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

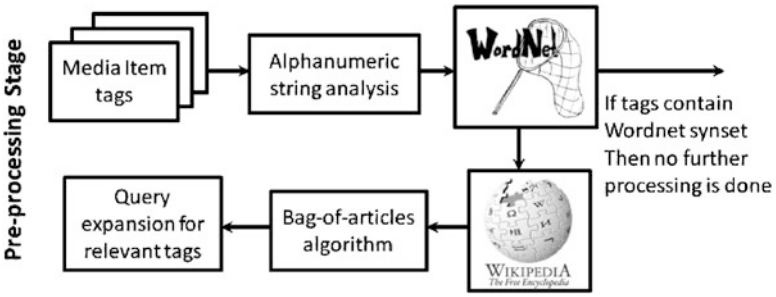

Fig. 1 Overview of the tag preprocessing phase

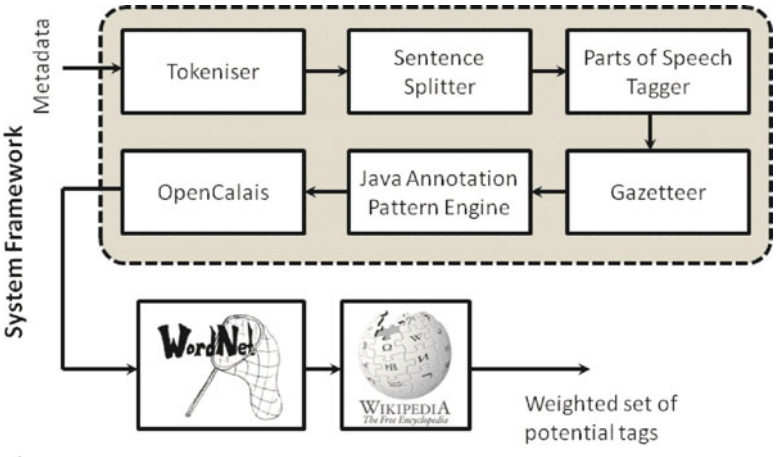

Fig. 2 Overview of the proposed system

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.

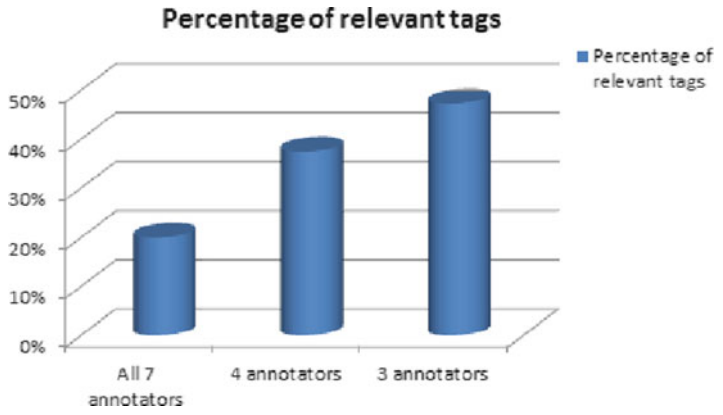

Fig. 3 Open-set annotation results

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)