice to speech and speech to text, stirring in image analysis, face detection, natural language processing, and a whole battery of machine learning algorithms give a systems perspective of today's web machinery, a toolbox of tricks.

Tomorrow's media world will work differently, and it requires bold imagination and risk-taking research and development effort to get there. As this book illustrates, many people are contributing to the effort, and there are some success stories from PetaMedia. The first ICT Knowledge and Innovation Communities to be selected by the European Institute of Technology had the Delft and Berlin partners. Also, the first Competitiveness and Innovation Programme grants for large-scale testing of semantic media access and search went to OpenSem.[8]

The 'visible web' of searchable data (what is today accessible to search engine crawlers, web information scrapers, and databases with structured metadata, including user tagging data) is only a part of the submerged dark web of data that future research needs to tackle. Search in unstructured or encrypted data will bring even more unexpected surprises and innovations. Nuggets of ambitious efforts and investment in this domain will generate over the next 5 years a major innovation. At a minimum, efforts to bring PhD students to a virtual centre of excellence would increase know-how and enhance competitiveness in the information retrieval and social computing business, creating jobs in the mobile and media sectors in particular. Read more of this book to see exactly what the last of the Mohicans discovered in the field of social media retrieval.

Loretta Anania

[7]The cordis.eu website is where the Work Programme is published: FP7 Call 10 WP2013 Cooperation Theme 3 ICT objective 1.1 future networks, 1.6 connected and social media, and other specific call texts such as collective awareness platforms.

[8]The CIP project OpenSem started in 2011 and includes relevant industrial and academic partners from two media search cluster projects PetaMedia and PHAROS.

Video Technology for Storage and Distribution of Personalised Media

Glenn Van Wallendael, Jan De Cock, Davy Van Deursen, Marta Mrak, and Rik Van de Walle

Abstract Enabling social interaction between people by means of high quality multimedia data imposes technical challenges on the adaptation of this data. The main factors causing the need for personalized adaptation are the high bitrate associated with video information, along with the complexity to decode it and the diversity in standards it is compressed with. Different generations of compression standards have been developed, for example, MPEG-2 Video, H.264/AVC, and HEVC, each generation improving compression efficiency at the cost of increased decoding complexity. Because of the bitrate constraints and complexity associated with video compression, adaptation of the multimedia content to the end user preferences, the end user device, and the network circumstances is necessary to guarantee a high quality of experience. In order to facilitate low-complexity adaptation of the previously mentioned video compression standards, scalable variations of these standards are developed. This chapter gives an overview of these techniques and describes how personalised multimedia delivery is assisted by structural and semantic metadata.

1 Introduction

The realisation of social multimedia systems depends on the availability of numerous technologies, including solutions for distribution, description, storage and personalisation. With multimedia (and in particular video) as a social medium,

G. Van Wallendael (✉) • J. De Cock • D. Van Deursen • R. Van de Walle
ELIS – Multimedia Lab, Ghent University – IBBT, Gaston Crommenlaan 8 bus 201, B-9050 Ledeberg-Ghent, Belgium
e-mail: glenn.vanwallendael@ugent.be

M. Mrak
British Broadcasting Corporation, Research and Development, Department, Centre House, 56 Wood Lane, W12 7SB London, UK
e-mail: marta.mrak@bbc.co.uk

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_16,

requirements on infrastructure are becoming more challenging compared to previous-generation systems based on textual communication. This is because rich multimedia, which includes video content, poses a much larger informational load on the system. For instance, if in Twitter each message consisting of up to 140 characters is replaced by a 1-minute video, it would take a massive infrastructure to support such functionality.

The full success of social multimedia distribution is only reached when everyone has access to the multimedia, irrespective of the circumstances and the capabilities of both the user's receiver device and its connection. Therefore, adaptation to these technical capabilities must be incorporated when trying to enable such services. The technical aspects that the system must take into account can be categorised as 'hard' constraints imposed on the system.

Besides the hard constraints, the user's needs should also be taken into account by the system. Depending on their location or device, users can have different preferences concerning the presentation of the multimedia stream. Although the user's device may have a 3D capable screen, for example, the user can still prefer to watch some content in 2D. For certain content, users tolerate inferior quality because they value the message more than the media itself. Therefore, it is more important for the end user to effectively receive the information on its device no matter what the quality is. All these decisions largely depend on the nature of the content and the user's preference. These non-technical requirements are categorised as the soft constraints on the system.

Irrespective of constraints applied to the system, technical aspects related to storage, personalisation and distribution should be considered when a true social multimedia system is being developed. Therefore, in this chapter, three key technology domains necessary for enabling universal multimedia distribution are described:

1. Video representation: Different technical solutions exist for enabling universally accessible multimedia. However, the most demanding content is video because high compression is necessary to make it widely accessible. Therefore, an introduction about different formats for video compression with the focus on different generations of video standards is given in this chapter. Based on a variety of compression algorithms, video coding standards have different compression capabilities and properties. Next, technical aspects on how a compression format is extended to enable scalable coding are explained. Such scalable functionality enables adaptation of a video stream to different requirements. Then, the complexity issues for the multimedia server related to the compression are explained.
2. Video description: Understanding multimedia at a high level requires knowledge of the available content. In the related section, different metadata representations of the video content are described, ranging from high-level container format metadata to low-level bitstream descriptions. Also, it is demonstrated how semantic and social metadata can assist in the video delivery and retrieval process.
3. Content adaptation and delivery: the possibilities to adapt multimedia content tailored to the characteristics of the usage environment are discussed. First, different structural axes are presented influencing the quality of service (QoS) of

a multimedia delivery session. Second, semantic axes influencing the quality of experience (QoE) are introduced. Subsequently, different adaptation techniques are discussed in order to exploit both the structural and semantic adaptation axes. Finally, different technologies that support adapted multimedia delivery are described.

2 Video Representation

2.1 Video Compression Formats and Performance

Any service that includes video is highly likely to use compression because of the high information load of visual information. The development of video compression as it is still used nowadays, goes back to the mid-1990s. A collaboration between the International Telecommunication Union Telecommunications sector (ITU-T), the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) established a compression standard called H.262/MPEG-2 Part 2 [13] (from here on denoted as MPEG-2 Video). Until today, this standard is still used in many popular applications, such as for DVD-Video. With a new collaboration between the aforementioned standardisation bodies, an improved video compression standard called H.264/MPEG-4 Part 10, also known as H.264/AVC (Advanced Video Coding) [14,31] was introduced in 2003. With this new generation of video compression, encoder and decoder complexity increased, but compression gains of around 50% were obtained [23]. Gaining 50% compression efficiency corresponds to achieving the same subjective video quality at half the required bitrate. At this moment, the same standardisation bodies are working towards a new generation of video compression standard called High Efficiency Video Coding (HEVC) [6]. Once again, it is expected that a 50% compression gain, but this time relatively to H.264/AVC, can be obtained at equal visual quality [18].

In this summary of dominant solutions for video, only a subset of existing compression standards is discussed. It should be noted that other video compression standards like H.261, H.263, H.263+(+), MPEG-4 Visual, VP8, VC-1 all have different characteristics, but in general, these standards show the same trend. With every new generation, tools are added or improved resulting in additional compression gain when these tools are properly applied.

2.2 H.264/AVC and HEVC Video Compression

The majority of video compression standards are based on a block-based hybrid video coding architecture. Block-based processing means that a video picture is divided into smaller blocks which form the basis for further compression operations.

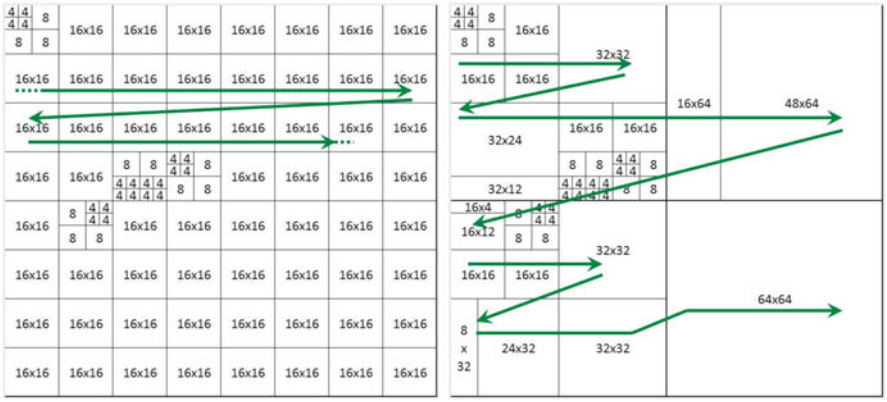


Fig. 1 Division and coding order of an example picture of size 128 × 128 with H.264/AVC 16 × 16 MBs (*left*) and HEVC 64 × 64 CTBs (*right*)

In H.264/AVC, these blocks are 16 × 16 pixels in size and are called macroblocks (MB). For HEVC, the basic blocks are up to 64 × 64 pixels in size and are called coded tree blocks (CTB). In both cases, such basic blocks can be further divided down to 4 × 4 blocks, where in HEVC both square and non-square blocks can be used. An example of the division of a 128 × 128 picture for H.264/AVC (left) and for HEVC (right) is given in Fig. 1. In this figure, arrows indicate the order in which the blocks are processed by the video decoder. For the HEVC example, only the order in which 32 × 32 coding units (CUs) are processed is shown for simplicity. In general, the CTBs are processed in a raster scan order, while subdivisions are processed in a Z-scan order.

The hybrid aspect of the block-based video codecs refers to the combination of a prediction step for each block, followed by a transformation step.

For the prediction stage, a block can be predicted from surrounding pixels (intra prediction) or from previously decoded pictures (inter prediction). Intra pictures (I-pictures) are defined as pictures which only contain intra-predicted blocks. In H.264/AVC, nine intra prediction modes for 4 × 4 blocks and four intra prediction modes for 16 × 16 blocks are available for all profiles, while nine additional intra prediction modes for 8 × 8 blocks have been included for the High profiles. For HEVC, angular intra prediction with up to 34 prediction modes has been included, along with planar intra prediction.

Inter prediction is also called motion-compensated prediction (MCP). Inter-predicted pictures (P-pictures) offer a higher flexibility and compression efficiency since blocks can contain both inter (motion-compensated) and intra-predicted blocks. B-pictures further generalise P-pictures by additionally allowing bidirectionally predicted blocks.

In H.264/AVC, macroblocks can be split into MB partitions (of 16 × 8, 8 × 16, 8 × 8 pixels), and 8 × 8 partitions can further be split into submacroblock partitions. Each partition is predicted based on one (in the case of P macroblocks) or two

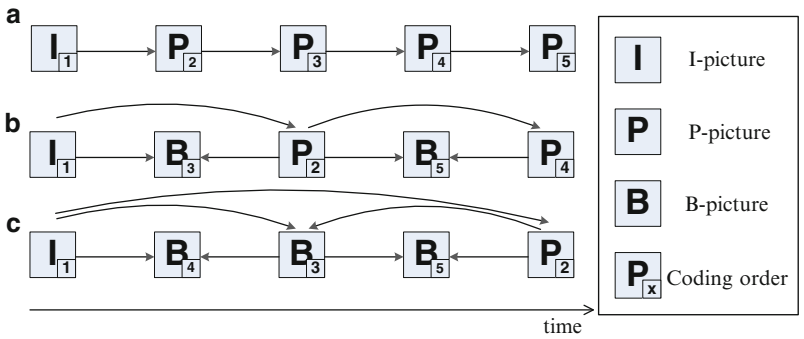


Fig. 2 Three different coding order schemes with increasing decoding delay: (**a**) no delay, (**b**) 1 picture delay and (**c**) 3 pictures delay. *Arrows* indicating the pictures used for prediction

(in the case of B macroblocks) motion vectors. In HEVC, starting at the CU size, information about the way the block is predicted is made clear by means of prediction unit (PU) information. A CU can be predicted entirely at once, resulting in a PU size equal to the CU size or it can be split into smaller square or rectangular PUs. On the PU level, motion information consisting of the chosen reference pictures, the motion vectors, and the motion vector predictors is indicated. It is on this level that a reference picture can be chosen.

Finally, after prediction of the different blocks in the video picture, a transformation and quantisation step is applied. Quantisation is the step introducing loss in the video stream, but also causing major compression efficiency gains. The quantisation phase is specifically designed to introduce losses which are as unnoticeable as possible to the human visual system (HVS). The more bitrate is saved by the quantisation process, the more visible the losses become for the end user. Therefore, the quantisation process takes care of the trade-off between visual quality and video bitrate.

H.264/AVC uses a 4×4 integer transform in all profiles, and adds an 8×8 integer transform for the High profiles. In HEVC, after prediction of the CU, a transform step is performed described on the transform unit (TU) level. A recursive tree can be created on this level reducing the transform from a maximum size of 32×32 down to a 4×4 transform.

Based on the available picture types, different coding configurations with varying coding efficiency and delay properties can be constructed. When the coding process of a video stream is started, no information from previous pictures is available. Therefore, only intra prediction can be applied to the first picture to be decoded. In Fig. 2, this can be observed by the I-picture that is located at the beginning of every coding configuration. I-pictures which provide the possibility to start decoding a video stream are also called random access points. When these pictures are present within a video stream, they provide the means to start decoding from that position. Therefore, these pictures facilitate random access within a video stream. In the first example, Fig. 2a, a decoding order with a low-delay characteristic is shown. After the I-picture, only P-pictures are included. All P-pictures have the property that only previous pictures in time can be used for predicting a picture being decoded.

To increase compression efficiency, a strategy as in Fig. 2b can be applied. In this example, a decoding configuration with a delay of one picture is shown. After decoding the first picture in time, the third picture is decoded. Although delaying the decoding of the second picture looks less efficient, overall better compression performance is typically obtained. This improved compression efficiency is achieved by using bipredicted pictures (B-pictures), which can be predicted from past and future pictures. Having the possibility to predict from past and future pictures introduces a trade-off between decoding delay and compression efficiency.

In the last example, Fig. 2c, the decoding delay is increased to three pictures. By arranging the temporal dependencies in a hierarchical way, different temporal levels are introduced. Since 'lower' temporal layers are not dependent on higher layers, the dependent layers can be removed without harming the decoding process of the lower layers. This introduces a first type of (temporal) scalability. Hierarchical coding not only offers temporal scalability, but also offers improved coding efficiency, as was demonstrated in [25]. Other types of scalability are discussed in the next section.

2.3 Scalable Video Coding

When different resolutions or quality versions of the same video are needed, the most straightforward way is to compress streams with different resolutions and qualities independently from each other. For example, one video track can be created having a low resolution and a separate track contains the higher resolution. If a single video is needed in several resolutions or quality versions, a coded representation of each version of the video must be retained. To tackle this problem in an efficient way, scalable compression has been introduced. Previous standards had some limited scalabilities. However, for the H.264/AVC compression standard, the scalable video coding (SVC) extension was specified in [14] as Annex G. The main principle of scalable coding is that lower fidelity versions of a video stream can be used during the prediction process of higher fidelity versions. In the context of SVC, the low-fidelity representation is called the base layer (BL), while the dependent high-fidelity version is called the enhancement layer (EL).

In Fig. 3, the principle of scalable compression is illustrated. From the prediction structure in this example, it can be observed that the high-fidelity representation can make use of inter-layer prediction from the low-fidelity version. Since additional prediction possibilities are available in the high-fidelity representation, higher compression efficiency can be obtained compared to the situation where both versions are encoded independently from each other. This construction is called scalable because of the low complexity involved to turn the high-fidelity version into a low-fidelity version. In practice, only a bitstream extraction process which removes the high-fidelity packets from the bitstream is required. The complexity needed for this bitstream extraction process is minimal, which makes scalable video coding very well suited for adaptation to highly personalised environments.

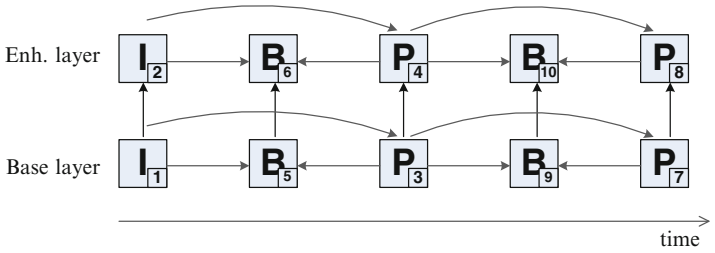


Fig. 3 Schematic illustrating principle of scalable compression

In the following subsections, the scalability opportunities provided by SVC are discussed followed by an overview of inter-layer prediction mechanisms, which result in improved rate-distortion performance over simulcast solutions.

2.3.1 Temporal Scalability

Using specific configurations defined by the picture types and prediction directions (as shown in Sect. 2.2), temporal scalability can be obtained in the bitstream. This form of scalability can be configured for single-layer H.264/AVC and, by extension, for scalable streams in SVC. Each picture in an SVC stream is provided with a temporal identifier, which makes it possible for a bitstream extractor to easily recognise and drop temporal layers from the original stream. In this way, the frame rate of a video bitstream can be reduced (e.g. from 60 to 30 or 15 Hz), depending on the requirements of the user and capabilities of the network and/or decoding device.

2.3.2 Spatial Scalability

Spatial scalability [27] allows a single scalable stream to support multiple resolutions, which are coded dependent on each other. In this way, a single bitstream can reach devices with varying resolution ranging, for example, from high-definition televisions through tablets to smartphones. In the network, only a simple packet removal operation is required to reduce the resolution to the desired characteristics.

In SVC, redundancy between scalable 'dependency' layers is exploited by predicting motion, residual and intra information based on lower layers. In every dependency layer, apart from the single-layer coding techniques (intra and inter prediction from Sect. 2.2), additional inter-layer prediction mechanisms are provided. Motion information has to be upscaled between dependent layers, and residual data has to be upsampled (using normative upsampling filters) to obtain the prediction signal. SVC not only supports typical resolution ratios such as 2:1 (dyadic scalability) or 1.5:1 (e.g. when using a 720p base layer and a 1080p enhancement layer) but also arbitrary ratios through the inclusion of extended spatial scalability (ESS) [9].

2.3.3 Quality Scalability

Different techniques for quality scalability (also denoted as fidelity or SNR scalability) have been included in SVC. We here give a brief overview of the available techniques. More information about the SNR scalability tools in the SVC specification can be found in [26].

1. Coarse-grain scalability: The CGS design uses techniques based on dependency layers, which are already available for spatial scalability. The major difference is that for CGS, no upsampling is required between successive dependency layers. By using inter-layer prediction mechanisms, enhancement layers can contain quality refinements of transform coefficients in lower layers by using a decreasing quantisation step size.
2. Medium-grain scalability: MGS uses techniques similar to CGS, but provides more flexibility. MGS allows the use of up to 16 quality levels per dependency layer, hereby significantly increasing the number of achievable rate extraction points. Also, the MGS quality levels can be removed at any point at the bitstream, while switching between CGS dependency layers is only possible at pre-defined points in the bitstream. Since there is no closed motion compensation loop provided for every quality level in MGS, drift will arise when MGS levels are dropped from the bitstreams. To prevent drift from spreading across the sequence, a key picture concept is used. In this way, the drift remains confined in time.
3. Fine-grain scalability: In addition to CGS and MGS, which are both part of the SVC standard, an FGS solution based on progressive refinement (PR) slices exists. It is not part of the SVC standard and currently not in use because of very limited practical value, the complexity of the design and the large syntax overhead. However, to achieve similar functionality as FGS, MGS can be used, allowing up to 128 quality extraction points. As an additional technique, slice header syntax elements allow for a progressive transmission of residual coefficients in the bitstream. These syntax elements indicate for the blocks of residual coefficients the start and end position of the coefficients which are actually transmitted in the current quality level. In this way, the lowest-frequency coefficients can be grouped together in a slice, while higher-frequency coefficients can be sent in refinement quality levels.

2.3.4 Inter-Layer Prediction

In the SVC design, three inter-layer prediction mechanisms were introduced in addition to single-layer prediction strategies (MCP and intra prediction). Inter-layer prediction results in improved coding efficiency of the SVC design over simulcast solutions. When compared to scalable extensions of earlier video coding standards, such as MPEG-2 Video, H.263 and MPEG-4 Visual, SVC inter-layer prediction mechanisms are significantly different. In those earlier standards, the inter-layer

prediction methods are based on reconstructed samples of the lower layer signal, similar to the inter-layer intra prediction technique in SVC.

1. Inter-layer intra prediction: If an enhancement layer macroblock is coded using base mode flag equal to 1, and the co-located macroblock in the reference layer is intra-coded, the prediction is formed by inter-layer intra prediction. For spatial scalability, a four-tap filter is used to upscale the reconstructed pixels in the reference layer to form the prediction for the current layer. In the case of quality scalability, no upscaling is required, and the reconstructed lower layer intra-coded macroblock is used for prediction.
2. Inter-layer residual prediction: In inter-layer residual prediction, the residual coefficients that are obtained by coding lower layers can be used to predict the coefficients of higher layers, as indicated by the residual prediction flag. When using spatial scalability, the corresponding residual data in lower layers is upsampled using a bilinear filter, and used as prediction for the residual signal of the current macroblock.
3. Inter-layer motion prediction: Each additional enhancement layer can reuse the motion information of underlying layers (reference picture indices and motion vectors), by simply setting a 'base mode' flag. When motion information from the reference layer is not exactly reused, it is also possible to form a motion vector predictor based on the MVs of the corresponding macroblock in the lower layers, as indicated by the motion prediction flag. In spatial scalability, the underlying MVs are scaled to form the predictor. In CGS, no scaling is required and the motion vectors can be reused as such.

SVC introduced a single-loop decoding concept, which implies that only one inter prediction loop is required at the decoder. Given the high complexity of inter prediction, single-loop decoding significantly reduces the complexity of an SVC decoder. Specifically, during decoding, motion information is 'propagated' (upscaled) from the lowest layer to the higher layers, and decoding is only executed in the highest layer based on this upscaled information. This reduction in decoding complexity, however, comes at the price of some loss in rate-distortion efficiency [26].

2.4 Towards Scalable HEVC

2.4.1 Architecture Supporting Quality Scalable Coding Within HEVC

In this section, the feasibility and usefulness of scalable HEVC is evaluated. For that purpose, HEVC has been extended to include quality scalability. The performance of such a solution has been evaluated and compared to simulcast HEVC solutions.

While the SVC extension of H.264/AVC is a single-loop scalable codec, the approach presented in this evaluation is a multi-loop scalable codec based on HEVC. In order to enable quality scalability based on HEVC, changes in picture coding

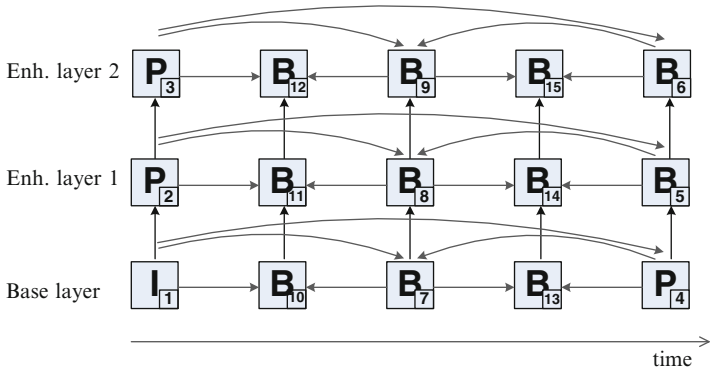


Fig. 4 Coding order for the presented scalable architecture

order are introduced and techniques to enable inter-layer prediction are devised. In the presented multi-loop approach, the lower layer is decoded entirely before the enhancement layer can be reconstructed. With multi-loop decoding, adaptive inter-layer prediction by means of motion compensation can be applied.

To obtain an efficient multi-loop scalable implementation of HEVC, the HEVC picture decoding order was modified to first process every quality layer of the same picture before advancing to following pictures in decoding order. Consequently, base layer pixel information can be made accessible for the higher layer decoding process. The coding order resulting from this decision is illustrated with the numbering in Fig. 4.

In this solution, inter-layer prediction is enabled entirely by means of motion compensation. As a result, all flexibility available in the HEVC specification can be used to adaptively enable inter-layer prediction. On two levels of granularity, inter-layer prediction can be controlled.

First of all, by means of HEVC reference picture list modification signalling, inter-layer prediction can be enabled or disabled on a picture level. Enabling inter-layer prediction can be done by signalling a message telling the decoder to include the lower layer in the reference picture list. In the absence of such a message, no inter-layer prediction will be used. This flexibility is beneficial, for instance, when coding a hierarchically encoded video stream with large quality difference between layers. In this scenario, inter-layer prediction can be disabled, for example, at the highest temporal layer. At these positions, pictures are closest to already decoded pictures, and therefore, temporal prediction is preferred to inter-layer prediction. Disabling inter-layer prediction for these pictures leaves room for extra-temporal predictions in the reference picture lists, optimising compression efficiency of the video stream. When inter-layer prediction is enabled at the picture level by including the lower layer in the reference picture list, additionally, inter-layer prediction can be adaptively signalled on the PU level. By using the reference index pointing to this layer, inter-layer prediction will be used instead of temporal prediction. In the encoder, this choice is made dependent on rate-distortion performance.

For insertion of the lower layer in the reference list of the currently processed picture, three different scenarios can occur depending on the prediction type of the base layer picture (namely, I, P or B). This is also illustrated with the prediction types indicated in the pictures in Fig. 4. When there is an I-picture in the BL, the Nth EL can only use layer N − 1 for prediction. As a result, in this scenario, EL pictures will be unidirectional P-pictures (pictures 2 and 3 in Fig. 4). Consequently, each reference picture list at level N contains the picture from layer N − 1 as reference. Layers enhancing a P-predicted BL will become B-predicted pictures (pictures 5 and 6 in Fig. 4), because the lower layer can be added to a reference picture list. In the presented approach, there has been chosen to add layer N − 1 as the first picture to one of the reference picture lists.

For B-predicted BL pictures, a similar procedure is followed, but it must be considered that the maximum number of reference pictures may already have been reached (since only a certain number of pictures can be contained in the reference picture lists). The lower layer is included as an additional reference at the cost of the last reference picture in this list.

2.4.2 Test Settings and Results for Scalable Coding

To evaluate the performance of scalable coding for HEVC, first a comparison is made with single-layer coding. The single-layer version has the same quality as the EL of the scalable variation. Therefore, this comparison represents the cost of introducing scalability to a video stream. As a second test, a comparison is made with simulcast encoding. With the simulcast configuration, two single-layer versions are encoded at the quality of the layers of the scalable configuration. This result represents the gain of scalable coding when the flexibility of having different bitrates is required. Afterwards, compression gains from scalable coding using HEVC over SVC are evaluated.

For this evaluation, conditions similar to the ones used during the HEVC standardisation process [4] are considered. For H.264/AVC and SVC, compression performance is measured using the Joint Scalable Video Model (JSVM) reference software version 9.19.14 [15]. Performance of HEVC and the proposed scalable HEVC will be evaluated using the HEVC test model (HM) version 4.0 [21]. All the defined tests are run on test sequences used in HEVC development.[1] As a metric for evaluating the results, Bjøntegaard Delta (BD) [2] measurements are used. This measure indicates the average bitrate change assuming a constant PSNR within the evaluated PSNR range. Because the PSNR can be considered constant with the BD rate measure, no PSNR numbers are reported as with a bitrate-PSNR evaluation.

[1] Details about the different classes of test sequences: class A (4 sequences, 2,560 × 1,600), class B (5 sequences, 1,920 × 1,080), class C (4 sequences, 832 × 480), class D (4 sequences, 416 × 240), class E (3 sequences, 1,280 × 720), class F (screen content; 4 sequences, from 832 × 480 to 1,280 × 720).

Table 1 BD rate results of SVC compared to single-layer H.264/AVC and compared to simulcast H.264/AVC

Class	Single-layer H.264/AVC (%)	Simulcast H.264/AVC (%)
B	24.4	−15.5
C	19.2	−21.6
D	18.7	−22.0
F	18.4	−26.9
Avg	20.4	−21.1

Table 2 BD rate performance of the proposed scalable HEVC solution compared to different competing technologies

Class	Sequence	Single-layer HEVC (%)	Simulcast HEVC (%)	SVC (%)
B	Kimono	16.7	−21.8	−62.2
B	ParkScene	18.1	−18.8	−43.3
B	Cactus	20.9	−18.5	−49.5
B	BasketballDrive	18.2	−20.6	−58.5
B	BQTerrace	13.6	−17.3	−44.0
Average B		17.5	−19.4	−51.5
C	BasketballDrill	22.0	−19.3	−42.7
C	BQMall	22.5	−19.4	−59.7
C	PartyScene	20.3	−18.6	−49.0
C	RaceHorses	17.0	−20.3	−49.0
Average C		20.5	−19.4	−50.1
D	BasketballPass	20.2	−20.4	−12.8
D	BQSquare	21.4	−18.9	−43.8
D	BlowingBubbles	22.5	−16.9	−23.0
D	RaceHorses	20.3	−19.4	−35.7
Average D		21.1	−18.9	−28.8
F	BasketballDrillTest	23.3	−19.1	−36.2
F	ChinaSpeed	19.9	−20.1	−33.4
F	SlideEditing	9.9	−36.4	−35.9
F	SlideShow	23.4	−24.7	−65.4
Average F		19.1	−25.1	−42.7
Average		19.4	−20.6	−43.8

To make a comparison possible, first the compression performance of SVC is evaluated. For this purpose, in Table 1, BD rate results of SVC compared to single-layer H.264/AVC and compared to simulcast H.264/AVC are shown. In this table, it can be observed that the cost of encoding a video stream as two scalable SVC layers adds a BD rate cost of 20.4%. The third column of the table shows that SVC reduces bitrate with 21.1% at equal PSNR compared to a simulcast H.264/AVC configuration.

Equivalently, the performance of the proposed solution is evaluated in Table 2. In this table, scalable coding of HEVC is compared to single-layer HEVC and

simulcast HEVC. On average, the BD rate cost of scalability applied on HEVC is 19.4%. It can be observed that introducing scalability to HEVC comes at an equivalent cost as when scalability was enabled for H.264/AVC (Table 1 column 2). In the next column of Table 2, the gain of scalable coding compared to simulcast HEVC is listed. Similar to H.264/AVC and SVC, it can be observed that a gain of 20.6% caused by scalable coding can be obtained compared to simulcast HEVC. In the last column of Table 2, a comparison between the proposed scalable implementation of HEVC and SVC is made. On average, 43.8% BD rate saving is realised by scalable coding based on HEVC compared to SVC. It can also be observed that BD rate savings are consistent over all tested sequences indicating that the proposed solution provides stable gains irrespective of the video content. From all these results, it can be concluded that the increased compression performance of single-layer HEVC still allows for additional gains resulting from scalability.

3 Video Description

Due to the growth of available (social) multimedia content on the Web in recent years, metadata has an increasingly important role in bringing order to the huge amount of multimedia content [28]. Metadata, which is generally defined as 'data about data', enables the effective organisation, access and interpretation of multimedia content. Searching, indexing, linking, sharing and presentation of multimedia content are example applications of multimedia metadata. Other important applications of metadata are adaptation and delivery of multimedia content. In this section, different kinds of metadata are introduced that are able to enhance and accommodate video processing (i.e. video adaptation and delivery).

Two kinds of metadata are introduced: structural and semantic. Structural metadata describes the low-level information about the multimedia content (e.g. the available scalability layers, the available tracks, and the location of the random access points), while semantic metadata describes high-level information about the multimedia content (e.g. depicted objects, person who shared the video, and ratings of the video). Both types of metadata can assist in the adaptation processes required for personalised video delivery, as discussed in Sect. 4.

3.1 Metadata for Video Adaptation

3.1.1 Bitstream Descriptions

Bitstream descriptions are automatically generated (mostly Extensible Markup Language (XML)-based) descriptions of the high-level structure of media bitstreams, called Bitstream Syntax Descriptions (BSDs). A BSD acts as an

additional layer on top of the media bitstream and describes how the bitstream is organised in scalability layers or packets of data. These layers and packets are identified through byte range offsets.

The main application of these BSDs is adaptation of video streams compressed using scalable video formats. More specifically, adaptation operations are performed based on the BSD, rather than directly on the binary bitstream, allowing the use of for instance XML transformation tools to implement the adaptation operations. Other applications include tracing (i.e. a detailed BSD can be seen as a trace file of the multimedia content) and automatic bitstream validation.

Within the MPEG-21 multimedia framework [11], Digital Item Adaptation (DIA) provides two standardised metadata formats for bitstream descriptions [24]: the Bitstream Syntax Description Language (BSDL) and the generic Bitstream Syntax Schema (gBS Schema). In Listing 1, an example BSD (using BSDL) is shown describing the structures of an SVC video stream (i.e. the Network Abstraction Layer Unit (NALU) and corresponding SVC header). As one can see, syntax elements relevant for video adaptation purposes (i.e. the SVC header elements indicating in which layer a certain NALU is located) are described in the BSD, while other information (e.g. slice payload) is referenced through a byte range.

3.1.2 Media Fragment Annotations

Ontologies have been introduced to link descriptions (which can exist in different metadata formats) of media resources and can assist in adaptation of these resources, as discussed below for video fragments. The Ontology for Media Resources 1.0 [17], a.k.a. the Media Annotation (MA) ontology, is both a core ontology for describing media resources and a mapping between the core ontology and a set of metadata and container formats currently describing media resources published on the Web (such as MPEG-7 or QuickTime). This core ontology provides the basic information needed by targeted applications for supporting interoperability among the various kinds of metadata formats related to media resources that are available on the Web.

The ontology divides its properties into a number of different categories: identification, content description, rights, distribution, fragments, technical properties, etc. An overview of the class diagram is depicted in Fig. 5. The base class is *ma:MediaResource*, representing an audio-visual resource on the Web. Two direct subclasses are defined: *ma:MediaFragment* and *ma:Image*. The former is a reference to a media fragment as defined in the media fragments URI specification [29], while the latter represents a still image. Further, the ontology provides four types of track fragments: video, audio, text and image.

In Listing 2, an example instance of the MA ontology is depicted. In this example, a spatio-temporal fragment of a video stream is described. More specifically, it is stated that the bounding box (69,47,6,11) over the period between 11 and 16 s depicts the person Elvis Presley. Note that these media fragment annotations are considered as semantic metadata: they provide information about the content of

```
<bitstream bs1:bitstreamURI="video.264">
  <!-- ... -->
  <byte_stream_nal_unit>
    <zero_byte>0</zero_byte>
    <startcode>000001</startcode>
    <nal_unit>
      <forbidden_zero_bit>0</forbidden_zero_bit>
      <nal_ref_idc>3</nal_ref_idc>
      <nal_unit_type>21</nal_unit_type>
      <nal_unit_header_svc_extension>
        <reserved_zero_two_bits>0</reserved_zero_two_bits>
        <priority_id>0</priority_id>
        <temporal_level>0</temporal_level>
        <dependency_id>0</dependency_id>
        <quality_level>1</quality_level>
        <!-- ... -->
      </nal_unit_header_svc_extension>
      <raw_byte_sequence_payload>
        <slice_layer_in_scalable_extension_rbsp>
          <slice_payload>1610 1552</slice_payload>
        </slice_layer_in_scalable_extension_rbsp>
      </raw_byte_sequence_payload>
    </nal_unit>
  </byte_stream_nal_unit>
  <byte_stream_nal_unit>
    <!-- ... -->
  </byte_stream_nal_unit>
</bitstream>
```

Listing 1 Excerpt from a BSD describing an SVC video stream

```
<http://example.com/video.mp4#t=11,16&xywh=69,47,6,11>
  a nsa:SpatialFragment, nsa:TemporalFragment;
  nsa:spatialUnit nsa:percent;
  nsa:spatialX "69"^^xsd:nonNegativeInteger;
  nsa:spatialY "47"^^xsd:nonNegativeInteger;
  nsa:spatialW "6"^^xsd:nonNegativeInteger;
  nsa:spatialH "11"^^xsd:nonNegativeInteger;
  nsa:temporalUnit nsa:npt;
  nsa:temporalStart "11.0"^^xsd:double;
  nsa:temporalEnd "16.0"^^xsd:double;
  foaf:depicts <http://dbpedia.org/page/Elvis_Presley>.
```

Listing 2 An example instance of the MA ontology (in Turtle syntax)

the video. For instance, temporal fragments do not indicate whether they start at a random access point or not, which is important when using these descriptions as input for fragment extraction algorithms (as explained in Sect. 4.2.).

Media fragment annotations also provide support to link ‘social metadata’ to multimedia content or fragments of multimedia content. Examples of social metadata properties are the owner of the multimedia content, his/her social network,

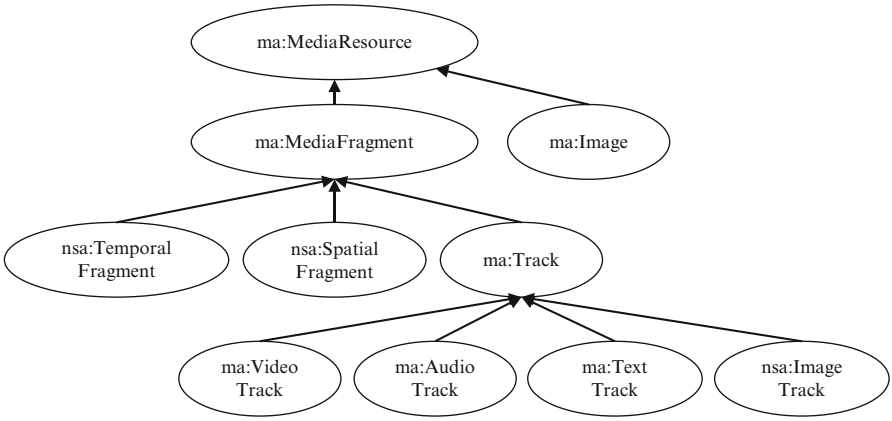


Fig. 5 Extended class diagram of the media annotations ontology

comments on particular multimedia content, ratings of particular multimedia content, etc.

For connecting the multimedia content to the social metadata, the Media Annotations ontology provides a number of hooks. To represent the social metadata itself, one can use the metadata schemes provided by for instance Flickr or YouTube. Otherwise, specific social ontologies can be used, depending on the application.

The Friend-Of-A-Friend (FOAF, [5]) ontology is able to describe characteristics of people and social groups that are independent of time and technology; as such they can be used to describe basic information about people in present day, historical, cultural heritage and digital library contexts. In addition to the FOAF core terms, there are a number of terms for use when describing Internet accounts, address books and other Web-based activities. Related to FOAF is the Semantically-Interlinked Online Communities (SIOC) ontology [3]. SIOC provides methods for interconnecting discussion methods such as blogs, forums and mailing lists to each other.

A number of schemes or ontologies exist to represent user ratings to Web resources (and thus for multimedia content in particular). The Trust ontology [10] is an extension of the FOAF ontology, defining properties about user profiles. The Review ontology [1] introduces the concept of a review, with corresponding rating assessment. Finally, the Rating ontology [19] is a lightweight ontology to store and share ratings on the Web.

3.2 Metadata for Video Delivery

A number of different approaches exist to deliver multimedia content to the end user through the Web. Within classic client-server architectures, three main media delivery methods exist: Hypertext Transfer Protocol (HTTP) download, real-time

```
<MPD>
   <Period start="PT0.00S" duration="PT50.00S">
      <Representation mimeType="video/mp4" bandwidth="1000000">
         <SegmentInfo duration="PT50.00S">
            <Url sourceURL="/MediaHigh/media1.mp4"/>
            <Url sourceURL="/MediaHigh/media2.mp4"/>
            <!--   -->
         </SegmentInfo>
      </Representation>
      <Representation mimeType="video/mp4" bandwidth="500000">
         <SegmentInfo duration="PT50.00S">
            <Url sourceURL="/MediaLow/media1.mp4"/>
            <Url sourceURL="/MediaLow/media2.mp4"/>
            <!--   -->
         </SegmentInfo>
      </Representation>
   </Period>
</MPD>
```

Listing 3 Example DASH manifest

streaming and HTTP adaptive streaming. The latter combines the advantages of both the HTTP download and the traditional streaming approach. HTTP adaptive streaming acts like streaming but is based on HTTP progressive download. More specifically, it uses HTTP as transmission protocol and performs the media delivery as a long series of very small progressive downloads (rather than one big progressive download). Note that, in contrast to real-time streaming, the media do not have to be transmitted at more or less the same rate as its bitrate, which results in less strong requirements in terms of bandwidth, delay and packet loss. Moreover, the protocol is very flexible towards bandwidth changes: the client can switch to a different quality version between two progressive downloads (i.e. the server has no notion of this since it is just serving static chunks of media).

Manifest files play a crucial role within HTTP adaptive streaming. A manifest file provides all the details necessary for a client to consume the multimedia content. More specifically, it contains information regarding the available quality versions, the different segments, the location of random access points, etc. The typical workflow for delivering media content through HTTP streaming is as follows: (1) the client fetches the manifest file of the corresponding requested media resource; (2) based on this manifest file, the client requests and downloads the segments (through HTTP).

Next to proprietary solutions (e.g. Apple's HTTP Live Streaming and Microsoft's Smooth Streaming), a standardised solution was developed within MPEG: Dynamic Adaptive Streaming over HTTP (DASH, [12]). DASH became an international standard in November 2011 and is the first HTTP adaptive streaming technology that also provides support for SVC. An example DASH manifest is depicted in Listing 3. The example shows two different quality versions, each with their corresponding temporal segments.

4 Content Adaptation and Delivery

In order to enable Universal Multimedia Access and thus provide each user the requested multimedia content, allowing the best possible experience, techniques and algorithms to adapt and deliver multimedia content in an efficient way is necessary. This section elaborates on different adaptation techniques and discusses possible adaptation axes based on the technologies and descriptions in the previous sections.

4.1 Adaptation Techniques

A multimedia adaptation system tries to meet the user needs by customising the content based on properties of the usage environment and user preferences. A wide variety of multimedia customisation approaches exist, as described by Magalhaes et al. in [20]. These approaches are also presented in Fig. 6. Two major categories can be distinguished: media bitstream selection and adaptation. Media bitstream adaptation can further be divided into two categories: low level and high level.

4.1.1 Media Bitstream Selection

One possibility to meet the usage environment constraints is to choose between several media bitstreams representing the same content. Media bitstream selection, also known as simulcast, corresponds to the identification of the most adequate media bitstream from those available to be sent to the end user. The selected

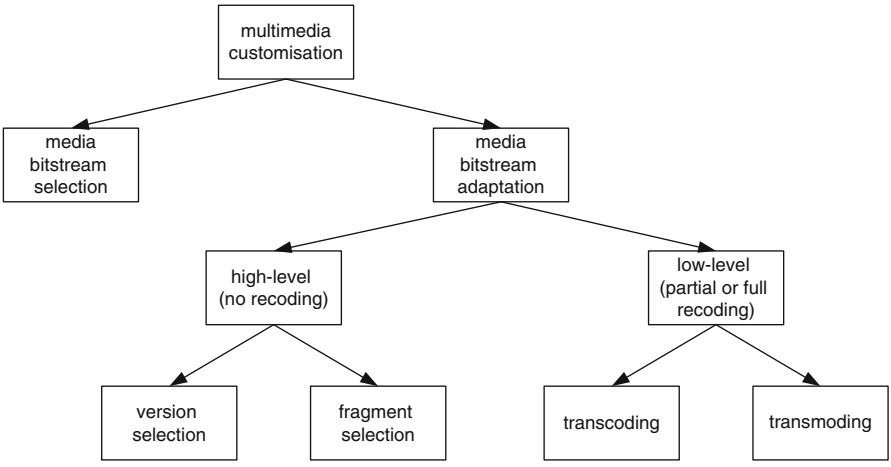


Fig. 6 Multimedia content customisation approaches

bitstream may already be adequate enough or may need further adaptation using techniques discussed in the following text. Content versions may include different media types (e.g. video or a thumbnail image), coding formats (e.g. H.264/AVC or MPEG-2 Video) or bitstream characteristics (e.g. resolution or bitrate).

4.1.2 Low-Level Adaptation

Low-level adaptation operations completely or partially recode the compressed media bitstream to customise it according to the constraints of the usage environment.

Transmoding is a low-level adaptation technique which is used when the usage environment conditions do not support the compressed bitstream in its original media type. In this case, a transformation can be used from one media type to another. Examples are the conversion of text to audio, video to key frames and large images to video.

Transcoding [30] is a popular low-level adaptation technique, which typically uses signal-processing operations. Examples of such operations are bitrate reduction, spatial scaling and coding format conversions. Transcoding solutions have been introduced, for e.g., MPEG-2 Video and H.264/AVC [8, 22], which allow significant reductions in computational complexity needed for video adaptation. For HEVC, transcoding has to be re-examined, given the changes in coding tools, such as large coding units, quadtree partitioning, angular and planar intra prediction, adaptive loop filtering, sample-adaptive offset filtering and advanced motion vector prediction. By reusing information from the incoming HEVC bitstream, such as quadtree partitioning information, motion data and loop filtering coefficients, a large complexity reduction can be obtained during re-encoding. For previous video coding standards, compressed-domain transcoding has been examined to further reduce complexity [7], hereby avoiding a full pixel-domain reconstruction. The impact of such compressed-domain operations in HEVC remains a topic for further study.

4.1.3 High-Level Adaptation

High-level adaptation operations typically perform the removal or modification of high-level bitstream structures (such as entire video/audio tracks or network abstraction layer packets in the case of H.264/AVC or HEVC video). Hence, compressed media bitstreams can be customised without the need of a complete or partial recode process. Two categories of high-level adaptation operations exist: version and fragment selection. Applying version selection results in a tailored version of the media bitstream based on the constraints of the terminal and network constraints of the end user. Hence, the resulting media bitstream will contain the same content, but the visual quality of this content will be different (e.g. lower resolution or lower frame rate). Scalable video coding (as explained in Sects. 2.3 and 2.4) is very well suitable for applications which allow version selection. A scalable media resource consists of different layers providing different

quality. Hence, multiple (lower quality) versions of the same media resource can be extracted by performing simple editing operations. The bitstream extraction process typically involves the removal of particular data blocks and the modification of the values of certain syntax elements. In contrast to transcoding, scalable coding intrinsically assumes that the content will be distributed through heterogeneous usage environments. This implies that the coding format already provides several layers, decoupling coding and adaptation processes.

On the other hand, fragment selection maintains the visual quality of the content of media bitstreams, but the resulting tailored media bitstream will contain only a subset of the original content (e.g. a specific temporal fragment or a specific spatial region). Hence, with fragment selection, the media bitstream is customised based on the semantics of the multimedia content and the users' preferences. The metadata discussed in Sect. 3 (bitstream descriptions, media fragment annotations) are very well suited to assist in the high-level adaptation process.

4.2 Enabling Quality of Experience

The previous section presented an overview of adaptation techniques that can adapt the video content for personalised delivery. In this section, the types of adaptation operations that can be performed are given. The types of adaptations are based on combining the video coding techniques discussed in Sect. 2, the descriptions in Sect. 3 and the adaptation operations in Sect. 4.1. The adaptations are discussed along three common 'structural' axes and the selection of fragments based on semantic and social metadata.

4.2.1 Structural Adaptation Axes

To anticipate the various heterogeneous usage environments of end users, the multimedia content can be adapted along three axes: resolution, frame rate and visual quality (see Fig. 7).

Reducing the resolution from a single-layer video bitstream is not straightforward in the compressed domain because of the dependencies between the blocks in which a video picture is divided [16]. Prediction information from a block in a video stream is mainly dependent on previously decoded blocks. Removing blocks from the video picture would break this dependency chain and turn the stream undecodable. A first solution to enable the provision of different resolutions is to maintain several video bitstreams with different resolutions of the same video content on the server. As usual, maintaining different versions results in an additional storage cost related to the number of resolutions provided. Scalable coding with the low resolution in the base layer combined with the high resolution in an enhancement layer can lead to lower storage requirements. A downside of scalable coding is its limited support of legacy decoder devices. Finally, resolution transcoding can

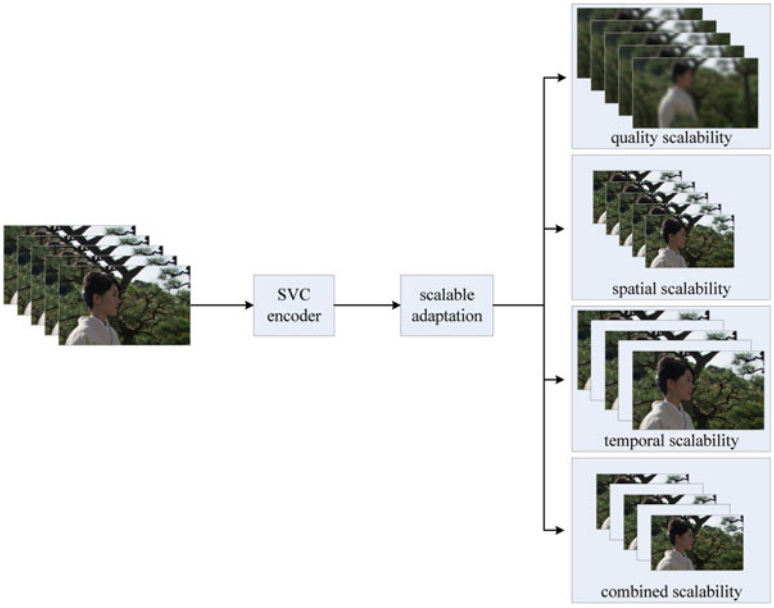


Fig. 7 Scalable dimensions in which a scalable video stream can be adapted

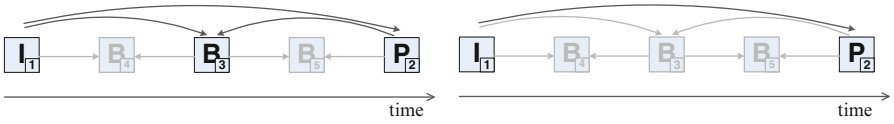


Fig. 8 Temporal scalable structure reducing to half frame rate (*left*) and quarter frame rate (*right*)

be applied, which shifts the processing requirements to the adaptation node (e.g. in a network gateway device). The choice between one of those solutions largely depends on the application for which the system is designed.

In contrast to resolution scaling, frame rate adaptation can be straightforward if the prediction structure of the video stream is thought through. When a hierarchical prediction structure is used, temporal scalability can be provided. When observing the example in Fig. 2a, it is inappropriate for frame rate reduction. Removing pictures from this video stream would break the prediction chain and make the video stream undecodable. Both other prediction structures in this figure can be made temporally scalable. In Fig. 2c, for example, all even pictures can be removed, resulting in the prediction structure of Fig. 8. Additionally, with low-complexity bitstream manipulation, the remaining B-pictures in this example can be removed as well, resulting in a quarter of the original frame rate.

The video bitrate can further be reduced by lowering the visual quality of the bitstream. To achieve this, the quantisation process is influenced to generate fewer bits, with a lower-quality reconstruction of the picture as a consequence.

As with other dimensions of adaptation, three technical solutions exist to change the quantisation quality from a pre-encoded video stream. First of all, different video streams can be stored providing each a different level of quantisation. Second, by transcoding the video stream to a suitable bitrate, precise network capacity variations can be adapted to. Finally, quality scalability techniques can be applied, as discussed in Sect. 2.

The quality of experience (QoE) delivered to the end user is highly influenced by the (combination of) adaptation operations that are applied when rate reduction is required. At which bitrate, which dimension for adaptation should be applied is a challenging problem. For example, it is known that for small variations in bitrate, adaptation by means of quantisation change is subjectively most appealing. When large bitrate variations should be covered, resolution change is advisable. When the purpose of the service is to provide optimal QoE with user preferences taken into account, the problem can get very complicated. For some content, users tolerate inferior quality because they value the message more than the media itself. And therefore it is more important for the end user to effectively receive the information on their device no matter what the quality is. In this scenario, users are willing to give up some temporal or spatial quality for improved error resilience. Metadata added to the video resources can aid in determining the preferred combination of adaptation operations for each individual user, leading to a personalised 'adaptation decision taking engine'.

4.2.2 Fragment Extraction Based on Semantic Annotations

Multimedia content can be further adapted assisted by semantic information that highlights specific fragments, based on track, spatial and temporal fragment descriptions. For each fragment type, we describe the possibilities regarding the extraction of these fragments from the multimedia content. The resulting semantic adaptations are often based on socially shared metadata, which describe, for example, a spatial or temporal 'region of interest' in video fragments, as shared on social networks.

The way in which relevant tracks are extracted from a media resource is dependent on the container format. In contrast to temporal and spatial fragment extraction, tracks are not 'encoded' but 'encapsulated' within a container format and can thus always be extracted without low-level transcoding operations. Based on the headers of the container format, it is possible to locate the proper byte ranges corresponding to the desired track. However, since tracks within a media resource are usually interleaved (e.g. interleaved audio and video, typically with intervals of 0.5–1 s), the number of byte ranges corresponding to one track becomes very high (e.g. a 30-min video track interleaved with a 1 s interval is represented by 1,800 byte ranges).

For extracting spatial regions without recoding, independently coded spatial regions are required. More specifically, (mostly rectangular) regions of the picture need to be coded independently from each other. For instance, regions of interest (ROIs) and a background are coded independently, which results in the possibility

to extract these ROIs. In case the underlying bitstream is not encoded in terms of independent spatial regions (or the coding format does not support this), transcoding operations are necessary to obtain the spatial region.

When trying to access a temporal fragment from a compressed video stream, problems related to random access occur. When a temporal fragment is selected, the decoder must be provided with a video segment starting from the preceding random access picture. Three different scenarios can be differentiated. In the first scenario, a trade-off is made between segment start time precision and compression efficiency by choosing the random access interval. Only one version of the video stream is stored and the start time of an arbitrary temporal segment is changed. Secondly, a more flexible solution can be provided with a scalable representation of the video stream. With a scalable video stream, one of the layers can be provided with a high frequency of random access pictures, but at a reduced quality. A second high quality layer can be added with a lower random access frequency. Consequently, chances are higher that the segment can be selected from a more precise position, but with a small quality loss until the random access picture from the high quality layer is decoded. Finally, there is still the possibility of transcoding the video stream. The video serving system then decodes the video stream from the position of the preceding random access picture and encodes starting from the requested start position of the video fragment. Regarding the complexity of transcoding combined with the loose constraint related to the position for random access, transcoding is certainly not advisable.

Once again, a combination of such fragment extraction operations can be performed, which can be based on system requirements, personal preferences and (socially shared) metadata.

5 Conclusion

In this chapter, an overview of video compression technologies that enable highly personalised content distribution is given. Three key technology domains necessary for enabling universal multimedia distribution are described: video representation, video description and video adaptation and delivery.

For video representation, state of the art compression techniques are discussed that offer efficient distribution and scalability options for personalised delivery. The principles of scalable video coding are described in terms of the different types of scalability and inter-layer prediction mechanisms. Furthermore, the feasibility and usefulness of scalable HEVC is evaluated. For that purpose, HEVC has been extended to include quality scalability. When adding a scalable layer to an HEVC video stream an extra BD rate cost of 19.4% must be taken into account. When compared to simulcast HEVC, average BD rate gains of 20.6% can be obtained. Furthermore, 43.8% BD rate saving is realised by scalable coding based on HEVC compared to SVC.

In the context of video description, two kinds of metadata are introduced: structural and semantic. Structural metadata describes the low-level information about the multimedia content (e.g. the available scalability layers, the available tracks, and the location of the random access points). On the other hand, semantic metadata describes high-level information about the multimedia content (e.g. depicted objects, person who shared the video, and ratings of the video). Both types of metadata can assist in the adaptation processes required for personalised video delivery.

Finally, in order to provide each user the requested multimedia content, enabling the best possible experience, techniques and algorithms to adapt and deliver the multimedia content in an efficient way is discussed. Different adaptation techniques and possible adaptation axes are presented. Based on the video technology and metadata descriptions, a wide range of adaptation options become available, creating a good base for numerous applications, including enriched social media retrieval.

References

1. Ayers, D.: Review vocabulary. http://purl.org/stuff/rev
2. Bjøntegaard, G.: Doc. VCEG-M33: calculation of average PSNR differences between RD-curves. Tech. rep., ITU-T and ISO/IEC, Austin, USA (2001)
3. Bojărs, U., Breslin, J.G.: SIOC core ontology specification (2010). http://rdfs.org/sioc/spec/
4. Bossen, F.: Doc. JCTVC-F900: common test conditions and software reference configurations. Tech. rep., ITU-T and ISO/IEC, Torino, Italy (2011)
5. Brickley, D.: FOAF vocabulary specification 0.98 (2010). http://xmlns.com/foaf/spec/
6. Bross, B., Han, W.-J., Ohm, J.-R., Sullivan, G.J., Wiegand, T.: Doc. JCTVC-H1003: High efficiency video coding (HEVC) text specification draft 6. Tech. rep., ITU-T and ISO/IEC, San Jose, USA (2012)
7. De Cock, J., Notebaert, S., Lambert, P., Van de Walle, R.: Requantization transcoding for H.264/AVC video coding. Signal Process. Image Commun. **25**(4), 235–254 (2010)
8. Fernandez-Escribano, G., Kalva, H., Martinez, J.L., Cuenca, P., Orozco-Barbosa, L., Garrido, A.: An MPEG-2 to H.264 video transcoder in the baseline profile. IEEE Trans. Circuits Syst. Video Technol. **20**(5), 763–768 (2010)
9. Francois, E., Viéron, J., Burdin, N.: Doc. JVT-Q013: additional results on ESS evaluation. Tech. rep., ITU-T and ISO/IEC (2005)
10. Golbeck, J.: The Trust Ontology (2010). http://trust.mindswap.org/trustOnt.shtml
11. ISO/IEC 21000-7:2004: Information technology – multimedia framework (MPEG-21) part 7: digital item adaptation. Tech. rep., ISO/IEC (2004)
12. ISO/IEC 23009-1 (DASH): Dynamic adaptive streaming over HTTP (DASH) – part 1: media presentation description and segment formats. Tech. rep., ISO/IEC (2011)
13. ITU-T Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2): Generic coding of moving pictures and associated audio information – part 2: video. Tech. rep., ITU-T and ISO/IEC JTC 1 (1994)
14. ITU-T Recommendation H.264 and ISO/IEC 14496-10 (AVC): advanced video coding for generic audiovisual services. Tech. rep., ITU-T and ISO/IEC (2003)
15. Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG: Joint scalable video model. Tech. rep., ITU-T and ISO/IEC (2011)

16. Karczewicz, M., Chen, P., Joshi, R., Wang, X., Chien, W.J., Panchal, R., Reznik, Y., Coban, M., Chong, I.S.: A hybrid video coder based on extended macroblock sizes, improved interpolation, and flexible motion representation. IEEE Trans. Circuits and Syst. Video Technol. **20**(12), 1698–1708 (2010)
17. Lee, W., Bailer, W., Bürger, T., Champin, P.-A., Malaisé, V., Michel, T., Sasaki, F., Söderberg, J., Stegmaier, F., Strassner J. (ed.): Ontology for Media Resources 1.0. 'W3C Recommendation'. World Wide Web Consortium (2012). http://www.w3.org/TR/mediaont-10/
18. Li, B., Sullivan, G.J., Xu, J.: Doc. JCTVC-G399: comparison of compression performance of HEVC working draft 4 with AVC high profile. Tech. rep., ITU-T and ISO/IEC, Geneva, Switzerland (2011)
19. Longo, C., Sciuto, L.: A lightweight ontology for rating assessments, In: Proceedings of the 4th Italian Workshop on Semantic Web Applications and Perspectives (SWAP 2007), Bari, pp. 11–20 (2007)
20. Magalhaes, J., Pereira, F.: Using MPEG standards for multimedia customization. Signal Process. Image Commun. **19**(5), 437–456 (2004)
21. McCann, K., Sekiguci, S., Bross, B., Han, W.-J.: Doc. JCTVC-F802: HM4: HEVC test model 4 encoder description. Tech. rep., ITU-T and ISO/IEC, Torino, Italy (2011)
22. Notebaert, S., De Cock, J., Vermeirsch, K., Van de Walle, R.: Complexity and quality assessment of MPEG-2 to H.264/AVC intra transcoding architectures. In: 9th International Symposium on Signal Processing and Its Applications (ISSPA 2007), Sharjah, pp. 1–4 (2007)
23. Ostermann, J., Bormans, J., List, P., Marpe, D., Narroschke, M., Pereira, F., Stockhammer, T., Wedi, T.: Video coding with H.264/AVC: tools, performance, and complexity. IEEE Circuits Syst. Mag. **4**(1), 7–28 (2004)
24. Panis, G., Hutter, A., Heuer, J., Hellwagner, H., Kosch, H., Timmerer, C., Devillers, S., Amielh, M.: Bitstream syntax description: a tool for multimedia resource adaptation within MPEG-21. Signal Process. Image Commun. **18**(8), 721–747 (2003)
25. Schwarz, H., Marpe, D., Wiegand, T.: Doc. JVT-P014: hierarchical B pictures. Tech. rep., ITU-T and ISO/IEC (2005)
26. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans. Circuits Syst. Video Technol. **17**(9), 1103–1120 (2007)
27. Segall, C.A., Sullivan, G.J.: Spatial scalability within the H.264/AVC scalable video coding extension. IEEE Trans. Circuits Syst. Video Technol. **17**(9), 1121–1135 (2007)
28. Smith, J.R., Schirling, P.: Metadata standards roundup. IEEE Multimed. **13**(2), 84–88 (2006)
29. Troncy, R., Mannens, E., Pfeiffer, S., Van Deursen, D. (ed.): Media Fragments URI 1.0. 'W3C Recommendation'. World Wide Web Consortium (2012). http://www.w3.org/TR/media-frags/
30. Vetro, A., Christopoulos, C., Sun, H.F.: Video transcoding architectures and techniques: An overview. IEEE Signal Process. Mag. **20**(2), 18–29 (2003)
31. Wiegand, T., Sullivan, G.J., Bjøntegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. IEEE Trans. Circuits Syst. Video Technol. **13**(7), 560–576 (2003)

Social Aware TV Content Delivery Over Intelligent Networks

Francisco Fraile, Pau Arce, Román Belda, Ismael de Fez, Juan Carlos Guerri, and Ana Pajares

Abstract As explained throughout the book, the analysis of social media has a great potential in improving the Quality of Experience (QoE) of media services. In this context, video delivery over IP networks is one of the processes that could benefit from the analysis of the social aspects associated to broadband video services. This chapter presents network intelligence as a technology enabler to integrate social media analysis in the management of video content delivery in IP networks. More precisely, this chapter presents dynamic reservations to manage network resources in an intelligent way, anticipated reservations to guarantee network resources during a time interval and content caching to maximize the usage of bandwidth across the network. All these techniques allow network operators to benefit from social media, improving the QoE of users considerably.

1 Introduction

In recent years, video applications over IP networks have experienced an increasing growth and the weight of video traffic in IP networks is expected to continue rising in the years to come. According to the Visual Networking Index (VNI) of Cisco, in 2010 Internet video (IPTV, Internet video and video on demand) represented 40% of consumer Internet traffic, reaching 62% by 2015, not including P2P video file exchange [7]. Accounting for video P2P file exchange, this percentage goes up to 90%. This growth is pushing network operators to reconsider how video services are provided today and how video traffic is treated across the networks.

F. Fraile (✉) • P. Arce • R. Belda • I. de Fez • J.C. Guerri • A. Pajares
Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
e-mail: ffraile@iteam.upv.es; paarvi@iteam.upv.es; robelor@iteam.upv.es; isdefez@iteam.upv.es; jcguerri@dcom.upv.es; ana-maria.pajares@unirioja.es

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_17,

In this sense, the Quality of Experience is an important parameter to take into account, since users expect a content transmission of high quality, without interruptions or losses and available as fast as possible. Consequently, the network must comply with a series of constraints that guarantee successful video delivery across the network. For instance, the DVB-IPTV standard defines a maximum packet delay jitter of 40 ms peak-to-peak, a recommended maximum packet loss equivalent to one noticeable artefact per hour and a recommended maximum time to join a channel of 500 ms. Other television services, like content download, have different user requirements and the network needs to guarantee other constraints for them.

However, it is not trivial to provide TV services with QoE guarantees to a massive amount of users. Traditionally, Internet Service Providers (ISPs) do not make any distinction between services across their network. Video traffic is treated with the same best effort delivery service as the rest of the services, not taking into account its characteristics or requirements. Best effort does not guarantee any of the network constraints required by video service. Instead, the network handles traffic in the best possible way given the traffic conditions. One way to overcome the limitations of best effort for video services is overdimensioning the network. This alternative, widely used among ISPs, consists of increasing the network capacity until the requirements of video services are met in peak traffic conditions. Nevertheless, in the long run, this approach would only be viable if network capacity could increase as rapidly as video traffic demands, which is not the case [21].

Another alternative, deployed by IPTV service providers, is creating static reservations for linear television content delivery. With static reservations, network managers manually define static network paths of fixed capacity, named tunnels, used exclusively to deliver television services with guarantees. This solution meets the requirements of linear TV services, at the expense of additional network management costs and network infra-utilization, since it is not possible to use the bandwidth reserved for other tunnels even if there are no linear TV services that use it.

Due to the limitations of static reservations, ISPs are looking into enabling technologies that could boost the network performance for video services. Technologies such as network intelligence or programming networks provide means to control network hardware with software, to let applications dictate how network devices should treat the data flows passing across them. The objective of these applications is to improve the performance of the network, providing features like virtualization of infrastructure, dynamic resourcing, end-to-end performance optimization or end-to-end security.

In this sense, the social features inherent to mass media can be used by media networking applications to improve network resource utilization. For instance, people living close by are interested in the same local news or sport events. Everybody tends to watch TV at the same time frames and certain content items are more popular than others. Service providers have used this information in the management of video services over the network, but thanks to these technologies, the network itself can use this information to improve the QoE of IP television services and to make a better use of network resources.

This chapter will present social aware IP networking as a set of specific mechanisms in intelligent IP networks designed to profit from social behaviours of TV viewers and improve the QoE in terms of video quality, service availability and waiting times. More precisely, the following section will present the related work regarding the use of social aspects of media in content delivery mechanisms. Later, this chapter will explain the main concepts related to intelligent networks, using a typical IPTV network as reference. Following, this chapter will highlight the relationship between television services management and the social aspects of television, regarded as the influence of the service in society. Then, this chapter will show how these aspects can be accounted for in the management of television services over intelligent networks, through different use cases. Finally, the main points presented in this chapter are summarized in the conclusions section.

2 Related Work

The impact of social media has grown significantly in the last years. As discussed in the previous section, there are some aspects of the natural social connections between users and the way they consume media content that can be exploited to better assign resources in the distribution system. Social applications have merged with TV contents to give rise to social TV. Cesar and Geerts [4] explain how people socialize around television content and review the main existing applications for social TV, identifying key aspects of social interaction which are content selection, communication between users, community building and user status. This kind of information has been demonstrated to be useful to improve the usability and performance of content delivery systems. Pouwelse et al. [23] propose TRIBLER, a peer-to-peer (P2P) transmission system aware of users relationships based on the idea of treating users as social partners. This paradigm exploits social phenomena (e.g. friendship, communities) in order to improve content discovery, recommendation and downloading. Some works have made use of peer information to provide a peer-assisted video on demand (VoD) content delivery service [15, 24, 17], but without taking into account any information about the kind of content or the consumer habits. In addition, social networks can be used to obtain data regarding the interests and preferences from groups of users [19] and, consequently, identify and recommend interesting IPTV contents for users of the same group [16].

One step further in peer-assisted VoD is Zebroid [6], which is an architecture capable of gathering information from the IPTV operational data (e.g. popular content, measured bandwidth consumption of subscribers, busy hours of network) in order to pre-position popular content in customer set-top boxes. In this sense, the utilization of information about consumer habits in groups of users and content popularity, which could not be addressed easily in traditional TV broadcasting services, is regarded nowadays as an interesting mechanism to provide applications and delivery networks with a sort of intelligence. Zheng et al. [26] present a network protocol model that introduces a social aware plane. This new plane uses social

relationships and topological characteristics of social networks to enhance network protocols in every layer (e.g. routing, congestion control, security).

Most research works point in this direction. This fact is reflected in the existence of several research projects about this topic. There are different meaningful European projects (within the 7th Framework Programme) focused on network intelligence [9]. One of the most representative projects is COAST. This project creates a content-aware network of intelligent nodes, which classify content and identify web services and discover on line where services are located and content is cached, in order to optimally match users' requests with availability. On the other hand, the ENVISION project promotes the cooperation between different service providers to create a cross layer solution where network resources are dynamically mobilized to where they are most needed. Moreover, the way the content is accessed is adapted on the fly to what the network is able to deliver.

Regarding network technologies, the NoTube project proposes the use of IPTV to develop a service architecture based on semantic technologies, for personalized creation, distribution and consumption of TV content. Also, HBB-Next and P2P-Next aim to create a personalized network using hybrid and P2P networks, respectively.

Other projects focused on social content awareness are SARACEN and My-Media. The former proposes collaborative, social/context aware and scalable media distribution for optimizing the Quality of Experience and providing personalized media streaming services. MyMedia proposes an open source software framework to dynamically personalize the delivery and consumption of multimedia.

The aforementioned studies and works evidence the importance that social media and social aspects regarding video consumption have achieved nowadays as a research field.

3 Network Intelligence

Routers are the main components of IP networks. As it is well known, their main functions are to listen for packets of any routing protocol associated with them and to forward the packets according to the information provided by routing protocols.

The functional architecture of modern routers consists of two different planes, the control plane and the data plane. The control plane manages messages of routing protocols and generates a routing table: the set of rules that determine how packets should be forwarded in the data plane. The data plane, or forwarding plane, is where packets are actually forwarded from one port to another. Due to this separation, the forwarding plane just needs to deal with real-time packet processing and can be implemented with specialized hardware, which improves performance to a great extent. The data plane is responsible for performing fast forwarding but can also provide other features like deep packet inspection (DPI). However, since it is built on application-specific hardware, development of new features to run on this plane remains very hardware dependent.

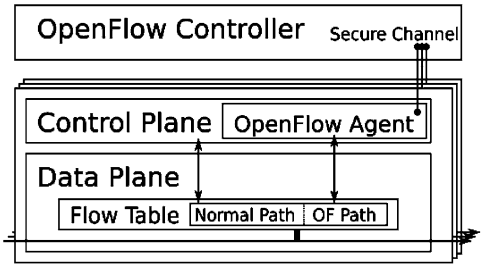


Fig. 1 OpenFlow Architecture

On the other hand, control planes can be built on general-purpose hardware. This characteristic greatly simplifies development of new features for the control plane. Nowadays, most routers (ranging from medium to high-end performance) implement this architecture for the sake of performance and easier software-hardware integration. Indeed, routers need to be able to handle different routing protocols, management tools and support added value services, and software integration in routers is a crucial design aspect. A relevant category of added value services focuses on traffic engineering (TE), i.e. improving the network resource utilization provided by common routing protocols.

Network intelligence provides extended functionality to networking infrastructure beyond packet forwarding. It is based on capabilities such as DPI and TE. Network intelligence is used to collect information about network utilization and perform bandwidth management, traffic shaping and usage-based billing among others.

Most manufacturers offer network intelligence products as dedicated hardware or software services. The possibilities of this technology are numerous, but commercial solutions are mainly limited to proprietary solutions that satisfy generic needs, common to many service providers. However, there are technologies that allow third-party developers to experiment with new concepts (OpenFlow) and to develop custom solutions (Junos SDK) such as the study cases described below.

3.1 OpenFlow

OpenFlow [18] is a protocol to remotely control routers through a secured connection. The remote controller machine (OpenFlow controller) manages all OpenFlow capable routers of the network by setting their flow tables. Each flow table entry contains a number of packet fields that packets have to match to pertain to the flow defined by the entry.

Figure 1 shows the OpenFlow architecture. The OpenFlow agent is an implementation of the OpenFlow protocol that runs on the control plane, and its main function is to translate OpenFlow protocol instructions to the flow table on the data plane. When the OpenFlow router receives a packet that does not match any flow entry in the flow table, the router inquires the controller for an action and the response is cached for future matching packets.

This architecture enables third-party controller development without compromising any implementation details of the router. The system is presented as a tool for researchers to test new protocols on production environments. OpenFlow makes it possible to split a production network into several independent slices, which enables testing new protocols and services. Each slice has its own view of the network, which can differ from one slice to the other. Hence, problems at any slice will not affect any other. Therefore, OpenFlow routers can handle production traffic with built-in network protocols on a slice separated from other potentially instable research or testing slices.

3.2 Junos SDK

Junos SDK (formerly the Partner Solution Development Platform-PSDP) is a development program that opens Junos OS (Junos Operating System) to third-party application development [8]. With Junos SDK, it is possible to develop applications that run side by side with built-in ones.

Junos SDK applications run on top of the Junos OS. Junos OS makes a logical division into three planes:

- The control plane, where the Junos OS runs. It runs on a hardware called Routing Engine (RE) and exposes APIs and services through the Routing Engine SDK.
- The data plane is built on ASIC-based hardware, referred to as Packet Forwarding Engine (PFE), and runs Junos OS microcode. To achieve the best performance, the data plane packet processing is generally *stateless*.
- The service plane is an extension of the data plane to perform *stateful* packet processing. The service plane runs on specialized hardware called Service Engine (SE). The Services SDK is used to develop applications than run at this plane. Some typical applications that run at service plane are *stateful* firewall or traffic monitoring.

Figure 2 shows the architecture of Junos OS, highlighting the three different kinds of traffic from the point of view of a router. Control traffic (CT) is the traffic associated with the router, consisting mostly of traffic generated by routing protocols. Transit traffic (TT) is composed by crossing traffic that is forwarded based on the routing table without any other processing. Finally, serviced traffic (ST) is forwarded through an SE for processing.

Applications running on the RE benefit from the rich environment that the Junos OS provides as it is based on FreeBSD. Any kind of network service or protocol can be developed as well as Command Line Interface (CLI) and Simple Network Management Protocol (SNMP) extensions among others.

Service applications run at SE. Service engines are hot-swappable modules without I/O ports. Their hardware consists of a multiprocessing and multithreaded CPU and sufficient memory to keep a high number of flow states. Service applications can monitor or even modify packets passing through the PFE.

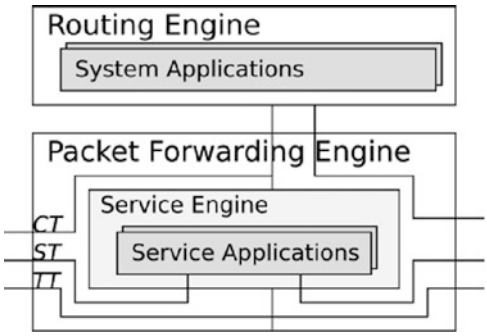


Fig. 2 Junos OS Architecture

3.3 Network Intelligence and QoE of Video Services

As described above, network equipment add-on applications can become important tools for network management. More precisely and regarding video services, intelligent routers can provide certain advantages for video management and add value to video services. It is worth noting that there are several kinds of video traffic and each one can have different network requirements. For instance, mean bandwidth is the most important parameter for progressive-download VoD services, whereas live TV streaming requires the maximum video rate to be guaranteed and minimize the jitter. A VoD client would not notice high jitter, as long as the mean bandwidth is sufficient to prevent buffer underrun.

Deep packet inspection provides routers with the ability to distinguish which kind of traffic is being forwarded, even the kind of video service. Moreover, traffic engineering allows for a more accurate treatment of packet flows and their respective priorities. These mechanisms turn into a proper management of video flows according to the service requirements, improving considerably the QoE perceived by the user. This way, network intelligence is useful not only to assign network resources efficiently and to improve the network performance but also to ameliorate the final user experience and satisfaction.

This video-aware IP networking framework can be further improved through the analysis of the social characteristics of the video (e.g. user ratings, popularity, etc.), in addition to the characteristics and the requirements of the video service. Provided that these social characteristics are available, as metadata associated to each video item, the network can use this information to better improve end-to-end network resource management. Additionally, the network can enrich this metadata since DPI allows tracking the audience of a content item across the network. This bidirectional exchange of information between the social plane of the content descriptions and the network is hereby referred to as social aware IP networking.

4 Social Aspects of TV Service Management

Originally, television was a highly time-dependent, hardly accessible medium. There were no means to store programs to watch them at a later time; broadcasting was expensive and so were television terminals. Therefore, viewers had limited and fleeting opportunities to watch the content they like and huge masses gathered to watch the few television programs available. These facts motivated the development of strong relationships between television as a service and the influence it had on society, known as social aspects of TV. Given the many ramifications of this area and the topic of this chapter, this section only focuses on social aspects accounted for in the management of television content delivery.

The most representative aspect is the relationship between the popularity of television content and its scheduling. Recalling that there were no other means to access content, popular programs gathered many simultaneous spectators and broadcasters used this to shape their program grids, placing popular shows in the day parts with more potential viewers and giving birth to the concept of prime time. Up to the date, popularity is the most important parameter for television service operators, since the viewership of programs has a direct impact on their incomes.

The way that popularity is measured depends on the topology of the service. Traditional television broadcast platforms lack a feedback channel from the user. For this reason, it was necessary to develop audience estimation systems based on other mechanisms, like telephone surveys or by installing audience measurement equipment on a representative sample of the population. Contrarily, IP-based TV services provide technical means to measure audience accurately. For instance, TV on demand portals run on content management systems (CMS) able to log the requests that are issued to each program. However, regardless of the service topology, audience tracking is used for positioning content: television operators must ensure that most popular content is easier to access. Hence, in linear TV program grids, most popular shows are scheduled on privileged time slots, whereas TV on demand services reserve the most visible areas of their sites for them.

Audience measurement is not only used to estimate the popularity of programs. Beside the audience ratings, audience measurement regards demographic aspects (e.g. gender, age, geographical proximity) and social aspects (e.g. level of education, cultural proximity) of the audience because these are also important business metrics for broadcasters, since they are helpful to better focus content to specific target groups. This practice, known as audience segmentation, improves the effectiveness of advertisement campaigns (including campaigns for related programs), increases user satisfaction and, in turn, helps at gaining audience in general terms.

In traditional broadcast TV services, audience segmentation can only be achieved by creating thematic channels or by configuring the program grid, for instance, to favour specific age groups. Then again, IP-based TV services offer more means to better implement audience segmentation. Since users do not necessarily share the same interface with the service, it is possible to adapt it according to the information about the user that is available to the service operator. There are different alternatives to obtain demographic information about a user, for instance,

explicitly when the user subscribes to the service. Social networks represent an outstanding source of information for TV service providers. Therein, users do not only provide explicit demographic or social information but in addition, there is a lot of implicit information, including information about social links between users. All this information can be used to narrow down the service segmentation to more homogeneous and smaller groups to which the content offer is adapted for.

Ultimately, when the segmentation strategy considers information about the users as individuals, service segmentation becomes service personalization. Personalization is nowadays a common feature in most popular video on demand websites, which implement recommenders to highlight content of interest for the user (as well as advertisements of interesting products). In this sense, the actual usage of the service is another relevant source of information for service operators. In fact, traditional audience surveys also regard aspects related to service usage, such as attention span or zapping habits. These data are taken into account when configuring program grids and even on content production. On the other hand, service usage information is becoming more and more important to assess the relevance of video on the Internet, which is very important for positioning. Also related to usage patterns, the increasing use of multiple screens and the interaction of television viewers with other services (multitasking) have become relevant socio-technical aspects for television service providers. In fact, television programs have additional services associated to the primary television service, for instance, social TV applications or complementary content like real-time statistics. These services are not necessarily bound to one terminal alone, resulting in additional requirements for television content delivery.

5 Social Aware IP Networking

The previous section provided an overview of different social aspects of television services and their relationship to television content delivery. As explained, knowing *who is watching* is very important for television service providers and social media is a great asset to get to know better the audience. This section will describe how social media analysis can improve the efficiency of television content delivery over intelligent networks. The section will cover three different applications of network intelligence aimed at improving the efficiency of video content delivery. In these practical cases, social media analysis can provide additional information to support the applications.

5.1 Dynamic Reservations

Nowadays, one of the most common technologies used by network operators in core networks is MultiProtocol Label Switching (MPLS) [2]. MPLS employs labels for configuring label-switched paths (LSP) and for forwarding IP packets at a very

high rate. In order to generate these paths, labels must be distributed among the routers. On a traditional IP/MPLS network, traffic will flow through the shortest path across the core network. If congestion arises, packets will be dropped, and the end user experience will be severely degraded. Since the shortest path is not always the best solution, the Resource Reservation Protocol with Traffic Engineering (RSVP-TE) [1] can be used instead as the signalling protocol that generates the end-to-end paths. RSVP-TE extends RSVP allowing the explicit establishment of LSPs taking into account network constraint parameters such as packet delay, jitter and available bandwidth.

In order to provide IPTV services and guarantee the video quality, network operators reserve bandwidth for the television services, by means of establishing static tunnels of fixed capacity for all the channels. Even though service providers are using RSVP-TE to configure paths, the mechanisms to guarantee video traffic requirements are based on manual and static configurations of traffic classification and queuing, with very basic admission control systems.

Murcia et al. [20], show that a reasonably better network bandwidth utilization in IPTV or general video delivery networks can be obtained with an automated reservation mechanism. In this sense, some additional intelligence is required so that the network can offer the necessary transport behaviour for video traffic and guarantee the Quality of Service (QoS). Therefore, Murcia et al. [20] propose a Video-Aware IP Architecture, which is able to manage MPLS tunnels according to actual video bandwidth needs and current traffic conditions. The solution presented in that paper differs from others in that the allocation (and also deallocation) of MPLS tunnels is based on an automatic and adaptive scheme driven by the application requirements and RSVP-TE. That is, by creating new tunnels automatically, this architecture can assure that new incoming video flows will have the required bandwidth guaranteed, as long as the total amount of bandwidth reservation does not exceed the total network capacity. The solution is focused on a practical environment such as IPTV networks, but the idea of MPLS Application-Aware TE is valid for any kind of content.

The dynamic reservation solution relies on the capacity to program the network intelligence. The key feature in that solution is to provide intelligence to the Provider Edge (PE) routers so they can be aware of video traffic. This solution makes use of a daemon installed in the ingress routers of the core network. This daemon provides PE routers with some knowledge of the incoming video flows entering the core network. In addition, it establishes a communication with a management proxy and dynamically reserves tunnels that fulfil the service requirements (such as bandwidth or packet jitter).

Figure 3 shows the Video-Aware IP Architecture components and the communication between them. The aim of this scheme is to define a session control that enables automatic resource reservation when a user requests a video. From this video request, which contains a session description, the session control starts the necessary processes to establish a QoS path in the MPLS network. This mechanism is automatic and considers the possibility to adapt the video flow to the available bandwidth if the usual process does not obtain a successful path.

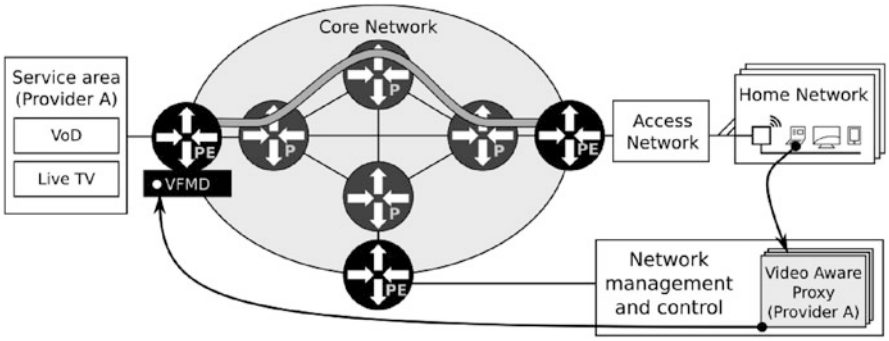


Fig. 3 Video-Aware IP Architecture

The process starts when the user requests media content to the Video Service Provider. This request is routed by the core network and arrives at the network management and control plane. In this case, this block is formed by one or more Video-Aware Proxies (VAPs). There can be as many as Video Service Providers.

The VAP behaves like a Real-Time Streaming Protocol (RTSP) relay, i.e. it handles the media request received from the client and forwards it to the streaming server. Bandwidth resources are allocated before the video server sends the media stream. For this purpose, the VAP sends a request to the PE router, which is able to allocate the resources pointed out in the exchange of RTSP messages. This router is the key element in the architecture, since in this router resides the main part of the video-aware intelligence. The router contains a Video Flow Management Daemon (VFMD) in charge of requesting bandwidth reservations when there is a request from the VAP. It is also in charge of the creation of firewall rules to send the video flows through the tunnel established with QoS guarantees.

This way, the PE router creates the label-switched paths and forwards video flows accordingly. These tunnels have been allocated with the necessary bandwidth so that the Video Service Provider can deliver the media stream to the client with the required QoS. Once the tunnels are created, the PE router builds a flow information database to manage all video flows. Thus, the router has an up-to-date picture of the resource reservations of the video traffic that enters to the core network through it.

Moreover, the platform can support simulcast operation mode. In contrast with normal operation where the client chooses the desired quality for media content, in simulcast mode, the quality is chosen by the VFMD according to the state of the network. Hence, if it is not possible to provide the best video quality, the VFMD will consider choosing a lower bit rate alternative of the video and allocate the resources according to the new requirements. This mechanism results in a better network utilization efficiency, since there might be resources that would not be reserved otherwise if video requirements were larger than the available resources. Consequently, more users can be served with minimum QoS guarantees, although there is still a trade-off between service availability and video quality.

Finally, it is worth mentioning that by managing resource reservation programmatically on PE routers within this scheme, new patterns regarding tunnel creation can be designed according to the social demand of certain video contents. As mentioned, audience measurement requires collecting usage data for a particular content. Through the use of proxy daemons on the edge routers, which inspect IP packets looking for video traffic, the network is able to provide very accurate audience measurement for video content, thus allowing the control plane to use this information to favour the delivery of popular content. Note that this information is complementary to any data obtained from the social networks and therefore, the decision-making process can be supported by the analysis of social media. In this sense, QoS tunnels can be created at certain time of the day with an approximate bandwidth allocation according to the estimated video demand. Due to the dynamism of this mechanism, non-priority resources can be deallocated to free resources for other kind of traffic, or reallocated later if needed. This concept is addressed in depth in the next subsection.

5.2 Advanced Resource Reservations

Throughout the years, numerous solutions have been proposed to offer QoS to real-time applications in computer networks. Nevertheless, recent developments, such as the massive consumption of multimedia contents, the possibility of accessing interactive services through hybrid TV networks (e.g. HbbTV) or the use of grid computing, bring attention back to the technique of advance reservations (AR) as a solution to provide a guaranteed QoS and to improve the QoE of the users. In addition, this technique could be greatly refined by using information about content popularity and consumption trends extracted from social networks, as proposed below.

Traditional transport services consider only immediate reservations (IR) requests. This approach is based on the reservation of resources along the path from the sender to the receiver just before the application starts sending data, i.e. reservations are immediate and remain established during the whole duration of the session. Furthermore, the admission control rejects the service request if there are not enough resources available to establish the session.

Although this model is suitable for many applications, many others, such as real-time applications, require a guaranteed transport service for a specific time and duration. For instance, there are different interactive TV applications that require bidirectional video communication with the spectators to let them participate in the shows from their homes. Unlike immediate reservations, with advance reservations, it is possible to pre-allocate network capacity for these scheduled events and provide the video with QoS guarantees.

On the other hand, the load of the core network in IPTV networks will greatly increase with the demand of unicast services such as VoD and nPVR (network-based Personal Video Recorder). In this scenario, it would be very interesting

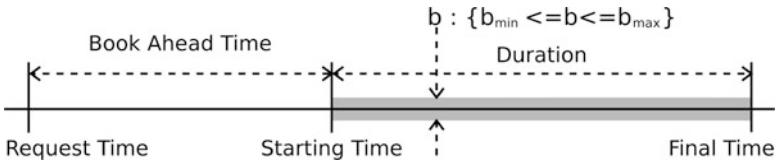


Fig. 4 Time parameters in advance reservations

to estimate usage patterns in advance and use this information for the network planning. In fact, it is only necessary to guarantee that the reservations are in place right before the service session starts. In order to guarantee these resources, it would be useful to allow reserving resources in advance for the required time through advance reservations (AR). This solution not only provides guaranteed resources but also allows service providers to make a better planning of the network resources. Furthermore, the analysis of the data available in social networks could be of great utility to estimate future usage patterns, because thereby users express their interests about content. Nowadays, it is quite common to analyse social networks in search of trending content. If these trends in social networks are mapped to the data managed by the network control plane, the utility of advanced reservations is improved.

Initial works, such as Ferrari et al. [12], Wolf and Steinmetz [25], Greenberg et al. [13] and Guerin and Orda [14], were focused on the fundamental requirements for supporting advance reservations in computer networks, whereas present works, such as Burchard [3], Escalona et al. [10] and Charbonneau and Vokkarane [5], are more focused on the use of AR in MPLS and optical networks. Independently of the technology on which the AR are applied, it is necessary to consider certain time parameters to carry out the reservation. Figure 4 shows the relationship between these times and AR.

The *request time* refers to the moment when a user requests a resource reservation, both with IR and AR. The *starting time* specifies the moment when the service session starts and therefore the deadline to make the reservations effective. The *duration* defines how long the session lasts and therefore, the minimum amount of time the reservation needs to be effective. The user has to specify an approximate time for this value. *Book ahead time* is the interval of time between the time when a user requests a reservation and the time when the resources are actually used. To avoid sessions starting far away from the time the reservation is set up or requesting resources for a very long time, the *book ahead time* and the *duration* have to be bounded. In this sense, there is a signalling overhead derived from the message exchange needed to keep the AR information updated across the network. Theoretically, longer *book ahead times* leave more room for traffic optimization, but there is a trade-off between the advantages of dealing with long *book ahead times* and the associated signalling overhead. Moreover, these temporal requirements are different depending on the applications. For example, a videoconference or a VoD session might be scheduled 1 week in advance, i.e. the *book ahead time* could be

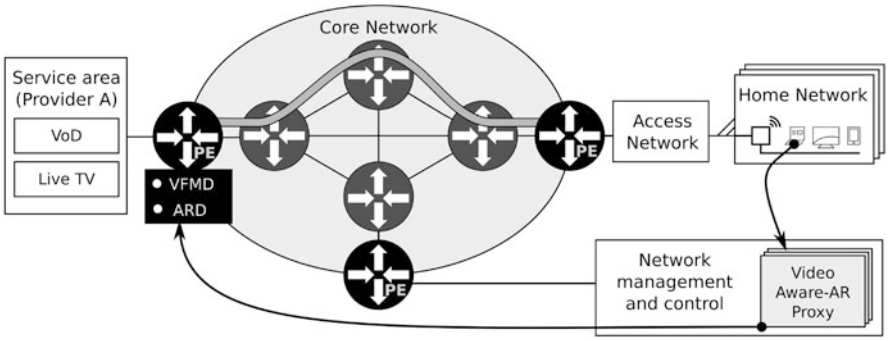


Fig. 5 Video-Aware Architecture with advance reservation

1 week and the duration could be 1 h. Also, *final time* indicates the instant when a session ends.

Parameter b represents the bandwidth assigned to the advance reservation. Scalable video codecs, such as H.264/SVC, allow adapting the bit rate based on the available bandwidth (rate-distortion parameter). The bandwidth range requirement ($b_{\min}$, $b_{\max}$) has to be included in the signalling interchanged between the user and the network.

On the other hand, Fig. 5 shows the architecture and the signalling needed between the user and the network. Most approaches use dedicated servers (centralized architectures) to handle the reservations in advance, and other proposals describe solutions based on session management in routers (distributed architecture). In router-oriented solutions, the state of the reservations of both IR and AR is stored in the routers and signalling (RSVP-TE), and routing (OSPF-TE) protocols must be extended to include time domain information and to maintain this state in the network.

The figure shows a centralized architecture in the control plane formed by a Video-Aware Proxy with Advance Reservation (VAP-AR) capabilities. The users or applications may communicate with this proxy through RTSP. A minimal modification in this protocol is needed. The RTSP message *SETUP Request* has been extended to incorporate the following parameters: type of reservation (IR, AR), *starting time*, *duration*, $b_{\min}$ and $b_{\max}$. The VAP-AR forwards the request to the Video Flow Management Daemon (VFMD) and the Advance Reservation Daemon (ARD), which have information about video flows and the Video Label-Switched Paths (VLSPs) that this MPLS router has currently established. The ARD uses the knowledge about IR and information about previous AR to accept the new request. Also, the ARD is in charge of sending RSVP-TE messages to reserve the bandwidth before the *starting time* and to tear down at *final time*.

In summary, in order to benefit from advance reservations, it is necessary to estimate the relevance of the content to be offered and the traffic that its delivery will generate in the network. This estimation can be made from the popularity of

the content in social networks and the feedback that the network can provide about the consumption of similar content across the network or the estimated prime times, as explained above.

5.3 Popular Content Caching

In the last sections, an estimation of the popularity of media content is used to infer the likelihood of (short-term) content consumption. This knowledge is used to improve network resource reservation management. This section will show how the same principle can be used to develop enhanced caching techniques. These caching techniques improve the network response for user requests and, in turn, improve the efficiency of network resources usage. Content caching can clearly benefit from information about content popularity and ratings extracted from social networks.

IPTV linear TV channels are multicast IP services, meaning that users watching the same channel are members of the same multicast group, forming a multicast tree from the service headend to the different egress routers that connect viewers to the core network. There are different technical solutions to implement such point-to-multipoint paths (P2MP) between the ingress router and the egress routers, mostly based on native IP multicast or MPLS traffic engineering methods. Regardless of the technology used to build the multicast tree and forward multicast traffic across the network, operators can use RSVP-TE to maintain the IPTV traffic encapsulated in MPLS tunnels.

Initially, the P2MP LSPs are statically configured to support the delivery of all IPTV channels with QoS guarantees. Consequently, the network manager configures the tunnels according to the channel offer, so that there is sufficient bandwidth to support the delivery of all channels in parallel. This way, the P2MP between the ingress router and the egress routers must be able to support a copy of each channel. However, the overall IPTV traffic does not always use the reserved capacity, leaving some residual capacity in the tunnels that can be exploited for other purposes.

First, the total number of channels requested by users may be smaller than the number of channels offered by the service provider. If this is the case, the multicast groups inside the tunnel do not use the reserved capacity to its full extent, leaving a residual reserved capacity in a P2MP between the IPTV source and all the egress routers in the core network. Normally, the probability of having all multicast groups inside the tunnels is very high, especially in prime time, but there are other times of the day, for instance, at night, when it is not likely that the reserved capacity is used to its full extent. At such times, it is also likely that the demand for best effort traffic is low, meaning that no other service could benefit from the available capacity in the core network.

Additionally, recall that with static reservations, the network manager needs to specify the amount of bandwidth that can be allocated for IPTV flows or for each flow in the tunnel. Given the variability that is characteristic of video traffic, in order

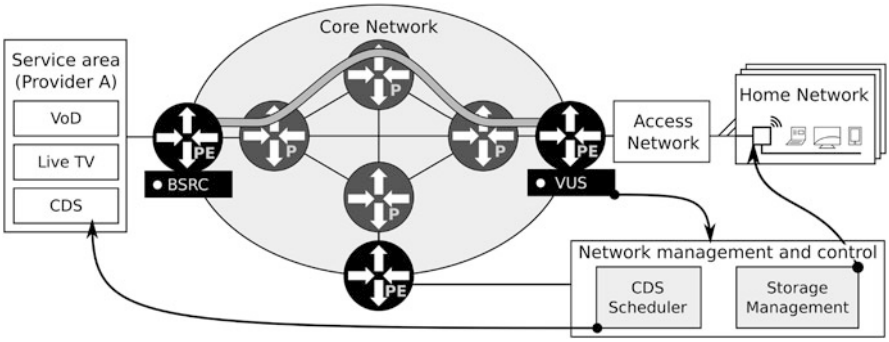


Fig. 6 CDS combined with network intelligence

to provide the QoS guarantees required by IPTV services – maximum packet jitter and packet loss – the allocated bandwidth is necessarily higher than the average bandwidth of IPTV traffic. Hence, there is some residual reserved capacity in the P2MP even if there is demand for all the IPTV channels offered.

This residual capacity left in the tunnels can be used to push television content from the service area to storage capacity closer to the user domain through a Content Download Service (CDS) bundled with the IPTV service, as suggested in Chen et al. [6]. Figure 6 shows how this can be implemented with intelligent networks. While in Chen et al. [6], the associated CDS uses a P2P content delivery protocol; this case study considers multicast content delivery, through the tunnels reserved for the IPTV service. As for the storage capacity, this proposal could use either network storage or remotely managed storage in the home network. For the sake of simplicity, the rest of the section will only regard the usage of network storage.

The key component in the architecture of the proposal is the multicast CDS used to fill the spare capacity in the tunnels. The multicast CDS in this proposal is based on the FLUTE protocol [22], which is the same protocol adopted by the DVB-IP standard [11] for the provision of multicast CDS. FLUTE is a multicast protocol for the delivery of files to massive amounts of users, and FLUTE content downloads are organized in sessions. The service provider needs to announce the parameters of the session (start time, end time, access information) in advance, since client applications need this information to initialize the download process. There are two different kinds of CDS sessions: scheduled sessions, which are bounded to a time period explicitly defined by the service provider; and carousel sessions, which are available in long periods of time, with no explicit indication about its start or end times. The latter seem more appropriate for background CDS because there is no need to define a time window for the session in advance.

As indicated in the diagram, the control of the CDS sessions is delegated to a CDS Scheduler function in the network management and control plane. The CDS Scheduler function determines the parameters of the session and, most importantly, the files to be delivered in the carousel session. Similarly, the control of the network

storage is delegated to the storage management function. This function determines which files must be kept at each network storage node.

In order to make the best out of the residual capacity in the tunnels, the management of the background CDS uses the information provided by the Video Usage Statistics (VUS) service running on PE routers. The VUS service provides information about the usage of IPTV services by customers, measuring the audience of the content offered through the television services. This audience measurement determines the popularity of content items behind each ingress router, taking into account the preferences of neighbouring users. With this information and applying information filtering techniques, the storage management function determines the content items that better match the preferences of clients and determines which files should be stored in each network storage node. Then, the CDS Scheduler includes the selected content items in the background CDS session so that they are delivered to the storage nodes using the residual capacity in the tunnels.

In this sense, the rate of the multicast CDS is adjusted by an application running on the egress PE routers, named Background Service Rate Control (BSRC) daemon. FLUTE sessions are organized in several multicast groups referred to as FLUTE channels. FLUTE channels are introduced to enable application level rate control. Thus, clients issue requests to join all the channels of a session and, if they detect congestion, they leave a number of channels, decreasing the overall rate of the session and consequently avoiding causing congestion. The same principle can be applied by the BSRC to control the rate of the background CDS session. This way, the BSRC daemon monitors the capacity available in the tunnel and adjusts the service rate accordingly. This avoids sending FLUTE channels that are later discarded by the admission control feature of RSVP-TE.

This covers the basics of managing a background CDS that uses the residual capacity left in the tunnels reserved for IPTV services. This kind of background service can have different applications. First, it can be used as a very efficient mechanism to preload a Content Delivery Network (CDN) for a VoD service within the ISP domain. By monitoring the usage of associated services, background CDS management functions make sure that the content pushed to the storage nodes is very likely to be requested to the origin server. Other possible application is audience segmentation, since the content pushed to network storage can be used as alternative content to the linear IPTV program grid. This use case requires that the media players running in customer premises equipment can seamlessly switch between the IPTV session and a session with a network storage node.

6 Conclusions

This chapter has shown how the use of information obtained from social media can improve IPTV content delivery over intelligent networks.

IPTV service providers aim to offer more interesting content with better video quality, thus improving the QoE of the service. This improvement must

be cost-efficient and profit from the available resources on the network, avoiding unnecessary investments or overdimensioning the core network.

In this sense, this chapter has described the operation of core routers, emphasizing the way they treat the video traffic nowadays. In addition, this chapter presented the concept of network intelligence and the benefits brought by the dynamic management of network resources, specifically for video distribution services.

Furthermore, three use cases that share a common architecture have been presented. All of them use network intelligence technology and programmable network mechanisms to improve video distribution over IP networks. These solutions can profit from input data regarding social aspects in order to make use of the available network resources in a more efficient manner.

By combining social awareness and network intelligence, a wide range of possibilities is opened for creating new services and improving the efficiency of existing ones.

References

1. Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., Swallow, G.: RSVP-TE: Extensions to RSVP for LSP tunnels. IETF RFC 3209 (2001). http://tools.ietf.org/html/rfc3209
2. Bayle, T., Aibara, R., Nishimura, K.: Performance measurements of MPLS traffic engineering and QoS. In: Proceedings of ISOC INET. Stockholm, Sweden (2001)
3. Burchard, L.O.: Networks with advance reservations: applications, architecture, and performance. J. Netw. Syst. Manag. **13**(4), 429–449 (2005). doi:10.1007/s10922-005-9004-7
4. Cesar, P., Geerts, D.: Understanding social TV: a survey. In: Proceedings of the NEM Summit 2011. Torino, Italy (2011)
5. Charbonneau, N., Vokkarane, V.: A survey of advance reservation routing and wavelength assignment in wavelength-routed WDM networks. IEEE Commun. Surv. Tutor. **99**, 1–28 (2011). doi:10.1109/SURV.2011.111411.00054
6. Chen, Y.F.R., Jana, R., Stern, D., Wei, B., Yang, M., Sun, H., Dyaberi, J.: Zebroid: using IPTV data to support STB-assisted VoD content delivery. Multimed. Syst. J. **16**(3), 199–214 (2010). doi:10.1007/s00530-010-0184-y
7. CISCO: Cisco visual networking index: forecast and methodology, 2010–2015. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html (2011). Accessed 29 Feb 2012
8. Clemm, A., Wolter, R.: Network-Embedded Management and Applications. Springer, New York (2012)
9. CORDIS (Community Research and Development Information Service): http://cordis.europa.eu (2012). Accessed 29 Feb 2012
10. Escalona, E., Spadaro, S., Comellas, J., Junyent, G.: Advance reservations for service-aware GMPLS-based optical networks. Comput. Netw. **52**(10), 1938–1950 (2008). doi:10.1016/j.comnet.2008.02.026
11. ETSI: TS 102 034 V1.4.1 Transport of MPEG-2 TS based DVB services over IP based networks (and associated XML) (2009). http://www.etsi.org/deliver/etsi_ts/102000_102099/102034/01.04.01_60/ts_102034v010401p.pdf
12. Ferrari, D., Gupta, A., Ventre, G.: Distributed advance reservations of realtime connections. In: Proceedings of NOSSDAV. Durham, New Hampshire, United States, pp. 16–27 (1995)
13. Greenberg, A.G., Srikant, R., Whitt, W.: Resource sharing for book-ahead and instantaneous-request calls. IEEE/ACM Trans. Netw. **7**(1), 10–22 (1999). doi:10.1109/90.759312

14. Guerin, R.A., Orda, A.: Networks with advance reservations: the routing perspective. Proc. IEEE INFOCOM **1**, 118–127 (2000). doi:10.1109/INFCOM.2000.832180
15. Janardhan, V., Schulzrinne, H.: Peer assisted VoD for set-top box based IP network. In: Proceedings of P2P-TV. doi:10.1145/1326320.1326327 (2007)
16. Kim, S.C., Kim, S.K.: Personalized IPTV content recommendation for social network group. In: Proceedings of IEEE ICCE, pp. 469–470. doi:10.1109/ICCE.2011.5722688 (2011)
17. Liu, B., Cui, Y., Chang, B., Gotow, B., Xue, Y.: BitTube: case study of a web-based peer-assisted Video-on-Demand system. In: Proceedings of IEEE International Symposium on Multimedia, pp. 242–249. doi:10.1109/ISM.2008.88 (2008)
18. McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S., Turner, J.: OpenFlow: enabling innovation in campus networks. ACM SIGCOMM Comput. Commun. Rev. **38**(2), 69–74 (2008). doi:10.1145/1355734.1355746
19. Merani, M.L.: How helpful can social network friends be in peer-to-peer video distribution? In: Proceedings of IEEE International Conference on ICPADS, pp. 799–804. doi:10.1109/ICPADS.2011.68 (2011)
20. Murcia, V., Delgado, M., Vargas, T.R., Guerri, J.C., Antich, J.: VAIPA: a video-aware internet protocol architecture. In: Proceedings of IEEE International Conference on HPSR, pp. 140–145. doi:10.1109/HPSR.2011.5986017 (2011)
21. Nguyen, K.K., Jaumard, B., Agarwal, A.: A distributed and scalable routing table manager for the next generation of IP routers. IEEE Netw. **22**(2), 6–14 (2008)
22. Paila, T., Luby, M., Lehtonen, R., Roca, V., Walsh, R.: FLUTE – File Delivery Over Unidirectional Transport. IETF RFC 3926 (2004). http://tools.ietf.org/html/rfc3926
23. Pouwelse, J.A., Garbacki, P., Wang, J., Bakker, A., Yang, J., Iosup, A., Epema, D.H.J., Reinders, M., van Steen, M.R., Sips, H.J.: TRIBLER: a social-based peer-to-peer system. J Concurr. Comput. Pract. Exper. **20**, 127–138 (2008). doi:10.1002/cpe.1189
24. Suh, K., Diot, C., Kurose, J., Massoulié, L., Neumann, C., Towsley, D., Varvello, M.: Push-to-peer Video-on-Demand system: design and evaluation. IEEE J. Sel. Areas Commun. **25**(9), 1706–1716 (2007). doi:10.1109/JSAC.2007.071209
25. Wolf, L.C., Steinmetz, R.: Concepts for resource reservation in advance. Multimed. Tool. Appl. **4**(3), 255–278 (1997). doi:10.1023/A:1009684906050
26. Zheng, Q., Zhu, P., Wang, Y., Xu, M.: EPSP: enhancing network protocol with social-aware plane. In: Proceedings of IEEE/ACM International CPSCom, pp. 578–583. doi:10.1109/GreenCom-CPSCom.2010.64 (2010)

Distributed Media Synchronisation for Shared Video Watching: Issues, Challenges and Examples

Fernando Boronat, Rufael Mekuria, Mario Montagud, and Pablo Cesar

Abstract Current societal changes are transforming the way people retrieve, annotate and share media. While in the past users gathered together around media content, this has become an exception rather than the norm. As demonstrated by the popularity of social networking and personal communication tools, people expect the development of novel technologies that: help them connect with others (e.g. by tagging images of a friend from high school). One key challenge in this respect is to support synchronous communication between people separated in space. This chapter focuses exactly on that technologies and infrastructures for supporting social interactions between people while apart. In particular, it discusses the synchronisation aspects of distributed media consumption (TV, YouTube videos, games, photo albums). As part of the quality of experience (QoE), synchronisation is a key requirement for ensuring consistency of the media experience across locations. Starting with an overview of the research problem, the contribution of this chapter is to detail current and envisioned architectures for achieving what is commonly known as Inter-Destination Media Synchronization (IDMS), a topic drawing the attention of academy and industry alike.

F. Boronat (✉) • M. Montagud
Universitat Politècnica de València (UPV), Calle Paraninfo 1, 46730, Grao de Gandia, Valencia, Spain
e-mail: fboronat@dcom.upv.es; mamontor@posgrado.upv.es

R. Mekuria • P. Cesar
Centrum Wiskunde & Informatica (CWI), Science Park 123, Amsterdam 1098 XG, The Netherlands
e-mail: R.N.Mekuria@cwi.nl; p.s.cesar@cwi.nl

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_18,

Acronyms and Abbreviations

3GP	Third Generation Partnership Project or 3GPP file format
AM	Amplitude modulation
AMP	Adaptive media playout
ATSC	Advanced Television System Committee
AVTCORE	Audio/Video Transport Core Maintenance working group
BBC	British Broadcasting Corporation
CMTS	Cable modem termination system
CRT	Cathode ray tube
DASH	Dynamic Adaptive Streaming over HTTP
DCR	Degradation category rating
DCS	Distributed control scheme
DMB	Digital Multimedia Broadcasting
DTH	(ATSC) Direct to Home (Satellite)
DSLAM	Digital subscriber line access multiplexer
DTMB	Digital Terrestrial Multimedia Broadcast
DVB	Digital Video Broadcasting
ETSI	European Telecommunications Standards Institute
FM	Frequency modulation
GPS	Global Positioning System
HD	High definition
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IDMS	Inter-Destination Media Synchronization
IETF	Internet Engineering Task Force
IPTV	Internet Protocol Television
ISDB	Integrated Services Digital Broadcasting
ISP	Internet service provider
ITU	International Telecommunication Union
M/S	Master/slave
MPD	Media Presentation Description
MDU	Media data unit
MMSH	Microsoft Media Server HTTP streaming
MOS	Mean opinion score (MOS)
MPEG	Moving Picture Experts Group
MSAS	Media Synchronization Application Server
NOS	Nederlandse Omroep Stichting
NTP	Network Time Protocol
P2P	Peer to peer
QoE	Quality of experience
QoS	Quality of service
RFC	Request for Comments
RR	(RTCP) receiver report
RTCP	Real-Time Transmission Control Protocol

RTMP	Real-Time Messaging Protocol
RTP	Real-Time Transmission Protocol
RTSP	Real-Time Streaming Protocol
SD	Standard definition
SDES	(RTCP) source description report
SMS	Synchronisation maestro scheme
SPST	Synchronization Packet Sender Type
SR	(RTCP) sender report
Sync	Synchronisation
TISPAN	Telecommunications and Internet converged Services and Protocols for Advanced Networking
TV	Television
XR	(RTCP) Extended Report
VoD	Video on Demand
VoIP	Voice over IP
WG	Working group
XML	Extensible Markup Language

1 Introduction

Changes in the way people organise as a society are creating a demand for better communication and sharing tools. According to Eurostat, 17.6% of the population in Europe is single adults living alone,[1] and 32.5 million non-nationals live on the territory of an EU member state.[2] Similar situations can be found in other developed areas, such as the USA or Japan, where people's mobility has resulted in dispersed families and groups of friends. This situation has paved the way to social media as evidenced by statistics: 40% of the Skype[3] calls now use video and the time spent viewing video on social networking sites increased 98% year over year from 503.8 million minutes in October 2008 to 999.4 million minutes in October 2009.[4]

In response to societal changes, a number of socially-aware media services have emerged. They can be divided into asynchronous solutions that rely on uploads, updates and comments (e.g. Facebook, YouTube, Facehulu[5]) and more synchronous solutions that provide support for real-time communications (e.g. Google+ Hangout[6] and Yahoo!'s Zync[7]). This chapter will focus on the latter,

[1] http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-10-024/EN/KS-RA-10-024-EN.PDF

[2] http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Migration_and_migrant_population_statistics

[3] http://www.skype.com/

[4] http://mashable.com/2010/03/19/global-social-media-usage/

[5] http://techcrunch.com/2011/09/22/social-tv-gets-real-with-hulu-on-facebook/

[6] http://www.google.com/+/learnmore/

[7] http://sandbox.yahoo.com/Zync

where the goal is to enable social communication and real-time interactions between separated users. In particular, this chapter concentrates on techniques that allow distributed users to retrieve and consume media content together.

Commercially speaking, many services for distributed media watching already exist:

- Instant messaging solutions that are capable of embedding videos, such as Yahoo!'s Zync [1] and Windows Messenger.[8] In this case, users can launch a video during a conversation, watching it together with others.
- Web-based virtual viewing rooms, such as Justin.tv,[9] YouTube Social[10] and ClipSync.[11] Using these applications, remote users can communicate with others, while synchronously watching media. In most of the cases, the communication channel is in the form of text chatting.
- Synchronous album sharing, such as Flickr photo session,[12] which allows users to share photo streams with friends in real time, thus simulating the old practice of inviting people home to show them photos from the past.
- And virtual living rooms, such as Alcatel's AmigoTV [2] and Motorola's Social TV [3]. These services enable a communication channel (audio, video or text) between distributed homes; so dispersed families can experience the important habit of watching television together while apart.

All the previous examples, and use cases, are representative of the radical transformation towards shared media consumption. They open new research questions related to the media and to the infrastructure. Regarding media, researchers are investigating content recommendation based on the social graph and on user interactions around media (e.g. likings, comments). Infrastructure-wise, research targeted to assure quality of experience (QoE) for distributed media consumption is gaining momentum. The contribution of this chapter is at the infrastructure level, focusing on one particular aspect key for assuring QoE: Inter-Destination Media Synchronization or IDMS. IDMS refers to the simultaneous synchronisation of the playout of one or several media streams at two or more geographically distributed receivers. While commercial solutions for remotely watching media together exist, they still use rudimentary IDMS solutions. Further research in this area is required to be able to provide users with virtual watching and living rooms, where they can naturally communicate with others, as they would do it if they were in the same location.

The structure of this chapter is the following: First, different technologies that can be used for delivering media content are described. Section 2, therefore, provides a complete overview regarding media delivery technologies including TV-based

[8] http://explore.live.com/messenger

[9] http://www.justin.tv/

[10] http://www.ytsocial.com/

[11] http://www.clipsync.com/

[12] http://www.flickr.com/photosession

transmission (e.g. Digital Video Broadcasting or DVB-T/C/S/H), streaming standards (e.g. Real-Time Transmission Protocol/Real-Time Streaming Protocol or RTP/RTSP) and Web-based solutions (e.g. Hypertext Transfer Protocol, HTTP, and Dynamic Adaptive Streaming over HTTP, DASH). This overview indicates, in each case, what the intrinsic limitations regarding IDMS are. Next, in Sect. 3, the user experience is discussed. Given that synchronisation in existing delivery mechanisms is poorly achieved, the effects of such de-synchronisation on the user experience are discussed. Based on a number of studies, it is shown that poor QoE leads to high rates of user dissatisfaction. Section 4, then, looks into possible solutions for assuring IDMS. In a survey-like style, it summarises existing research efforts in this direction, indicating in which situations a given mechanism is better suited. Section 5 describes an IDMS solution proposed by the authors for allowing concurrently synchronised media presentations across multiple locations. That solution is based on the use of RTP/RTCP standard protocols [4], and currently it is in the process of becoming an IETF standard [5]. Finally, the chapter ends with the conclusions of the work, providing some future challenges.

2 Technologies Used in Shared Social Media Retrieval

According to CISCO's predictions[13] (June 2011), by 2012 video will be up to 50% of the consumer Internet traffic, and by 2015 three trillion Internet video minutes per month will be viewed. Similar conclusions can be seen from other sources, such as Global Internet Phenomena's[14] study on Netflix's rising. The reality is that online video consumption is increasing at unprecedented rates. IPTV (Television services over IP), Web-based video-sharing systems such as YouTube and Web-streaming services such as Netflix[15] and Hulu[16] are becoming instrumental for allowing people to consume content outside the traditional television flow. Nevertheless, it does not seem that TV is dead (or dying). According to a recent Nielsen's report on consumer's habits,[17] watching TV at home continues to be the preferred way to consume video.

Given the diversification of options for consuming videos, this section provides an overview of existing delivery technologies, focusing on the limitations regarding

[13] http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html

[14] http://www.sandvine.com/downloads/documents/05-17-2011_phenomena/Sandvine%20Global%20Internet%20Phenomena%20Spotlight%20-%20Netflix%20Rising.pdf

[15] https://www.netflix.com

[16] http://www.hulu.com/

[17] http://blog.nielsen.com/nielsenwire/online_mobile/what-consumers-watch-nielsens-q1-2010-three-screen-report/

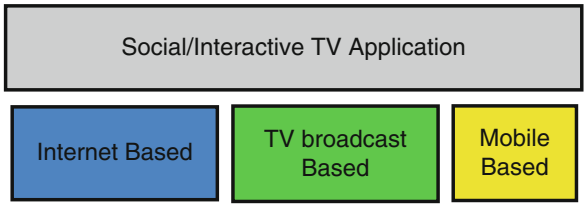


Fig. 1 Social TV for different types of broadcast media

synchronisation of the streams in distributed viewing situations. In particular, this section discusses media delivery technologies for digital TV broadcasting, for media streaming in the Web (e.g. YouTube) and for video delivery to smart phones and tablets. The assumption is that independently of the distribution technology used, shared experiences should be supported. Figure 1 shows this assumption, where shared video watching is possible using different delivery technologies. To evaluate the impact of these different types of media distribution, particularly on synchronisation at the receiver side, this section describes the different technologies and highlights possible causes of delay. Previous research [25, 27] shows that these delays can accumulate to a significant difference (up to 5 or 8 s, depending on the technology). These results demonstrate that some form of IDMS is a requirement for supporting distributed video watching.

2.1 Digital TV Broadcast

Television, the traditional video distribution medium, has lately undergone some changes. In particular, most countries have switched from analogue to digital broadcast. The main advantages of digital TV broadcasting include better picture quality, better coverage, ease of use and reliable reception. Moreover, digital broadcasting is a fast, efficient and scalable way to distribute high-quality media to many recipients. However, some disadvantages exist regarding social and distributed interactive applications. First, a return channel is not always available, and if so, the capacity tends to be small. Second, digital TV receiver modules are not yet common on laptops, smart phones and computers. A third disadvantage of digital TV broadcasting is that different operators generally use separated networks and different technologies. When content is distributed in different ways, synchronisation between recipients cannot be guaranteed. The lack of a common platform/network also restricts a possible synchronisation solution.

To better understand how delay is introduced in digital TV, a look at the technical characteristics is needed. The main aspects of digital TV are digital source coding (compression) of the audio and video data (i.e. MPEG-2) and the use of digital modulation techniques for better and more robust spectrum utilisation. This constitutes a difference with analogue broadcasting, where the audio/video signal is directly modulated on the carrier wave using amplitude modulation (AM)

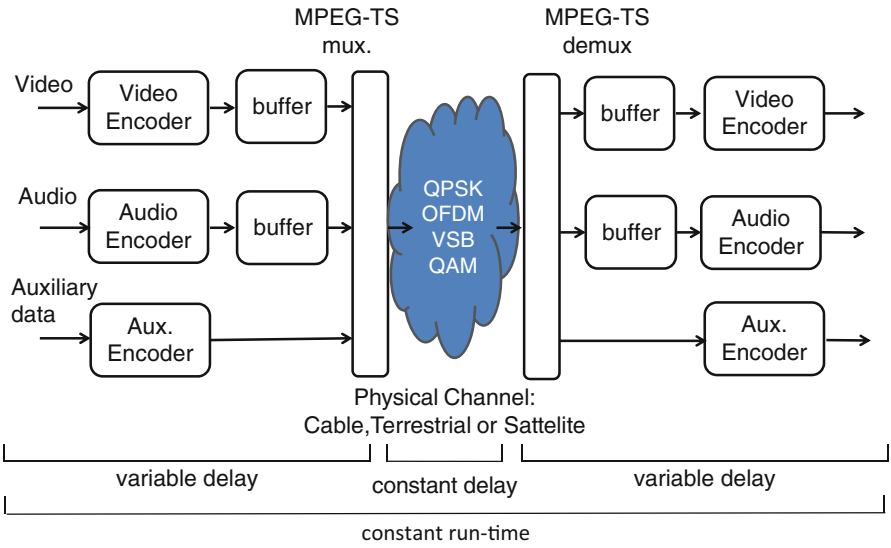


Fig. 2 End-to-end delay of typical digital TV system based on [9]

or frequency modulation (FM) techniques. To outline digital TV technology, a schematic diagram of the end-to-end delay in a typical digital TV system is shown in Fig. 2. The video and audio signals are first encoded using, respectively, MPEG-2 video [6] and audio [7] encoding techniques and are multiplexed into an MPEG-2 transport stream (TS) [8]. MPEG-2 TS is a stream of bytes that can contain multiple TV programs, audio and video media types. Auxiliary data such as program guides, information about the current program, are also multiplexed into the MPEG-2 TS. Subsequently the MPEG-2 TS is error coded, modulated and propagated on its physical medium, illustrated in the middle block. The received transport stream is demultiplexed so that audio, video and auxiliary data can be decoded and output synchronously according to the transmitted time. The mechanisms to synchronise the audio, video and auxiliary data are defined in MPEG-2 part 1 [6]. Figure 2, based on the DVB standard, shows that the total broadcasting and channel delay can be considered constant, even when delays at encoder and decoder may vary [9]. The delays in digital processing can accumulate to a significant difference compared to a signal transported using analogue techniques. While the initial digital TV standards (DVB, Integrated Services Digital Broadcasting or ISDB and Advanced Television System Committee or ATSC) used MPEG-2 video coding, in 2003 the Advanced Video Coding standard H.264/AVC [10] was released that offers similar quality compared to MPEG-2 with bit-rate savings up to 50% [11]. This standard is now commonly employed, as H.264 provides a compatible transmission format to adapt to the MPEG-2 TS format from digital TV (network abstraction layer units).

Table 1 Overview of digital TV standards worldwide

Organisation	Native source coding	Terrestrial	Cable	Satellite	Region
DVB	MPEG-2	DVB-T DVB-T2 DVB-H	DVB-C DVB-C2	DVB-S DVB-S2	Europe, India, Africa, Australia
ISDB	MPEG-2	ISDB-T ISDB-H	ISDB-C	ISDB-S	Japan, South America
ATSC	MPEG-2	ATSC Terrestrial ATSC-M/H	ATSC Cable	ATSC Direct-to-Home (DTH) Satellite	North America
DTMB	MPEG-2, H.264	DTMB-T/H	N/A	N/A	China
DMB	MPEG-2	T-DMB	N/A	S-DMB	South Korea

The H.264 standard defines different profiles offering a trade-off between latency, coding efficiency and error resilience suited to the particular application. For broadcasting applications over satellite, cable or terrestrial, the main profile of H.264 is recommended [11], which is not optimised for low delay. Apart from delay in the H.264 main profile, other coding delays can be introduced in the process in format or resolution conversion and in insertion of logos and subtitles. Other delays are introduced in the transmission, modulation and error correction. Such delays can accumulate, resulting in a delayed presentation time.

An overview of current digital TV standards worldwide is given in Table 1, based on information from worldwide standards from [12], DVB [9] and the Chinese standard Digital Terrestrial Multimedia Broadcast (DTMB) described in [13]. In Europe, a family of standards, DVB was developed. The ISDB standard originates in Japan, and it is also deployed in South America. ATSC is used in North America, while South Korea has its own TV standard, Digital Multimedia Broadcasting (DMB). The terrestrial standards generally have both a T version and an H version. DVB-H is geared at receivers that are mobile and require low power, while DVB-T aims at stationary receivers with sufficient power. It is difficult to predict the extra (relative) delay incurred in a DVB-H scheme compared to a DVB-T scheme based on the technology alone. Therefore, measurements of such relative delay in a real scenario are presented at the end of this section.

The current standards, while optimised for various goals such as compression, quality, coverage area and mobility, do not envision end-to-end synchronisation between different recipients. The advanced compression, digital modulation and error coding techniques can introduce delay that accumulates and has a significant effect on the QoE for distributed shared video watching.

A pilot study of the synchronisation between recipients of different broadcasting technologies can be found in [25, 27]. The results show local differences up to 5 s in one single geographic region. While such measurements may vary between countries' and regions' technologies, the results show that relative differences occur and can influence QoE in distributed video watching scenarios.

2.2 *Internet-Based Technologies for Watching Video*

While digital TV is still popular, digital video is watched online as well. Therefore, this section describes some of the related technologies: HTTP, RTP and Adobe Flash Video. Even though the Web infrastructure is mostly geared towards text documents, the HTTP protocol [14] can also be used for downloading videos. One can download a video file from an HTTP server, and when a certain part of the file is complete, start playing it. Systematically applying such an approach is referred to as progressive download and is employed by some media players/websites. While simple, progressive download bears some disadvantages. First, a live experience cannot be provided, due to the needed download time. Second, storing entire files on disk can be undesirable, especially if small storage capacity is available and the video file is large. Third, in case the bitrate goes down, content adaption is not supported, causing buffer underflow and paused video (i.e. playout discontinuities or interruptions), degrading the QoE. In order to address these disadvantages, novel adaptive HTTP streaming standards have been developed, such as HTTP streaming from Apple[18] and Microsoft.[19] This section focuses on 3GP-DASH (Third Generation Partnership Project or 3GPP file format DASH) [15], since it is an open standard that provides similar functionality. HTTP-based streaming solutions have some advantages. They avoid problems with firewalls and allow HTTP-based caching to be reused. Another advantage is that functionality can be moved to the client, reducing the server load and improving scalability. Figure 3 shows the basic operations supported by 3GP-DASH. The recipients initiate the session by requesting the media presentation description (MPD) file from an HTTP server. The MDP file is an Extensible Markup Language (XML) file that contains a hierarchic summary of the contents of the presentation. In 3GP- DASH, the presentation is composed of different periods that, in turn, are composed of different representations (codec, language, resolution), which again consist of different segments that can be individually downloaded using HTTP requests. Since the client now has the MPD file, it can, depending on its preference (seeking, available bandwidth, losses), request the segments using HTTP. By buffering at the client, the media can be continuously rendered. 3GP-DASH also supports live transmission, since the contents of the MPD file can be updated in real time. For more details about 3GP-DASH, the reader can consult the survey in [15]. It is important to highlight that the use of HTTP combined with TCP protocol generally introduces some extra delay due to TCP retransmissions and HTTP-based overhead.

Regarding Internet-based streaming, the most popular protocol is RTSP, as defined in [16]. RTSP provides support for controlling a media session, including setup, pause, play and stop functions (see Fig. 4). First, a description file is requested possibly via an HTTP GET command, similar as in 3GP-DASH. Then, from that

[18] http://developer.apple.com/resources/http-streaming/

[19] http://www.iis.net/download/smoothstreaming

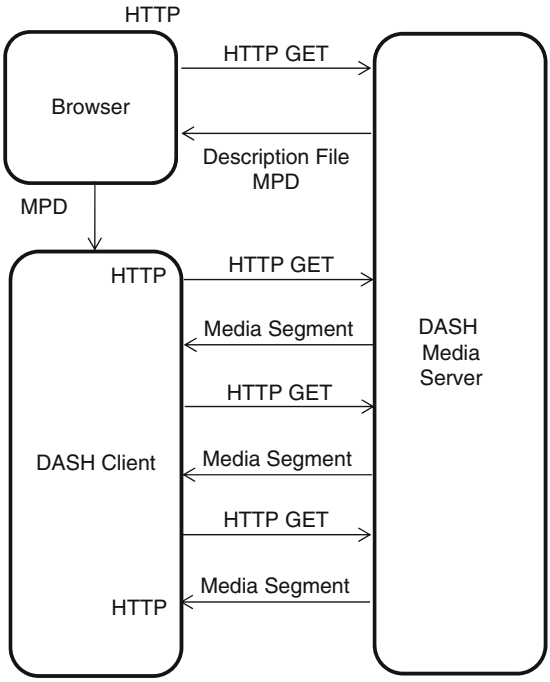


Fig. 3 3GP-DASH sequence diagram

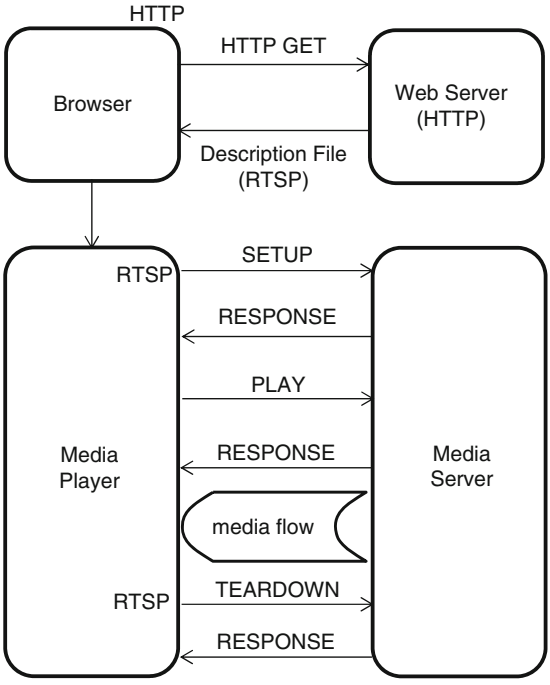


Fig. 4 RTSP sequence diagram

presentation description file, a URL to a single RTSP location at a media server is obtained. By issuing an RTSP setup request to that media server, a session with the server is started, and the media streams flow continuously from the server. In practice, the real-time protocol RTP [4] is often used to deliver these streams. The client can now issue RTSP play and pause commands to the RTSP media server and control the media session. The main advantages of RTP/RTSP based streaming are the following: less delay and more efficient bandwidth usage as no retransmission occurs. In case of distribution over RTP, the delay is mainly originated from IP network delay and jitter. Values for delay in the Internet from Internet service provider (ISP) data tend to range between 20 and 500 ms and between 0 and 500 ms jitter (variations in delay) as reported in the International Telecommunication Union (ITU) standard G.1050 [17]. Commonly in RTP, a buffer is used to smooth out the effects of jitter, at the cost of some extra delay. In the case of distributed video-sharing applications, the RTSP pause/play control functionality should be implemented to maintain synchronisation between receivers. Regarding distributed shared video watching, the use of RTP/RTSP is desirable as lower delays are expected.

A popular website on the Internet for watching video clips is YouTube. YouTube uses the Adobe Flash Video player and, more recently, HyperText Markup Language (HTML) 5 with WebM (VP8) or H.264. The delivery mechanisms employed in Adobe Flash Videos are based on the proprietary protocol RTMP (Real-Time Messaging Protocol) that bears similarities to RTP and RTSP. The specification of this protocol is available on the Adobe site.[20] Earlier Flash video versions support progressive download techniques, as described in the beginning of this section. Another alternative for watching online videos is to use Peer-to-Peer services, such as SopCast [18].

2.3 Mobile Media Delivery Technologies

Mobile video, using small portable devices, emerged in a later stage. It uses either Internet protocols or TV-like reception techniques. In mobile communications, limited bandwidth is often available that tends to fluctuate, making bit-rate adaptation desirable. 3GP-DASH or similar technologies, such as Apple's live HTTP streaming (employed on the iPhone YouTube app), provide such adaptation. Examples of mobile standards that allow broadcast-based watching of television on a (moving) mobile device are DTMB and DVB-H/ISDB-H, which offer high quality of reception. For example, using mobile broadcast, devices only receive media during slotted intervals of time for saving power and use different coding techniques. It is expected that such solutions introduce extra delay, but, based on technology, it cannot be predicted. Instead, in the measurement subsection DVB-H is compared with other DVB technologies.

[20] http://www.adobe.com/devnet/rtmp.html

3 User Experience

The motivation of this chapter is the support of distributed shared video watching or synchronous Social TV [19]. Some examples include virtual living room applications (e.g. Alcatel's AmigoTV or Motorola Social TV) and Web-based virtual viewing rooms. In all these cases, people in different locations watch the same TV or video content together, interacting with each other via voice or text chat. The next question to be answered is how delivery delays affect the QoE. This section aims to answer this question by focusing on three main topics:

1. The difference that people will notice or get annoyed by lack of synchronisation in a Social TV application and on what factors this is based.
2. If people that are synchronously connected feel more together as when they are not and on what factors they are based.
3. The importance of this effect compared with other factors relevant in such applications.

To answer these questions, a number of Social TV tests are presented, both in the home environment, with different TV types and based on online video. Subsequently, the results of a 36 people user trial carried out in a controlled environment that focused explicitly on testing the effect of de-synchronisation are discussed as well. By combining these results with the previous results from Sect. 2, the critical cases where IDMS is required for shared distributed watching are defined.

3.1 Benefits of Watching Together While Apart

Motorola performed an early experiment that investigated the benefits of shared distributed watching at home [3]. In a long field trial, friends living in five homes got installed one Social TV device, which allowed them to have voice and text communications while watching TV. QoE was evaluated by conducting multiple interviews with the participants. One key conclusion was that the bond between the friends had strengthened by the end of the study and that they felt more connected to each other because of using this application. A larger in-home field trial for Social TV conducted in 2007 [20] confirmed similar conclusions. Comparison of data from 50 households without Social TV functionality (baseline) with households with Social TV functionality showed that Social TV users felt more connected. Such results were in a later stage confirmed to hold for friends and strangers by tests in a controlled environment [21].

Oehlberg et al. presented a practical study comparing watching TV together in two distributed groups with watching in one large group [22]. They observed that distributed group viewing bore many similarities with co-located shared viewing. They observed that in both cases, conversations evolved mainly around program

Table 2 Questionnaire feeling of togetherness and connection between participants

1. The contact with my conversation partner was superficial (1–7)
2. I felt together with my conversation partner (1–7)
3. I felt that my conversation partner did not understand me well (1–7)
4. I felt connected to my conversation partner (1–7)
5. I got little satisfaction out of the conversation (1–7)
6. I felt my conversational partner and I could talk well to each other (1–7)

Table 3 Questionnaire ITU-T DCR adapted for synchronisation

Score	Impairment
5	The playout difference is not perceptible
4	The difference is perceptible but not annoying
3	The difference is perceptible and slightly annoying
2	The difference is perceptible and annoying
1	The difference is perceptible and very annoying

contents and occur at silent periods. The amount of conversation was also similar in both cases, and in both cases visual interaction was limited. Such observations hinted in the direction that synchronisation is important to guarantee the QoE in shared TV watching, as communication patterns could be broken.

3.2 User Test: Synchronisation in Distributed Shared Video Watching

In this section we report the results from a controlled user study to test the effects of de-synchronisation, when users watch a video from remote locations. Developing such test poses some challenges as many aspects can have an effect on the user experience. A first consideration was to choose the appropriate genre that encourages conversations. In this respect, a locally popular quiz was chosen (encourages conversation according to [23]). To artificially change the synchronisation level, an algorithm capable of synchronizing within 60-ms bounds was used. The specific synchronisation conditions were defined as 0, 500, 1,000, 2,000 and 4,000 ms representing the range of values in digital TV. Moreover, it was expected that the threshold value when people start to notice should be within this range. To reduce habituation effects, the order of different conditions was randomised. A broad audience of participants was recruited: 18 couples of friends, partners or family of various ages and educational backgrounds. Moreover, as in distributed shared watching the use of chat or voice changes the user experience (QoE), both were included in the tests. To assess the QoE we used both a standard QoE DCR (degradation category rating) scale defined by ITU [24] and a list of six questions on the feeling of togetherness experienced by the participants. The questionnaires employed in the subjective tests are shown in Tables 2 and 3.

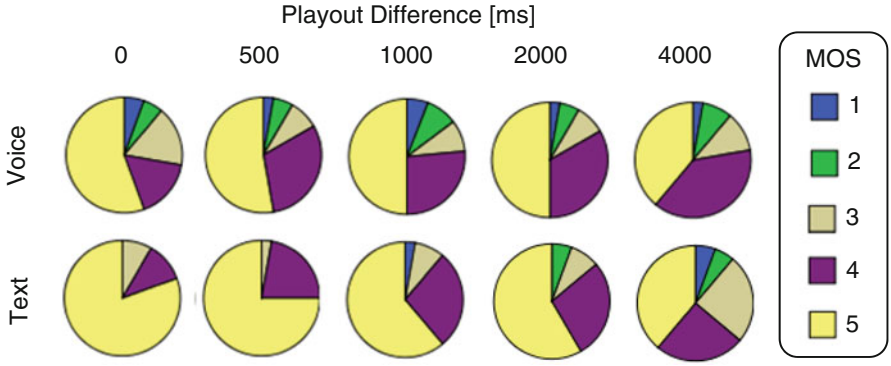


Fig. 5 Preliminary results ITU-T DCR opinion scores

Table 4 Mean and standard deviation of the results regarding the synchronisation condition (text chat)

	Synchronisation difference (ms)				
Parameter	0	500	1,000	2,000	4,000
Mean	4.0256	4.2894	4.1578	4.2368	4.0789
Standard deviation	1.1893	1.0109	1.1974	0.9982	1.0496

3.3 Watching Video Together: User Test Results

The experiments were carried out during 1 week at the *user experience lab* at the Catholic University of Leuven, Belgium. Figure 5 illustrates the results of the questionnaire from Table 4. It shows that most people did not notice the difference independently of the synchronisation difference. The mean opinion score (MOS) values were 4.1 for text chat and 4.5 for voice chat. To adapt to this data spread, it was aimed to find the perceptual thresholds of when people noticed or got annoyed by such differences.

The results based on whether people noticed or got annoyed are shown in Figs. 6 and 7. Voice chatters seemed to get annoyed and noticed, while the text chatters' responses looked random because in different synchronisation conditions, both the means and the standard deviations did not change much. The overall standard deviation was 1.084, while the average standard deviation per synchronisation condition was 1.089. The means in each condition did not change much compared to the overall mean of 4.17 as shown in Table 4. Statistical analysis regarding this experiment showed that overall text chatters do not notice synchronisation [25, 26]. In the studies performed, no differences based on factors like age, liking of the show, chat experience and frequency of chatting were found. Only the participants that were more active chatters (15 participants), with more than 400 words during the whole session, gave responses that correlated with the synchronisation condition compared to less active chatters (21 participants with less than 400 words). Figure 8 shows the differences between the groups.

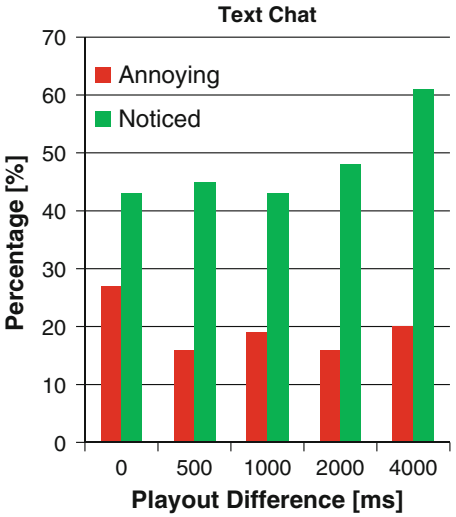


Fig. 6 Percentages of people either getting annoyed or noticing the playout differences (text chat)

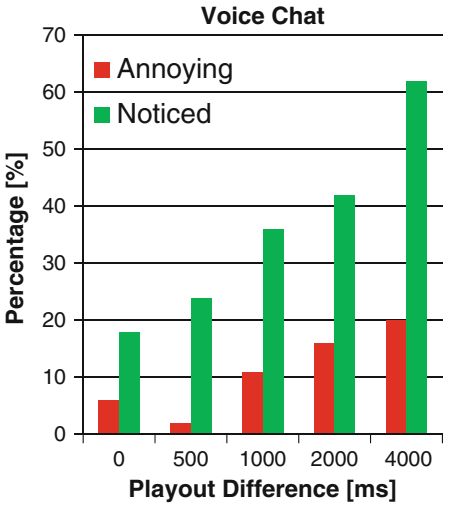


Fig. 7 Percentages of people either getting annoyed or noticing the playout differences (voice chat)

To check if the six answers of the togetherness questionnaire were answered coherently, consistency tests were used. Such tests measured if the answers to the different questions had a large amount of correlation, that is, if the same property was measured. The results were the following: Cronbach's alpha is 0.852 and Guttman's split half is 0.807 [27]. The range of alpha is 0–1, and values above 0.7 generally represent high consistency of answers in social science [28]. From the answers to the six questions on togetherness, on average voice chatters felt more together (5.1) than text chatters (4.3) on a seven-point scale. When chatting, the active chat group also felt more together (4.9) than the non-active chat group (3.9) on the same seven-point scale. The results are illustrated in Fig. 9.

Statistical analysis confirms the observations in Figs. 6, 7, 8 and 9 and can be found in [25, 26]. Voice chatters and active text chatters felt more together and

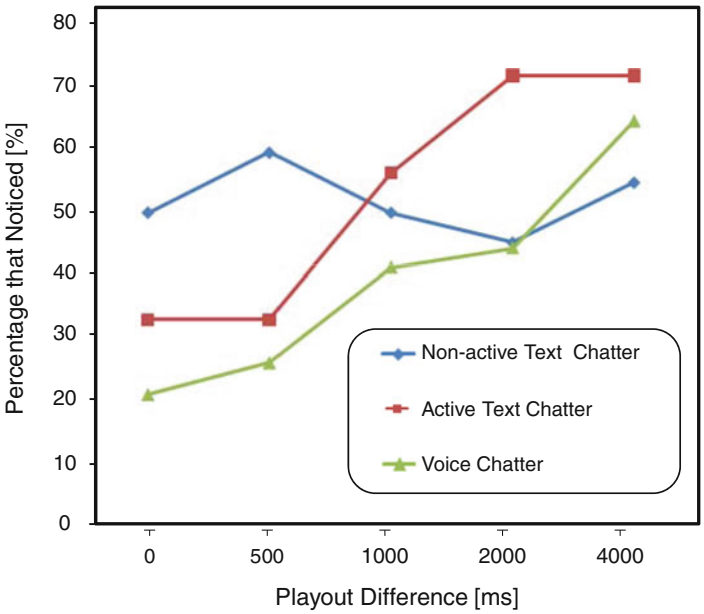


Fig. 8 Noticing playout differences, active and non-active text and voice chatters

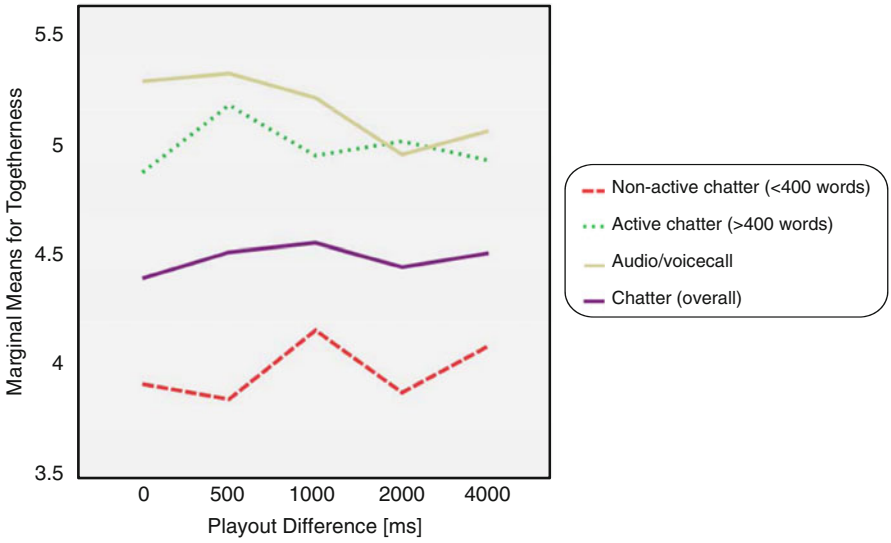


Fig. 9 Togetherness, active and non-active chatters

noticed de-synchronisation (1 s or more for voice, 2 s or more for active chat). Non-active chatters on the other hand felt less together and did not notice it when their playout was not synchronised.

4 Distributed Media Synchronisation Techniques

After motivating the research question-based measurements on how well technologies synchronise (objective testing) and how much de-synchronisation is acceptable for users (subjective testing), this section surveys the most relevant existing IDMS solutions. A classification is presented taking into account three different perspectives: the adopted architectural scheme for exchanging synchronisation information (Sect. 4.1), the strategies that can be employed to select a master reference playout point to synchronise with (Sect. 4.2) and a compilation of adjustment techniques than can be used to achieve IDMS (Sect. 4.3).

4.1 IDMS Control Scheme Architectures

Despite the increasing relevance of IDMS for realizing social shared experiences, exhaustive surveys about this kind of synchronisation are scarce. While most of the previous works on multimedia synchronisation have focused on intra-stream and inter-stream synchronisation techniques, this section solely focuses on IDMS solutions for assuring concurrently synchronised playout points at different locations.

In the context of IDMS, four synchronisation (sync) entities can be distinguished:

1. *Media Server*: The sender of the media stream(s).
2. *Sync Clients*: The sync entities that must render the media content in a synchronous way. They must also send control messages, including their current timing information to allow IDMS control.
3. *Master*: A specific Sync Client whose playout process has been selected as the IDMS reference.
4. *Sync Managers*: The sync entities responsible for collecting IDMS reports from the Sync Clients. Then, they calculate the differences between Sync Clients' playout timings and, if needed, send back to them new IDMS control messages, including IDMS setting instructions, to inform them about the required adjustments to get synchronised. These instructions will refer to the playout point of a selected master Sync Client for IDMS, which all the slave Sync Clients' playout processes must match. This entity is also referred to as *maestro* in some IDMS works.

When implementing IDMS, the first architectural decision is to determine the most appropriate location for the above sync entities. Two main approaches can be followed: *network-based* and *terminal-based*. In the first case, the sync entities are deployed at the network side, and the IDMS control processes are managed by those network entities, under control of the service provider or the operator (Fig. 10a). This way, the end-users' terminals do not have to implement any IDMS

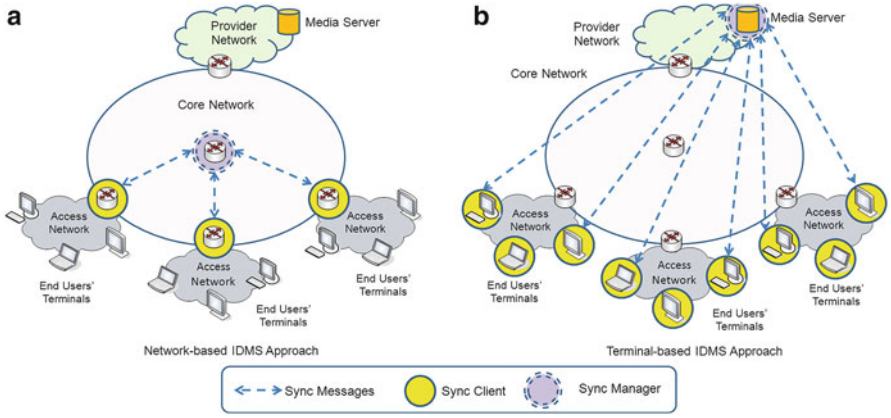


Fig. 10 Classification of IDMS based on the functionality location. (**a**) Network-based approach. (**b**) Terminal-based approach

functionality. The other approach consists of deploying the sync functionality at the end-users' terminals (Fig. 10b). By using this approach, the Sync Client entities are located in the end-users' terminals, and the Sync Manager functionality can be deployed as a separate independent entity, or either as part of a Sync Client or of a media server (as in Fig. 10b). In both approaches, the Sync Manager and the Sync Clients exchange control messages in order to maintain an overall synchronisation status in the group shared media experience.

Network-based IDMS solutions have been neither deployed nor tested in any scenarios yet. Only a design approach was proposed in [29] to meet the need of IDMS in advanced and large-scale IPTV services. So, this chapter mainly focuses on terminal-based IDMS solutions, which have been more extensively used up to date.

Independently of the architectural decision, three common structural schemes can be also considered depending on the distribution of the sync entities and on the communication processes used for exchanging synchronisation information (Fig. 11): two centralised schemes, master/slave (M/S) scheme and synchronisation maestro scheme (SMS), and one distributed scheme, distributed control scheme (DCS).

In the M/S scheme (Fig. 11a), only a master Sync Client is responsible for multicasting feedback reports about playout timing to the rest of (slave) Sync Clients (unidirectional communication). Accordingly, each slave Sync Client has to adapt its playout timing to match the one reported by the master Sync Client. In the SMS (Fig. 11b), all the Sync Clients send, in a unicast way (point-to-point), IDMS reports to a centralised Sync Manager (or *maestro*), which can be a completely separate sync entity (Fig. 11b1), the media server (Fig. 11b2), or one of the Sync Clients (Fig. 11b3). If the Sync Manager detects an asynchrony situation (i.e. when the maximum time discrepancy between the playout states of two Sync Clients exceeds

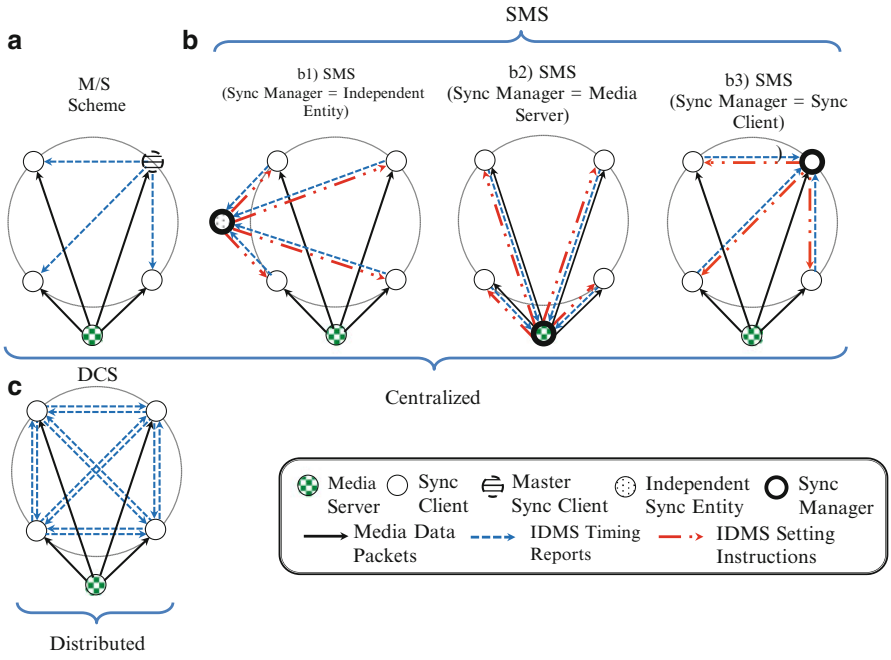


Fig. 11 Control schemes for IDMS. (**a**) M/S scheme. (**b**) SMS. (**c**) DCS

an allowable threshold) once it has collected the overall playout information, it will multicast a new control message to the distributed Sync Clients, including IDMS setting instructions, to make them enforce the required timing adjustments to get in sync (bidirectional communication).

Finally, in the DCS (Fig. 11c), all the distributed Sync Clients send in a multicast way (point-to-multipoint) IDMS reports to all the other Sync Clients. Therefore, each Sync Client can locally decide the IDMS reference from among its own playout timing and the ones received from the other Sync Clients.

4.1.1 Comparison of IDMS Control Schemes

Each one of the IDMS schemes has its own strengths and weaknesses, and therefore, their choice is largely application dependent. Here a qualitative comparison among them is presented, in order to reveal their suitability and effectiveness for distributed shared video watching scenarios. This comparison takes into account some key aspects for IDMS, such as location of the control nodes, flexibility, robustness, traffic overhead, scalability, interactivity, fairness, consistency, coherence, causality and security.

– *Location of the Control Nodes*. Centralised schemes are more sensitive to the location of the sync entities. Under heavily loaded network conditions, the IDMS performance (i.e. the level of synchrony among Sync Clients) using the SMS can be lower compared to the other two schemes if the media server is selected as the Sync Manager (Fig. 11b2). This is due to the fact that IDMS control messages sent by the Sync Manager are (or could be) sent through the same path as that of media data units[21] (MDUs), such as video frames, encapsulated in data packets. Thus, although IDMS control messages hardly increase the network traffic, it could cause some (data or control) packets to be dropped when bandwidth availability is scarce. Conversely, in the M/S scheme, if the most heavily loaded Sync Client is selected as the master, the data packets are less likely dropped on the intermediate links, as it does not need to receive IDMS reports, and its own sent IDMS reports may be transmitted in the opposite path to the MDUs. Therefore, in congestion situations, the M/S scheme may perform better than the SMS.
– *Flexibility*. In the M/S scheme, the reference playout timing for IDMS can only be selected from the IDMS reports sent by the master Sync Client. However, the Sync Manager, in the SMS, and the Sync Clients, in the DCS, can employ several dynamic policies for selecting the reference IDMS playout timing from all the collected IDMS reports (described later in Sect. 4.2). Furthermore, in the DCS, the Sync Clients can be divided into independent clusters, as in [30], so that they can only monitor the IDMS reports from the other Sync Clients belonging to the same cluster. Hence, the DCS outperforms the other schemes in terms of flexibility.
– *Robustness*. Generally centralised schemes are less robust than distributed schemes and so are here. In the former schemes, if the Sync Manager (in the SMS) or the master Sync Client (in the M/S scheme) fails to communicate with the Sync Clients, the latter cannot carry IDMS control and therefore will lose synchronisation. Nonetheless, the failure of any of the Sync Clients in a distributed architecture (DCS) has a minor effect on the other Sync Clients, because each one of them is independent and has locally all the necessary information for synchronising.
– *Traffic Overhead*. In the M/S scheme, only the master Sync Client sends IDMS reports to the slave Sync Clients. Therefore, the network load will not be significantly increased when including IDMS control. In the SMS, all the Sync Clients send unicast IDMS reports to the Sync Manager. If needed, the Sync Manager also sends new multicast IDMS control messages to the Sync Clients, to make them enforce IDMS adjustments, increasing the network load a bit more. In the DCS, the IDMS reports are exchanged in a multicast way between Sync

[21] Multimedia information can be modelled as streams that are made up of a time sequence of finite MDUs (also called in other works Media Units, MU; Information Units, IU; or Logical Data Units, LDU).

Clients. So, the traffic overhead may be higher in the DCS than in the SMS and higher in the SMS than in the M/S scheme.

- *Scalability*. The SMS may present higher scalability constraints because it requires that the Sync Manager gathers all the IDMS reports received from all the Sync Clients. If the control messages are generated at a nonadaptive rate, multiple Sync Clients may send IDMS reports almost simultaneously, thus originating a feedback-implosion problem. Even though this failure also applies to the DCS, here the saturation of the entities' computational resources is lower. As discussed, in both centralised and distributed schemes, the Sync Clients can be divided into independent logical subgroups (clusters), which can be synchronised separately thus improving the above scalability constraints. This is particularly beneficial in the DCS because the Sync Clients only have to process the IDMS reports from the Sync Clients belonging to the same cluster.
- *Interactivity*. On the one hand, lowest delays can be achieved by using the M/S scheme because, unlike the other two schemes, each slave Sync Client can compute the detected asynchrony every time it receives an IDMS report from the master Sync Client. Delays in the DCS might be larger than in the M/S scheme because each Sync Client must wait to gather the overall reports with the playout status of all the active Sync Clients (which can be sent and received at different instants). On the other hand, highest delays appear when using the SMS because, depending on the network topology and on the routing tree structure, the delay could increase significantly (the Sync Manager must collect all the IDMS reports from the Sync Clients and then send back new control messages including IDMS setting instructions to them). So, asynchrony situations (over a threshold) will be detected and corrected earlier using the M/S scheme than using the DCS and earlier using the DCS than using the SMS.
- *Fairness*. The M/S scheme is suitable for applications in which a single Sync Client has a certain priority level over the other Sync Clients. For example, in e-learning scenarios the Sync Client of the teacher can be selected as the master Sync Client, so all the Sync Clients of the students must be synchronised to it. However, this scheme cannot treat all the Sync Clients fairly. This problem is minimised when the SMS or the DCS schemes are employed, because the reference playout timing for IDMS is selected after a comparison among the IDMS reports sent by all the Sync Clients. As an example, the study in [31] concluded that the effectiveness of the IDMS control in competitive networked environments, in terms of fairness between the Sync Clients, could be improved by adjusting the overall playout timing to the slowest (most lagged) one. The DCS may outperform the SMS in terms of fairness because asynchrony situations, which can cause an annoying effect to de-synchronised Sync Clients, can be corrected earlier due to smaller delays.
- *Consistency*. In centralised schemes, inconsistency between Sync Clients' states occurs less likely, since all of them always receive the same control information about IDMS timing from the Sync Manager (in the SMS) or the master Sync Client (in the M/S scheme). However, in a distributed scheme (DCS), there is no guarantee that the same reference IDMS timing, from among all the collected

Table 5 Qualitative comparison among IDMS control schemes

		Factors									
		Robustness	Flexibility	Traffic overhead	Scalability	Interactivity	Fairness	Consistency	Coherence	Causality	Security
Scheme	DCS	1	1	3	2	2	1	3	2	3	3
	SMS	2	2	2	3	3	2	1	1	2	1
	M/S	3	3	1	1	1	3	2	3	1	2

1 best scheme, *2* good scheme, *3* worst scheme

IDMS reports, will be selected in all the distributed Sync Clients, since each one takes its own decisions locally, leading to a more probable potential inter-client inconsistency.

– *Coherence*. This concept refers to the ability to synchronously (and simultaneously) coordinate the media playout timing. Unlike in the DCS and the SMS schemes, in which the maximum asynchrony between Sync Clients can be estimated, in the M/S scheme, each Sync Client can only know the asynchrony between its local playout process and that of the master Sync Client. Using the M/S scheme, Sync Clients adjust their playout timing every time they detect a situation of asynchrony (regarding the playout timing of the master Sync Client). Hence, reactive synchronisation actions will not be performed simultaneously in all the Sync Clients [32]. Consequently, despite the fact that the M/S and SMS schemes are most adequate in terms of consistency, the latter outperforms the others (the M/S and the DCS Schemes) in terms of coherence.
– *Causality*. Causality in media synchronisation is required for preserving and/or restoring the original media timing and the chronological order of specific actions. Previous work [33] concluded that the SMS is slightly superior to the DCS in terms of causality, mainly due to minor traffic overhead. Similarly, it can be deduced that the performance in terms of causality provided by the M/S scheme is better than the one in the other IDMS schemes due to the same reason.
– *Security*. In the DCS, each Sync Client takes its own decisions, resulting in a lack of control of what each one is doing and whether a Sync Client is malicious or not. In the M/S scheme, this problem can be minimised if the IDMS operation of the master Sync Client is under control. In the SMS, the Sync Manager can use some mechanisms to check the validity of the arriving control packets and guarantee the overall synchronisation status. Hence, cheating is more difficult in centralised schemes than in the DCS.

To summarise, a ranked comparison among them is presented in Table 5. This is not an arithmetic weighting, but the numbers 1–3 are employed to classify the appropriateness of the IDMS schemes regarding each one of the considered factors. Note that the weight of the relative importance of each of these factors will depend on the context and space in which a specific video-sharing application is deployed.

Depending on the targeted goals, an implementer or application developer can choose to give more preference to interactivity than to traffic overhead, or more to flexibility and robustness than to security, or more to coherence than to scalability, etc. Therefore, no definitive rules can be given, but only indicative guidelines that can be followed in the design of an IDMS solution.

As it can be appreciated in the table, the M/S scheme can provide the best performance in terms of scalability, traffic overhead, interactivity and causality, but presents serious drawbacks if some features such as robustness, coherence, flexibility and fairness must be provided.

The DCS is a suited option for IDMS if performance in terms of robustness, fairness, flexibility, scalability and interactivity (i.e. achieving stringent synchrony levels) is desirable, despite of a slight cost in terms of traffic overhead, consistency or security.

An important limiting factor for the DCS and the M/S schemes is the support of multicast feedback capabilities (i.e. the ability to exchange useful information for IDMS in a point-to-multipoint way) among the distributed Sync Clients in most media streaming technologies, for example, those in which single-source multicast (SSM) is employed. In such cases only the media server can transmit data in a multicast way. So, it could prevent the deployment of an IDMS solution based on the DCS or the M/S schemes in some actual large-scale environments, such as IPTV broadcast distribution channels. In other controlled scenarios, where small groups of users are watching video content synchronously, independently of other groups of users, then the adoption of a the DCS or the M/S schemes may be an option.

As an additional drawback of the DCS, it requires that the distributed Sync Clients implement the functionality of processing the incoming IDMS reports from all the other Sync Clients and calculating the required IDMS adjustments to keep an overall synchronisation status. So, it implies additional complexity to the Sync Clients, which can result in an increase of the development costs of the IDMS solutions based on this signalling scheme.

Finally, we can observe that the SMS is the best IDMS control scheme in terms of consistency, coherence (all the Sync Clients need to be almost simultaneously synchronised to the same reference timing) and security, which are important aspects for distributed shared video watching scenarios.

On the other hand, the main weaknesses of using the SMS for IDMS are scalability and interactivity. The first disadvantage is not very different in the SMS compared to the DCS, and it can be significantly solved by using two control mechanisms: either dividing the session into logical groups (clusters), which may facilitate the IDMS management to the Sync Manager, or dynamically adjusting the transmission interval for the IDMS reports according to the number of active Sync Clients in the session and the available bandwidth. Interactivity, however, is not a crucial constraint for distributed shared video watching since, as demonstrated in Sect. 3, such scenarios do not require stringent synchrony levels. Also, some previous work has demonstrated the feasibility of an SMS for IDMS to keep the synchronicity within allowable limits (even more stringent levels that the ones required for Social TV were accomplished) in real scenarios [34].

Also, in some media streaming technologies, such as the ones using RTP/RTCP, distributed receivers send regularly feedback messages including QoS metrics (e.g. delay, jitter, packet loss information, etc.) to the Media Server, who can react accordingly (e.g. by adjusting its transmission timing or the media coding mechanism). If those feedback messages are extended to include useful information for IDMS, it would facilitate the deployment of an IDMS solution (as will be discussed in Sect. 5). This makes the SMS the most practical alternative for distributed shared video watching, especially if the Sync Manager functionality is incorporated into the media server resources.

Therefore, taking into account all the above features, it can be concluded that the SMS, in general, is best ranked for distributed video watching.

4.2 Master Reference Selection Policies

As discussed above the use of the SMS is preferable for distributed shared media watching, and, therefore, this section is focused on the SMS, in which the Sync Manager is responsible for performing the main IDMS control tasks. First, it must gather the IDMS reports from the Sync Clients. Once the overall playout information has been collected and if an asynchrony situation is detected, the Sync Manager can employ some dynamic policies to select the master reference playout point to synchronise with [30]. Possible policies include synchronisation to the slowest Sync Client, synchronisation to the fastest Sync Client, synchronisation to the mean playout point and synchronisation to the nominal rate of the Media Server.

The first strategy consists of selecting the playout timing of the most lagged (slowest) Sync Clients as the IDMS reference. Using this method there will not be any skips in the Sync Clients' playout, avoiding the consequent discontinuities, since (faster) slave Sync Clients will be forced to pause their playout processes (waiting for the slowest one). This policy is suitable for multimedia applications with flexible delay requirements, and it enables the inclusion of interactive error recovery techniques through retransmission requests. Besides, in distributed scenarios, where users compete with each other (e.g. network quizzes shows), this policy will be appropriate to guarantee fairness. However, if the playout process of the master Sync Client is extremely lagged (e.g. due to any problem, such as congestion), it could result in the progressive filling of the Sync Clients' playout buffers, which could eventually overflow. This way, loss of real-time sensation would be noticed, affecting the overall QoE. To avoid such situations, additional adjustment techniques, such as buffer fullness control, should be employed.

In the second method, the playout timing of the fastest (most advanced) Sync Client is selected as the IDMS reference. As an example, in collaborative networked scenarios, the efficiency of the overall work may be improved by adjusting the lagged playout timings to the earliest one. Nevertheless, if there is any slow Sync Client under bad conditions (e.g. long delays, jitter, network load, CPU overload),

it will be constantly skipping MDUs to get in sync, and the continuity of its playout process will be seriously affected. In such a case, if the playout process of the master Sync Client is extremely advanced, the playout buffers of all the Sync Clients may suffer underflow (progressive emptying of their buffer occupancy), as the session advances in time. Hence, some additional adaptive buffering control techniques should be also included when using this policy.

Note that both policies (sync to the fastest/slowest Sync Client) are dynamic processes, since the master and slave roles of the Sync Clients can be exchanged during the session lifetime, depending on the aforementioned uncontrollable factors, allowing M/S switching [35].

Another solution for selecting the IDMS reference is to define a virtual playout point (i.e. a fictitious Sync Client), obtained as the mean point of all the gathered playout processes. Using this method, the playout processes of the Sync Clients will be more continuous and smoother, since the values of the sync adjustments will be lower than in the previous policies. However, its use does not guarantee playout buffer overflow or underflow avoidance, because playout rate imperfections and end-system situations (bandwidth availability, network load, CPU congestion, etc.) are in fact unpredictable. Also in this case, and in order to avoid this, some techniques to control the buffers occupancy should be employed.

In all the previous discussed policies, the reporting of an erroneous playout point, either accidental or malicious, may lead to undesired behaviour. According to the adopted model, extremely advanced/lagged playout points will produce high adjustments on the Sync Clients' playout processes with the consequent significant loss of real-time/continuity perception. Therefore, in any implementation, with the aim of avoiding faulty behaviour it would be advisable that the Sync Manager in the SMS, or the distributed Sync Clients in the DCS and the M/S, consider inconsistent playout information (exceeding configured limits) as a malfunction service and reject that information in the calculation of the IDMS target playout point (i.e. the playout timing of the master Sync Client).

All the above master reference selection policies could also be employed by the distributed Sync Clients if a DCS for IDMS is adopted. Additionally, a fourth strategy can be adopted, but only when using the SMS and if the Sync Manager and the media server are implemented in the same sync entity. It consists of the synchronisation to the nominal rate of the Media Server. In such a policy, the Sync Manager will act as a virtual master Sync Client with an ideal target playout timing, which is defined as the ideal playout timing when there are no network delay jitter nor playout rate imperfections. Therefore, the maximum playout asynchrony will be calculated taking into consideration the playout point of this virtual ideal Sync Client as another Sync Client in the IDMS session. Using this policy, if network conditions are quite stable, underflow/overflow situations will be avoided. This is because the playout states of the possible deviated Sync Clients would be adjusted to this ideal playout point every time an asynchrony situation is detected. Furthermore, this technique is beneficial for accurate Sync Clients because the smaller the deviations are in their playout processes states, the lower the adjustments that would be needed to acquire IDMS.

Also, note that all the policies presented above are appropriate for scenarios with controlled or bounded network delays. Otherwise, some additional synchronisation techniques should be adopted to dynamically adjust the playout processes of the Sync Clients according to potential fluctuations of the end-to-end delays, in order to avoid buffer underflow/overflow situations. The discussion of such techniques is out of the scope of this section.

4.3 Synchronisation Adjustment Techniques

As explained above, the IDMS reference can be obtained following several policies. Once selected an IDMS reference, the next step is to perform in the Sync Clients the required media timing adjustments so that an overall synchronisation status is achieved. This subsection introduces a number of potential adjustment techniques employed in the most representative IDMS solutions. They are classified into four main categories according to their purpose (basic, preventive, reactive and common adjustment techniques), which in turn are divided into two groups, depending on the location at which they are executed (server or client side) . This classification is presented in Table 6. Readers are referred to [35] to find a more detailed explanation of those adjustment techniques.

Basic control techniques are the essential ones to preserve the temporal relationships of media streams. *Preventive control techniques* attempt to avoid situations of asynchrony before they occur, while *reactive control techniques* are used to recover synchronisation after asynchrony has been detected. Additionally, there are other techniques called *common techniques* that can be used as a means to prevent (preventive) or correct (reactive) situations of asynchrony. Those adjustment techniques are usually complemented in existing IDMS algorithms. Since preventive adjustment techniques cannot usually completely avoid asynchrony, the combination with reactive adjustment techniques is needed.

On the one hand, some techniques are always needed at the Sync Clients' side because of the existence of jitter. On the other hand, techniques at the media server side need to gather monitoring information from the Sync Clients. Based on such information, the media server can react accordingly. In some cases, the Media Server(s) and the Sync Clients cooperate for controlling the adjustment processes. A typical example is the combination of the attachment of useful information for synchronisation in the headers of the media data packets (sequence numbers, timestamps, identifiers, etc.), buffering techniques at the Sync Client side to smooth out the effect of the network jitter and either reactive skipping and pausing or reactive playout rate adjustment techniques to restore the synchronicity. Furthermore, several techniques for the same purpose can be used simultaneously at the same location (e.g. several reactive control techniques at the Sync Client side). Some other techniques cannot, however, be used cooperatively, such as adjustment of the playout rate and interpolation of media data.

Table 6 Synchronisation adjustment techniques (Adapted from Boronat et al. [35])

Technique's purpose	Location	Technique
Basic control	Server	Add information useful for synchronisation (timestamps, sequence numbers, source or group identifiers, markers, event information, etc.)
	Client	Buffering techniques
Preventive control	Server	Initial playout instant calculation
		Deadline-based transmission scheduling
		Interleave MDUs of different media streams in a single transport stream
	Client	Preventive skips of MDUs (eliminations or discardings) and/or preventive pauses of MDUs (repetitions, insertions or stops)
		Change the buffering time of the MDUs
Reactive control	Server	Adjust the transmission timing
		Decrease the number of transmitted media streams
		Drop low-priority MDUs
	Client	Reactive skips (eliminations or discardings) and/or reactive pauses (repetitions, insertions or stops)
		Playout duration extensions or reductions (playout rate adjustments)
		Use of a virtual time with contractions or expansions
		Master/slave scheme (switching or not)
		Late events discarding (Event-based)
		Rollback techniques (Event-based)
Common control	Server	Skip or pause MDUs in the transmission process
		Advance the transmission timing dynamically
		Adjustment of the input rate
		Media scaling
	Client	Adjustment of the playout rate
		Data interpolation

4.4 Comparison Among IDMS Solutions

Several IDMS solutions have been found following the above IDMS control and adjustment techniques. Table 7 shows a classification of them, regarding the adopted control schemes, the synchronisation information included in the media data packets and the adjustment techniques they implement. This classification has been adapted from the more extended one presented in [35]

It can be seen from Table 7 that each IDMS solution presents a different combination of synchronisation adjustment techniques and is based on a particular architectural scheme. Therefore, it is not easy to make a comparison between them even qualitatively. In order to clarify the relationships among them, their effectiveness, strengths and weaknesses and adjustment techniques used should be assessed, both objectively and subjectively, under the same conditions.

Besides, it is important to emphasise that some of the above synchronisation adjustment techniques are only applicable to specific use cases. As an example, some of them are only valid for stored media content, but not for live streams. Likewise, some other techniques, such as media scaling or data interpolation, are conditioned to the use of specific media coding mechanisms.

As an example, some previous studies have compared different reactive adjustment techniques for a specific synchronisation algorithm. In [36] we can find some results and conclusions focused on the inter-stream synchronisation quality, which can be extrapolated to IDMS solutions, because in both cases receivers are forced to dynamically adjust their playout timing to keep synchronisation if an asynchrony situation is detected.

In [36], it was concluded that adaptive adjustments (i.e. smooth variation in the media playout rate) were preferable to aggressive playout adjustments (i.e. reactive skips and pauses) in distributed shared video watching scenarios. This is because aggressive adjustments can originate a noticeable degradation of the user perception as regards the quality of the received media stream because, on one hand, some important information may not be presented to the users (due to the skipped MDUs) and, on the other hand, a sensation of loss of continuity may also be noticed (due to the paused MDUs). In [30], authors showed the feasibility of an adaptive media playout (AMP) technique for IDMS. It was shown that an overall synchronisation status can be achieved, while avoiding long-term playout discontinuities, by smoothly varying the media playout of the receivers, within perceptually tolerable ranges that would be imperceptible to users, every time an asynchrony situation was detected.

The assessment of a quantitative comparison among different IDMS solutions in order to clarify which one produces the best performance becomes even much more complicated. One of the reasons is that the motivation, background and environment in which each media synchronisation algorithm has been proposed can largely vary from each other. Furthermore, in the compiled IDMS works, there are no commonly used QoS metrics to measure the synchronisation performance. Consequently, a comparison among different IDMS solutions would require their

Table 7 Classification of IDMS solutions

IDMS solution	IDMS control scheme	Synchronisation information	Adjustment techniques
Virtual time rendering	M/S scheme [35]	Timestamps, and sequence numbers	Compilation of adjustment techniques used in different versions of the VTR-based IDMS algorithm:
(VTR)-based IDMS algorithms [31, 33, 37–44]	DCS [37, 38, 42]		Change of the buffering time according to the network delay estimation
			Decreasing the number of media streams
			Preventive pauses
			Reactive skips and pauses
	SMS [33, 38]		Skips at the media server side
			Playout duration extensions and/or reductions
			Virtual local time expansions and/or contractions
			Media scaling
			Interleaving MDUs
			Playout rate adjustments
			Initial transmission instant
	Enhanced SMS [31, 41, 43]		VTR techniques (upper cell)
			Event-based synchronisation control
			Combination of the selection of two reference output timings (the most advanced/lagged playout points)
[45]	DCS	Timestamp in first packet	Initial transmission and playout instant
			Playout rate adjustments (receiver's clock)
			Master/slave receiver switching (chairman)
Bucket synchronisation [46]	DCS	Timestamps and sequence numbers	Skips (discardings) and pauses (duplicates)
			Late events are dropped
Local lag and time warp (LL-TW) [47] and evolved versions [48, 49]	DCS	Timestamps	Event-based synchronisation control
			Playout duration extension
			Rollback-based techniques

(continued)

Table 7 (continued)

IDMS solution	IDMS control scheme	Synchronisation information	Adjustment techniques
Trailing state synchronisation (TSS) [50]	DCS	Timestamps	Event-based synchronisation control
			Playout duration extension
			Rollback-based techniques
Interactivity-loss avoidance (ILA) [51]	DCS	Timestamps	Event-based synchronisation control
			Playout duration extension
			Preventive MDU/event discarding
			Reactive event discarding
RTP-based feedback global protocol (RTP-FGP) [34]	SMS	Timestamps, sequence numbers and source identifier	Initial playout instant
			Reactive skips and/or pauses at the client side
			Virtual time expansion
			Master/slave receiver switching
Enhanced RTP-FGP [30, 32]	SMS [30]	Timestamps, sequence numbers, source identifier and group identifier	Techniques in RTP-FGP (upper cell)
			Smooth playout rate adjustments
			Control of buffer fullness level
	M/S scheme [32]		Independent synchronisation for different clusters (logical synchronisation groups)

evaluation under the same conditions, which would imply to implement them in the same media sharing applications and to make that performance comparison under the same (network or simulated) environments, which would be greatly difficult and time-consuming. Also, subjective assessment should be conducted to analyse the benefits on the user experience (QoE) in each one of the synchronisation solutions under comparison, which is the key factor to determine the suitability of a specific IDMS solution. That is the reason why only qualitative comparisons among existing synchronisation solutions have been traditionally performed in previous works.

5 RTP/RTCP-Based IDMS Approach

Based on the previous discussion, this Section presents the technology aspects of a promising IDMS solution the authors have been working on over the last years. The solution provides flexibility to adopt all the considered IDMS control schemes for a targeted application, although it has mainly adopted an SMS [5, 30, 34]. It implements a wide range of adjustment techniques (summarised in the two last rows in Table 7) to make it more robust and efficient, thus providing all the benefits discussed in the previous section. The approach does not define a new proprietary protocol. Instead, it is based on the use and extension of the RTP/RTCP standard protocols, which may facilitate the implementation, the deployment and the adoption in actual distributed multimedia systems. The solution presented in this chapter is being standardised under the umbrella of the ETSI (European Telecommunications Standards Institute) TISPAN (Telecommunications and Internet converged Services and Protocols for Advanced Networking) and the IETF (Internet Engineering Task Force) AVTCORE (Audio/Video Transport Core) Maintenance WG (working group).

5.1 Suitability of RTP/RTCP for Media Synchronisation

As highlighted in Sect. 2, RTP and RTCP protocols are extensively used in streaming media services such as Voice over IP (VoIP), videoconferencing applications, Video on Demand (VoD) or TV services over IP (IPTV). The timestamps, sequence numbers and payload (content)-type identification provided by RTP packets are useful for reconstructing the original media timing and for reordering and detecting packet losses at the client side. Moreover, the reporting features (transport and management of feedback control messages) provided by RTCP are useful to obtain quality feedback about data delivery. First, service providers can use the QoS metrics included in RTCP reports, such as round-trip time and loss rate, for troubleshooting and fault tolerance management. Second, the mapping time information and source-identification parameters provided by each RTCP sender report (SR) and source description report (SDES), respectively, are useful to allow inter-stream synchronisation (e.g. lip-sync) [4].

Using RTP/RTCP, the optimum transmission rate for the feedback control messages does not need to be computed, as required by most of the IDMS solutions presented in the previous section. This is because RTCP feedback reports are exchanged regularly between all the participants, and the report interval period is dynamically adjusted according to the active number of media servers and clients, and to the session bandwidth, as specified in RFC 3550 [4].

Additionally, further extensions to these protocols are accepted in RFC 3550, to include profile-specific information required by particular applications, as specified in RFC 5968 [52]. Likewise, RFC 3611 [53] allows the definition of new RTCP Extended Report (XR) blocks useful for exchanging additional QoS metrics as required by specific purposes, such as IDMS.

Finally, according to RFC 3550, the maximum fraction of the total amount of control traffic added by RTCP must be limited to 5% of the RTP session bandwidth, so the traffic overhead added by an IDMS based on the RTCP capabilities will not be very high.

This section presents a proposal to extend these standard protocols with additional RTCP block reports and packet types, including information about reception and playout timing of specific RTP data packets, to be exchanged between the sync entities described in previous section to acquire IDMS.

5.2 RTP/RTCP-Based IDMS Approach Under Standardisation Process

As stated above, the solution presented in this chapter is being standardised under the umbrella of the ETSI TISPAN project and is currently a milestone for the IETF AVTCORE WG. Both specifications have followed the initial RTCP route in [34] to exchange useful information for IDMS.

The ETSI TISPAN is a major European-based organisation that mainly considers the standardisation of next-generation networks (NGN) and its associated services. Whereas the first ETSI TISPAN IPTV standards mainly focused on regular TV services, such as broadcast TV and content on demand, the recent ETSI TISPAN Release 3 standards [54] contain a series of specifications for advanced large-scale IPTV services, including personalisation, Social TV and IDMS features.

After standardisation in ETSI, a new standardisation activity has been undertaken within the IETF, the standardisation body where Internet standards, such as Request for Comments (RFCs), are developed. There were multiple reasons for bringing IDMS work towards the IETF. The IETF AVTCORE WG is responsible for standardisation of RTP/RTCP functionalities. Even though the earlier work within ETSI was based on the RTCP-based IDMS approach, further work on this solution seems more suitable within the IETF, where most RTCP extensions are developed. Also, the ETSI proposal is a dedicated solution for use in large-scale IPTV deployments, with low to medium level synchronisation requirements. Other services such as

Internet-based video streaming, synchronous e-learning or networked loudspeakers may also benefit from IDMS, but they are not supported by the ETSI solution. Thus, an Internet Draft on IDMS was proposed in order to design a more general applicable IDMS solution [5].

5.2.1 IDMS Architecture

Both ETSI TISPAN and IETF proposals have mainly adopted an SMS for IDMS. The Sync Manager is called here Media Synchronization Application Server (MSAS), and it is responsible for collecting IDMS reports from the distributed Sync Clients, calculating the temporal discrepancy between their IDMS timing and sending to them new control messages including IDMS setting instructions. Based on such control messages, the Sync Clients will perform the required adjustments to acquire IDMS. The operational aspects for selecting a Sync Client as the IDMS master reference (Sect. 4.2) and the reactive adjustment techniques to be implemented (Sect. 4.3) are not specified, but left to vendor-specific implementations. This allows vendors to differentiate their solution from that of other vendors.

ETSI TISPAN considers various mappings of the IDMS functional architecture onto the sync entities in a given IPTV architecture. In one mapping, aimed at small-scale deployments, the Sync Clients are located in the end-users' terminals, also called user equipments (UEs), and the Sync Manager (MSAS) is implemented in the network, as a functional entity separated from the media distribution function (MDF). For synchronisation using a direct communication channel between multiple end users, the Sync Manager (MSAS) can also be co-located with a specific Sync Client (in a UE). In another mapping, aimed at large-scale deployments, the Sync Client can be located within an edge node of the transport network (network-based approach), for example, a DSLAM – digital subscriber line access multiplexer – or CMTS – cable modem termination system – or even higher up in the network hierarchy. Furthermore, at a higher level (e.g. in the core network), a Sync Manager must be used to control the IDMS timing of the Sync Clients [29]. In such a case, the Sync Client is selected such that further downstream delays are considered acceptable for IDMS (i.e. any delay differences introduced from the edge node to the end users cannot be controlled by the IDMS approach).

The architecture proposed in the IETF Internet Draft is based on that of ETSI, but it has been simplified. In it, the functionality of Sync Client is defined as part of an RTP receiver, and the functionality of Sync Manager (MSAS) is defined as part of the RTP sender (Fig. 12).

5.2.2 IDMS Protocols

After the configuration of the sync entities and the setup of the session, IDMS control messages must be exchanged between Sync Clients and the Sync Manager. Sync Clients must send IDMS status information (indicating arrival and/or presen-

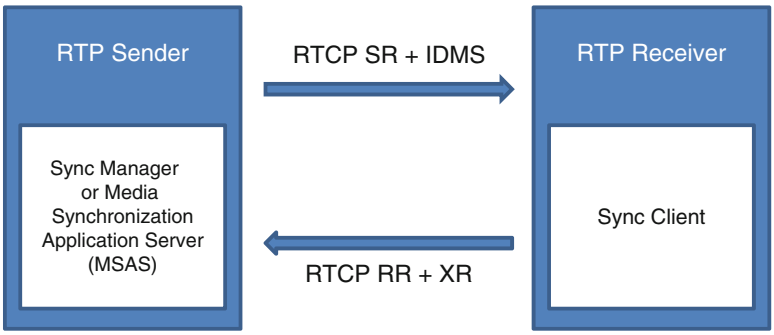


Fig. 12 Functional entities for IDMS (IETF Internet Draft)

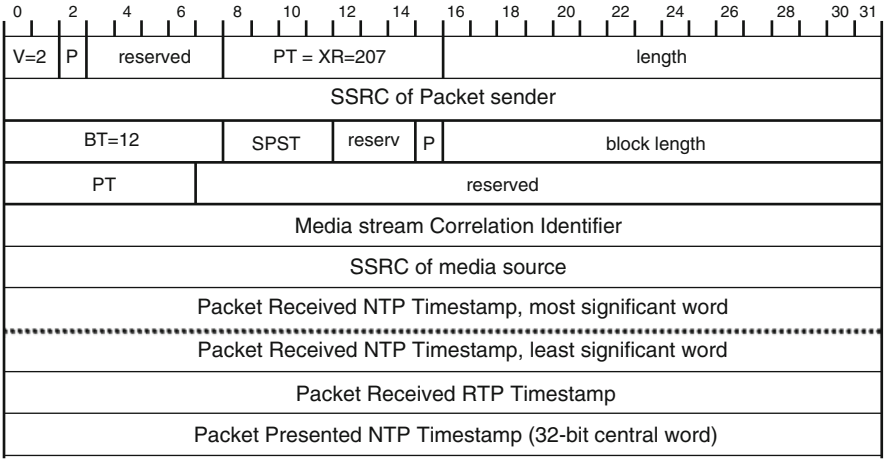


Fig. 13 RTCP XR block for IDMS (ETSI TISPAN & IETF Internet Draft)

tation timing information of RTP media packets) to the Sync Manager, whereas the Sync Manager must monitor the overall synchronisation status and send to the Sync Clients new control messages including IDMS setting instructions, if needed. Those control messages must inform about a target playout point (which is taken from the IDMS timing of the selected master Sync Client) to which all the Sync Client must synchronise.

As indicated in RFC 3611, *RTCP XR blocks are suited for transporting new metrics regarding media transmission or reception quality*. Consequently, a new RTCP XR block type has been specified for providing feedback about receipt and presentation times for RTP packets (Fig. 13). The explanation of their fields can be found in the publicly available Internet Draft [5].

In the ETSI solution, both the Sync Clients and the Sync Manager use the same RTCP XR block for IDMS. If sent by a Sync Client, the Sync Manager interprets it

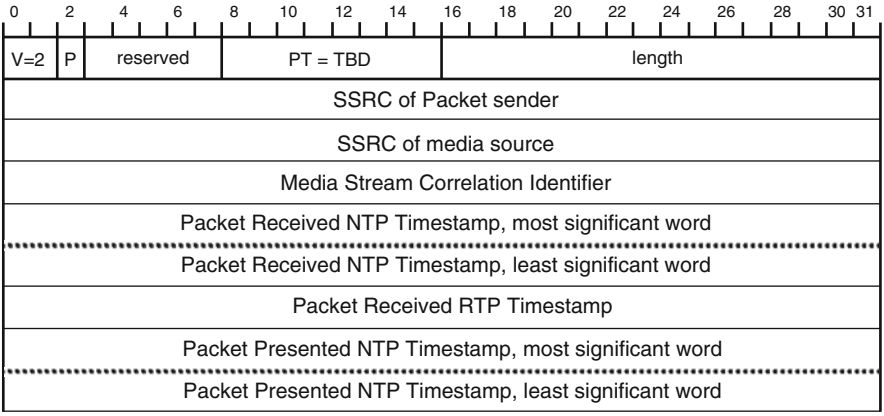


Fig. 14 RTCP settings packet type for IDMS (IETF Internet Draft)

as the current IDMS timing (reception and/or playout time of a specific RTP packet) for that Sync Client. If sent by a Sync Manager, it is interpreted as the IDMS timing of the selected master Sync Client. This way, all the Sync Clients belonging to the specific group indicated in this XR block must match this reference timing. In order to identify if an incoming XR was sent by a Sync Client or by a Sync Manager, the SPST (Synchronization Packet Sender Type) field of that XR block should be inspected [5].

The same policy was also adopted in the initial versions of the Internet Draft. However, the feedback messages sent by the Sync Manager are not used for monitoring purposes, but for control purposes, providing guidance for all the Sync Clients about when to playout the media. Thus, for indicating IDMS setting instructions rather than conveying a metric on how well media playout is doing, a new RTCP packet type, called RTCP IDMS settings packet, was defined (Fig. 14).

During the IDMS session, the Sync Manager is continuously monitoring the reception of RTCP XR blocks for IDMS from the Sync Clients. Once a new report is received, its IDMS parameters (i.e. timestamp of the RTP packet to which the report refers, the reception and playout times for that RTP packet and the group or cluster identifier to which the Sync Client belongs) are locally registered. If the overall playout information in a specific group has already gathered, the Sync Manager will compute the maximum asynchrony in that group. If the detected asynchrony exceeds an allowable threshold, the Sync Manager will select one Sync Client as the master reference for IDMS (following one of the policies introduced in Sect. 4.2) and will consider its IDMS timing parameters for calculating a target playout point (the timestamp of a specific RTP packet and its proper playout time) for all the Sync Clients in that group . This target playout point will be communicated to the Sync Client by sending them a new RTCP IDMS settings packet. Otherwise, if the detected asynchrony is lower than the allowed threshold, the IDMS parameters from all the Sync Clients will be erased, and a new IDMS cycle will be initiated.

Regarding the current state of our RTP/RTCP-based IDMS proposal in practice, a small-scale trial was performed in [34], which proved the feasibility of a preliminary version of our RTP/RTCP-based IDMS proposal, using the SMS, to keep the asynchrony among distributed receivers within allowable limits in a real WAN scenario between the campuses of the Polytechnics University of Valencia (Spain). Then, the IDMS solution was implemented in network simulator 2 (NS-2) in order test its performance in various network environments, under different conditions, and to easily include and test some enhancements and extensions to our IDMS solution [30], such as the implementation of different control schemes (SMS, DCS and M/S), IDMS management for independent groups of consumers and adaptive media playout techniques to minimise long-term playout discontinuities.

Despite the simulation results show that the IDMS solutions is valid and its performance is satisfactory, we also pretend, as a future work, the implementation of the proposal in an actual synchronous media sharing application in order to perform large-scale trials and real-world assessments, analyzing the effects over the user experience (QoE) of different levels of out of sync situations and how they are avoided by using our adaptive IDMS solution.

A proof-of-concept work was also undertaken and is described in [55], which shows the validity of the ETSI IDMS approach.

6 Conclusions and Future Challenges in Synchronous Shared Media Consumption

Online media retrieval and consumption are becoming social activities. While in the past services were designed for one specific network and device, now users in different domains are expecting to enjoy shared experiences. They want, for example, to be able to converse while watching multimedia content together (Social TV). Enabling such services faces a number of technological (e.g. synchronisation, universal session handling) and perceptual (e.g. presence awareness, QoE) challenges.

This chapter has focused on one specific challenge ahead: Inter-Destination Media Synchronisation (IDMS). IDMS studies the synchronisation of different media streams across multiple locations, and it is essential for enabling shared experiences. The contributions of this chapter are fourfold.

First, it reports on realistic experiments measuring the synchronisation differences between delivery technologies. Results show long delays of up to 6 s for Internet-based video watching (including P2P) and of up to 5 s for TV-based reception. Second, it highlights results from subjective user tests, indicating that playout differences of over 2 s become annoying for distributed consumption of media. These two contributions motivate the overall objective of the chapter: better synchronisation mechanisms are needed for enabling shared experiences. Third, an exhaustive surveys of existing synchronisation solutions have been qualitatively

compared, showing their advantages and disadvantages. Nevertheless, standardisation is essential for avoiding the proliferation of proprietary closed solutions. The fourth contribution is the description of the IDMS standardisation efforts by ETSI TISPAN and IETF.

Regarding synchronisation, future research is needed for better understanding user requirements for different application scenarios. Moreover, the deployment of working systems will possibly challenge the applicability of some of the concepts included in this chapter. Multimedia shared experiences as a research topic is still in its infancy. The intention of this chapter is to exhaustively explore one of the challenges ahead, paving the path for further research.

Acknowledgments UPV: This work has been financed, partially, by Universitat Politècnica de Valencia (UPV), under its R&D support programme in PAID-05-11-002-331 Project and in PAID-01-10.

CWI: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. ICT-2011-7-287723 (Reverie project). The authors would like to thank the fruitful collaboration with David Geerts (K.U. Leuven) and with Ishan Vaishnavi (CWI & Huawei).

UPV and CWI: The authors would like to thank the following people working at TNO: Hans Stokking, Ray van Brandenburg and M. Oskar van Deventer.

References

1. Shamma, D., Bastea-Forte, M., Joubert, N., Liu, Y.: Enhancing online personal connections through synchronized sharing of online video. CHI '08: extended abstracts on human factors in computing systems (2008). doi:10.1145/1358628.1358786
2. Coppens, T., Trappeniers, L., Godon, M.: AmigoTV, towards a Social TV experience. In: Proceedings of EuroITV, Aalborg, Denmark (2004)
3. Huang, E., et al.: Of social television comes home: a field study of communication choices and practices in TV-based text and voice chat. In: CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2009). doi:10.1145/1518701.1518792
4. Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V.: RTP: a transport protocol for real-time applications, RFC-3550, July 2003
5. Brandenburg, R. Van, Stokking, H., Deventer, M.O. Van, Boronat, F., Montagud, M., Gross, K.: RTCP for inter-destination media synchronization, draft-brandenburg-avtcore-rtcp-for-idms-07.txt, IETF Audio/Video Transport Core Maintenance working draft, 11 October 2012
6. Motions Pictures Experts Group (MPEG): MPEG-2 part 2, video, standard ISO/IEC 13818-2007 (2007)
7. Motions Pictures Experts Group (MPEG): MPEG-2, part 3, audio ISO/IEC 13818-3 (1998)
8. Motions Pictures Experts Group (MPEG): MPEG-2 part 1, systems, standard ISO/IEC 13818-1 (2007)
9. Reimers, U.: DVB The Family of International Standards for Digital Video Broadcasting. Springer, Braunschweig (2006)
10. ITU-T: H.264: advanced video coding for generic audiovisual services (2003)
11. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. IEEE Trans. Circuit Syst. Video Technol. **13**(7), 560–576 (2003). doi:10.1109/TCSVT.2003.815165
12. Yiyan, W., Hirakawa, S., Reimers, U., Withaker, J.: Overview of digital television development. Proc. IEEE **94**(1), 8–21 (2005). doi:10.1109/JPROC.2005.861000

13. Kumar, A.: Implementing Mobile TV. Elsevier, Burlington (2010)
14. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: RFC 2616: hypertext transfer protocol – HTTP/1.1, IETF (1999)
15. Stockhammer, T.: Dynamic adaptive streaming over HTTP- standards and design principles. In: MMSys '11: Proceedings of the Second Annual ACM Conference on Multimedia Systems (2011). doi:10.1145/1943552.1943572
16. Schulzrinne, H., Rao, A., Lanphier, R.: Real-Time Streaming Protocol (RTSP). IETF (1998)
17. ITU-T: G.1050 network model for evaluating multimedia transmission performance over internet protocol (2007)
18. Lu, Y., Fallica, B., Kuipers, F.A., Kooij, R.E., Van Mieghem, P.: Assessing the quality of experience of SopCast. Int. J. Internet Protoc. Technol. **4**(1), 11–23 (2009, March)
19. Cesar, P., Geerts, D.: Past, present, and future of social TV: a categorization. In: CCNC: Proceedings of the IEEE Consumer Communications and Networking Conference (2011). doi:10.1109/CCNC.2011.5766487
20. Boertjes, E., et al.: ConnecTV: share the experience. In: Proceedings of EuroITV, Amsterdam, the Netherlands, pp. 139–140 (2007)
21. Weisz, D., et al.: Watching together: integrating text chat with video. In: CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2007). doi:10.1145/1240624.1240756
22. Oehlberg, L., Duchenaut, N., Thornton, J.: Social TV: designing for distributed, sociable television viewing. In: Proceedings of EuroITV, Athens, Greece (2006)
23. Geerts, D., Cesar, P., Bulterman, D.: The implications of program genres for the design of social television systems. In: UXTV'08: Proceedings of the International Conference on Designing Interactive User Experiences for TV and Video, Mountain View (CA), USA (2008)
24. ITU-T: Recommendation P.800: methods for objective and subjective assessment of transmission quality (1996)
25. Mekuria, R.: Inter-destination synchronization for TV-Broadcasts, Delft University of Technology (2011)
26. Geerts, D., Vaishnavi, I., Mekuria, R., Deventer, M.O., Cesar, P.: Are we in sync? Synchronization requirements for watching online video together. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver (BC), Canada, pp. 311–314 (2011)
27. Cortina, J.M.: What is coefficient alpha? An examination of theory and applications. J. Appl. Psychol. **78**, 98–104 (1993)
28. Osborne, J.W.: Effect sizes and the disattenuation of correlation and regression coefficients: lessons from educational psychology practical assessment. Res. Eval. **8**(11) (2003)
29. Stokking, H., Van Deventer, M.O., Niamut, O.A., Walraven, F.A., Mekuria, R.N.: IPTV inter-destination synchronization: a network-based approach. ICIN'2010, Berlin, October 2010
30. Boronat, F., Montagud, M., Vidal, V.: Smooth control of adaptive media playout to acquire IDMS in cluster-based applications, IEEE LCN 2011, pp. 617-625. Bonn, October 2011
31. Hashimoto, T., Ishibashi, Y.: Group synchronization control over haptic media in a networked real-time game with collaborative work, Netgames'06, Singapore, October 2006
32. Montagud, M., Boronat, F.: Implementation and evaluation of an M/S scheme for inter-destination multimedia synchronization (IDMS). Netw. Protoc. Algorithm J. **3**(3), 80–98 (2011, December)
33. Ishibashi, Y., Tomaru, K., Tasaka, S., Inazumi, K.: Group synchronization in networked virtual environments. In: Proceedings of the 38th IEEE International Conference on Communications, pp. 885–890, Alaska, May 2003
34. Boronat, F., Guerri, J.C., Lloret, J.: An RTP/RTCP based approach for multimedia group and inter-stream synchronization. Multimed. Tool Appl. J. **40**(2), 285–319 (2008, June)
35. Boronat, F., Lloret, J., García, M.: Multimedia group and inter-stream synchronization techniques: a comparative study. Inf. Syst. **34**(1), 108–131 (2009, March)
36. Ishibashi, Y., Tasaka, S., Ogawa, H.: Media synchronization quality of reactive control schemes. IEICE Trans. Commun. **E86-B**(10), 3103–3113 (2003, October)

37. Ishibashi, Y., Tsuji, A., Tasaka, S.: A group synchronization mechanism for stored media in multicast communications. In: Proceedings of the INFOCOM '97, Washington, DC, April 1997
38. Ishibashi, Y., Tasaka, S.: A group synchronization mechanism for live media in multicast communications. In: IEEE GLOBECOM'97, Phoenix (AZ), USA, pp. 746–752, November 1997
39. Ishibashi, I., Tasaka, S.: A distributed control scheme for group synchronization in multicast communications. In: Proceedings of International Symposium Communications, pp. 317–323, Kaohsiung, November 1999
40. Ishibashi, Y., Tasaka, S.: A distributed control scheme for causality and media synchronization in networked multimedia games. In: Proceedings of the 11th International Conference on Computer Communications and Networks, pp. 144–149, Miami, October 2002
41. Ishibashi, Y., Hasegawa, T., Tasaka, S.: Group synchronization control for haptic media in networked virtual environments. In: Proceedings of the 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp. 106–113, Chicago, March 2004
42. Nunome, T., Tasaka, S.: Inter-destination synchronization quality in a multicast mobile ad hoc network. In: Proceedings of IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1366–1370, Berlin, September 2005
43. Kurokawa, Y., Ishibashi, Y., Asano, T.: Group synchronization control in a remote haptic drawing system. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp. 572–575, Beijing, July 2007
44. Hosoya, K., Ishibashi, Y., Sugawara, S., Psannis, K.E.: Group synchronization control considering difference of conversation roles. In: IEEE 13th International Symposium on Consumer Electronics, ISCE '09, Kyoto, Japan, pp. 948–952, May 2009
45. Akyildiz, I.F., Yen, W.: Multimedia group synchronization protocols for integrated services networks. IEEE J. Sel. Area Commun. **14**(1), 162–173 (1996, January)
46. Diot, C., Gautier, L.: A distributed architecture for multiplayer interactive applications on the internet. IEEE Netw. **13**(4), 6–15 (1999, July/August)
47. Mauve, M., Vogel, J., Hilt, V., Effelsberg, W.: Local-Lag and timewarp: providing consistency for replicated continuous applications. IEEE Trans. Multimed. **6**(1), 47–57 (2004, February)
48. Hesselman, C., Abbadessa, D., Van Der Beek, W., et al.: Sharing enriched multimedia experiences across heterogeneous network infrastructures. IEEE Commun. Mag. **48**(6), 54–65 (2010, June)
49. Vaishnavi, I., Cesar, P., Bulterman, D., Friedrich, O., Gunkel, S., Geerts, D.: From IPTV to synchronous shared experiences challenges in design: distributed media synchronization. Signal Process. Image Commun. **26**(7), 370–377 (2011, August)
50. Cronin, E., Filstrup, B., Jamin, S., Kurc, A.R.: An efficient synchronization mechanism for mirrored game architectures. Multimed. Tool Appl. **23**(1), 7–30 (2004, May)
51. Palazzi, C.E., Ferretti, S., Cacciaguerra, S., Roccetti, M.: On maintaining interactivity in event delivery synchronization for mirrored game architectures. In: IEEE Global Telecommunications Conference Workshops, pp.157–165, Dallas, December 2004
52. Ott, J., Perkins, C.: Guidelines on extending the RTP control protocol (RTCP), RFC 5968, September 2010
53. Friedman, T., Caceres, R., Clark, A.: RTP control protocol extended report (XR), RFC 3611, November 2003
54. ETSI TISPAN, "IMS-based IPTV stage 3 specification", TS 183 063 v3.4.6 (2010–12)
55. Löbner, T.: Implementing ETSI standardised RTCP-based inter-destination media synchronization. Master thesis, Hamburg, Diplomica Verlag (2011)

eGuided: Sharing Media in Academic and Social Networks Based on Peer-Assisted Learning e-Portfolios

Paulo N.M. Sampaio, Rúben H. de Freitas Gouveia, and Pedro A.T. Gomes

Abstract Over the past years, different methodologies have been applied within classrooms in order to provide students with a successful learning process, leading thus to the proposal of different methodologies, paradigms and tools, such as blended learning, peer-assisted learning (PAL) and e-portfolios. In particular, teachers can use e-portfolios to understand students' learning needs and provide them with an individualised learning approach. In this chapter, we introduce eGuided, a peer-assisted learning-based e-portfolio system that customises students' learning experiences based on their academic and professional background. eGuided provides a dynamic environment, supported by an academic and social network, where individuals can share their academic and professional goals and respective multimedia content. In this chapter, we analyse the implementation of this dynamic environment, as well as how sharing media among different individuals can improve students' learning experiences.

1 Introduction

Currently, high-level education faces a number of issues relating to the diversity of students and to their increasing degree of mobility. As discussed in [1], these issues are due to a set of factors, such as students' different cultural and educational background, isolation, the need for more engaging learning experiences, lack of

P.N.M. Sampaio (✉)
Computing and Systems Research Center – NUPERC, Salvador University – UNIFACS, Salvador, Bahia, Brazil
e-mail: pnms.funchal@gmail.com

R.H. de Freitas Gouveia • P.A.T. Gomes
Madeira Interactive Technologies Institute (M-ITI), Universityof Madeira (UMa), Funchal, Madeira, Portugal
e-mail: rubahfgouveia@gmail.com; patg@sapo.pt

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_19, © Springer-Verlag London 2013

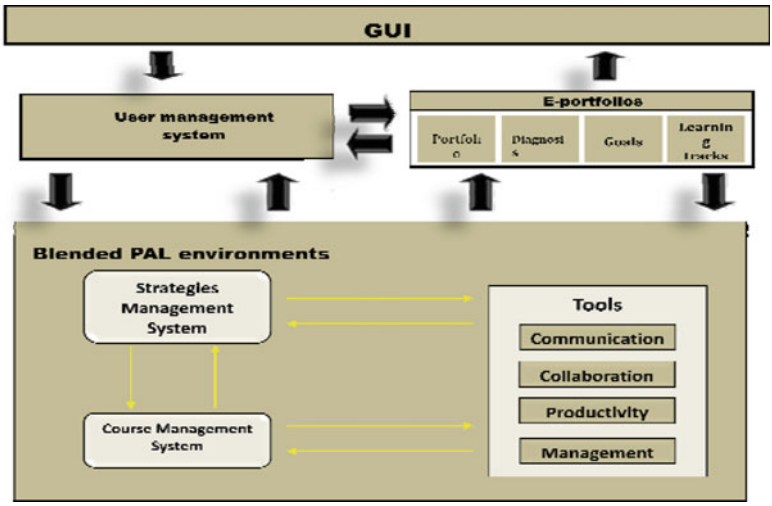


Fig. 1 Architecture of the blended peer-assisted learning platform

sense of belonging and lack of collaboration among students, among others. These issues result in a growing demand of new methodologies, approaches and tools that foster students' socialisation and improve their learning outcomes. In this context, methodologies, techniques and innovating tools, supported by Information and Communication Technologies (ICTs), have been developed to promote better educational experiences. Among these solutions are blended learning, peer-assisted Learning and e-portfolios.

In particular, a blended peer-assisted learning (ePAL) platform [2] has been developed at University of Madeira which integrates different e-learning tools and platforms. This platform is composed by two main applications, which provide a key role in improving students' learning outcomes: an e-portfolio and a blended PAL platform (see Fig. 1).

These applications are integrated in the ePAL platform as follows: Based on a student's e-portfolio analysis, the e-portfolio application determines which PAL strategies are appropriate to improve the student's learning experience. This analysis is based on all assessments a student has received in his e-portfolio, where his skills (e.g. communication, teamwork) are graded, helping to identify eventual assets and weaknesses he possesses. The average grade of each skill will determine which PAL strategies are appropriate to students' learning experience. For instance, based on the assessments performed on a certain student's e-portfolio, it was found that he possesses strong communication and problem-solving skills. Therefore, the systems diagnosis is that reciprocal teaching is the most appropriate PAL strategy in order to optimise his learning outcomes. This diagnosis is achieved based on an in-depth study of learning outcomes, validated through the analysis of several inquiries that were carried out in order to validate the method proposed in [3].

With the feedback provided by the e-portfolio application, professors can adapt their courses to the learning needs of their students (intentional learning), implementing, through the ePAL platform, for instance, reciprocal teaching activities for student who are supposed to learn better with this strategy. In this chapter, we do not address the development aspects of the blended peer-assisted learning platform, in order to further explore the development aspects of the e-portfolio tool.

The blended PAL platform allows a professor to set up and configure PAL strategies, indicating all activities that should be carried out by students during a PAL session, and which ICTs (e.g. videoconference, whiteboard) are available to carry out these activities. The process of selecting strategies and tools will be important in the future of PAL sessions, contributing to a more effective and customisable learning process.

In this chapter, we study and analyse the referred paradigms and define a comprehensive relationship between e-portfolios and PAL learning strategies, identifying how an e-portfolio can be assessed and consequently guide students towards a learning strategy that satisfies their learning needs. In particular, we present eGuided, a peer-assisted learning-based e-portfolio system that customises students' learning experience based on their academic and professional background. eGuided provides a dynamic environment supported by an academic and social network, where individuals can share their academic and professional goals and their respective multimedia content. We aim at analysing the main aspects concerning the implementation of a Social and Academic Network in eGuided and also determining how students' learning experience can be improved through content sharing.

2 From e-Learning to e-Portfolios (Blended Learning, PAL, e-Portfolios)

The proposal of new solutions and methodologies that improve students' learning experience and engagement with their learning process has been a major challenge to the scientific community and teachers. With this in mind, and in order to provide an insight about students' academic and professional development, different methodologies, paradigms and tools, such as blended learning, peer-assisted learning (PAL) and e-portfolios, have been proposed.

2.1 Blended Learning

Blended learning combines different learning methodologies and technologies, relying on different online educational environments, providing thus a more effective learning process. In general, blended learning can be adapted to solve practical constraints that may arise during the deployment of a course (e.g. physical facilities,

schedule, budget, etc.). Blended learning improves the quality and diversity of human interaction in a learning environment, providing realistic practical opportunities for learners and teachers to make learning independent, useful and sustainable.

2.2 Peer-Assisted Learning

Peer-assisted learning (PAL) is a learning paradigm in which students that share the same academic level interact with each other, under the supervision of students with a higher academic level. This approach aims at facilitating the process of knowledge acquisition, since each student is able to call upon other students to express doubts regarding to the subjects being studied or when experiencing difficulties to accomplish certain tasks.

Students that have completed a course can volunteer to help other students who are undergoing the same course in the following years, helping them to improve their understanding and boost their learning. PAL can also help first year students to integrate themselves in a new academic environment.

2.2.1 Types of PAL

Several different PAL learning strategies have been proposed; however, four of the most applied and generic strategies were adopted in this project. Each of these strategies portray different learning approaches, therefore each student (assisted by his teacher) can choose the strategy that better suits him. These strategies are:

- Role Playing – In this strategy students have to simulate real situations, improving thus their communication and improvisational skills.
- Reciprocal Teaching – Strategy that helps students gain new forms of comprehension, namely, formulate questions, summarise information, make predictions and clarify problems.
- Peer Tutoring – In this strategy one of the senior students with distinguished learning outcomes assumes the role of the tutor while other students assume the role of tutees.
- Cooperative Learning – A learning methodology where students with different academic levels work together in cooperative groups.

2.3 e-Portfolios

An e-portfolio is an electronic repository of acquired learning – knowledge, skills and abilities acquired through formal/informal, accidental and incidental learning [4]. An e-portfolio is, as a pedagogical tool, composed of an online repository of

work, organised and structured throughout a certain period of time. One of the outcomes of an e-portfolio is to provide a wider and more detailed view of the student's learning experience, and also explore different aspects of his cognitive, metacognitive and affective development [5].

e-Portfolios are currently used in different contexts, from recruitment to assessment, and can be grouped into the following categories: student e-portfolios, teacher e-portfolios and institutional e-portfolios [6]. Nowadays, most e-portfolios are hybrid, since users, in general, do not create an e-portfolio strictly for assessment, development or showcase purposes, but instead try to combine these into a richer e-portfolio.

The main goal of this project is to study and analyse the referred paradigms and build a comprehensive relationship between e-portfolios and PAL learning strategies, in order to identify how an e-portfolio can be assessed and consequently guide students to a learning strategy that meets their learning needs. In the next section, we present methodology proposed to correlate an e-portfolio with PAL strategies.

3 Steering Students to ePAL Strategies Based on e-Portfolio Assessments

The process of adding/receiving comments and evaluations to e-portfolios allows users to obtain a better perception of "where and how their learning experience can be improved"; these elements, however, do not provide enough information in order to direct students straightforwardly to appropriate learning strategies.

Therefore, a detailed analysis was carried out to the structure of e-portfolios, providing an understanding of which data can be collected from e-portfolios, in order to identify user's learning profiles and consequently direct them to appropriate learning strategies.

3.1 Correlating e-Portfolio with Skills

In general, e-portfolios are composed of personal data (academic, professional, etc.) and artifacts (which feature a student's work exhibits). Personal data are generic, which hinder a straightforward correlation between users' profiles and their appropriate learning strategy. As for artifacts, these are exposed in e-portfolios since they demonstrate skills [7] possessed by the users.

In this case, a hybrid assessment approach was adopted by the combination of skills with evaluations. The result represents a change in the traditional evaluation system, since evaluations become directly related to skills. In other words, when evaluating, users firstly identify the skills they believe to find in an e-portfolio

Table 1 Minimal requirement for PAL learning strategies related to essential skills

PAL learning strategies/ skills	Adaptability	Self-discipline	Communication	Leader-ship	Organisation	Problem solving	Team-work
Cooperative learning	Good	Good	*Very good*	Irrelevant	Sufficient	Sufficient	*Very good*
Peer tutoring	Irrelevant	Irrelevant	Irrelevant	Irrelevant	Irrelevant	Irrelevant	Irrelevant
Reciprocal teaching	Good	Good	Good	Irrelevant	Sufficient	*Very good*	Good
Role playing	*Very good*	Good	*Very good*	Irrelevant	Irrelevant	Good	Good

and then assess them. Besides obtaining a list of possessed skills, users are also able to determine their skills' level (insufficient, sufficient, good, very good or excellent) [8].

In order to determine which skills a user has to possess so that a given learning strategy shall be recommended to him, a literature review was carried out [9–14]. In the course of this study, a generic list of seven essential skills was obtained: *auto-discipline*, *adaptability*, *communication*, *teamwork*, *organisation*, *leadership* and *problem resolution*.

3.2 Determining PAL Strategies Based on Skills

After obtaining the "seven essential skills", it became essential to understand the relation between these skills and the PAL learning strategies.

It is important to recall that skills are assessed not only by the owner of an e-portfolio but also by other users, determining an average level of a user in a given skill. A low level denotes that a user has a weak level, or weakness, in a certain skill. On the other hand, a high level indicates a strength possessed in a certain skill. Thus, we concluded that:

- A given learning strategy shall be recommended to a user if his strong skills (assets) meet this strategy's requirements.
- A strategy will be recommended to a user if this strategy proves to help reducing possible weaknesses (weak skills).

More than a learning strategy may be recommended to a user at a time, being up to him to follow one that potentiates eventual strengths or one that helps addressing his weaknesses. The four PAL learning strategies were analysed, ending up in a comprehensive idea of their requirements related to the seven essential skills, as presented in Table 1.

Considering cooperative learning, this strategy will be recommended to a student if his strong skills are teamwork and communication. As for the student's weaknesses, cooperative learning shall be recommended to a student if he possesses low grades (weak skills) in communication and workgroup and in opposite needs to improve them.

As for peer tutoring, the only requirement students have to fulfil to follow this learning strategy is to be willing to learn. Therefore, peer tutoring is a generic strategy which will always be recommended to students.

Reciprocal teaching relies upon problem-solving skills; therefore, this strategy shall be recommended to a student if he has a very good level in this skill. On the other hand, if a student possesses a low score in the problem-solving skill, and needs to improve it, this strategy should be recommended to him.

Similarly to the previous PAL learning strategies, if a student has a very good grade in adaptability and communication skills, then role playing will be recommended to him. Nevertheless, if he reveals a weakness in these skills and needs to improve them, role playing should be recommended to him.

Besides the seven essential skills, users can also identify and assess any other skills they deem necessary. However, only the seven essential skills are considered by the system when recommending appropriate PAL learning strategies to students.

4 eGuided: A Peer-Assisted Learning-Based e-Portfolio

eGuided is an e-portfolio management system that allows the personal follow-up of students' assets and recommends appropriate PAL learning strategies to students according to their educational needs. eGuided aims at supplying a peer-assisted learning-based e-portfolio in order to provide development and maintenance of an e-portfolio composed of academic, professional and personal skills/competences; 360° evaluation of an e-portfolio (auto, lower, peers and superiors); evaluation of students' skills (adaptability, communication, discipline, leadership, organisation, problem solving, teamwork, etc.); recommendation of PAL learning strategies (cooperative learning, peer tutoring, role playing and reciprocal teaching); academic and social networking; academic follow-up; and a virtual space for professional opportunities.

eGuided's main features include My e-Portfolio, Skills, Friends, Messages and Notifications, and User Profile. "My e-Portfolio" is the main section of eGuided, in which users can manage their e-portfolio. An e-portfolio is divided into three sections: Academic (education background), Professional (professional experience) and Personal Skills and Competences (which are skills that a user has developed along his lifetime and are not necessarily educational or professional). Figure 2 illustrates the "My e-Portfolio" section.

If a user selects, for example, the "Academic" section, all his respective Academic Institutions will be shown, with the following options (see Fig. 3): "Add Institution(s)", "Assess", "Delete" and "Grant Access".

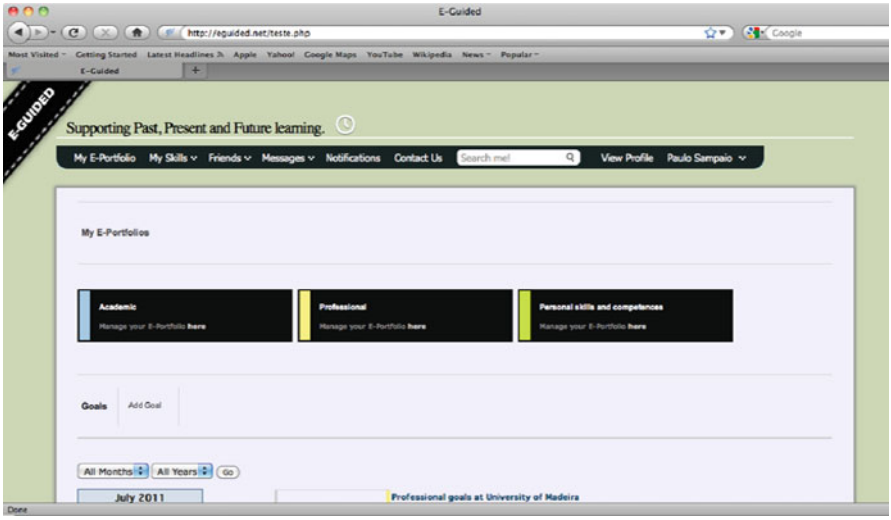


Fig. 2 My e-Portfolio management

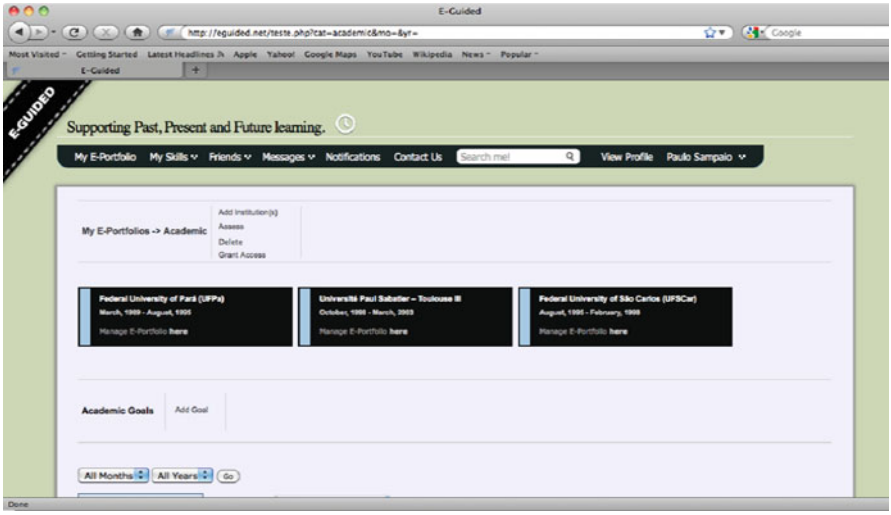


Fig. 3 Managing academic section

By selecting the option "Add Institution(s)", users will be presented with a list of existing institutions. Alternatively, users can add new institutions to this list. When adding a new Academic Institution, users may also add new programs (such as BSc in Computer Science) and new disciplines (such as Data Structures) as well as their period of attendance.

Any content in an e-portfolio can be subject of assessment. This assessment can be carried out by the owner of the e-portfolio or by other users whose access where granted as a peer colleague by a hierarchically superior person (employer or

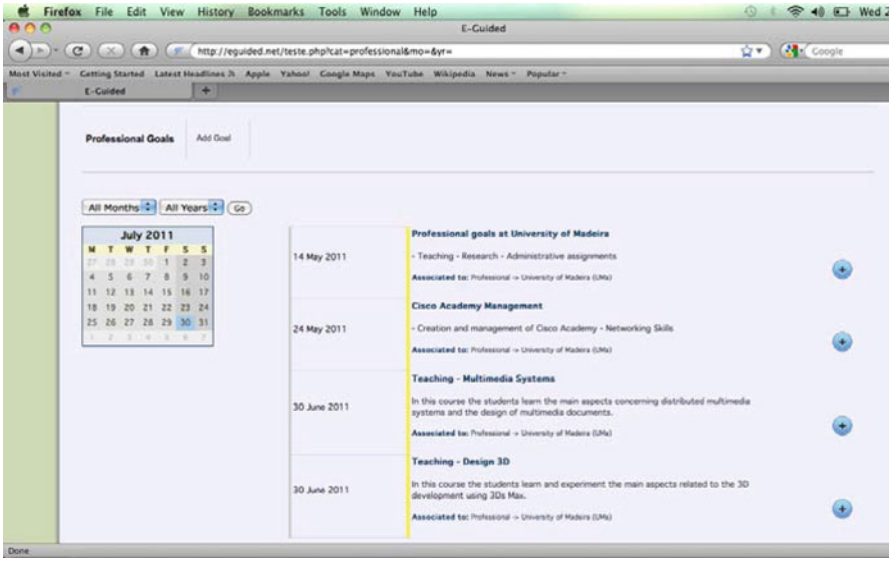


Fig. 4 Professional goals

a teacher) or by an inferior person (employee or a student, if the owner is, e.g. a teacher). In order to assess an element of an e-portfolio, evaluators must first identify the skill(s) – predefined or not – they believe to find in an e-portfolio (e.g. adaptability, communication, discipline). After choosing the skill(s), the user must grade each of the identified skill.

Users may also add goals to their e-portfolios. A goal is an objective the user wants to achieve by adding a particular artifact in his e-portfolio. An academic goal may be associated with a course, a program or an institution. For instance, a user can add a goal "Gain programming skills" in his Academic Institutions. A goal is composed by a name, a short description and its period of accomplishment. Adding goals allow users to register and keep track of deadlines, while allowing users to organise artifacts by goals. Figure 4 illustrates some goals included in the professional section of a certain user's e-portfolio.

"My Skills" is one of the main sections of eGuided, distinguishing eGuided from existing e-portfolio platforms. This section allows the owner of an e-portfolio to follow the progress he has made, regarding to the collaborative evaluations performed on his e-portfolio. It is divided into two subsections: My Skills and My Learning Strategies.

"My Skills" presents users with all skills identified in their e-portfolio and their average grades, along a period of time. These include the seven essential skills, as well as any other skill identified. Skills are presented in a temporal chart, ranking from weak to excellent, as depicted in Fig. 5.

The "My Learning Strategies" section presents users with their recommended PAL learning strategies, in the form of a temporal chart, as depicted in Fig. 6. Each learning strategy is displayed in a different colour, and by hovering a mouse cursor

Fig. 5 My skills

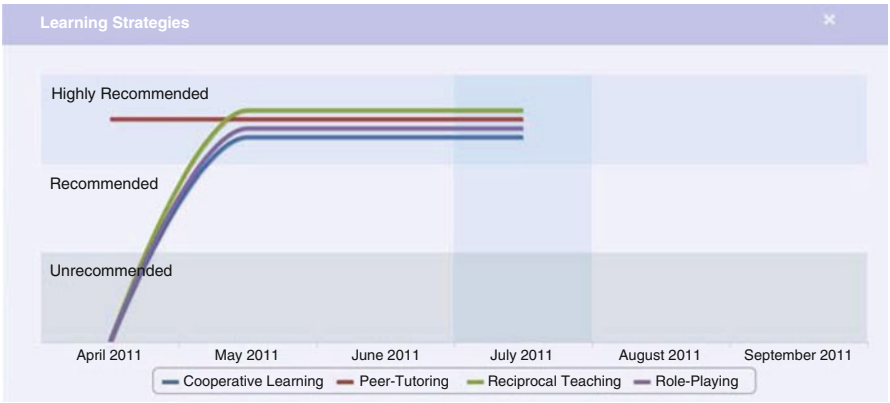


Fig. 6 My learning strategies

over a respective learning strategy's line, the diagnostic proposed for this particular recommendation (based on the e-portfolio's evaluation) will be presented.

The "View Profile" section displays a user's personal information in a *resume* format (Fig. 7), which is available for other users to view. Through the "View Profile" section, users can display the following information: Follow me, which display users e-mail, social networks, etc.; Personal Information, such as nationality and birth date, among others; Areas of Interest; Spoken Languages; Skills; Educational background; Work Experience; and Personal Skills and Competences.

In particular, displaying skills together with areas of interest, spoken languages, academic data and professional data is important and reveals a distinction compared to other existing e-portfolio platforms, since it enables potential employers or collaborators to browse and locate user's e-portfolios easier, by targeting, for example, users with particular personal interests.

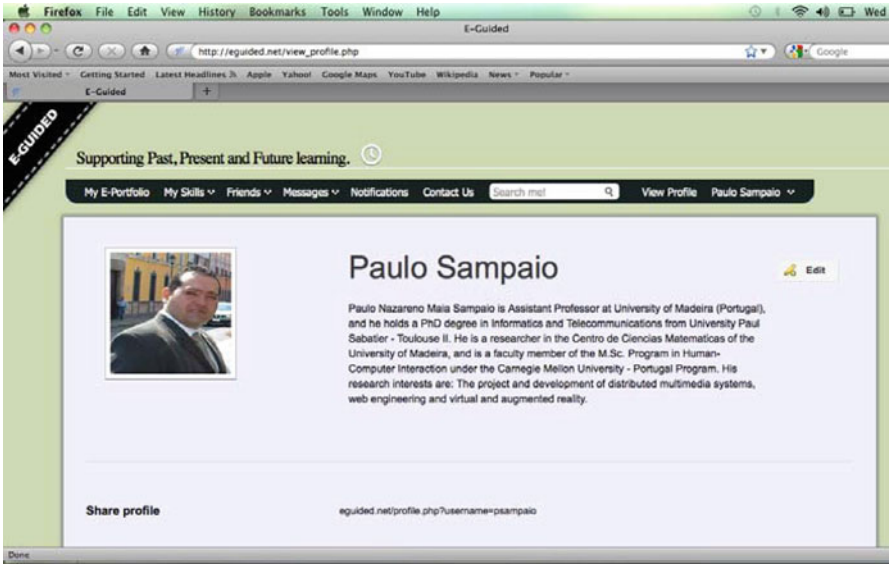


Fig. 7 Presentation section

After having presented the main features of eGuided, the next section addresses the features of eGuided that enable social networking and sharing of multimedia content.

5 Sharing Social and Academic Media in eGuided

The Social Network approach aims at understanding social interaction conceived and investigated through the properties of relations between and within individuals [15]. Similarly, an Academic Network allows students, teachers and other academic staff to share their profile and achievements while sharing similar goals. Currently, social and academic activities are developed when participants can contact each other using synchronous (e.g. chat) and asynchronous (e.g. posts) resources. Also, an important feature of Social and Academic Networks, generally not explored in existing platforms, is the ability to share and assess documents and media files (e.g. text, images, video, audio, etc.). This feature enriches Social and Academic Networks, allowing participants to share common goals through a collaborative approach.

Through eGuided, users can upload documents and files, associating them with goals. These documents and files are called artifacts and illustrate a user's achievements related to a certain goal in his e-portfolio. For instance, a student can add a goal "Exhibiting personal projects and assignments related to the conclusion

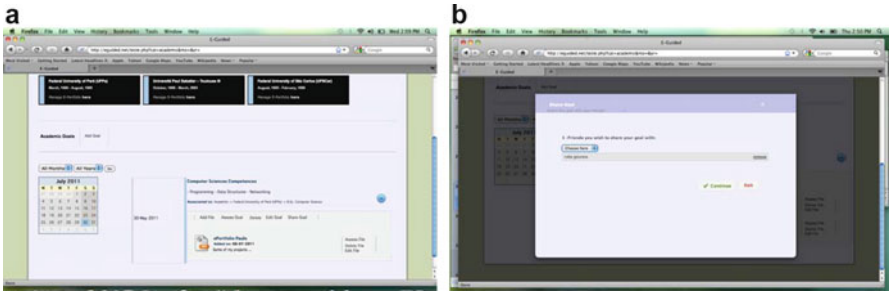


Fig. 8 Sharing goals and artifacts (**a**) Creating a new goal and adding an artifact (**b**) Sharing a goal with contacts

of his Masters course", associating further on all files relevant to expose this goal. As discussed before, a goal is defined as a task or competence a user wants to accomplish or acquire related to a specific entry in his e-portfolio.

In general, users can share common goals, such as in a group of colleagues that are working on the same assignment or in a class when a professor needs to share documents with his students. In these cases, eGuided allows participants to create their goals, upload any type of documents and media (artifacts) associated with that goal and share the goal with any number of participants of the platform (friends), as illustrated respectively in Fig. 8a, b. It is important to note that, besides the user who created a goal, anyone with whom he shares the goal can add, assess, edit or delete existing files in the goal.

Besides the ability of sharing goals with other eGuided users, a user may also share his e-portfolio on the web, if he wishes, for example, to exhibit his skills and competences to a potential employer.

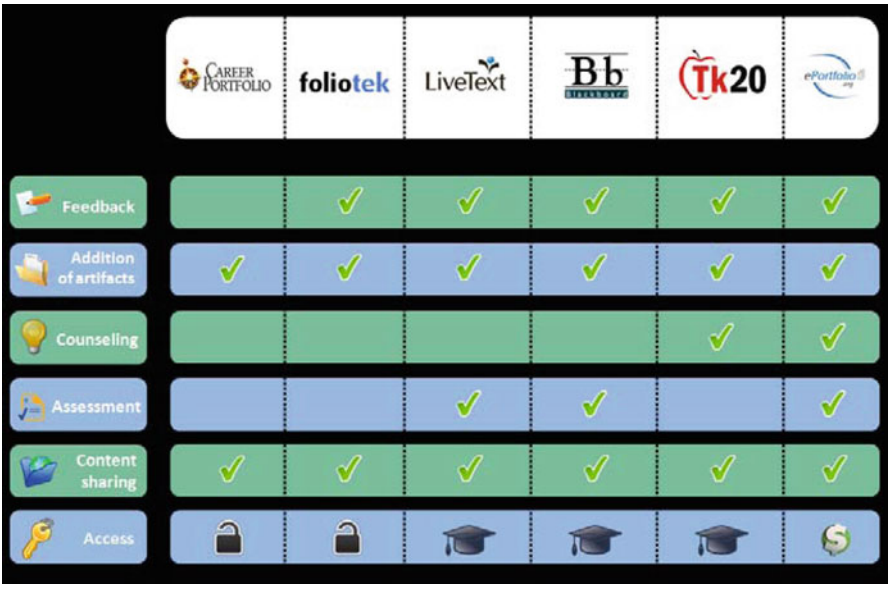
By providing the ability of adding personal artifacts (e.g. images, videos, etc.), relating these to specific goals and sharing and assessing goals (by the owner of the e-portfolio, a colleague, a superior or a subordinate), eGuided represents an important step towards a collaborative view of e-portfolios. As a consequence, these can be more complete, impartial and display a person's competences and skills.

6 Related Works

Currently, there is a large number of existing e-portfolio platforms, most of them commercial or private academic solutions not allowing the external access [16]. Some of the existing e-portfolios platforms are Career Portfolio Program (CPP) [17], foliotek [18], LiveText [19], Blackboard [20], Tk20 [21] and ePortfolio [22].

Based on these platforms, a comparative presentation is proposed in Table 2, with comparative criteria such as Collaborative Feedback, Addition of artifacts, Counseling (identification of educational needs), Assessment, Content sharing and Access.

Table 2 Comparison of the e-portfolio platforms



this figure will be printed in b/w

By analysing Table 2, we can conclude that platforms Career Platform and foliotek have free access, but lack relevant features only present in paid platforms or in those limited to academic communities. Most of all, none of these platforms provide an effective mechanism for counselling and identifying educational needs, and collaborative assessment of e-portfolios. eGuided proves to be valuable since it provides a collaborative approach for building and assessing e-portfolios under a PAL perspective, and has free access. In fact, the collaborative assessment of e-portfolios in eGuided provides an insight about a user's skills and competences, steering users (students, teachers and potential employers) to a more reliable academic and professional environment.

7 Conclusions

This chapter presented the main aspects regarding to the development of an e-portfolio platform, called eGuided. eGuided is an e-portfolio management system that allows the personal follow-up of students' assets and the recommendation of appropriate PAL learning strategies to them according to their educational needs.

Based on the literature review, it was possible to verify that most existing e-portfolio platforms provide different functionalities, including the addition of academic and professional records, the inclusion of multimedia content, sharing content among users and assessment. However, the main difference between

eGuided and existing platforms relies on the collaborative approach for building and assessing an e-portfolio, which allows the generation of more reliable exhibits, providing a useful tool for teachers to direct their students to appropriate PAL learning strategies, as well as providing valuable information for potential employers who are seeking professionals with specific skills and competences.

References

1. Ng, E.M.W.: Engaging student teachers in peer learning via a blended learning environment. Issue. Inform. Sci. Inf. Technol. **5**, 325–334 (2008)
2. Teixeira, J.M., Camacho, M.F., de Gouveia, R.H., Sampaio, P.N.M.: Blended peer-assisted learning platform: improving learning outcomes with a collaborative environment. J. Educ. Technol. Syst. **39**(4), 371–395 (2010–2011)
3. Gouveia, R.H.deS.: e-guided: a platform for the generation of e-portfolios. M.Sc. dissertation, Informatics Engineering, University of Madeira, Portugal (2011)
4. Baker, K.C.: ePortfolio for the assessment of learning. FuturEd White Paper. Canada. http://www.futured.com/documents/FuturEdePortfolioforAssessmentWhitePaper.pdf (2005)
5. Wang, L.-C., Chen, M.-P.: Enhancing ICT skills learning through peer learning: perspectives 390 of learning style and gender. Int. J. Educ. Inf. Technol. **2**(1), pp. 18–23 (2008)
6. Billings, et al.: e-Portfolio basics: types os e-Portfolios. In: Regis Electronic Portfolio Project. Available http://academic.regis.edu/LAAP/eportfolio/index.html (2003)
7. Definition of Skill. Wikipedia – The Free Encyclopedia. Retrieved October (2011). http://en.wikipedia.org/wiki/Skill
8. Academic Grading in Portugal. Wikipedia – The Free Encyclopedia. Retrieved October (2011), grading in Portugal http://en.wikipedia.org/wiki/Academic grading in Portugal
9. 15 personal skills you need on the job. Training On-Line Magazine. Retrieved October (2011). http://www.trainingmag.com/article/15-personal-skills-you-need-job
10. Handsal, R.: Your job skills portfolio: giving you an edge in the marketplace. Quintessential Careers. Retrieved October (2011). http://www.quintcareers.com/career_doctor_cures/skills_employers_seek.html
11. Carnevale, A.: Employability skills – an employer perspective getting what employers what out of the too hard basket. Skilling Solutions Queensland – Department of Education, Training and Employment. Retrieved October (2011). http://www.skillingsolutions.qld.gov.au/resources/employability%20skills.acci.pdf; http://www.skillingsolutions.qld.gov.au/resources/employability_skills.acci.pdf
12. Careers4Graduates. What skills do employers want? Retrieved October (2011). http://www.careers4graduates.org/changing/employerswant.phtml
13. Carnevale, A.: Workplace basics: the essential skills that employers want. ERIC – Education Resources Information Center. Retrieved from October (2011). http://www.eric.ed.gov/PDFS/ED319979.pdf
14. Askov, E.N., Aderman, B., Sherow, S., Hemmelstein, N., Clark, C.: Upgrading Basic Skills for the Workplace. Institute for the Study of Adult Literacy, Pennsylvania State University, University Park (1989)
15. Definition of Social Network. Wikipedia – The Free Encyclopedia. Retrieved October (2011). http://en.wikipedia.org/wiki/Social_network
16. Summers, T., Zaldivar, M.: ePortfolio growth across the university and the country. Learning Technologies – Virginia Tech. Retrieved October (2011). http://www.lt.vt.edu/LT_Update/2010_LT_Update.pdf
17. University of Saint Joseph: Career ePortfolio program. http://www.usj.edu.mo/en/students/career/career-portfolio-program (2011)

18. The Foliotek System – Helping Creating Passionate and Competent Professionals. Retrieved January (2012). www.foliotek.com/
19. LiveText. Retrieved January (2012). https://www.livetext.com
20. Blackboard – Technologies and Solutions Build for Education. Retrieved January (2012). http://www.blackboard.com
21. TK20 WebSite – Comprehensive Outcomes-based Assessment and Reporting. Retrieved January (2012). http://www.tk20.com/
22. ePortfolioWebSite. Retrieved January (2012). http://www.eportfolio.org/

Exploiting Social Media for Music Information Retrieval

Markus Schedl

Abstract This chapter will first provide an introduction to information retrieval (IR) in general, before briefly explaining the research field of music information retrieval (MIR). Hereafter, we will discuss why and how social media mining (SMM) techniques can be beneficially employed in the context of MIR. More precisely, motivations for the common MIR tasks of *music similarity computation*, *music popularity estimation*, and *auto-tagging music* will be provided, and the current state-of-the-art in employing SMM techniques to these three tasks will be elaborated.

Developing music similarity measures is an important task in MIR as such measures are a key ingredient for music recommendation systems, automated playlist generators, and intelligent browsing interfaces, among others. In this chapter, it will be shown how to infer music similarity information from microblogs, collaborative tags, web pages, playlists, and peer-to-peer networks. Estimating the popularity of a music item is obviously important for the music industry but also to create serendipitous music retrieval and recommendation systems. Therefore, approaches that derive such information from web page counts, geo-located microblogs, a peer-to-peer network, and a social music platform will be reviewed. Eventually, different music auto-tagging methods that assign semantic labels to music pieces will be presented. In particular, computational approaches that rely on machine learning techniques as well as human-centred strategies that infer tags directly from some kind of user input (e.g. "games with a purpose") will be addressed.

M. Schedl (✉)
Department of Computational Perception, Johannes Kepler University, Altenberger Straße 69, 4040 Linz, Austria
e-mail: markus.schedl@jku.at

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_20, © Springer-Verlag London 2013

1 Introduction to Information Retrieval

The discipline of information retrieval (IR) is a mature field of research as early work dates back to the 1950s, for instance [59]. Since I can only give a very brief introduction to this exciting field here, the interested reader is referred to one of the many excellent books that offer comprehensive coverage of IR. I personally recommend [22] for an introduction and [3] and [8] for a more comprehensive coverage.

Broadly speaking, IR is concerned with elaborating and testing methods to uncover information from potentially large corpora of text (traditional IR) or (more recently) multimedia, in response to the user's expression of an *information need*. This information need is usually given as a text *query*, the classical example being a user who types in a query string into his or her preferred *search engine*. Texts are most frequently organised in the form of *documents*, although other representations exist. Hence, it is usually also documents which are returned as response to a query to a search engine.

In order to be able to promptly provide search results for millions of queries issued every hour to major search engines, enormous amounts of computational power are required. But of no lesser importance are highly efficient representations of the documents. For this purpose, an *inverted index* is commonly created from the documents. Such an inverted index stores, for each term t, a list of documents in which t occurs or a list of documents and the precise positions of t within each document. The former is referred to as *document-level inverted index*, *record-level inverted index*, *inverted file index*, or just *inverted file*; the latter is typically named *full inverted document index*, *word-level inverted index*, *full inverted index*, or *inverted list*. The major advantage of a full inverted index is that it allows for *phrase search*, that is, finding an exact phrase within a document, not only a single term. In a regular expression notation, the two variants of the mapping implemented by the two flavours of indexes can be written as follows:

document-level index:	`term`$\mapsto$`document*`
world-level index:	`term`$\mapsto$`(document,position+)*`

If the user now wants to search for a particular topic, expressed as a query q, the retrieval system computes a matching score between q and the indexed documents D. A common approach is to compute term weights $w(q, d)$ between q and each document d, which estimate the importance of the document for the query. The documents are then ranked with respect to $w(q, d)$ and displayed to the user in descending order of term weight. This classical retrieval approach is often called *term vector model* or *vector space model*. Since its proposal in 1975 by Salton et al. [71], many extensions as well as alternative retrieval approaches have been suggested. More recent methods include *probabilistic retrieval* [45] and *graph-based models* [7].

2 Music Information Retrieval at a Glance

Unlike traditional IR, music information retrieval (MIR) is a relatively young field of research, dating back only about a decade. An early and quite general definition of MIR, which highlights the multidisciplinarity of the field, is given by Downie in [17]:

> MIR is a *multidisciplinary* research endeavour that strives to develop innovative *content-based searching schemes*, *novel interfaces*, and *evolving networked delivery mechanisms* in an effort to make the worlds vast store of music accessible to all.

A later definition given by Schedl [72] focuses on extracting and processing musical information on different levels and modalities:

> MIR is concerned with the *extraction*, *analysis*, and *usage* of information about any kind of music entity (for example, a song or a music artist) on any representation level (for example, audio signal, symbolic MIDI representation of a piece of music, or name of a music artist).

Due to recent developments, such as audio and music streaming services (e.g. `Spotify` [41]), personalised web radio (e.g. `last.fm` [27]), and increasing use of multimedia data in social media, MIR has gained considerably in importance as a research field.

Although MIR is a highly multidisciplinary research field, including areas as diverse as music theory, library science, psychology, law, and artificial intelligence, one of its key goals is to better understand how humans perceive, create, process, and interact with music. Given its strong connection to computer science, MIR approaches to achieve this broad goal typically involve elaborating computational models of music perception. These approaches commonly take as input the audio signal or other modalities of a music item and compute *features* that strive to describe particular aspects of the music item, for example, rhythm, harmony, or timbre. Figure 1 depicts a schematic and simplified illustration of how a signal-based (content-based) audio feature extractor works. First, the audio signal is sampled and digitised, yielding a representation as *pulse code modulation* (PCM). For instance, when producing a compact disc, the sampling frequency is typically 44,100 Hz, and each sample is described via 16 bits. For a stereo recording, the data volume hence amounts to 176,400 bytes per second. The PCM representation is then split into (often overlapping) frames with a typical length of between 2^8 and 2^{12} samples. Low-level features in the *time domain* can then be computed directly on these frames. To capture frequency information, alternatively, it is very common to apply a *windowing function* to each frame and subsequently compute the *fast Fourier transform* (FFT) [13], which converts the data from a time–amplitude representation into a frequency–magnitude system. Hereafter, several post-processing steps are commonly performed, for instance, employing some psychoacoustic model of human auditory perception. Eventually, one regularly has to decide how to combine the features computed for each frame of a piece of music to create a global representation. Methods range from computing simple statistical moments to complex time-series modelling via *hidden Markov models* (HMM) [5].

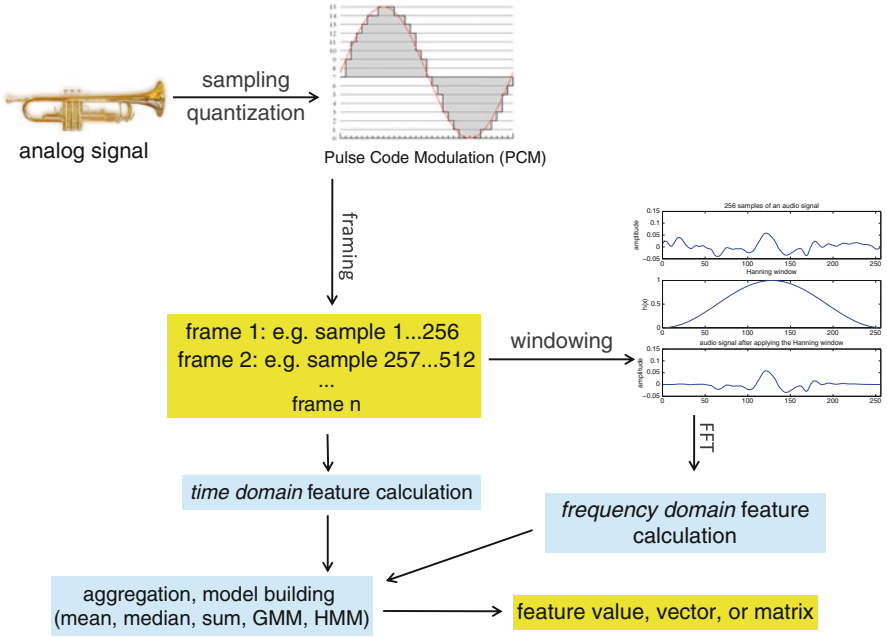


Fig. 1 Basic scheme of an acoustic feature extractor

The computational features extracted via algorithms similar to the one just described can be used for a wide range of MIR tasks, for instance, to *estimate similarity between music items* which in turn enables the creation of *music recommendation systems*, of *playlists automatically generated*, and of clustering-based *user interfaces* to music collections. If semantic labels describing the music items are available, another popular task is to automatically learn relations between audio features and semantic descriptors. This task is commonly referred to as *auto-tagging*.

The content-based feature extraction framework described above represents the traditional MIR strategy to computationally grasp aspects of a music item that should relate to human music perception. In the past few years, however, MIR has seen a paradigm shift to incorporate additional factors into computational models of music perception and description. In particular, contextual aspects of the music items and of the listener are increasingly taken into account. Integrating these with traditional content-based methods, Fig. 2 shows the three broad pillars from which perceptual music information can be extracted, according to [76].

Music content feature extractors derive information directly from the audio representation of a piece of music, by applying signal processing techniques. A typical example are features inferred from time-invariant *Mel frequency cepstral coefficients* (MFCC) representations of the audio signal, which serve to some extent to describe the coarse timbre of an audio signal. Overviews of common content-based extraction techniques are provided, for instance, in [9, 20, 57].

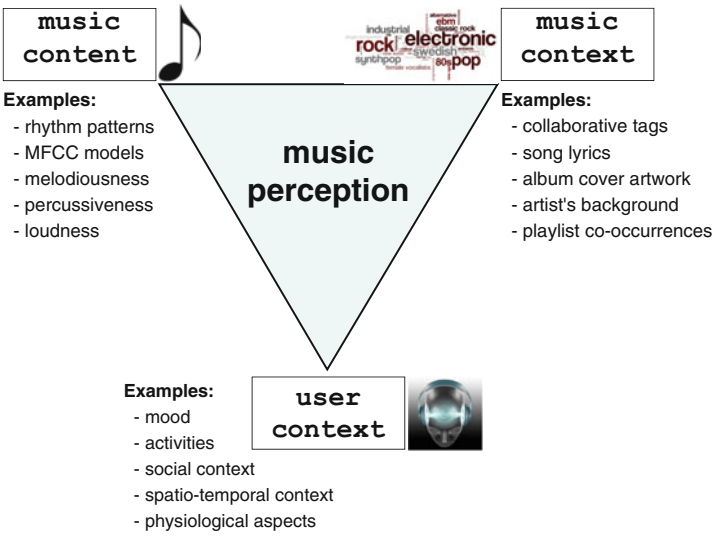


Fig. 2 Categorisation of computational aspects that influence music perception

Music context refers to aspects that are not encoded in the audio signal (or cannot be extracted with current methods), nevertheless are related to a music item. For instance, collaborative tags about a performer, semantic meaning of song lyrics, or the political background of an artist fall into this category. More details on feature extraction and similarity estimation from the music context can be found, for example, in [75].

User context relates to personal properties, preferences, and feelings of the music listener. The user context hence includes the user's mood, activities, friends, or level of musical training. Although these highly individual factors are obviously influential on music perception, MIR literature centred around the user is relatively sparse. Among the existing work, I would like to highlight the following: Cunningham et al. present an interesting study on why people dislike particular music [15]; Lee conducted a thorough analysis of natural language music queries [55] and personalised and user-aware music retrieval and recommendation are treated, for example, in [4, 10, 81].

Finally, it is noteworthy that some aspects fall into more than one category. For example, song lyrics might be seen to belong to the *music content* as they are obviously encoded in the audio signal. However, with current MIR techniques, it is impossible to extract and convert them to a semantically meaningful textual representation. On the other hand, many web pages list huge amounts of song lyrics, which make it easy to extract them from a *contextual* data source. I therefore predominantly see them in the *music context* category. A similar overlap might occur for collaborative tags. One can argue that such tags are the outcome of many users, hence would count them to the *user context*. However, according to my categorisation, the *user context* refers to individual, personal factors of the user, not to user groups.

3 Social Media Mining in Music Information Retrieval

Usage of social media has seen a tremendous increase during the past couple of years. People create, modify, and most importantly share massive amounts of multimedia data (text, images, music, videos) on platforms such as `Twitter` [42], `Facebook` [34], `last.fm` [27], and `YouTube` [43].

As music plays a vital role in many human lives and everyone has an opinion about music, user-generated content related to music items such as artists, performers, songs, albums, or music videos is available in abundance. Given the remarkable commercial interest in music distribution and delivery, innovative music retrieval systems are becoming increasingly important. Such systems include personalised, user-aware music recommenders [4], automated playlist generators [64], or intelligent browsing interfaces [48] that transcend the traditional filtering-based browsing scheme according to an artist–album–track hierarchy.

Given the huge amount of user-generated data and the broad interest in music, elaborating sophisticated methods to mine social media content in order to derive semantic information about music and other media items is an ongoing research endeavour, which is currently pursued quite actively. In the following, we will hence discuss the state of the art in three key areas of MIR, where social media mining (SMM) can help improve upon traditional solutions. More precisely, the topics covered are how to compute similarities between music items such as songs or artists, how to estimate the popularity of a music item, and how to tag music items, that is, assign semantic labels to a piece of music, album, or artist.

3.1 Music Similarity Estimation

Computing similarity estimates between two music items (e.g. songs or artists) is an important task in MIR as it enables, among others, automated creation of playlists, recommending items similar to the favourites of a user, or applying clustering techniques and consequently creating user interfaces that foster browsing music collections in an intuitive way.

An example for automated music playlist generation is [68], where content-based data and contextual data (extracted from music-related web pages) are combined to create seamless playlists. Pohle et al. aim at creating playlists in which consecutive tracks sound as similar as possible. Figure 3 shows a music browser entitled `Traveller's Sound Player`, which allows to interact with the generated playlists.

A user interface to music collections, named `nepTune`, is presented in [48], where Knees et al. extract audio features from digital audio files to train a *self-organising map* (SOM) [51]. The SOM uses similarities between feature representations of songs to cluster the music collection under consideration. The clusters are then visualised via first estimating the distribution of the data items over the map and subsequently using the estimated densities as height values to create a

Fig. 3 Screenshot of the `Traveller's Sound Player` interface for automated playlist generation

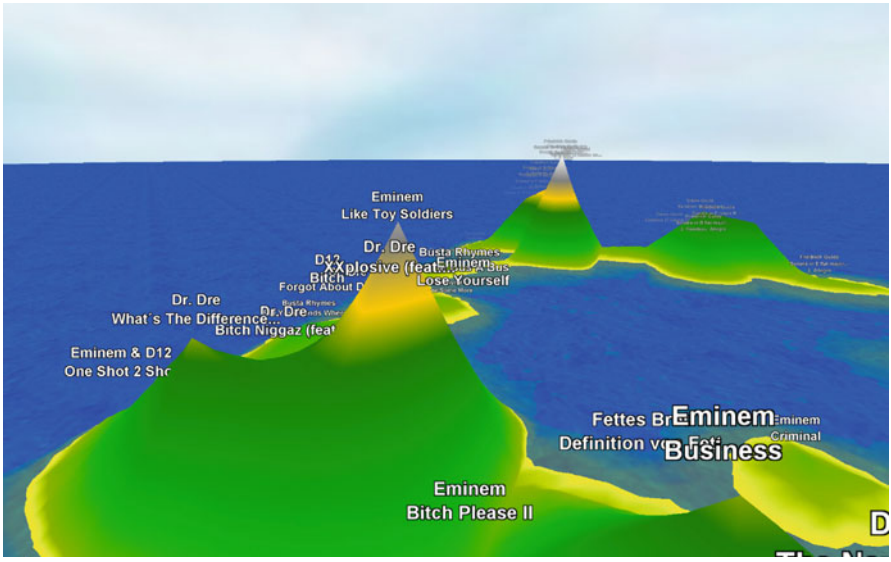


Fig. 4 Screenshot of the `nepTune` browsing interface for music collections

virtual landscape of the music collection. The landscape generated in this way can then be navigated through in the manner of a computer game. Figure 4 shows a screenshot of the `nepTune` interface.

Various kinds of social media have been used to derive similarity scores between music items. In the following, we will particularly focus on methods that construct a similarity measure from user-generated shared *playlists* (e.g. available from `Art of the Mix` [30]) [2], *shared folders in P2P networks* [58], *microblogs* [80], and *collaborative tags* [19,56]. Social media sources for collaborative tags include dedicated platforms such as `last.fm` or the recently quite popular "games with a purpose" [54,61,90].

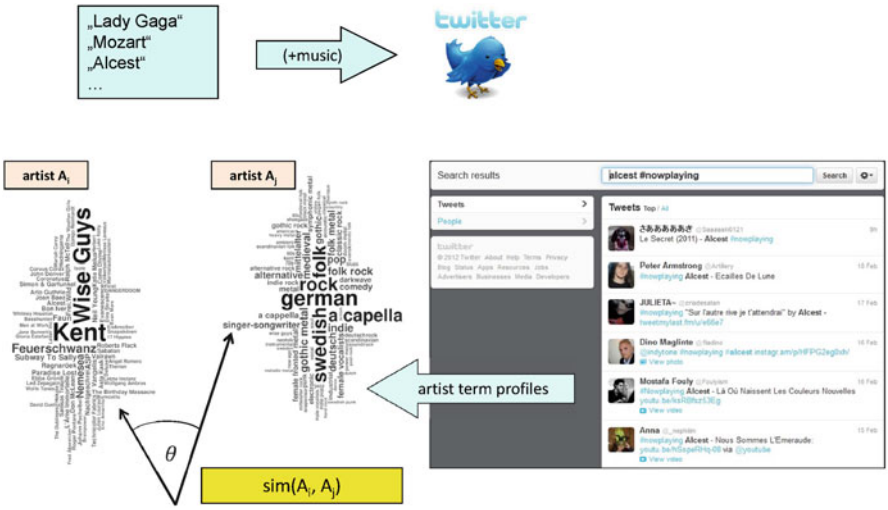


Fig. 5 Artist similarity estimation from microblogs

According to the exploited data source and similarity computation strategy, the methods under discussion can be categorised into *text-based* (microblogs and collaborative tags) and *co-occurrence-based* (web pages, playlists, and shared folders in P2P networks), each of which requires different algorithms to construct a similarity measure. Evaluation, on the other hand, can be performed using the same techniques; most common are *genre classification* and comparison against *human similarity judgements*.

3.1.1 Text: Microblogs and Collaborative Tags

Text-based approaches to music similarity estimation typically approximate the similarity by employing the vector space model, which was introduced in Sect. 1. In the following, we will discuss how to derive similarity information from microblogs and from collaborative tags extracted from `last.fm` or gathered from “games with a purpose”.

Microblogs

A comprehensive study of different aspects in estimating music artist similarity from microblogs is presented in [74]. For the experiments in this chapter, the *vector space model* was applied, that is, each artist is modelled as a *term weight vector* in a high-dimensional *feature space* and similarities between these term vector representations are calculated. An overview of the basic approach is depicted in Fig. 5. First, the `Twitter` API is used to retrieve microblogs for

each artist in a given list of 3,000 music artists. The returned tweets for each artist are then concatenated, resulting in a *virtual artist document*, and a term vector representation for each artist is computed. The actual similarity estimate between two artists A_i and A_j is eventually obtained by calculating a similarity1 function S_{A_i,A_j}.

In the study presented in [74], several thousand combinations of the following single aspects have been assessed:

- Query scheme
- Index term set
- Term frequency (TF)
- Inverse document frequency (IDF)
- Normalisation with respect to document length
- Similarity function

Evaluating different *query schemes* is motivated by the fact that earlier work in web-based MIR has shown an improvement in the accuracy of similarity estimates when adding music-related keywords to the search query (e.g. "music" or "music review") [47, 78, 92]. Different *index term sets*, that is, lists of terms used to filter the microblogs and create the term weight vectors, have been assessed as well. The number of terms in the index term set corresponds to the dimensionality of the respective feature vectors (TF · IDF vectors). The *term frequency* $r_{d,t}$ of a term t in a virtual artist document d estimates the importance t has for document d, hence for the artist under consideration. The *inverse document frequency* w_t estimates the overall importance of term t in the whole corpus and is commonly used to weight the $r_{d,t}$ factor, that is, downweight terms that are important for many documents and hence less discriminative for d. Performing this calculation for all terms in the used index term set and each virtual artist document results in one TF · IDF vector per artist. It is common to subsequently *normalise* the TF · IDF vectors with respect to document length. Finally, different *similarity functions* S_{d_i,d_j} to estimate the proximity between the term vectors of two virtual artist documents d_i and d_j are examined.

As for evaluation, *mean average precision* (MAP) scores are computed on genre labels predicted by various classifiers. More precisely, given a query or seed artist, the retrieval task is to find artists of the same genre via similarity. MAP is simply computed as the arithmetic mean of the $precision@k$ scores, that is, the average precision of k-nearest neighbour (kNN) classifiers for varying values of k.

Although reporting all results of [74] is way beyond the scope of this chapter, I would like to summarise the most important findings in the following.

Query Scheme: When dealing with microblogs, it is preferable to use only the artist name (no additional keywords) to query the `Twitter` API or more general to select the tweets relevant to the artist under consideration.

Index Term Sets: Even though using all terms in the corpus yields the highest MAP values, results are by far the most unstable ones. This means that slightly modifying a single other aspect can cause a significant decline in accuracy when

using all terms in the corpus. Given the high computational complexity due to feature spaces of dimensionality greater than one million, employing no particular index term set is not favourable. Best and most robust results were achieved on average using a dictionary of musical genres, musical instruments, and emotions, which was gathered from `Freebase` [35].

Term Frequency: A simple binary match TF formulation should not be used. The most favourable algorithmic variants are logarithmic formulations and an adapted *Okapi BM25* formulation [69,70].

Inverse Document Frequency: Among the IDF formulations, binary match yields the worst results. Also signal estimates and signal-to-noise ratios do not perform much better. Again, logarithmic formulations and the modified *Okapi BM25* formulation yield top results.

Normalisation: Performing no normalisation for document length performs best, both in terms of accuracy and robustness. This is presumably due to the special characteristics of tweets, which are limited to 140 characters, a limit commonly exhausted by `Twitter` users. Normalisation hence does not improve results but increases computational costs.

Similarity Function: Among the similarity functions under estimation, the *Jeffrey divergence*-based function performs very well, while at the same time maintaining a reasonable stability level. Also the *Jaccard coefficient* performs remarkably well. *Euclidean similarity* performed inferior in all combinations.

Overall, the best performing variants found in the experiments are given by the three term-weighting functions in Eqs. 1–3, in combination with the *Jaccard coefficient* similarity function (Eq. 4). In these equations, N represents the total number of documents in the corpus, $f_{d,t}$ is the number of occurrences of term t in document d, f_t denominates the total number of documents containing term t, W_d is the document length of d, and $\mathcal{T}_{d_1,d_2}$ denotes the set of distinct terms in documents d_1 and d_2.

$$w_{d,t} = \log_e(1 + f_{d,t}) \cdot \log_e \frac{N - f_t}{f_t} \tag{1}$$

$$w_{d,t} = \log_e(1 + f_{d,t}) \cdot \log \frac{N - f_t + 0.5}{f_t + 0.5} \tag{2}$$

$$w_{d,t} = (1 + \log_e f_{d,t}) \cdot \log_e \frac{N - f_t}{f_t} \tag{3}$$

$$\text{sim}(d_1, d_2) = \frac{\sum_{t \in \mathcal{T}_{d_1,d_2}} (w_{d_1,t} \cdot w_{d_2,t})}{W_{d_1}^2 + W_{d_2}^2 - \sum_{t \in \mathcal{T}_{d_1,d_2}} (w_{d_1,t} \cdot w_{d_2,t})} \tag{4}$$

Table 1 Most popular tags and their artist frequencies, among a set of 1,995 artists

Tag	Frequency
Jazz	809
Seen live	658
Rock	633
60s	623
Blues	497
Soul	423
Classic rock	415
Alternative	397
Funk	388
Pop	381
Favourites	349
American	345
Metal	334
Electronic	310
Indie	309

Table 2 Tags assigned by `last.fm` users only once, among a set of 1,995 artists

Crappy girl singers
Stuff that needs further exploration
Disco noir
Knarz
Lektroluv compilation
Gdo02
Electro techo
808 state
Good gym music
Techno manchester electronic acid house
Music i tried but didnt like
American virgin festival

Collaborative Tags

User-generated tags that are assigned to music items are a valuable, albeit noisy source for different MIR tasks, not least for similarity estimation and music retrieval purposes.

Geleijnse et al. gather tags from `last.fm` to generate a "tag ground truth" on the artist level [19]. The authors first filter redundant and noisy tags using the set of tags associated with tracks by the artist under consideration. Similarity between two artists is then estimated as the number of overlapping tags. Evaluation on a set of 1,995 artists, using `last.fm`'s similar artist function as ground truth, shows that the number of overlapping tags between similar artists is much larger than the overlap between arbitrary artists (about 10 vs. 4 tags after filtering). Another interesting observation is that the tags assigned to the largest number of artists fall into only three semantic categories – genres, personal references, and time periods (Table 1). The least frequent tags are shown in Table 2. Often they are more prosaic, represent specific personal notes, or simply contain typos.

Another work on collaborative tags is [56], where Levy and Sandler construct a semantic space for music pieces based on tags retrieved from `last.fm` and `MusicStrands` [28], a web service (no longer in operation) that allowed users to share playlists. To this end, all tags found for a specific music piece are tokenised, and a document-term matrix based on TF · IDF weighting is created. Each track is hence represented by a term vector. For the TF part of the weighting, three different approaches are considered: using the number of users that applied the tag, ignoring the number of users (performing no TF weighting at all), and restricting the terms to adjectives by employing a part-of-speech (POS) tagger. Levy and Sandler further analyse the influence of applying *latent semantic analysis* (LSA) [16] to reduce the dimensionality of the feature space. The authors then compute the similarity between the resulting feature vectors using the cosine measure. For evaluation, the authors employ a retrieval scenario and report *average precision* values. They judge the relevance of retrieved terms as having assigned the same genre or artist label as the seed. Levy and Sandler find that using all terms (not only adjectives) is preferable. They also found the incorporation of the number of users that applied the tag into the TF score superior.

It was in 2007 when the MIR community recognised the value of "games with a purpose" for MIR tasks. In this very year, three papers proposing different music-tagging games were found in the proceedings of the annual "International Society for Music Information Retrieval" (ISMIR) conference, the main scientific venue for MIR research. The principal motivation for such games is to let users solve tasks that are hard or even infeasible to perform for a computer, while at the same time being entertaining enough to attract and keep many users playing. In the music domain, Law et al. present `TagATune`, a game for semantic annotation of music and audio [54]. Two players are paired and are then listening to the same piece of audio. They can describe the audio by entering words but are rather told to guess what their partners are thinking, because both players will only score points if their tags match. If this is the case for one tag, the game will proceed to the next track. Even though `TagATune` was originally designed to harvest semantic descriptions for music and audio, it also implements a "comparison round", where users are presented three songs – one seed track and two alternatives to choose from. They then have to decide which of the alternatives sound more similar to the seed song. From this kind of information, relative similarity judgements and in turn a similarity measure can be derived, as done by Law and von Ahn [53], Stober [88], and Wolff and Weyde [93], for instance.

A similar game, called `Listen Game`, is presented by Turnbull et al. in [90]. It aims at uncovering semantic relationships between words and music. Again, players are grouped and listen to the same songs. They subsequently have to choose from a list of words the one that best and the one that worst describes the song. Users get immediate feedback about which tags other players have chosen. From the data collected, Turnbull et al. employ the *mixture hierarchies expectation maximisation* (MH-EM) [91] algorithm to learn semantic associations between words and songs. These associations are weighted and can therefore be used to construct tag weight vectors for songs and in turn to define a similarity measure for retrieval [89].

Table 3 Most popular tags assigned by players of a "game with a purpose" on music annotation

Tag	Frequency
Drums	793
Guitar	720
Male	615
Rock	571
Synth	429
Electronic	414
Pop	375
Bass	363
Female	311
Dance	297
Techno	224
Electronica	155
Piano	153
Rap	140
Synthesizer	136

Mandel and Ellis present another game for music annotation in [61]. It differs from the other games presented so far in that it uses a more fine-grained scoring scheme. Players receive more points for new tags to stimulate the creation of a larger semantic corpus. More precisely, a player who first uses a tag t to describe a particular song scores two points if t is later confirmed (used again) by another player. The third and subsequent players that use the same tag t do not receive any points. Thus, players who are the first to use a word t for tagging a particular song do not receive an immediate reward but will score two points as soon as another player will have used t. The authors report the most popular tags confirmed by at least one user. They are summarised in Table 3. Compared to the top tags extracted from `last.fm` (Table 1), the tags originating from the tagging game more often describe instruments and gender of the main performer.

3.1.2 Co-occurrences: Web Pages, Playlists, and P2P Networks

The family of *co-occurrence approaches* to music similarity estimation is based upon the assumption that *two music items are more likely to be similar if they co-occur in the same document*, for instance, a playlist, a web page, or a tweet.

Web Pages

In this vein, [78] defines the similarity of two artists as the conditional probability that one artist is to be found on a web page that is known to mention the other artist. This conditional probability can either be calculated on *crawled web pages* that relate to the artists under consideration or heuristically approximated using *page count information* from major search engines.

The former strategy, performing web crawls to infer similarities, is followed in [12] and [72][Chap. 3]. To this end, a certain amount of top-ranked web pages returned by a search engine is retrieved for each artist A_i. Subsequently, all pages fetched for A_i are searched for occurrences of all other artist names in the collection. The number of page hits represents a co-occurrence count that equals the document frequency of the artist term "A_j" in the corpus of web pages for artist A_i. This count is expressed by the asymmetric function $\text{cooc}(A_i, A_j)$. A similarity score is then computed by relating this count to the total number of pages successfully fetched for artist A_i. Symmetrising these scores for all pairs of artists eventually leads to the similarity function shown in Eq. 5. Please note that $\text{cooc}(A_i, A_i)$ and $\text{cooc}(A_j, A_j)$ refer to the total number of web pages successfully crawled for artists A_i and A_j, respectively.

$$\text{sim}(A_i, A_j) = \frac{1}{2} \cdot \left[\frac{\text{cooc}(A_i, A_j)}{\text{cooc}(A_i, A_i)} + \frac{\text{cooc}(A_j, A_i)}{\text{cooc}(A_j, A_j)} \right] \tag{5}$$

The heuristic solution referred to in the beginning of this section is proposed in [78]. It relies solely on the page count estimates provided by a search engine. In short, these page count estimates for queries like `"artist name i"` or `"artist name i"+"artist name j"` are used to infer the relative frequency of both artists' co-occurrence and in turn the conditional probability as indicated above. Equation 6 gives a formal representation of the symmetrised similarity function.

$$\text{sim}(A_i, A_j) = \frac{1}{2} \cdot \left[\frac{pc(A_i, A_j)}{pc(A_i)} + \frac{pc(A_i, A_j)}{pc(A_j)} \right] \tag{6}$$

Comparing the two strategies (web crawls and page count estimates) in terms of computational complexity, it is obvious that the former one requires fewer requests to the search engine. The number of queries to the search engine grows indeed linearly with the number of music items in the collection. In contrast, the second approach that entirely relies on page count estimates from search engines grows quadratically in the number of queries. It is hence less suited for mid- and large-size music collections.

Playlists

Exploiting co-occurrence information from playlists to derive a similarity estimate between music items was probably first suggested in [66]. Pachet et al. consider radio station playlists from a French radio channel and compilation CDs from

CDDB[1] to extract co-occurrences between tracks and between artists. The authors count the number of co-occurrences of two artists (or pieces of music) A_i and A_j in the radio station playlists and compilation CDs. They define the co-occurrence of an entity A_i to itself as the number of A_i's occurrences in the considered data source. To account for different frequencies, that is, popularities, of songs or artists, the co-occurrence counts are normalised. Assuming that co-occurrence is a symmetric function, the similarity measure used by the authors is the same as given1 by Eq. 5.

Focusing on social media data, Baccigalupo et al. present an approach to derive artist similarity information from playlists shared by members of a web community [2]. The authors look at more than one million playlists made publicly available by MusicStrands [28]. The authors extract the 4,000 most popular artists from the playlist set, measuring popularity as the number of playlists in which each artist occurs. They further take into account that two artists consecutively occurring in a playlist are probably more similar than two artists occurring farther away in a playlist. To this end, the authors define a distance function $d_h(A_i, A_j)$ that counts how often a song by artist A_i co-occurs with a song by A_j at a distance of h. Thus, h is a parameter that reflects the number of songs in between the occurrence of a song by A_i and the occurrence of a song by A_j in the same playlist. The distance between two artists A_i and A_j is defined by Eq. 7, where the playlist counts at distances 0 (two consecutive songs by artists A_i and A_j), 1, and 2 are weighted with factors β_0, β_1, and β_2, respectively. The authors empirically set the weights to $\beta_0 = 1$, $\beta_1 = 0.8$, and $\beta_2 = 0.64$.

$$\mathrm{dist}(A_i, A_j) = \sum_{h=0}^{2} \beta_h \cdot \left[d_h(A_i, A_j) + d_h(A_j, A_i) \right] \tag{7}$$

$$|\mathrm{dist}|\,(A_i, A_j) = \frac{\mathrm{dist}(A_i, A_j) - \widehat{\mathrm{dist}}(A_i)}{\left| \max\left(\mathrm{dist}(A_i, A_j) - \widehat{\mathrm{dist}}(A_i) \right) \right|} \tag{8}$$

$$\widehat{\mathrm{dist}}(A_i) = \frac{1}{n-1} \cdot \sum_{j \in X} \mathrm{dist}(A_i, A_j) \tag{9}$$

To account for the *popularity bias*, that is, very popular artists co-occur with a lot of other artists in many playlists simply because they are well known and often listened to by the average music listener, the authors perform normalisation according to Eq. 8, where $\widehat{\mathrm{dist}}(A_i)$ denotes the average distance between A_i and all other artists (Eq. 9) and X is the set of the $n-1$ artists other than A_i.

[1] CDDB is a web-based album identification service that returns, for a given unique disc identifier, meta-data like artist and album name, tracklist, or release year. This service is offered in a commercial version operated by Gracenote [38] as well as in an open source implementation named freeDB [36].

Peer-to-Peer Networks

Peer-to-peer (P2P) networks represent another source of music-related data since users of this kind of network are commonly willing to reveal meta-data about their shared content. For music files, meta-data typically shared is filenames and ID3 tags. By analysing which items co-occur in a user's shared folder, researchers have created music similarity measures.

Among early work that makes use of data extracted from P2P networks is [18, 58, 92], and [6]. These papers all extract data from the P2P network `OpenNap` to derive music similarity information.[2]

Logan et al. [58] and Berenzweig et al. [6] report on having determined the 400 most popular artists on `OpenNap` in mid-2002. The authors harvested meta-data on shared content, which yielded about 175,000 user-to-artist relations from about 3,200 shared music collections. Logan et al. [58] especially highlights the sparsity in the `OpenNap` data, in comparison with music content data. Logan et al. compare similarities defined by artist co-occurrences in `OpenNap` collections, by expert opinions from `allmusic.com` [29], by playlist co-occurrences from `Art of the Mix`, by data gathered from a web survey, and by audio feature extraction (MFCCs) [1]. They calculate a "ranking agreement score" by comparing the top N most similar artists according to each data source and calculating the pairwise overlap between the sources. Their main findings are that the co-occurrence data from `OpenNap` and from `Art of the Mix` show a high degree of overlap, the experts from `allmusic.com` and the participants of the web survey agree moderately, and the signal-based measure has a rather low agreement with all other sources (except for comparison to the allmusic.com data).

Whitman and Lawrence use a software agent to retrieve from `OpenNap` a total of 1.6 million user–song relations over a period of 3 weeks in August 2001 [92]. To alleviate the *popularity bias*, the authors use a similarity measure as shown in Eq. 10, where $C(A_i)$ denotes the number of users that share songs by artist A_i, $C(A_i, A_j)$ is the number of users that have both artists A_i and A_j in their shared collection, and A_k is the most popular artist of the whole data set. The second factor (in the right-hand part of the equation) downweights the similarity between two artists if one of them is very popular and the other is not.

$$\text{sim}(A_i, A_j) = \frac{C(A_i, A_j)}{C(A_j)} \cdot \left(1 - \frac{\left|C(A_i) - C(A_j)\right|}{C(A_k)}\right) \tag{10}$$

In [18], Ellis et al. aim to build a ground truth for artist similarity estimation. The authors report on having extracted from `OpenNap` about 400,000 user-to-song relations, covering about 3,000 unique artists. Again, the co-occurrence

[2] It is not clear whether the four mentioned publications make use of exactly the same data set. In any case, the authors emphasise that they only extract meta-data from `OpenNap`, but do not download any files.

data is compared with artist similarity data gathered by a web survey and with `allmusic.com` data. In contrast to Whitman and Lawrence, Ellis et al. take indirect links in `allmusic.com`'s similarity judgements into account. To this end, Ellis et al. propose a transitive similarity function on similar artists from the `allmusic.com` data, called "Erdös distance". More precisely, the distance $d(A_i, A_j)$ between two artists A_i and A_j is measured as the minimum number of intermediate artists needed to form a path from A_i to A_j. As this definition also allows to derive information on dissimilar artists (with a high minimum path length), it can be employed to obtain a complete distance matrix.

A recent approach by Shavitt and Weinsberg derives similarity information on the artist and on the song level from the `Gnutella` file-sharing network [84]. The authors collected meta-data of shared files from more than 1.2 million `Gnutella` users in November 2007. They restricted their search to music files (MP3 and WAV). The crawl yielded a data set of 530,000 songs. Information on both users and songs are represented via a 2-mode graph showing users and songs. A link between a song and a user is created if the user shares the song. Analysing the resulting network, it turned out that most users of the P2P network share similar files.

Shavitt and Weinsberg further propose an approach to *artist recommendation*. To this end, they construct a user-to-artist matrix V, where $V(i, j)$ gives the number of songs by artist A_j that user U_i shares. They subsequently perform direct clustering on V using the k-means algorithm [60] with the Euclidean distance metric. Artist recommendation is then performed using either data from the centroid of the cluster to which the seed user U_i belongs or information about the nearest neighbours of U_i within the cluster to which U_i belongs.

In addition, Shavitt and Weinsberg address the problem of *song clustering*. Accounting for the *popularity bias*, the authors define a distance function that is normalised according to song popularity, as shown in Eq. 11, where $uc(S_i, S_j)$ denotes the total number of users that share songs S_i and S_j. C_i and C_j denote, respectively, the popularity of songs S_i and S_j, measured as their total occurrence in the data set:

$$\text{dist}(S_i, S_j) = -\log_2\left(\frac{uc(S_i, S_j)}{\sqrt{C_i \cdot C_j}}\right) \tag{11}$$

3.2 Music Popularity Estimation

Estimating the popularity of a music artist or song in a certain region of the world is an important task, not only for the music industry. Also the cosmopolitan and culturally aware music aficionado is likely to be interested in which music is currently "hot" in different parts of the world. Not least artists are interested to know where in the world their music is particularly (un)popular. Furthermore, popularity information can serve as an important component for *serendipitous music retrieval* systems [10, 81].

An artist's or song's popularity can be estimated via a wide variety of predictors, such as traditional charts (e.g. "Billboard Hot 100" released weekly for the United States of America by the Billboard Magazine [26]), microblogging activity, playcounts (e.g. from last.fm or YouTube), occurrences on web pages, and shared folder analysis in P2P networks.

Scientific work on this topic includes [73], where Schedl et al. compare different data sources for artist popularity estimation on a per-country basis. In [50], Koenigstein et al. analyse search queries issued within a P2P network to infer music popularity. Grace et al. compute popularity rankings from user comments in a social network [21].

Given the large interest record companies, producers, and artists have in this kind of information, it is not surprising that there also exist businesses specialised on music popularity measurement. Examples are Band Metrics [31] or BigChampagne Media Measurement [32]. Even though they obviously do not reveal details of their algorithms, it can be reasonably assumed that these companies harvest multiple data sources to create their predictors. The music information platform Echo Nest [25] even offers a public API function to retrieve a ranking based on the so-called "hottness" of an artist [24]. This ranking is based on editorial, social, and mainstream aspects [23]. However, this web service does not provide country-specific information, and Echo Nest is known to have a strong focus on the USA.

In the following, approaches that make use of social media to predict the popularity of an artist or a song will be presented and discussed. Also properties of the data sources, such as particular biases, availability, noisiness, and time dependence, will be addressed.

3.2.1 Data Sources for Popularity Estimation

The popularity of an artist or track can be defined on different levels of granularity (e.g. individual user, peer group, country, or cultural region). Incorporating previous approaches presented in [49, 50], Schedl et al. compare different ways to derive popularity information from various social media sources [79] on the level of countries. To this end, a framework is established that uses the following proxies for popularity:

- Page counts of web pages
- Artist occurrences in geo-located microblogs
- Meta-data from folders shared in the Gnutella P2P network
- Playcount data from the social music platform last.fm

The approaches proposed to compute popularity rankings from each data source are detailed below.

Another work that infers popularity information from social media is [21], where Grace et al. compute popularity rankings from user comments in the social network MySpace [40]. To this end, the authors apply various annotators to crawled

MySpace artist pages in order to spot, for example, names of artists, albums, and tracks, sentiments, and spam. Subsequently, a data hypercube (OLAP cube) is used to represent structured and unstructured data and to project the data to a popularity dimension. A user study showed that the list generated by this procedure was on average preferred to Billboard charts.

Web Page Counts

Page counts are gathered by querying the web search engines Google [37] and Exalead [33] for ⟨artist, country⟩ tuples. To guide the search towards musically relevant web pages and avoid distortions caused by artist names that equal common speech words (e.g. "Bush", "Kiss", "Hole"), the query scheme `"artist name" "country name" music` is employed. Furthermore, a factor resembling *inverse document frequency* (IDF) is used to downweight popularity of artists that are popular everywhere in the world since the aim is to uncover popular artists specific to each country. The final ranking score is calculated according to Eq. 12, where $pc_{c,a}$ is the page count value returned for the country-specific query for artist a and country c, $|C|$ is the total number of countries for which data is available, and df_a is the number of countries in which artist a is known according to the data source (i.e. the number of countries with $pc_{c,a} > 0$).

$$\text{popularity}_{c,a} = pc_{c,a} \cdot \log_2\left(1 + \frac{|C|}{df_a}\right) \tag{12}$$

Geo-Located Microblogs

Microblogs are retrieved from Twitter using the search API and are then narrowed in two ways. First, only posts containing the hashtag #nowplaying are considered. This filtering is directly supported by the Twitter API. Secondly, the search is narrowed to a specific country. To this end, posts are categorised according to their location within a certain radius around the major cities of the world. Tweets are then aggregated to the country level. Scanning the retrieved microblogs for occurrences of the artists of interests and counting the number of their appearances for a given country c eventually yield a count equal to the term frequency ($tf_{c,a}$) of artist a in an aggregated document comprising all tweets gathered for cities in country c. Equation12 again gives the ranking score, when $pc_{c,a}$ is replaced with $tf_{c,a}$.

P2P Network

Shared folder data from the P2P network Gnutella is extracted employing a two-stage process, similar to [49]: a *crawler* component discovers the highly dynamic

network topology; a *browser* queries the active nodes – corresponding to users – for meta-data of files in their shared folders. The crawler treats the network as a graph and performs *breadth-first exploration*. Discovered active nodes are enqueued in a list that is processed by the browser. Shared digital content is associated with artists by matching the artist names of interest against ID3 tags of shared music files. Occasionally ID3 tags are missing or misspelled. Artists names are therefore also matched against the filenames. Creating popularity charts for specific countries requires determining the geographical location of the users. The necessary geo-identification process is based on IP addresses. First, a list of all unique IP addresses in the data set – typically over a million – is created. IP addresses are then geo-located using the commercial `IP2Location` [39] database. Each IP address is hence attached a country code, a city name, and latitude–longitude coordinates. The geographical information obtained in this way pinpoints fans and enables tracking spatial diffusion of artists popularity [50]. Aggregating the amount of digital content associated with each artist for the country under consideration yields the final ranking score.

Social Music Platform

As last data source, artist popularity based on the user community of the social music platform `last.fm` is considered. Despite the issues of *hacking and vandalism* and a certain *community bias* [75], which are inherent to collaborative music information systems, the playcounts of `last.fm` users can be expected to reflect which music is currently popular in this community. First, the top 400 listeners of each country are gathered via the `last.fm` API. The most frequently played artists for each of these listeners are extracted subsequently.[3] Aggregating these playcounts for each ⟨artist, country⟩ pair finally yields a popularity ranking.

3.2.2 A Multifaceted Comparison of Different Data Sources

It was shown in [79] that the popularity charts obtained from the different, inhomogeneous data sources do not correlate highly. Each data source hence covers different aspects of popularity, which indicates that the quest for artist popularity is a multifaceted and challenging task, especially in the era of multichannel music distribution.

Trying to uncover the different dimensions of the five data sources (the four web and social media sources and traditional music charts), Table 4 compares

[3] In the meantime, `last.fm` has extended its API with a `Geo.getTopArtists` function that returns the top-played artists in a particular country.

Table 4 Comparing different social media sources to infer popularity information

Source/aspect	Bias	Availability	Noisiness	Time dependence
Web page counts	Web users	Widespread	High	Accumulating
Twitter	Community	Country-dependent	Medium	Current
P2P	Community	Country-dependent	Low–medium	Accumulating
Last.fm	Community	Widespread	Medium–high	Accumulating
Traditional charts	Music industry	Country-dependent	Low	Current

Table 5 Availability of data for popularity estimation

Data source	Countries
Web page counts	240
Twitter	155
P2P	86
Last.fm	240

them according to several criteria relevant to the task of popularity estimation. One issue is that certain approaches are prone to a specific *bias*. The average `last.fm` user, for instance, does not represent the average music listener of a country, that is, `last.fm` data is distorted by a *community bias*. The same holds for `Twitter`, which is biased towards artists with very active fans. On the other hand, some very popular artists may have fans who use `Twitter` to a much lower degree. Traditional charts are frequently biased towards the record sales figures the music industry commonly uses as proxy.

Another aspect is data *availability*. While page count estimates are available for all countries of the world, the approaches based on P2P and `Twitter` data suffer from a very unbalanced coverage for different countries. Also traditional music charts vary strongly in terms of availability between countries. Table 5 shows the number of countries for which data could be extracted for each approach, as presented in [79]. Please note that these results are based on a list of 240 countries retrieved from `last.fm`.

A big advantage of traditional charts is their robustness against noise. In contrast, *page count estimates* are easily distorted by ambiguous artist or country names. `Last.fm` data suffers from hacking and vandalism [11], as well as from unintentional input of wrong information and misspellings.

According to the dimension of *time dependence*, the data sources can be categorised into "current" and "accumulating", relating to whether they reflect an instantaneous popularity or a general, time-independent popularity. The largest overlap in popularity rankings between the investigated data sources can be explained by the dimension of time dependence. It is present between the output of the page count predictor and the P2P rankings, the data sources behind both of which share an accumulating strategy of data storage. `Twitter` and `last.fm` on the other hand are more time dependent in that they reflect better the current "hotness" of an artist than his or her overall popularity.

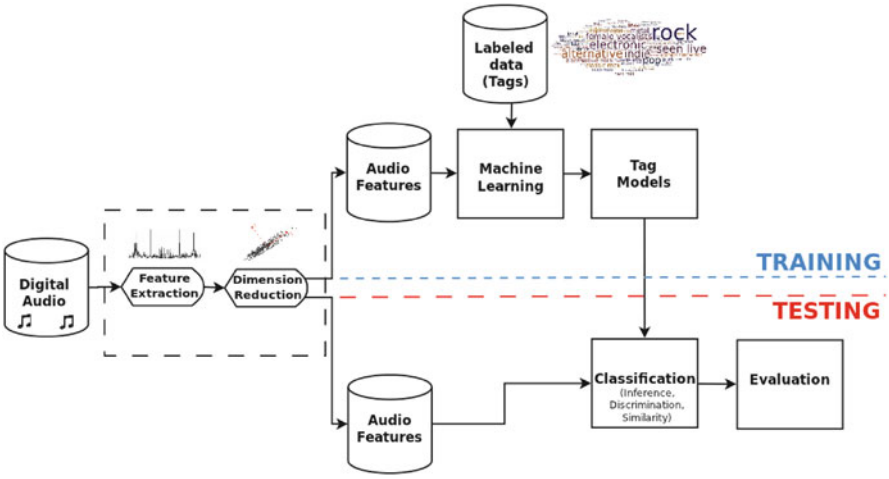


Fig. 6 Basic scheme of a music auto-tagger

3.3 Auto-tagging Music

Semantic labels attached to multimedia items, such as images, music pieces, or videos, have become an important means to categorise and describe such items and to communicate particular opinions or feelings about them by users of social media. The process of automatically attaching semantic labels, or tags, to music pieces is referred to as *auto-tagging* and is a rather recent research endeavour in MIR. Typically, first, a machine learning approach, a *supervised learner* to be more precise, is employed on a training data set that associates feature representations (commonly music content or music context features) with semantic tags. After training is finished, the classifier is used to predict labels to previously unseen music items. In order to increase computational efficiency, optionally some *feature selection* or *dimensionality reduction* technique might be employed to the input feature vectors before training the classifier. This is of particular importance when dealing with high-dimensional representations of music items, which are typically present when modelling music items via a multimodal approach, for instance, via a feature vector describing aspects of the music content and the music context [76]. It was shown by Sordo in [86] that a dimensionality reduction of 95% by applying *principal components analysis* (PCA) [44] to the `CAL500` data set [89] and 600-dimensional audio feature vectors does not significantly decrease accuracy but decreases computational costs considerably. Figure 6 depicts the general framework of an auto-tagger [86].

One can broadly categorise music-tagging efforts into approaches that learn relations between feature representations of music files and semantic tags, henceforth referred to as "computational approaches" and strategies to infer tags directly

Table 6 Comparing different approaches to tag music

Source/approach	Advantages	Disadvantages
Human surveys	Well-defined vocabulary, high-quality annotations, strong labelling	Restricted vocabulary, yields only small data sets, high human effort, time-consuming
Social tags	Unrestricted vocabulary, incorporates social and cultural context, wisdom of the crowds	Popularity bias, community bias, weak labelling
Games with a purpose	Wisdom of the crowds, entertainment factor yields high-quality and fast annotations	Cheating, tags valid only for short segments (incentive for quick skipping)
Web pages	Incorporates cultural context, large corpus available, no immediate human involvement necessary	Noisy annotations, weak labelling, sparseness in the "long tail"
Auto-tagging	Not affected by cold-start problem, no immediate human involvement necessary, strong labelling	Computationally expensive, limited by training data

from some kind of user input, referred to as "human-centred approaches", in the following. Both strategies will be addressed below. As for the former one, methods that learn tags from *co-occurrence data* (*collaborative filtering*), *audio features*, and *web pages* will be introduced. For the category of human-centred approaches, another *game with a purpose* will be presented, and the use of *music folksonomies* to infer tags and associate them to semantic categories will be discussed.

Turnbull et al. in [52] compare various data sources and corresponding algorithms (computational and human-centred) for the task of tagging music: *human surveys*, *social tags*, *games with a purpose*, *web pages*, and *auto-tagging*. They elaborate on advantages and disadvantages of each, which are summarised from [52] and extended by the author in Table 6. *Weak labelling* refers to the fact that one cannot infer from the absence of a tag t that t does not apply to the item. Users might simply not have thought of the tag in such a case. In contrast, a *strongly labelled* data set is complete in the sense that the absence of tag t does mean that t is not suited to describe the item under consideration. For an explanation of *popularity bias* and *community bias*, please refer to Sects. 3.1.2 and 3.2.1, respectively.

3.3.1 Computational Approaches

As described above, the most common approach to auto-tagging music is to train a classifier on music content features and learn relations between them and a set of

tags that are known to relate to the corresponding music items. As features typically rhythm and/or timbre descriptors are used [63], sometimes high-level features are included in addition [86].

Sordo proposes a simple and efficient algorithm based on a weighted vote k-nearest neighbour (kNN) classifier to propagate tags from training data to unseen music items [86]. Given a training set of PCA-compressed feature vectors of a song collection together with a set of labels for each piece, the proposed *weighted vote kNN* algorithm first determines the k closest neighbours to the seed song s, which should be tagged. The frequency of all tags assigned to s's neighbours are then summed up per tag, and a threshold relative to k ensures that only frequently used tags are predicted for s.

An approach similar to Sordo's is suggested in [46]. Kim et al. also employ a kNN classifier to address the problem of auto-tagging artists, but they analyse different artist similarity functions. They compare similarities derived from artist co-occurrences in `last.fm` playlists ("scrobbles"), from `last.fm` tags, from web pages about the artists, and from music content features. Using as ground truth tags manually assigned by music experts, Kim et al. found that the similarity measure based on `last.fm` co-occurrences performed best, both in terms of precision and recall. When using a kNN classifier, it is crucial to carefully select the similarity measure to determine the nearest neighbours. Depending on the origin of the features, a common choice is *cosine similarity* (for term-weighting features) or one of the distances/divergences *Mahalanobis*, *Manhattan*, or *Kullback–Leibler* (for music content features).

In their proposed algorithm to extract tags from artist-related web pages, Schedl et al. use a dictionary of musically relevant terms to filter the textual content of the pages under consideration [77]. The authors propose three different term-weighting functions to score the extracted tags per artist and predict the resulting top-ranked tags. A user survey was conducted to evaluate the quality of the suggested tags in terms of descriptiveness. Quite surprisingly, participants in the study found tags suggested by a simple document frequency function superior to those proposed by TF $\cdot$ IDF-based term scoring.

Mandel et al. [63] use *conditional restricted Boltzmann machines* [85] to learn tag language models over three sets of vocabularies: annotations by users of `Amazon`'s `Mechanical Turk`, of the tagging game `MajorMiner` [62], and of `last.fm`. The models are learned on the level of song segments. Optionally different "contexts" are included, that is, track level and user level annotations are factored in.

Seyerlehner et al. in [82] use a combination of different audio features described within their block-level framework [83]: *spectral pattern* (SP), *delta spectral pattern* (DSP), *variance delta spectral pattern* (VDSP), *logarithmic fluctuation pattern* (LFP), *correlation pattern* (CP), and *spectral contrast pattern* (SCP). Associations between songs and tags are then learned using a *random forest* classifier.

Recently, two-stage algorithms have become popular. In the first stage, they infer higher-level information from music content features, such as term weight vector representations. These new representations are then fed into a machine learning

Table 7 Most frequently used tags in the `TagATune` game, in descending order of frequency

classical, guitar, piano, violin, slow, strings, rock, techno, opera, drums, same, flute, fast, diff, electronic, ambient, beat, yes, harpsichord, indian, female, vocal, no, synth, quiet, no vocals, soft, sitar, no vocal, classic, male, singing, solo, vocals, cello, loud, woman, pop, male vocal, choir, violins, new age, beats, no voice, harp, voice, weird, instrumental, dance

algorithm to learn semantic labels [14, 65]. Sometimes this second stage is said to incorporate contextual aspects since correlations between tags are frequently considered. Alternatively, the term weight vectors inferred in the first stage can also be used as input to a music similarity measure [82]. As an example for a two-stage auto-tagger, Miotto et al. in [65] first model *semantic multinomials* over tags based on music content features. In order to account for co-occurrence relations between tags, they subsequently learn a *Dirichlet mixture model* of the semantic space, which eventually yields a *contextual multinomial*.

3.3.2 Human-Centred Approaches

Games with a purpose have already been introduced in Sect. 3.1.1, where it was shown how to use their results for similarity estimation. In [53], Law and von Ahn present another interesting game with a purpose that focuses on *input-agreement*. Users have to agree whether they are listening to the same piece of music or not, that is, they have to agree on the input. To this end, they can exchange any free-form text that helps to reach the goal. Usually players enter descriptive tags in an effort to quickly choose the correct one of the two classes "same" or "different". If they agree on the correct class, both players are awarded points and the next round starts. Each game lasts for a total of 3 min.

According to Law and von Ahn, this input-agreement mechanism offers the advantage of being more popular and producing a higher number of tags than other, similar games. Analysing the most frequently used tags (Table 7), most of them describe genre, instrumentation, and properties of the music. Due to the very nature of the game design, the top list also includes *communication tags* that are unsuited to describe the music itself ("yes", "same", "no", "diff"). Furthermore, *negation tags* are frequently used to indicate the absence of a particular musical aspect ("no vocal", for example).

Music *folksonomies* present another valuable source for musical information. They are created by large numbers of users via tagging particular music items with their own, specific vocabulary. Although this vocabulary is probably not as precise as the one employed by music experts, the wisdom of the crowds is potentially able to cover more diverse aspects of human music perception than experts can think of.

Sordo et al. [87] present a method to automatically categorise tags extracted from `Wikipedia` into semantically meaningful groups, which they call "facets".

Table 8 Top facets of music extracted from Wikipedia

Music genres
Music geography
Musical groups
Musicians
Musical culture
Occupations in music
Music people
Record labels
Music technology
Sociological genres of music

To this end, starting at the most generic page about "music", the authors extract links from DBpedia, a machine-readable knowledge base created from Wikipedia pages. Applying some heuristics, pages not related to music are filtered out. To the remaining nodes, the PageRank algorithm [67] is applied to determine a relevance score for all nodes/pages in the network. The top facets Sordo et al.'s method found on a data set of about 600,000 artists and 400,000 tags are given in Table 8. The facets and tags extracted in this way are particularly interesting for music retrieval systems, where the user might want to restrict the results to a search query to tracks that are similar according to a specific facet.

4 Conclusions and Research Directions

In this chapter, we have discussed how various kinds of social media can be used for common music information retrieval tasks. More precisely, approaches to infer *music similarity* from text and co-occurrence information were presented, strategies to estimate the *popularity* of a music item from social media were elaborated, and recent methods to automatically assign semantically meaningful tags to music items, a process also known as *auto-tagging*, were discussed.

I am sure that future research in music information retrieval has to strive for a holistic perspective in a sense that information is not only derived from the audio signal and from meta-data but from many different types of multimedia material. Given the spiralling success of social media, analysis and data mining of the respective sources will open unprecedented opportunities to elaborate truly personalised and context-aware music retrieval and recommendation systems. Some concrete challenges in the context of social media mining for music information retrieval are *analysing music video clips* (official music videos as well as user-generated versions) to infer descriptive information; *processing images* of album covers, of band photographs, and of concert snapshots taken by enthusiastic music aficionados; and *making sense of textual data* about music items (for instance, microblogs). In addition, *data fusion* techniques are required to build multifaceted models that describe both music items and listeners in order to eventually enable the next generation of intelligent music retrieval systems.

References

1. Aucouturier, J.-J., Pachet, F., Sandler, M.: "The way it sounds": timbre models for analysis and retrieval of music signals. IEEE Trans. Multimed. **7**(6), 1028–1035 (2005)
2. Baccigalupo, C., Plaza, E., Donaldson, J.: Uncovering affinity of artists to multiple genres from social behaviour data. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR), Philadelphia (2008)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval – The Concepts and Technology Behind Search,2nd edn. Pearson, Harlow (2011)
4. Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Lüke, K.-H., Schwaiger, R.: InCarMusic: context-aware music recommendations in a car. In: International Conference on Electronic Commerce and Web Technologies (EC-Web), Toulouse (2011)
5. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**(6), 1554–1563 (1966)
6. Berenzweig, A., Logan, B., Ellis, D.P., Whitman, B.: A large-scale evaluation of acoustic and subjective music similarity measures. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR), Baltimore (2003)
7. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Inf. Retr. **15**(1), 54–92 (2012)
8. Büttcher, S., Clarke, C.L.A., Cormack, G.V.: Information Retrieval: Implementing and Evaluating Search Engines. MIT, Cambridge (2010)
9. Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M.: Content-based music information retrieval: current directions and future challenges. Proc. IEEE **96**, 668–696 (2008)
10. Celma, O.: Music Recommendation and Discovery – The Long Tail, Long Fail, and Long Play in the Digital Music Space. Springer, Berlin (2010)
11. Celma, O., Lamere, P.: ISMIR 2007 tutorial: music recommendation. http://mtg.upf.edu/~ocelma/MusicRecommendationTutorial-ISMIR2007. Accessed Dec 2007. September 23–27 2007
12. Cohen, W.W., Fan, W.: Web-collaborative filtering: recommending music by crawling the web. WWW9/Comput. Netw. **33**(1–6), 685–698 (2000)
13. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**(90), 297–301 (1965)
14. Coviello, E., Chan, A.B., Lanckriet, G.: Time series models for semantic music annotation. IEEE Trans. Audio Speech Lang. Process. **19**(5), 1343–1359 (2011)
15. Cunningham, S.J., Downie, J.S., Bainbridge, D.: "The pain, the pain": modelling music information behavior and the songs we hate. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR), London (2005)
16. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**, 391–407 (1990)
17. Downie, J.S.: The scientific evaluation of music information retrieval systems: foundations and future. Comput. Music J. **28**, 12–23 (2004)
18. Ellis, D.P., Whitman, B., Berenzweig, A., Lawrence, S.: The quest for ground truth in musical artist similarity. In: Proceedings of 3rd International Conference on Music Information Retrieval (ISMIR), Paris (2002)
19. Geleijnse, G., Schedl, M., Knees, P.: The quest for ground truth in musical artist tagging in the social web era. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), Vienna (2007)
20. Gouyon, F., Herrera, P., Gomez, E., Cano, P., Bonada, J., Loscos, A., Amatriain, X., Serra, X.: Content processing of music audio signals. In: Polotti, P., Roccheso, D. (eds.) Sound to Sense, Sense to Sound: A State-of-the-Art in Sound and Music Computing, pp. 83–160. Logos Verlag, Berlin GmbH (2008)

21. Grace, J., Gruhl, D., Haas, K., Nagarajan, M., Robson, C., Sahoo, N.: Artist ranking through analysis of on-line community comments. In: Proceedings of the 17th ACM International World Wide Web Conference (WWW 2008), Bejing (2008)
22. Grossman, D.A., Frieder, O.: Information Retrieval: Algorithms and Heuristics. Kluwer International Series on Information Retrieval. Springer, Dordrecht (2004)
23. http://developer.echonest.com/docs/method/get_hotttnesss. Accessed Jan 2010
24. http://developer.echonest.com/docs/method/get_top_hottt_artists. Accessed Jan 2010
25. http://echonest.com. Accessed Feb 2012
26. http://en.wikipedia.org/wiki/Billboard_Hot_100. Accessed May 2009
27. http://last.fm. Accessed Jan 2012
28. http://music.strands.com. Accessed Nov 2009
29. http://www.allmusic.com. Accessed Jan 2010
30. http://www.artofthemix.org. Accessed Feb 2008
31. http://www.bandmetrics.com. Accessed May 2010
32. http://www.bigchampagne.com. Accessed May 2010
33. http://www.exalead.com. Accessed Aug 2010
34. http://www.facebook.com. Accessed Feb 2012
35. http://www.freebase.com. Accessed Jan 2010
36. http://www.freedb.org. Accessed Feb 2008
37. http://www.google.com. Accessed Mar 2010
38. http://www.gracenote.com. Accessed Feb 2008
39. http://www.ip2location.com. Accessed Mar 2010
40. http://www.myspace.com. Accessed Nov 2009
41. http://www.spotify.com. Accessed Feb 2012
42. http://www.twitter.com. Accessed Jan 2012
43. http://www.youtube.com. Accessed Feb 2012
44. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
45. Jones, K.S., Walker, S.S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manag. **36**, 779–808 (2000)
46. Kim, J.H., Tomasik, B., Turnbull, D.: Using artist similarity to propagate semantic information. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR), Kobe (2009)
47. Knees, P., Pampalk, E., Widmer, G.: Artist classification with web-based data. In: Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR), Barcelona, pp. 517–524 (2004)
48. Knees, P., Schedl, M., Pohle, T., Widmer, G.: An innovative three-dimensional user interface for exploring music collections enriched with meta-information from the web. In: Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara (2006)
49. Koenigstein, N., Shavitt, Y.: Song ranking based on piracy in peer-to-peer networks. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), Kobe (2009)
50. Koenigstein, N., Shavitt, Y., Tankel, T.: Spotting out emerging artists using geo-aware analysis of P2P query strings. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas (2008)
51. Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences, vol. 30, 3rd edn. Springer, Berlin (2001)
52. Lanckriet, G., Turnbull, D., Barrington, L.: Five approaches to collecting tags for music. In: Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR), Philadelphia (2008)
53. Law, E., von Ahn, L.: Input-agreement: a new mechanism for collecting data using human computation games. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI), Boston, pp. 1197–1206 (2009)
54. Law, E., von Ahn, L., Dannenberg, R., Crawford, M.: Tagatune: a game for music and sound annotation. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), Vienna (2007)

55. Lee, J.H.: Analysis of user needs and information features in natural language queries seeking user information. J. Am. Soc. Inf. Sci. Technol. (JASIST) **61**, 1025–1045 (2010)
56. Levy, M., Sandler, M.: A semantic space for music derived from social tags. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), Vienna (2007)
57. Li, T., Ogihara, M., Tzanetakis, G. (eds.): Music Data Mining. CRC/Chapman Hall, Boca Raton (2011)
58. Logan, B., Ellis, D.P., Berenzweig, A.: Toward evaluation techniques for music similarity. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR): Workshop on the Evaluation of Music Information Retrieval Systems, Toronto (2003)
59. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM J. **1**, 309–317 (1957)
60. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Statistics, vol. I, pp. 281–297. University of California Press, Berkeley/Los Angeles (1967)
61. Mandel, M.I., Ellis, D.P.: A Web-based game for collecting music metadata. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), Vienna (2007)
62. Mandel, M.I., Ellis, D.P.W.: A web-based game for collecting music metadata. J. New Music Res. **37**(2), 151–165 (2008)
63. Mandel, M.I., Pascanu, R., Eck, D., Bengio, Y., Aiello, L.M., Schifanella, R., Menczer, F.: Contextual tag inference. ACM Trans. Multimed. Comput. Commun. Appl. **7S**(1), 32:1–32:18 (2011)
64. McFee, B., Lanckriet, G.: The natural language of playlists. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), Miami (2011)
65. Miotto, R., Barrington, L., Lanckriet, G.: Improving auto-tagging by modeling semantic co-occurrences. In: Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht (2010)
66. Pachet, F., Westerman, G., Laigre, D.: Musical data mining for electronic music distribution. In: Proceedings of the 1st International Conference on Web Delivering of Music (WEDELMUSIC), Florence (2001)
67. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. In: Proceedings of the Annual Meeting of the American Society for Information Science (ASIS), Pittsburgh, pp. 161–172 (1998)
68. Pohle, T., Knees, P., Schedl, M., Pampalk, E., Widmer, G.: "Reinventing the wheel": a novel approach to music player interfaces. IEEE Trans. Multimed. **9**, 567–575 (2007)
69. Robertson, S., Walker, S., Hancock-Beaulieu, M.: Large test collection experiments on an operational, interactive system: Okapi at TREC. Inf. Process. Manag. **31**, 345–360 (1995)
70. Robertson, S., Walker, S., Beaulieu, M.: Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In: Proceedings of the 7th Text REtrieval Conference (TREC-7), Gaithersburg, pp. 253–264 (1999)
71. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
72. Schedl, M.: Automatically extracting, analyzing, and visualizing information on music artists from the world wide web. Ph.D. thesis, Johannes Kepler University Linz, Linz (2008)
73. Schedl, M.: On the use of microblogging posts for similarity estimation and artist labeling. In: Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht (2010)
74. Schedl, M.: #nowplaying madonna: a large-scale evaluation on estimating similarities between music artists and between movies from microblogs. Inf. Retr. **15**, 183–217 (2012)
75. Schedl, M., Knees, P.: Context-based music similarity estimation. In: Proceedings of the 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS), Graz (2009)
76. Schedl, M., Knees, P.: Personalization in multimodal music retrieval. In: Proceedings of the 9th Workshop on Adaptive Multimedia Retrieval (AMR), Barcelona (2011)

77. Schedl, M., Pohle, T.: Enlightening the sun: a user interface to explore music artists via multimedia content. Multimed. Tools Appl. Spec. Issue Semant. Digit. Media Technol. **49**(1), 101–118 (2010)
78. Schedl, M., Knees, P., Widmer, G.: A web-based approach to assessing artist similarity using co-occurrences. In: Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI), Riga (2005)
79. Schedl, M., Pohle, T., Koenigstein, N., Knees, P.: What's hot? Estimating country-specific artist popularity. In: Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht (2010)
80. Schedl, M., Knees, P., Böck, S.: Investigating the similarity space of music artists on the micro-blogosphere. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), Miami (2011)
81. Schedl, M., Hauger, D., Schnitzer, D.: A model for serendipitous music retrieval. In: Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI 2012): 2nd International Workshop on Context-Awareness in Retrieval and Recommendation (CaRR 2012), Lisbon (2012)
82. Seyerlehner, K., Schedl, M., Knees, P., Sonnleitner, R.: A refined block-level feature set for classification, similarity and tag prediction. In: Extended Abstract to the Music Information Retrieval Evaluation eXchange (MIREX 2011)/12th International Society for Music Information Retrieval Conference (ISMIR 2011), Miami (2009)
83. Seyerlehner, K., Widmer, G., Pohle, T.: Fusing block-level features for music similarity estimation. In: Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10), Graz (2010)
84. Shavitt, Y., Weinsberg, U.: Songs clustering using peer-to-peer co-occurrences. In: Proceedings of the IEEE International Symposium on Multimedia (ISM): International Workshop on Advances in Music Information Research (AdMIRe), San Diego (2009)
85. Smolensky, P.: Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281. MIT, Cambridge (1986)
86. Sordo, M.: Semantic annotation of music collections: a computational approach. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona (2012)
87. Sordo, M., Gouyon, F., Sarmento, L.: A method for obtaining semantic facets of music tags. In: Proceedings of the Workshop on Music Recommendation and Discovery (WOMRAD), Barcelona (2010)
88. Stober, S.: Adaptive methods for user-centered organization of music collections. Ph.D. thesis, Otto-von-Guericke-University, Magdeburg (2011)
89. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Towards musical query-by-semantic-description using the CAL500 data set. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Amsterdam (2007)
90. Turnbull, D., Liu, R., Barrington, L., Lanckriet, G.: A game-based approach for collecting semantic annotations of music. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), Vienna (2007)
91. Vasconcelos, N.: Image indexing with mixture hierarchies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Kauai (2001)
92. Whitman, B., Lawrence, S.: Inferring descriptions and similarity for music from community metadata. In: Proceedings of the 2002 International Computer Music Conference (ICMC), Göteborg, pp. 591–598 (2002)
93. Wolff, D., Weyde, T.: Adapting metrics for music similarity using comparative ratings. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), Miami (2011)

Index

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4, © Springer-Verlag London 2013