



INTELLIGENT SYSTEMS REFERENCE LIBRARY
Volume 50

Gabriella Pasi
Gloria Bordogna
Lakhmi C. Jain (Eds.)

Quality Issues in the Management of Web Information

 Springer

Intelligent Systems Reference Library

Volume 50

Series Editors

J. Kacprzyk, Warsaw, Poland
L. C. Jain, Adelaide, Australia

For further volumes:
<http://www.springer.com/series/8578>

Gabriella Pasi · Gloria Bordogna
Lakhmi C. Jain
Editors

Quality Issues in the Management of Web Information

 Springer

Editors

Gabriella Pasi
Dipartimento di Informatica
Sistemistica e Comunicazione
Università degli Studi di Milano Bicocca
Milano
Italy

Gloria Bordogna
CNR - IDPA – Istituto per la
Dinamica dei
Processi Ambientali
Dalmine
Italy

Lakhmi C. Jain
University of Canberra
Canberra
Australia
and
University of South Australia
South Australia
Australia

ISSN 1868-4394

ISBN 978-3-642-37687-0

DOI 10.1007/978-3-642-37688-7

Springer Heidelberg New York Dordrecht London

ISSN 1868-4408 (electronic)

ISBN 978-3-642-37688-7 (eBook)

Library of Congress Control Number: 2013934982

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Foreword

Professor Dr. Carlo Batini
Department of Informatics, Systems and Communication
University of Milan Bicocca,
Milan
Italy

A book on Quality issues in the management of Web Information has to deal with a potentially wide number of issues. The concept of quality is pervasive, so pervasive that it is even difficult to provide a shared and usable definition of the concept of quality. The difficulty is reduced (not too much. . .) if we delimit the area of considered technologies and related resources. This book is focused on web retrieval technologies and on the information resource, that web retrieval technologies access and manipulate to provide knowledge access services to human beings and computer applications.

Information is in turn a pervasive concept, that is inherently related to other two concepts, data and knowledge. We can say with Boisot [1999] that “data is discrimination between physical states of things (black, white, etc.) that may convey or not convey information to an agent. Whether it does so or not depends on the agent’s prior stock of knowledge. . . . thus, whereas data can be characterized by a property of things, knowledge is a property of agents. . . information establishes a relationship between things and agents.”

The world is dirty, and also the Web, often a too vivid and faithful representation of the world, is dirty. Yet, the Web is and will be more and more in the future, the most accessed source of knowledge for the human beings. From this scenario, we can understand why the issue of information quality is of growing relevance in Computer Science literature and Information Systems applications, and in a wide spectrum of research areas and real life applications.

The issue of quality has been historically investigated first in the simplest case, data stored in databases, structured in domains and tables, and managed in transactional applications, under the rigid control of the organization. The second

age corresponds to the dispersion of data of interest for the organization in a multiplicity of databases, heterogeneous in format, content and semantics, that lead typically to represent in the information systems of the organization the same entity of the real world with multiple heterogeneous representations, characterized usually by different levels of quality.

A number of dimensions and related metrics have been proposed to formally characterize quality of data in these two scenarios. An analysis of the literature on data quality (see Batini and Scannapieco [2006] and Batini et al. [2009]), reports more than 50 dimensions and about 100 metrics, and at least 12 methodologies for the assessment and improvement of data quality in Information Systems using the database technology. Among dimensions, the most relevant are accuracy, currency, completeness and consistency, for the definitions see Batini and Scannapieco [2006]. Techniques range from record linkage and entity identification, to data cleansing, quality driven query answering, edit imputation and correction, outlier identification. With the advent of networks and the planetary diffusions of the Web, new types of information systems and data access and usage paradigms had to be considered. Among information systems, cooperative information systems allow different autonomous organizations to share data, applications and services, while peer-to-peer systems are characterized by higher autonomy and heterogeneity and absence of common management of data. Among new data access and usage paradigms, the evolution of information systems to cover a wide range of information representations, such as semi structured texts, unstructured texts, maps, images, videos, sounds, lead to develop access mechanisms where searches are based on metadata, tags and full-text indexing, giving rise to the Information retrieval research discipline.

In the area of data and information quality, the above diversification resulted in the investigation of dimensions, methodologies and techniques that cover all of the above mentioned types of information representations, previously in the world of single-organization information systems and cooperative information systems, and now in the different articulations of peer-to-peer information systems and the immense world of the Web. And while a fil rouge can be identified among dimensions defined in the different types of information representations (see Batini et al. [2012] for a discussion), when the other coordinates are considered (types of information systems and the Web), the need arises to consider new dimensions and new techniques. Among dimensions and related determinants, due to the uncontrolled and “anarchic” character of the Web, the attention is shifted to dimensions such as trustworthiness, provenance, authority, age and popularity (see e.g. for a discussion on Ramachandran et al. [2009]) that refer to quality of sources, besides the data and information they convey.

Focusing on the main theme of this volume, techniques are a wide range and cover issues such as quality driven retrieval, quality aware similarity search, quality of volunteered geographical information systems, quality based knowledge discovery in specific domains, quality of web engines. Such techniques are investigated in several papers of the present volume.

References

1. Batini, C., Scannapieco, M.: *Data Quality: Concepts, Methodologies and Techniques*. Springer (2006)
2. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: *Methodologies for data quality assessment and improvement*. ACM Computing Surveys (2009)
3. Batini, C., Palmonari, M., Viscusi, G.: *The many faces of information and their impact on information quality*. In: *Proceedings of International Conference on Information Quality*, Paris (2012)
4. Boisot, M.: *Knowledge Assets: Securing Information Advantage in the Information Economy*. Oxford University Press (1998)
5. Ramachandran, S., Paulraj, S., Joseph, S., Ramaraj, V.: *Enhanced Trustworthy and High-Quality Information Retrieval System for Web Search Engines*. IJCSI International Journal of Computer Science Issues 5 (2009)

Preface

This main focus of this book is on the quality issue in the management of information used in Web applications. A variety of tasks are concerned with and affected by the assessment of quality. The chapters included in this book are related to the tasks of Information Retrieval, Geographic Information Retrieval, Information Filtering and Knowledge Extraction. These areas demonstrate that by modelling and exploiting the quality dimensions of the information objects considered, it is possible to improve systems' effectiveness.

The problem of assessing the quality of textual information has been investigated for a long time. Several distinct proposals have been formulated. There is not a single unifying consensual definition of a texts' suitability for the task in hand. The problem of texts' quality assessment may be considered in relation to the information content itself (objective criteria), or from the user point of view. For example, in the context of Information Retrieval it is clear that the relevance of the documents to a request depends on several aspects. These are related to the distinct properties of the documents, the search, the user who formulated the query and the users who accessed the documents previously. It may include other information such as user ratings and tags, and the context of both documents and queries. One of the relevance dimensions may be related to the quality of documents. In case of Web pages well known algorithms (such as PageRank) have been defined.

This book has been organised into nine chapters. It includes recent contributions related to quality-based information management on the Web. Academic and applied researchers working on the issue of information quality will find the book a valuable reference resource. The methods, models and systems proposed in this book can inspire and motivate further research on important issues. It is hoped that final year undergraduate, masters and PhD students in computer science, and information systems will find in this book an excellent compiled reference text for their future studies.

We wish to thank all the contributors and referees for their excellent work and assistance and Springer-Verlag in producing this publication.

Gabriella Pasi, Italy
Gloria Bordogna, Italy
Lakhmi C. Jain, Australia

Editors



Gabriella Pasi

Gabriella Pasi graduated in Computer Science at the Università degli Studi di Milano, Italy, and took a PhD in Computer Science at the Université de Rennes, France. She worked as a researcher at the National Council of Research in Italy till 2005. Since 2005 she is Associate Professor at the Università Degli Studi di Milano Bicocca, Milano, Italy, where, within the Department of Informatics, Systems and

Communication she leads the Information Retrieval research Lab. Her research activity mainly concerns the modelling and design of flexible and personalised systems for the management and access to information (such as Information Retrieval Systems, Information Filtering Systems and Data Base Management Systems).

She is member of organizing and program committees of several international conferences. She has co-edited eight books and several special issues of International Journals. She has published more than 180 papers on International Journals and Books, and on the Proceeding of International Conferences. She is involved in several activities for the evaluation of research, in particular, she was appointed as an expert of the Computer Science panel for the Starting Grants of the Programme Ideas at the European Research Council. She is the Vice-President of the European Society for Fuzzy Logic and Technologies (EUSFLAT).

She is a member of the Editorial Board of the international journals *Fuzzy Sets and Systems*, *Journal of Computational Intelligence Systems*, *Web Intelligence and Agent Systems*, *Intelligent Decision Technology: An International Journal*, and *ACM Applied Computing Review*.

She was the coordinator of the European Project *PENG* (Personalized News Content Programming), a STREP (Specific Targeted Research or Innovation Project), within the VI Framework Programme, Priority II, Information Society Technology.

She organized several International events among which the IEEE / WIC / ACM International Joint Conference on Web Intelligence and Intelligent Agent

Technology, Università degli Studi di Milano Bicocca, 15–18 September 2009 (in the role of General Chair), the PhD School on Web Information Retrieval (WebBar 2007), the Seventh International Conference on Flexible Query Answering Systems (FQAS 2006), the European Summer school in Information Retrieval (ESSIR 2000), and she co-organizes every year the track “*Information Access and Retrieval*” within the ACM Symposium on Applied Computing.



Gloria Bordogna

Gloria Bordogna received the Laurea degree in Physics from the University of Milano in 1984. After a research contract at Politecnico of Milano, in 1986 she joined the National Research Council of Italy (CNR) where she currently holds the position of senior research scientist within the Institute for the study of the Dynamics of Environmental Processes (IDPA) located in Dalmine (BG), Italy.

From 2003 to 2010, she was adjunct professor of Information Retrieval and Geographic Information Systems at the Faculty of Engineering of Bergamo University. Her research interests mainly focus on *soft computing methods for the management of imprecision and uncertainty of textual and geographic information, multimedia information retrieval models, Flexible Query Languages for Information Retrieval Systems and Data Base Management Systems, Decision making for Environmental applications.*

Her current research is in the field of Web information retrieval, environmental knowledge acquisition by fuzzy data mining, and multisource geographic information fusion.

She participated in several European funded projects (*e-court, Peng, Ide-Univers*) international projects *EXIRPTS*, and Italian projects among which the current ones *RITMARE* and *Sistemati*.

She is in the editorial boards of *ACM SIGAPP – Applied Computing Review* and of the *Scientific World Journal*.

She also serves as a reviewer for several international journals such as *IEEE Trans. on Fuzzy Systems*, *IEEE Trans. on Sys. Man & Cyb.*, *FS&S*, *Inf. Proc. & Manag.*, *JASIST*, *JGIS*. She edited three volumes published by Springer and a special issue of *JASIST* and published over 150 scientific articles in journals, books, and proceedings of international conferences. She co-organizes every year the track “*Information Access and Retrieval*” within the ACM Symposium on Applied Computing.



Lakhmi C. Jain

Dr. Lakhmi C. Jain serves as Adjunct Professor in the Faculty of Information Sciences and Engineering at the University of Canberra, Australia. Dr. Jain founded the KES International for providing a professional community the opportunities for publications, knowledge exchange, cooperation and teaming. Involving around 5000 researchers drawn from universities and companies world-wide, KES facilitates international cooperation and generate synergy in teaching and research. KES regularly provides networking opportunities for professional com-

munity through one of the largest conferences of its kind in the area of KES. www.kesinternational.org

His interests focus on the artificial intelligence paradigms and their applications in complex systems, art-science fusion, e-education, e-healthcare, unmanned air vehicles and intelligent agents.

Contents

1 An Introduction to Quality Issues in the Management of Web Information	1
<i>Gabriella Pasi, Gloria Bordogna, Lakhmi C. Jain</i>	
1 Introduction	1
2 Chapters Included in the Book	2
References	3
2 Inverted File-Based General Metric Space Indexing for Quality-Aware Similarity Search in Information Retrieval	5
<i>Daniel Blank, Andreas Henrich</i>	
1 Introduction	5
1.1 Indexing Feature Objects for Quality-Aware IR	7
1.2 Applying Quality Measures as Filters	8
1.3 Contribution	8
2 Related Work	9
2.1 Purely Object-Pivot Based MAMs	10
2.2 Tree-Based MAMs	10
2.3 Inexact MAMs with Inverted Files	12
3 Outline of IF4MI	13
4 Experiments	17
4.1 Retrieval Performance of IF4MI	17
4.2 Comparing the Retrieval Performance of IF4MI with the M-Tree and PM-Tree	22
4.3 Memory Requirements of the Different Approaches	25
4.4 Using the Space Partitioning of the Metric Index	26
4.5 Improvements to Pivot Filtering	28
4.6 Processing Filter Queries	30
5 Conclusion and Outlook	31
References	32

3	Performance Prediction for Quality Recommendations	35
	<i>Josephine Griffith, Colm O’Riordan, Humphrey Sorensen</i>	
1	Introduction	35
2	Previous Work	37
	2.1 Evaluating Information Retrieval Quality	38
	2.2 Evaluating Collaborative Filtering Quality: Predictive Accuracy Focus	38
	2.3 Evaluating Collaborative Filtering Quality: User-Centric Focus	39
	2.4 Performance Prediction in Information Retrieval	40
	2.5 Performance Prediction in Collaborative Filtering	40
3	Datasets	41
4	Performance Prediction Approach	44
	4.1 Learning the Performance Prediction Rules	44
5	Evaluation: Testing the Rules	48
6	Results	49
	6.1 Movielens	49
	6.2 Lastfm	50
	6.3 Bookcrossing	50
	6.4 Performance Prediction Scenario	51
7	Conclusions	51
	References	52
4	Automated Cleansing of POI Databases	55
	<i>Guy De Tré, Daan Van Britsom, Tom Matthé, Antoon Bronselaer</i>	
1	Introduction	55
2	Related Work	57
	2.1 Coreference Detection	58
	2.2 Merging of Coreferent Data	58
3	Some Preliminaries	59
	3.1 Basic Concepts on Objects	59
	3.2 Basic Concepts on POIs	60
4	Detection of Coreferent POIs	62
	4.1 Elementary Evaluators for Atomic Objects	63
	4.2 Evaluators for Co-location	64
	4.3 Evaluators for Complex Objects	69
5	Merging of Coreferent POIs	71
	5.1 Merge Functions for Atomic Objects	71
	5.2 Merge Functions for Complex Objects	74
	5.3 Merging of Coreferent POIs	74
6	An Illustrative Example	76
	6.1 Illustration of Coreference Detection	76
	6.2 Illustration of Merging	80
7	Conclusions and Further Work	87
	7.1 Contribution	87

7.2	Context	88
7.3	Further Work	89
	References	89

5 A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web 93

Andrea Ballatore, David C. Wilson, Michela Bertolotto

1	Introduction	93
2	Survey of Open Linked Geo-Knowledge Bases	96
3	Open Geo-Knowledge Bases in Action	101
3.1	Mapping, Aligning, and Merging Geo-Knowledge Bases.....	101
3.2	Ontology-Powered Geographic Information Retrieval (GIR)	103
4	The OSM Semantic Network	106
5	The Quality of Crowdsourced Geo-Knowledge Bases	108
6	Current Limitations of Geo-Knowledge Bases and GIR	110
7	Conclusions and Future Work	112
	References	113

6 Fact Based Search Engine: News Fact Finder Utilizing Naive Bayes Classification 121

Ricardo Salmon, Cristina Ribeiro, Swathi Amarala

1	Introduction	121
2	Sentence Search Engine	123
2.1	Sentence Search	123
2.2	Search Engine Data Structure	124
3	News Fact Finder Naive Bayes	126
3.1	Algorithm	126
3.2	Data Structure	127
4	Fact Extraction	127
4.1	Opinion Sentence Search	128
4.2	Fact Finder	129
5	Experimentation	131
5.1	Experimental Results	132
6	Conclusion	141
	References	142

7 Quality-Based Knowledge Discovery from Medical Text on the Web 145

Andreas Holzinger, Pinar Yildirim, Michael Geier, Klaus-Martin Simonc

1	Introduction	145
2	Web-Based Tools for Analyzing Biomedical Literature	147
3	Pointwise Mutual Information	152
4	Symmetric Conditional Probability	154

- 5 FACTAs Scoring Methods: Frequency, PMI, and SCP 155
- 6 Conclusion 156
- 7 Future Work 157
- References 157

- 8 An Information Reliability Index as a Simple Consumer-Oriented
Indication of Quality of Medical Web Sites 159**
- Federico Cabitza*
- 1 Background and Motivations 159
- 2 The Medical Reliability Index Score 161
- 3 The Evaluation Process Related to the MIR Index 163
- 4 Validation of the MIR Index in the Medical Domain 166
 - 4.1 Validation Method 167
 - 4.2 Results 168
- 5 Discussion and Concluding Remarks 174
- References 175

- 9 Challenges for Search Engine Retrieval Effectiveness Evaluations:
Universal Search, User Intents, and Results Presentation 179**
- Dirk Lewandowski*
- 1 Introduction 179
- 2 Search Engine Evaluation 182
- 3 User Intents 184
- 4 Results Presentation in Web Search Engines 185
 - 4.1 Organic Results 185
 - 4.2 Sponsored Results 186
 - 4.3 Shortcuts 186
 - 4.4 Results from Special Collections 187
- 5 Users' Results Selection 187
 - 5.1 Results Position 188
 - 5.2 Results Description Content 189
 - 5.3 Results Description Size 189
 - 5.4 Results Description Design 189
- 6 Approaches to Weighting Results on the SERPs 190
 - 6.1 Results to Be Considered 190
 - 6.2 Position 190
 - 6.3 Screen Real Estate 191
 - 6.4 Click-Through Rates 191
- 7 Combining Universal Search SERPs and User Intent 191
- 8 Conclusion 192
- References 193

- Author Index 197**

Chapter 1

An Introduction to Quality Issues in the Management of Web Information

Gabriella Pasi, Gloria Bordogna, and Lakhmi C. Jain

1 Introduction

Since a long time information Quality Assessment is a crucial issue in organizations, where it relates to the ability of the organization to adequately fulfil the needs and expectations of its customers and users (Batini et al. 2008). In this context the evaluation of data quality has been the main concern (Batini and Scanapieco 2006). In recent years the quality issue has been increasingly concerned with the evaluation of information contained in several media sources such as

Gabriella Pasi
Università degli Studi di Milano Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Viale Sarca 336
20126 MILANO
Italy

Gloria Bordogna
CNR - IDPA -- Istituto per la Dinamica dei Processi Ambientali
c/o POINT, via Pasubio 5, 24044 Dalmine (BG)
Italy

Lakhmi C. Jain
University of Canberra
Canberra, ACT 2601
Australia
and
University of South Australia
Mawson Lakes Campus
South Australia SA 5095
Australia

texts, videos, images and sound. The advent of the Web has in fact produced a massive quantity of multimedia documents, the quality of which is not guaranteed due to the uncontrolled method used in their generation. This has raised the problem of determining how to assess the quality of information on the Web, which includes the important aspect of evaluating the information sources. With this being the aim, the assessment of information quality should include all of the above mentioned aspects with appropriate dimensions and metrics.

In the literature, several quality dimensions have been identified for both the texts and web pages but an adequate unifying vocabulary is not yet available. Many synonyms and distinct terms are used to identify the same dimensions (Eppler and Witting, 2000). For example, by the expression “contextual quality” one may indicate the source reputation, and the accessibility of the web site, and the in-link and out-link structure of the web page. The intrinsic quality is often named “representative quality”, which is dependent on concepts such as conciseness, the depth, the readability, grammar correctness and other factors such as timeliness.

Last but not least, the ISO defines the concept of information quality as “*the totality of the characteristics of an entity that bear on its ability to satisfy both the stated and implied needs*” [ISO 8402] (Devillers and Jeansoulin, 2010). This means that the quality of information is strictly related to the use of the information itself. Then according to (Devillers and Jeansoulin, 2010) it may be seen as an external measure. Thus, when evaluating the quality of documents it is essential to take into account both the purpose and the use for which the information is needed. That is, both the task and the query context.

For example, a mobile user searching on the Internet for “Japanese Restaurants in Milan” would be willing to first retrieve the web pages dealing with highly regarded Japanese restaurants close to the current location, as detected by any mobile device with which (s)he is equipped. To answer this user's need an information retrieval system, besides the restaurant quality assessment should also take into account a number of aspects such as opinions of other users who have already visited the restaurant, and the evaluations of specific journals/associations. If in fact the restaurant is evaluated by trusted journals and associations, these opinions are also an important factor.

Another example is the case of a user looking for a suitable camera to purchase. (S)he will probably trust more “higher quality” web pages. These will compare the prices of the cameras which have the particular wanted characteristics if the cameras have been evaluated by many knowledgeable users and, more specifically, those who are professional photographers. These are examples in which the quality assessment of web information represents an especially important issue.

2 Chapters Included in the Book

In this volume several recent contributions related to the issue of quality-assessment for Web Based information are included.

Chapter 2 presents an approach for a quality based indexing data structure. Metric space access methods are also reviewed. It is shown how they may be applied in the context of Quality Aware Information Retrieval.

In Chapter 3 the issue of quality related to the task of collaborative filtering is considered. Quality is considered from the perspective of performance prediction. In particular, the performance of a collaborative filtering system is predicted, based on rules learned by using a machine learning approach.

Chapter 4 presents a novel soft computing technique for the semi-automated cleansing of Points of Interest (POIs) in spatial databases with Volunteered Geographic Information (VGI), to obtain spatial information of a higher quality.

Chapter 5 addresses the issue related to the integration and usage of Volunteered GI within collaborative open datasets to attain spatially-aware Information Retrieval. This goes beyond the traditional Gazetteers-based approach. This Chapter pays particular attention to the crucial issue of quality assessment of open ontologies, as well as that of crowd sourced data.

Chapter 6 proposes the use of a novel approach to perform quality-based indexing and searching of facts present in news sources available on the Web. News-FactFinder is defined and developed to provide users with factual information relevant to a particular query, where a sentence level search is applied.

Chapter 7 presents a selection of quality-oriented Web-based tools for analyzing biomedical literature. Which includes PolySearch, FACTA and Kleio. Point-wise Mutual Information (PMI), which is a measure used to discover the strength of a particular relationship, is also analysed to provide an indication of the strength of users' queries and a concept in a knowledge source.

Chapter 8 proposes an intentionally simple composite index of information quality, in the so called Medical Information Reliability (MIR) index. This index considers the attitudes of potential and actual consumers towards taking into account information quality. It is intended as "trust indicator" of online sources of medical information.

Chapter 9 presents some techniques for evaluating the quality of Web search engines' ability when used to effectively retrieve information. It identifies those factors that lead to a need for new evaluation methods besides the traditional one based on Recall and Precision.

References

- [Batini and Scannapieco, 2006] Batini, C., Scannapieco, M.: *Data Quality: Concepts, Methodologies and Techniques*. Springer, Berlin (2006)
- [Batini et al., 2008] Batini, C., Barone, D., Cabitza, F., Ciocca, G., Marini, F., Pasi, G., Schettini, R.: *Toward a Unified Model for Information Quality*. In: *QDB/MUD 2008*, pp. 113–122 (2008)
- [Eppler and Witting, 2000] Eppler, M.J., Witting, D.: *Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years*. In: *Proceedings of the 2000 Conference on Information Quality*, pp. 83–96. Massachusetts Institute of Technology, Cambridge (2000)
- [Devillers and Jeansoulin, 2010] Devillers, R., Jeansoulin, R.: *Fundamentals of spatial data quality*. Wiley Online Library (2010) ISBN: 9781905209569, doi:10.1002/9780470612156

Chapter 2

Inverted File-Based General Metric Space Indexing for Quality-Aware Similarity Search in Information Retrieval

Daniel Blank and Andreas Henrich

Abstract. The notion of quality in its broadest sense is central to information retrieval (IR) where a user's information need is to be fulfilled as good as possible. A user searching for *cars on sale in Bamberg* might be interested in car dealers geographically close to Bamberg with high user ratings. The buyer might already know or trust a person who trusts the particular dealer. Furthermore, the cars which are sold by the dealer should offer a high quality on different levels—the type of car in general as well as the car to be bought. If the buyer can only travel to Bamberg on weekends, availability of the car dealer becomes another important factor. As this example shows, the integration of various quality aspects in IR is challenging but essential.

Thus, there is a need for scalable and efficient indexing and retrieval techniques which can cope with such search situations. Here, metric space access methods (MAMs) present a flexible indexing paradigm. We will briefly review these techniques and show how they can be applied in the context of quality-aware IR. Furthermore, we will present IF4MI which is purely based on the inverted file concept and thus inherently provides a multi-feature MAM. It can make use of extensive knowledge in the field of inverted file-based indexing and represents a versatile indexing technique for quality-aware IR.

1 Introduction

The relevance of documents (we use images as a running example) to information needs depends on various aspects and many different context factors are applied in IR, e.g. the geographic location *where* and the time *when* an image was taken, EXIF tags describing the exposure time and thus *stylistic properties* of an image, to name only a few. Some aspects directly refer to

Daniel Blank · Andreas Henrich

Media Informatics Group, University of Bamberg, 96052 Bamberg, Germany
e-mail: {daniel.blank, andreas.henrich}@uni-bamberg.de

Table 1 Access methods for different types of information

type of information	representation	similarity	access method
unstructured text	textual	IR model	inverted file
structured text	textual	IR (XML) model	inverted file
numbers, date & time	numeric 1-d	difference	e.g. B ⁺ -tree
(geo-)spatial	numeric 2-d	specific distance	SAM
spatial	numeric δ -d	specific distance	SAM
complex objects	black box	distance metric	MAM
complex objects	black box	distance	NAM

the quality of an entity involved in the IR process such as user ratings, trust measures, and for example page rank values [6] as an indicator for page quality. Considering and applying these aspects during search can enhance the search process and thus better result quality. From an indexing perspective it is thus worthwhile to design comprehensive access methods which can deal with the heterogeneity of these quality and/or context factors and which also keep track with the requirements according to efficiency and scalability.

Table 1 gives a brief overview of different types of information involved in a retrieval process and how they are typically indexed. The dominating index structure for unstructured text which can be optimized according to the applied IR model is the inverted file (cf. [50]). Retrieval systems for structured text are also often based on the inverted file (cf. [24]). Date & time information is usually represented as 1-dimensional numeric values. A concrete time stamp can for example be represented as an integer denoting the difference between the current time and midnight, January 1, 1970 in days, hours, minutes, or seconds. Thus, corresponding information can be indexed in a B⁺-tree (cf. [1]) or any similar structure. In opposition, geospatial information is frequently represented in terms of 2-dimensional spatial coordinates which are typically indexed in multi-dimensional (spatial) access methods (SAMs; for an overview cf. [35]). Interestingly, IR libraries such as Apache Lucene (<http://lucene.apache.org/>) relying on the inverted file concept have been extended to support the indexing of 1-dimensional numeric information in order to perform e.g. numeric range queries [36]. In addition, geospatial extensions for Apache Lucene are available.

What is missing in order to build a comprehensive access structure for exact search which inherently supports the indexing of all the above-mentioned types of information in an inverted file is the ability to index and search δ -dimensional feature vectors (in many scenarios $\delta \gg 2$) and even more complex objects (e.g. features for content-based image retrieval (CBIR), music IR, video retrieval, retrieval of DNA sequences, etc.).

In some settings, the representation of a feature object is not necessarily a feature vector. More complex objects may arise from feature modeling. Thus, metric access methods (MAMs) treat the feature representation as a black box and only rely on a distance metric which is capable of comparing different

objects. But also in situations where feature objects can be modeled as vectors (e.g. in CBIR) it might be desirable to use a certain distance metric which cannot be indexed by SAMs. There are also scenarios where the similarity of feature objects can best be modeled by non-metric distances (cf. [38]). While we will discuss and present MAMs in the following, we should stress that some of the approximation techniques for MAMs (cf. Sect. 2.3) can also be successfully applied in the context of non-metric access methods (NAMs) [38].

Our approach IF4MI is not only a single-feature MAM, but inherently provides a multi-feature MAM as will be explained in the following. It can be helpful in the context of quality-aware IR along two possible directions. We will describe these directions in sections 1.1 and 1.2 with the help of a small example scenario when searching for cars. The contribution of our chapter will in more detail be outlined in Sect. 1.3.

1.1 Indexing Feature Objects for Quality-Aware IR

MAMs can be applied for indexing tasks in various domains. In Sect. 4 we will use IF4MI to build a CBIR system. Within the thematic focus of this book chapter MAMs can also be used to index different kinds of quality profiles in the form of structured objects. For example, we could assume a database filled with car quality profiles (cf. Fig. 1). We thus can retrieve the k most similar profiles according to a given query profile when the similarity between different profiles can be modeled by a metric distance function.

<pre> <car_quality> <manufacturer> <certificates> <production> <ISO_TS_16949/> <ISO_9001/> <VDA_6.4/> </production> <environment> <ISO_14001/> </environment> </certificates> </manufacturer> <milage> <longlasting/> </milage> </car_quality> </pre>	<pre> <car_quality> <manufacturer> <certificates> <production> <ISO_TS_16949/> </production> </certificates> </manufacturer> <milage> <shortlasting/> </milage> </car_quality> </pre>
---	---

Fig. 1 Car quality profiles which could be indexed by MAMs

Distance metrics for tree-structured data are described in [2, 13]. In [10] for example user profiles are modeled as concept trees which are matched against document profiles in order to recommend relevant research papers to interested authors. In general, examples for quality-aware IR could be the indexing of quality profiles, user profiles, trust profiles, etc. MAMs are also used for the indexing of graph-based models supporting for example search for similar business process models[23], 3D object models [8], video models [25], or function-call graphs to detect malware programs [21].

1.2 Applying Quality Measures as Filters

IF4MI inherently provides a multi-feature MAM grounded on the inverted file concept. Additional filters can be used in combination with traditional query processing as described in Sect. 3. If we for example index the above-mentioned car quality profiles or car images, we could use a special tag such as *cabrio* as additional textual filter criterion (cf. Sect. 4.6). Quality measures for images—which can also be used as filter criteria—are for example described in [15] where the authors try to capture aesthetics and emotions and in [22] where page rank ideas are applied to identify the most characteristic image(s) for product search. These quality aspects can be incorporated into the inverted file structure similar to traditional page rank values (cf. [14]).

By doing so, we can build a multi-feature MAM which is able to index many different types of information as well as context and quality factors in a single structure.

1.3 Contribution

The contribution of this chapter is as follows:

- We give a brief overview on MAMs.
- We outline IF4MI—a flexible MAM suitable for quality-aware IR—and describe its main building principles and algorithms at heart.
- We compare IF4MI with two existing MAMs—an M-tree [12] and a PM-tree [40] implementation—showing that IF4MI can outperform these approaches in certain scenarios according to the number of necessary distance computations.
- We apply the space partitioning technique of the Metric Index—which we consider a current state-of-the-art MAM—showing that its pruning power can be brought to IF4MI and thus inverted files without relying on a mechanism which maps feature objects to one-dimensional values for storing them in adequate data structures such as a B⁺-tree.
- We present a heuristic to boost runtime performance of pivot filtering—frequently used by and applicable to many MAMs in order to avoid unnecessary distance computations.

- IF4MI can benefit from the extensive knowledge in the field of inverted file-based query processing. To give just one example, we show how the execution of multi-feature queries combining similarity search with a specific filter criterion can benefit from the inverted file concept. This opens doors for the application of quality-based filters essential in the context of quality-aware IR.

2 Related Work

Many similarity search problems can be modeled in general metric spaces where no assumption is made about the representation of the feature objects. MAMs only rely on the use of a metric distance. Thus, they provide a flexible basis for a quality-aware IR framework. On the one hand, *approximate* MAMs have been proposed which are based on the inverted file—the de facto standard index structure for text retrieval. On the other hand, there are many *exact* hierarchical and multi-step MAMs. We try to connect these two worlds and present IF4MI. In this section, we will further introduce MAMs and give an overview on related work in the field of MAMs (cf. also [11, 20, 35, 49]).

A metric space \mathcal{M} is defined as a pair $\mathcal{M} = (\mathbb{D}, d)$. \mathbb{D} represents the domain of objects $o \in O$ with $O \subset \mathbb{D}$ and $d : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}$ corresponds to a metric distance function which satisfies the metric postulates $\forall x, y, z \in \mathbb{D}$ [49]:

$$\begin{aligned}
 d(x, y) = 0 &\iff x = y && \textit{identity} \\
 d(x, y) > 0 &\iff x \neq y && \textit{non-negativity} \\
 d(x, y) &= d(y, x) && \textit{symmetry} \\
 d(x, y) + d(y, z) &\geq d(x, z) && \textit{triangle inequality}
 \end{aligned}$$

Range and k -nearest neighbor (k -NN) queries which are defined in the following are amongst the most popular types of similarity queries in IR. A query object $q \in \mathbb{D}$ provides the basis for retrieving similar objects from the database O .

A *range query* with query object $q \in \mathbb{D}$ and search radius $r \in \mathbb{R}^+$ retrieves all database objects from O which are within distance r from q , i.e. $\{o \in O \mid d(q, o) \leq r\}$. The subspace $\mathbb{V} \subset \mathbb{D}$ for which $\forall v \in \mathbb{V} : d(q, v) \leq r$ and $\forall v' \in \mathbb{D} \setminus \mathbb{V} : d(q, v') > r$ is called the *query ball* [38].

In some scenarios it might be difficult to explicitly specify the search radius r . Instead, one might be interested in the k closest database objects from q .

A *k -nearest neighbor query* (k -NN query) retrieves the k closest database objects from q , i.e. a set $K \subset O$ with $\forall o \in K, o' \in O \setminus K : d(q, o) \leq d(q, o')$ and $|K| = k$. In the following $|O| \gg k$ is assumed.

Many MAMs rely on a set $C = \{c_i \mid 1 \leq i \leq n\}$ of reference objects (also called pivots or centers) in order to structure the feature space. Within ball partitioning methods [49] the feature space is partitioned by usually multiple hyperspheres. In contrast, many structures relying on generalized hyperplane

partitioning [49] compute the n_{idx} ($1 \leq n_{idx} \leq n$) closest reference object(s) and assign o to the corresponding cluster(s).

In the field of MAMs, it is often assumed that the number of necessary distance computations at query time is the dominating cost factor [37]. To reduce these costs, the triangle inequality is usually used in combination with distances which have been precomputed at indexing time in order to prune regions of the feature space or individual feature objects from search. Various pruning constraints are applied [11, 20, 35, 49]. We describe them in the context of range queries in Sect. 3 following the notation of [49]. How these pruning criteria are used for k -NN queries in case of IF4MI is also outlined in Sect. 3.

2.1 *Purely Object-Pivot Based MAMs*

The Approximating and Eliminating Search Algorithm (AESA) [45, 46] is an indexing technique storing all $\mathcal{O}(|O|^2)$ object-to-object distances. During query processing, database objects are iteratively selected as pivots (one per round) according to certain criteria and the pre-computed distances are used for the pruning of irrelevant database objects by applying pivot filtering (cf. Sect. 3 and formula 1 on page 15). Those database objects which cannot be pruned have to be evaluated against the query. Besides its $\mathcal{O}(|O|^2)$ space complexity, the construction time complexity of AESA is also $\mathcal{O}(|O|^2)$ [11]. For a long time, AESA has been considered to require the least number of distance computations amongst all MAMs [18, 41]. Thus, it can be considered as a best case comparison baseline for various MAMs. Only recently, two approaches have emerged which claim to be able to outperform AESA in this regard (cf. [18, 41]).

In order to overcome the quadratic space complexity of AESA, Linear AESA (LAESA) [28] has been proposed which applies a set of n pivots with $n \ll |O|$. Only object-pivot distances are stored and used in order to prune irrelevant database objects. LAESA-based query processing is similar to object pruning in the posting lists of our approach (cf. Sect. 4). In future work we will take a closer look at improvements which have been proposed in the context of LAESA (for references cf. [11, 20]) to further improve our approach—for example by indexing the pivots themselves.

2.2 *Tree-Based MAMs*

Amongst the most prominent tree-based MAMs is the M-tree [12]. Feature objects are administered at the leaf nodes. Inner node entries consist of a routing object, a pointer to a subtree, a maximum distance of objects in the subtree from the routing object, and a distance from the routing object to the routing object of the parent node entry. While traversing the balanced tree structure during query processing, irrelevant subtrees are pruned if they

do not intersect with the query ball. The PM-tree [40] can be envisaged as an extension of the M-tree. Additional preselected pivots are applied in order to support more restrictive pruning (cf. Sect. 4.2). Therefore, subregions of the feature space are in case of the PM-tree represented more precisely by intersections of a hyper-sphere and multiple hyper-rings (in contrast to the M-tree where they are represented only by a hyper-sphere). Pivot filtering is additionally applied at the leaf level where database objects are administered.

SSSTree [7] is a recent approach which applies generalized hyperplane partitioning. The heart of SSSTree is a dynamic pivot selection technique which adapts to the intrinsic dimensionality (cf. Sect. 4) of the metric space. In experiments in [7], the SSSTree outperforms other techniques based on generalized hyperplane partitioning such as different variants of GNAT [5].

Recently, the Metric Index has been proposed [31, 32]. Similar to IF4MI, the Metric Index combines ball partitioning and generalized hyperplane partitioning. The Metric Index adopts the idea of permutation-based indexing (PBI) where lists of cluster IDs i sorted according to ascending $d(c_i, o)$ distances provide the basis for representing the clusters. A small number of m reference objects together with an additional parameter l ($1 \leq l \leq m$) denoting that the l closest centers are used for computing the PBI-based representations is sufficient to achieve a very fine-grained cluster structure with a high number of cluster cells.

Figure 2 outlines the space partitioning technique of the Metric Index where $m = 4$ reference objects are used and only the two closest ($l = 2$) centers are applied for computing the PBI-based representations. Cluster $\Gamma_{x,y}$ hereby denotes that database objects within this cluster are closest to cluster center c_x and second closest to center c_y . Theoretically, at most $m^l = m \cdot (m - 1) \cdot \dots \cdot (m - l + 1)$ clusters are thus possible. Usually, not all of them exist, e.g. cluster $\Gamma_{1,4}$ and $\Gamma_{4,1}$ are missing in Fig. 2. Based on the permutation-based representations and thus assignments to clusters, database objects are mapped to one-dimensional keys which are then indexed in a B^+ -tree. If a cluster cannot be pruned during query processing, the search radius is mapped to a key interval of the underlying B^+ -tree which is then queried. Experiments in [31, 32] show better performance of the Metric Index compared to the PM-tree according to the number of necessary distance computations.

IF4MI, presented in more detail in Sect. 3, is in some respect similar to the Metric Index which, however, uses a smaller number of centers ($m \ll n$) in combination with PBI techniques. In case of the Metric Index, the cluster to which an object o belongs can thus be computed with only m distance computations whereas IF4MI requires—without further optimization— n distance computations. At first glance, the permutation-based space partitioning seems beneficial. But, compared to the Metric Index IF4MI is able to apply different pruning constraints more extensively. This might lead to better pruning of irrelevant clusters. $\mathcal{O}(2n^2)$ additional space seems affordable in our case (cf. Sect. 4). We will evaluate the influence of individual

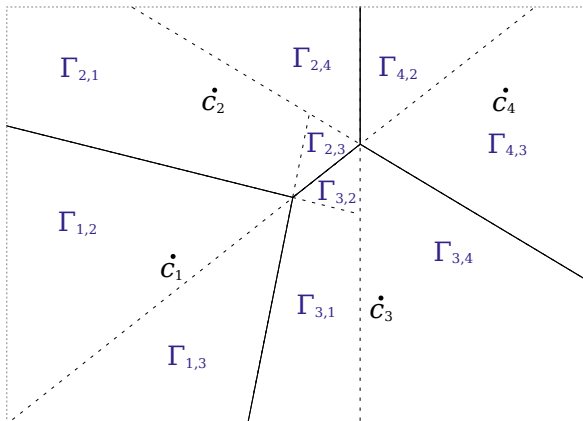


Fig. 2 Outline of the space partitioning applied by the Metric Index with $m = 4$ and $l = 2$. Cluster $\Gamma_{1,4}$ and cluster $\Gamma_{4,1}$ do not exist. Lines are drawn for visualization purpose. They cannot always be determined in a metric space.

pruning constraints on the retrieval performance in more detail in future work. In Sect. 4.4 we evaluate the application of a cluster structure similar to the one of the Metric Index within the inverted file-based IF4MI approach. It appears that IF4MI without PBI is competitive with the Metric Index and that furthermore PBI can also be used with IF4MI.

2.3 Inexact MAMs with Inverted Files

There are different approaches for fast approximate similarity search in general metric spaces. A recent approach offering probabilistic guarantees proposes locality sensitive hashing for metric spaces [43]. Many other approaches follow the idea of PBI assuming that two objects are similar if their permutations of cluster IDs ordered according to ascending $d(c_i, o)$ are similar. In [17], permutations are indexed in a prefix tree. Query processing relies on matching the prefix of the query permutation to subtrees. Other inexact approaches applying PBI are based on inverted files. These approaches are outlined in the following in more detail.

Gennaro et al. [19] maps every center c_i to a posting list. Object IDs are added to the posting lists of the n_{idx} closest cluster centers. A posting stores a virtual term frequency similar to the term frequency in traditional text retrieval. To compute the virtual term frequency, all n centers are sorted by ascending distance $d(c_i, o)$. A virtual term frequency of n_{idx} is assigned to the posting of o in the posting list of the closest center; a value of $n_{idx} - 1$ is assigned to the posting of o in the posting list of the second closest center, etc. Query processing is similar to text retrieval purely based on term frequency. The query is transformed so that the closest center gets a weight n_q , the

second closest a weight $n_q - 1$, etc. Different numbers of reference objects can be used to compute the permutation of cluster IDs (n), term frequency during indexing (n_{idx}), and query weights (n_q) with $1 \leq n_q \leq n_{idx} \leq n$.

Sznajder et al. [42] presents the Metric Inverted Index. Multiple features are indexed in a more independent way compared to [19]. Clustering is used to determine a set of n lexicon entries (i.e. cluster centers) per feature. For every feature, object references of o are added to the posting lists of the closest centers. Similarly, a set of matching posting lists is computed for the query. Query modes based on Boolean retrieval are then applied. Feature-specific scores are combined with an aggregate function. The same authors also propose Pivots Crossing Approximation [26] being conceptually similar to [42].

It should be noted that the use of inverted files for content-based media retrieval has been introduced many years ago (cf. e.g. [29]). Nevertheless, the approaches presented in this section cannot guarantee exact results. Heuristics are applied to find a balance between adequate retrieval quality and acceptable computational complexity. In this regard, we see IF4MI as a first step toward a scalable and efficient retrieval framework which allows both—exact as well as approximate query processing. It is part of future work to extend it with heuristics to further minimize computational complexity while at the same time ensuring adequate retrieval quality.

3 Outline of IF4MI

The basic idea of IF4MI is rather simple. We use n pivot objects and maintain a list for each pivot containing references to objects for which this pivot is the closest among all pivots. On this basis, IF4MI uses different pruning criteria applicable in the context of MAMs. These criteria are used to prune certain regions of the feature space (i.e. complete lists of the structure) or individual feature objects from search. We briefly outline these criteria in the context of range queries following the notation of [49]. They are in detail described in [11, 20, 35, 49]. How these pruning criteria are used for k -NN queries in case of IF4MI is also shown in this section.

If a query lies in the cell of center c^* , i.e. reference object c^* is the closest center out of the set C of all available reference objects according to a given query object q , we can prune—by exploiting the triangle inequality—any cluster Γ_i (and hence all objects within the very cluster) if $d(c_i, q) - d(c^*, q) > 2r$, where r corresponds to the search radius (*double-pivot distance constraint*).

If a maximum cluster radius r_i^{max} for a cluster Γ_i is given, i.e. the maximum distance of any object o in the cluster from its center c_i , the very cluster can be pruned if $d(c_i, q) - r > r_i^{max}$ (*range-pivot distance constraint*). A similar condition can be applied according to the minimum cluster radius r_i^{min} , i.e. the minimum distance of any object o within the cluster from its center c_i . We can prune cluster Γ_i if $d(c_i, q) + r < r_i^{min}$.

The range-pivot distance constraint can also be used in an inter-cluster way. Two matrices MAX and MIN are applied to store maximum and minimum cluster radii $r_{i,j}^{max}$ and $r_{i,j}^{min}$ respectively for $i, j \in \{1, \dots, n\}$ where $r_{i,j}^{max}$ represents the maximum distance from any object out of cluster Γ_i to cluster center c_j and $r_{i,j}^{min}$ represents the minimum distance from any object out of cluster Γ_i to cluster center c_j . Elements $r_{i,i}^{max}$ and $r_{i,i}^{min}$ on the diagonal of the matrices MAX and MIN thus capture the maximum cluster radius r_i^{max} and minimum cluster radius r_i^{min} of cluster Γ_i respectively as described above. Cluster Γ_i can be pruned if there exists a cluster Γ_j for which $d(c_j, q) + r < r_{i,j}^{min}$ or $d(c_j, q) - r > r_{i,j}^{max}$ [47].

Fig. 3 visualizes a search situation performing a range query with query radius r where cluster Γ_2 can be successfully pruned. By solely using the double-pivot distance constraint, cluster Γ_2 cannot be pruned since the query ball \mathbb{V} intersects cluster Γ_2 . (For visualization purposes the cluster border between Γ_1 and Γ_2 is drawn as a solid line.) If we administer for every cluster only the minimum and maximum cluster radius of objects in the cluster (shown by the hyper-ring $\mathbb{H}_{2,2}$ around cluster center c_2 in Fig. 3), cluster Γ_2 can still not be pruned. The matrices MIN and MAX are thus necessary to successfully prune cluster Γ_2 . If we also apply the radii $r_{2,1}^{min}$ and $r_{2,1}^{max}$, i.e. the minimum and maximum distance of feature objects in cluster Γ_2 from c_1 , it can be determined that there are in fact no relevant feature objects in the intersection area of the query ball \mathbb{V} and the hyper-ring $\mathbb{H}_{2,2}$. The region of possible feature objects is thus limited to the two intersection areas of $\mathbb{H}_{2,2}$ and $\mathbb{H}_{2,1}$ and since the query ball \mathbb{V} does not intersect any of these regions, cluster Γ_2 does not contain any database objects relevant to the query.

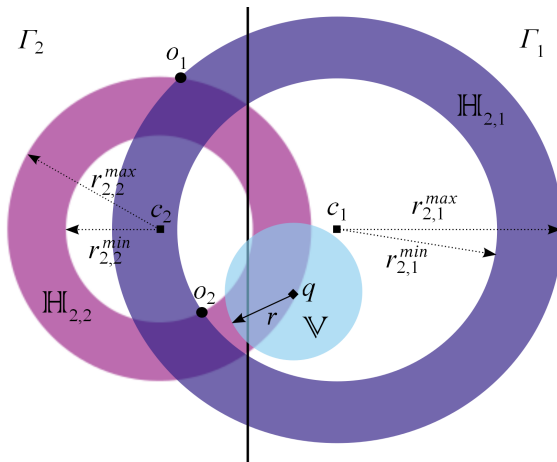


Fig. 3 Cluster pruning example demonstrating the usefulness of MIN and MAX

A further pruning constraint can be applied on an object level rather than a cluster level. If $|d(c_i, q) - d(c_i, o)| > r$, object o can be pruned without computing $d(q, o)$. This is called the *object-pivot distance constraint*. Hence, distance computations between the query object q and a feature object o can be avoided at the price of storing distance values $d(c_i, o)$ which are often anyway computed during the insertion of o into the index structure. Usually, $d(c_i, o)$ values are stored for multiple cluster centers c_i . Hence, the computation of $d(q, o)$ can be skipped if the condition in formula 1 is fulfilled. This so called *pivot filtering* is a direct application of the object-pivot distance constraint.

$$\max_{c_i} |d(c_i, q) - d(c_i, o)| > r \quad \text{pivot filtering} \quad (1)$$

IF4MI makes use of all the above-mentioned pruning criteria. We use a set of n reference objects and assign a feature object o to its closest cluster center $c^* = \arg \min_{c_i \in C} d(c_i, o)$. Cluster IDs $cid(c_i)$ are used as virtual terms. Hence, we obtain a lexicon of size n (cf. Fig. 4). During insertion, an object reference $oid(o)$ is only inserted into the posting list of c^* . Note that every object ID is thus contained only in a single posting list of the inverted index. By default, the postings of a posting list are sorted by $oid(o)$ in increasing order.

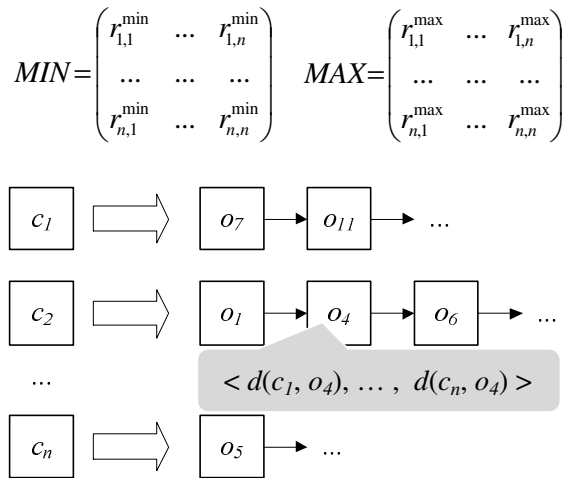


Fig. 4 Outline of IF4MI structure

Two additional matrices MAX and MIN as explained before are administered in main memory and used for cluster pruning, i.e. pruning of posting lists without processing them. This requires $\mathcal{O}(2n^2)$ additional space. At the object level, i.e. for each posting, we maintain up to n object-pivot distances $d(c_i, o)$. By doing so we can apply pivot filtering when traversing posting lists which could not be pruned before. An algorithm for k -NN queries with q as the query object is outlined in Fig. 5.

L: list of $\langle cid(c_i), d(c_i, q) \rangle$ pairs sorted by ascending $d(c_i, q)$
topk[]): array of length k with $\langle oid(o), d(q, o) \rangle$ pairs,
initially: $d(q, o) = \infty$ for all entries
n: total number of centers used
pos: position counter for *L*, initially set to 0

```

while topk[k - 1].d ≥ (L[pos].d - L[0].d)/2 do
  | if !clusterPruningPossible(L[pos].cid) then
  | | processPostingList(L[pos].cid, topk)
  | end
  | pos = pos + 1
  | if pos == n then
  | | break
  | end
end

```

Fig. 5 Algorithm for k -NN queries

The while-condition is a direct application of the double-pivot distance constraint. From the fact that L is ordered by ascending distance values $d(c_i, q)$ it follows that the search can stop as soon as the condition is not fulfilled for the first time since $L[pos].d$ monotonically increases with increasing pos . In the next line it is checked if the cluster with ID $L[pos].cid$ can be pruned. Here, the range-pivot distance constraint is tested with the help of MAX and MIN in an intra- as well as inter-cluster fashion. If a cluster cannot be pruned its posting list is processed and pivot filtering is applied to the objects whenever possible. If an object cannot be pruned $d(q, o)$ has to be computed and $topk$ is possibly updated. Hereby, the current search radius (i.e. $topk[k-1].d$) used for object and cluster pruning might also be updated. The algorithm for range queries is straightforward.

As mentioned before, we consider the Metric Index [31, 32] as a current state-of-the-art MAM. Its superior performance according to the number of necessary distance computations is influenced by the fact that it applies all the pruning constraints outlined above. IF4MI uses the same pruning constraints. However, an obvious difference between IF4MI and the Metric Index is the space partitioning used by these approaches. Thus, we extend IF4MI by applying the space partitioning of the Metric Index (cf. Sect. 2.2 and Fig. 2) and compare it with the initial space partitioning of IF4MI. Results are shown in Sect. 4.4. We use the algorithm outlined in [30] to map l -permutations which identify the clusters of the Metric Index to integers in the range of $[0, m^l - 1]$ in order to be able to store minimum and maximum cluster radii in the two matrices MAX and MIN as before. Query processing is implemented in form of a k -NN algorithm similar to the one described in [32] which relies on

a priority queue ordering clusters by a penalty score. This penalty score captures the “proximity” between a cluster and the query object.

4 Experiments

Retrieval performance of IF4MI is analyzed in detail in Sect. 4.1. Sect. 4.2 shows that IF4MI can in certain scenarios outperform the M-tree [12] and the PM-tree [40]—two alternative MAMs—according to the number of necessary distance computations. The memory requirements of the different approaches are analyzed in Sect. 4.3. In Sect. 4.4, we compare our approach with the Metric Index [32] and apply its space partitioning technique to our inverted file-based approach. Many MAMs such as the PM-tree, the Metric Index, and IF4MI rely on pivot filtering. In Sect. 4.5, we present a heuristic in order to reduce the number of centers c_i which need to be tested when applying pivot filtering. This heuristic is not limited to IF4MI and can be applied within various MAMs. Finally, Sect. 4.6 analyzes the benefits of IF4MI when performing multi-feature queries comprised of a k -NN similarity query and an additional filter criterion.

In the following, we assume a CBIR task. Experiments are based on the CoPhIR dataset [4]. Our collection consists of the first 100,000 images from CoPhIR archive no. 1. Twenty runs with varying sets of cluster centers are performed which are randomly chosen from the remaining images of the first archive. In each run we use the same set of 200 query objects randomly selected from CoPhIR archive no. 106. We thus perform 200 20-NN queries per run. We use the features which come with the CoPhIR dataset (distances in brackets according to [27]): ScalableColor (L1), ColorStructure (L1), ColorLayout (weighted L2), EdgeHistogram (variant of L1), and Homogeneous-Texture (variant of L1).

The difficulty of an indexing task in metric spaces can be characterized by the intrinsic dimensionality defined as $\rho = \mu^2 / (2\sigma^2)$ where μ represents the mean and σ^2 corresponds to the variance of the histogram of pairwise distances [11]. We prepared four feature combinations with different values of ρ : FC₁ (ColorLayout only, $\rho = 4.4$), FC₂ (EdgeHistogram only, $\rho = 7.7$), FC₃ (ColorLayout and EdgeHistogram, $\rho = 10.9$), FC₄ (all features, $\rho = 13.7$). In case of FC₃ and FC₄ features were normalized and weighted equally.

4.1 Retrieval Performance of IF4MI

At first, we consider the number of distance computations as the dominating cost factor. This is in correspondence with much of the literature in the field of MAMs [37]. In addition, it is motivated by the fact that IF4MI can be used as a main-memory indexing technique following the general trend toward in-memory databases [34], because of IF4MI’s potential for a relatively small index size. (The memory requirements of IF4MI will be analyzed in Sect. 4.3.)

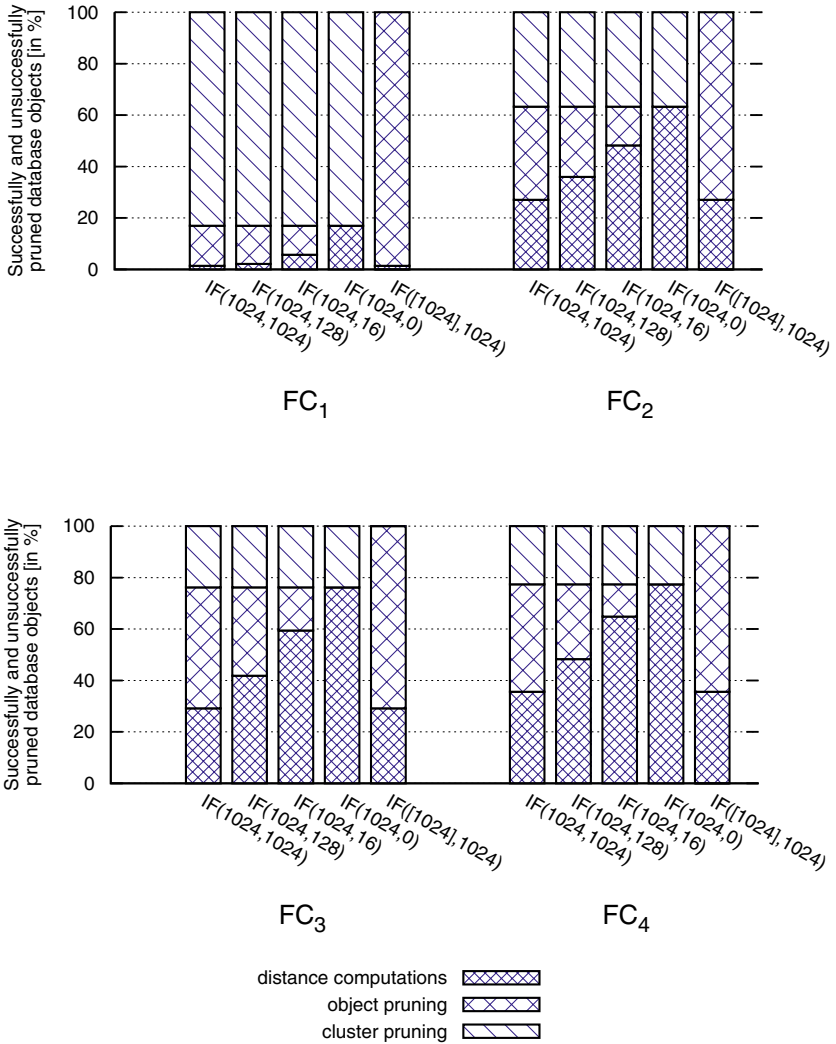


Fig. 6 Necessary distance computations and pruned feature objects

Of course, IF4MI can also be applied as a disk-based index. In this case, cluster pruning is beneficial in order to avoid the unnecessary scanning of posting lists which are stored on disk. When clusters cannot be pruned from query processing and posting lists are accessed, object pruning comes into place in order to further reduce the number of database objects for which distances $d(q,o)$ have to be computed.

Fig. 6 shows the average fraction of distance computations, the average fraction of pruned objects via cluster pruning, and the average fraction of pruned objects via object pruning. 100% correspond to $|\mathcal{O}| + n = 100,000 + 1024 = 101,024$ because the number of distance computations is always at least $n = 1024$ since the list L (cf. the algorithm in Fig. 5) is computed during query processing in any case. $\text{IF}(n, n')$ will in the following denote the parameter values of IF4MI with n indicating the total number of reference objects used and n' ($n' \leq n$) indicating the number of $d(c_i, o)$ distances stored and used for pivot filtering (cf. Sect. 3).

In case of $\text{IF}(1024, 1024)$ based on 1024 cluster centers 1024 precomputed $d(c_i, o)$ distances stored in the postings are used for object pruning (cf. Fig. 6). For $\text{IF}(1024, 128)$ cluster pruning is still based on the whole set of $n = 1024$ centers while only $n' = 128$ randomly chosen centers and thus 128 precomputed $d(c_i, o)$ distances are applied for pivot filtering. Of course, reducing the number of distance values stored in the postings for example from 1024 to 128 reduces the memory size of the index and I/O cost in case of the posting lists being stored on secondary memory. But, at the same time, the number of distance computations increases since fewer objects can be pruned. Both, n' and n are tuning parameters of IF4MI and a more detailed analysis in this regard is part of the remainder of this section.

$\text{IF}([1024], 1024)$ —corresponding to the rightmost bar in each group in Fig. 6—is a parameter setting similar to LAESA (cf. Sect. 2.1) where no cluster pruning is applied at all although the cluster structure based on 1024 reference objects is used for indexing. All posting lists are in this case processed in order to prune objects based on $n' = 1024$ precomputed $d(c_i, o)$ distances. From Fig. 6 it can be observed that pivot filtering at the posting level is very effective. $\text{IF}(1024, 1024)$ requires the same amount of distance computations as $\text{IF}([1024], 1024)$. Nevertheless, the latter is more expensive w.r.t. disk I/O costs since no database objects are pruned via cluster pruning. Of course, especially in a setting where the posting lists are administered on secondary memory, we would like to reduce the number of pruned objects through more effective cluster pruning instead of massively relying on object pruning and thus disk I/O reads. The parameter settings in Fig. 6 with $n' < n$ represent compromises reducing disk I/O costs while increasing the number of distance computations. An optimal value for n' can be found based on the computational complexity of the distance measure, the available storage space, and the general scenario of a main memory or secondary memory based index.

To give an impression, Fig. 7 shows—for a randomly selected run—the number of postings per posting list. For this particular run only a single cluster does not contain any postings. We can see that the distribution of the number of postings per cluster is skew. It is likely to be influenced by the technique used for choosing the reference objects. Thus, we will focus on this issue in more detail in future work when analyzing different pivot selection techniques (cf. e.g. [7, 9, 49]).

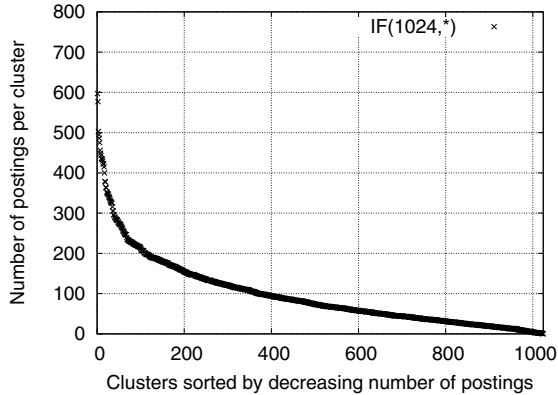


Fig. 7 Number of postings per posting list for IF(1024,*) [100,000 postings in total]. The placeholder * for n' indicates that these measurements are independent of n' .

In order to further analyze the characteristics of IF4MI we vary the parameters n and n' . Fig. 8 thus shows the fraction of necessary distance computations compared to a sequential scan for different parameter values of n and n' . Results are shown for FC₄ only. It can be observed that increasing n and thus the number of clusters only pays off—according to the number of necessary distance computations—if n' is relatively small (e.g. $n' \leq 40$). In these situations there is a steady decrease in the fraction of distance computations with increasing n , especially when n' is very small such as in case of IF(n ,1). For bigger numbers of n' (e.g. $n' = 128$) pivot filtering is able to prune large amounts of irrelevant database objects and thus a very small $n = 256$ already offers the best retrieval performance amongst the measured parameter settings. The increase in the fraction of distance computations for $n' = 128$ when n becomes large can be explained by the fact that n distance computations are already performed per query before actually entering the pruning process when computing the list L outlined in the algorithm presented in Fig. 5 on page 16. Here, the space partitioning used by the Metric Index is a suitable alternative when a larger number of clusters is desired (cf. Sect. 2.2 and Sect. 4.4).

As can be observed from Fig. 8, a large number of clusters is inevitable to reduce the number of distance computations when memory requirements do not permit the use of a sufficient number of reference objects for pivot filtering. Otherwise, whenever it is affordable to apply a rather large number of reference objects (e.g. $n' = 128$), the influence of choosing an adequate n is not that crucial, and even a small $n = 256$ leads to a reduction in the total number of distance computations; IF(256,128) performs best amongst the different parameter setting displayed in Fig. 8.

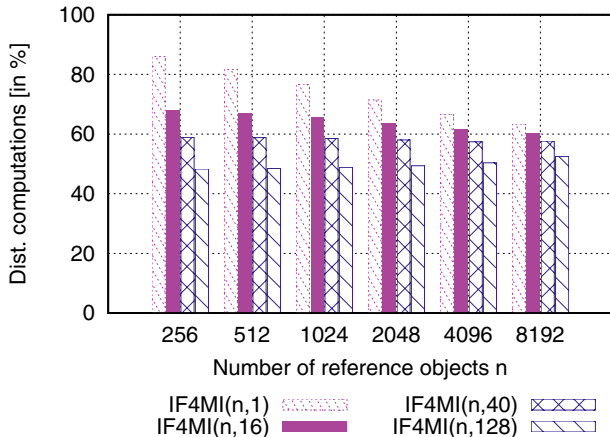


Fig. 8 Fraction of distance computations compared to a sequential scan when varying n and n' for IF4MI(n, n')

In order to reduce disk I/O costs when storing the posting lists of IF4MI on secondary memory, it becomes crucial to prune as many clusters as possible from search. Tab. 2 shows for a randomly selected run the average number of pruned clusters for different values of n . In case of many clusters (i.e. $n = 8192$) almost 50% of the clusters are pruned. However, the remaining 50% have to be read from disk.

Table 2 Number of pruned clusters of IF4MI for different values of n

n	empty clusters	pruned empty clusters	pruned clusters total
256	0	0.00	39.53
512	0	0.00	104.22
1024	1	0.97	270.58
2048	15	13.85	682.24
4096	76	70.87	1677.94
8192	395	369.14	3948.21

Fig. 9 shows for IF($n,40$)—when varying the number n of reference objects used—the average number of necessary distance computations as well as the average number of database objects being pruned from search by cluster and object pruning. It can be observed that larger numbers of clusters lead to more objects being pruned by cluster pruning. On the other hand, less database objects are to be pruned by object pruning. A larger number of clusters might thus justify the use of fewer reference objects for object pruning.

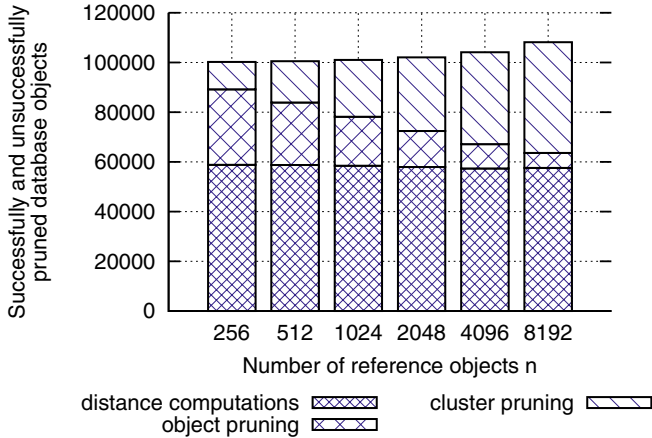


Fig. 9 Successfully and unsuccessfully pruned database objects for IF($n,40$)

4.2 Comparing the Retrieval Performance of IF4MI with the M-Tree and PM-Tree

In order to compare IF4MI with alternative approaches we use the M-tree library from <http://mufin.fi.muni.cz/trac/mtree/> and acknowledge its contributors. This library already provides some improvements to the original M-tree [12] such as a PM-tree implementation [40], a multi-way insertion algorithm [39], and an improved split policy [44]. The node size of the M-tree is set to 4096 byte which is the default block size of many current file systems, since we assume a secondary memory based index which is in large parts stored on disk and thus disk I/O might become a cost factor that cannot be ignored in addition to the number of necessary distance computations.

In contrast to the M-tree, the PM-tree relies on two parameters n_{hr} and n_{pd} . Inner node entries of the PM-tree apply a set of n_{hr} reference objects in order to be able to trim the covering region of a subtree through n_{hr} hyper-rings. A hyper-ring is hereby described by a reference object c_i and a pair of minimum and maximum distances of the database objects in the subtree from c_i . Therefore, $\mathcal{O}(n_{hr})$ additional space has to be administered per node entry in the inner nodes of a PM-tree. The second parameter n_{pd} affects the representation of the leaf nodes. A set of n_{pd} reference objects is used in order to be able to apply pivot filtering. Hence, n_{pd} additional $d(c_i, o)$ distance values are stored per feature object in a leaf node entry.

Fig. 10 visualizes the fraction of necessary distance computations compared to a sequential scan for different M-tree and PM-tree versions (assuming a block size of 4096 byte) as well as for IF4MI. Different approaches use the same set of x reference objects for pivot filtering in order to make results more comparable. The PM-tree, denoted as PM(n_{hr}, n_{pd}), corresponds to an

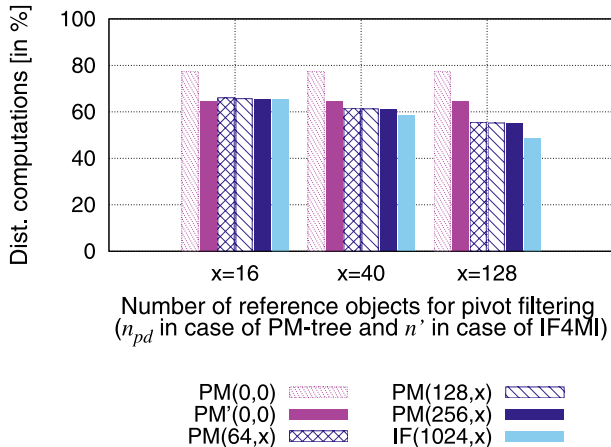


Fig. 10 Distance computations compared to a sequential scan for $\text{PM}(n_{hr},x)$ [block size set to 4096 byte] and $\text{IF}(n,x)$

M-tree if both $n_{pd} = 0$ and $n_{hr} = 0$. $\text{PM}'(0,0)$ refers to the M-tree using the multi-way insertion algorithm (cf. [39]).

We can observe the benefits of applying pivots since from Fig. 10 the differences in the number of distance computations between the M-tree approaches which do not apply pivots— $\text{PM}(0,0)$ and $\text{PM}'(0,0)$ —and the approaches that do so (PM-tree and IF4MI) are clearly noticeable in case of for example $x = 128$. Furthermore, under the current parameter setting, $\text{IF}(1024,x)$ with $x = 40$ and $x = 128$ is able to outperform the corresponding M-tree and PM-tree approaches when considering the necessary distance computations. It was not feasible to include results for e.g. $\text{PM}(1024,x)$ in Fig. 10 because of the dramatically increased memory requirements.

Fig. 11 visualizes the fraction of distance computations in relation to a sequential scan for different tree-based approaches when varying the node/block size of the trees. In case of $\text{PM}(128,128)$ ¹ and a block size of 2048 byte, there are many nodes with only a single entry since $128 \cdot 8$ byte (i.e. half of the block size) is already occupied for storing the hyper-rings. In Fig. 11, according to the number of necessary distance computations, larger block sizes than 2048 byte are more promising. It can be observed that when solely trying to reduce the total number of distance computations an intermediate block size performs best amongst the measured block sizes for the M-tree approaches

¹ We included measurements for $\text{PM}(128,128)$ since measurements for $\text{PM}(256,128)$ and a block size of 2048 byte failed because of increased memory requirements. Nevertheless, for block sizes of 4096 byte and larger, these measurements as well as those of the other PM-tree settings displayed in Fig. 10 show a similar trend with less distance computations the higher the block size.

(16384 byte for PM(0,0) and 8192 byte for PM'(0,0)). Thus, from this perspective, there seems to be an adequate clustering, i.e. assignment to subtrees, where feature objects are grouped under a common node entry which can successfully be pruned during search. With larger block sizes, a high branching factor might lead to a less adequate clustering and thus an increase in the number of necessary distance computations (cf. 32768 or 65536 for PM'(0,0) in Fig. 11). When applying a PM-tree, i.e. PM(128,128), the number of distance computations does not increase again in case of larger block sizes. Pivot filtering applied at the leaf level seems capable of still pruning many feature objects from search. A similar effect has been observed in Fig. 6 for IF4MI (e.g. IF([1024],1024)). Also the hyper-rings maintained in the inner nodes might lead to a more effective pruning of sub-trees.

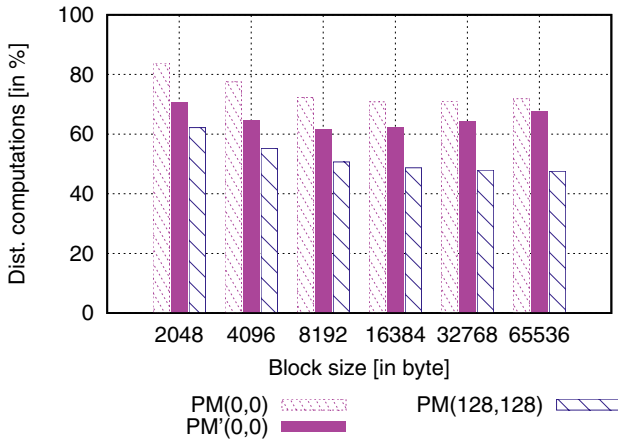


Fig. 11 Influence of different block sizes on the number of distance computations in relation to a sequential scan for the tree-based approaches

It might be argued that especially in case of a main memory-based indexing scenario, a PM-tree approach with a big enough block size and an adequate number of feature objects used for pivot filtering (e.g. $n_{pd} = 128$) might be used instead of IF4MI as well, since the performance of IF4MI according to the number of distance computations can be achieved (cf. Fig. 8). In this context, it is however important to notice that also an approach purely based on pivot filtering such as e.g. LAESA (cf. Sect. 2.1) might fit the needs when the only goal is to reduce the number of distance computations and other costs such as disk I/O are ignored. Here, the memory requirements of the different approaches become important. They are analyzed in the following Sect. 4.3.

4.3 Memory Requirements of the Different Approaches

IF4MI is currently built as a disk-based Lucene index which can be switched toward a main-memory index. The space complexity of IF4MI consists of several parts. $n' \cdot |O|$ precomputed distance values are stored in the posting lists of the inverted index since only n' instead of n distance values are stored per indexed database object. Besides this, $\mathcal{O}(2n^2)$ additional space is used for storing *MIN* and *MAX*. Furthermore, a negligible small amount of directory information is maintained in main memory. Finally, the database objects are stored in a separate field of the index (e.g. approx. $\omega = 0.5$ KB per database object in case of FC₄). The total memory requirements of IF4MI can thus be determined by formula 2, where α represents the storage space needed for a single distance value, e.g. 4 bytes when no binning technique is applied. This leads for example to approx. 110 MB in case of IF(1024,128) which is also measured in Tab. 3.

$$\mathcal{S}_{\text{IF4MI}(n,n')} = \alpha \cdot (n' \cdot |O| + 2n^2) + \omega \cdot |O| \quad (2)$$

Table 3 Memory requirements of IF4MI as measured during the experiments

memory requirements	
IF(1024,16)	65.1 MB
IF(1024,40)	74.8 MB
IF(1024,128)	110.0 MB

The memory requirements of IF4MI can easily be estimated in advance with the use of formula 2 as well as formula 3 when adapted to the space partitioning of the Metric Index (cf. Sect. 4.4). Thus, depending on the runtime complexity of the distance measure and the available main memory, adequate numbers of applied reference objects (i.e. parameter values for n and n') can be determined. It should also be noted that with an unoptimized implementation, query response time for the hardest feature combination FC₄ is on average 0.5 s in case of IF(1024,1024) (single CPU on Intel i7 860, 8MB cache, 2.8 GHz).

When estimating the memory requirements of the M-tree the following statistics are used:

There are $|O| = 100,000$ database objects to be indexed. The size of an inner node entry is considered to be 512 byte and a leaf node entry is estimated to occupy 508 byte (cf. Tab. 4). If a block size of 4096 byte is assumed, 8 entries fit in inner as well as leaf nodes. Of course, not all of the nodes are fully occupied. In the following calculations, an average load factor of $\ln(2) \approx 69.3\%$ is assumed (cf. [48]). This results—on average—in 5.54 and

Table 4 Memory requirements for M-tree node entries

<i>internal node entry</i>		<i>leaf node entry</i>	
routing object	500 byte	routing object	500 byte
reference to subtree	4 byte	distance to parent	4 byte
radius	4 byte	object identifier	4 byte
distance to parent	4 byte		
sum:	512 byte	sum:	508 byte

5.59 node entries for inner and leaf nodes respectively². Using these numbers, the outcome is an estimated M-tree index size of 89.4 MB. So, more memory is needed in comparison with IF(1024,40) (cf. Tab. 3) while at the same time leading to more distance computations (cf. Fig. 10).

Of course, the PM-tree approaches need more memory compared to the M-tree for representing the hyper-rings in the inner nodes and storing the object-pivot distances in the leaf nodes. Memory requirements considerably increase with increasing n_{hr} since n_{hr} additional distance pairs have to be administered for every entry of a PM-tree inner node.

In large-scale scenarios with potentially millions of database objects where it is inevitable to store parts of the index on secondary memory, there might be the need for adapting the node size of the tree structures to the physical block size of the underlying file system in order to reduce disk I/O costs. As shown before, for a typical block size of 4096 byte, when analyzing only the number of necessary distance computations, IF4MI can outperform the PM-tree. Furthermore, the storage space required by the reference objects within the nodes of a tree-based index structure might become a serious problem. When applying MAMs, no assumption is made about the representation of the database objects, thus they might become arbitrary complex according to memory requirements. Of course, references could be used instead of the objects themselves. However, this would require additional disk accesses.

4.4 Using the Space Partitioning of the Metric Index

Table 5 compares different space partitioning approaches. $MI_{\text{mod}}(m,l)$ hereby denotes the variant of IF4MI where the space partitioning of the Metric Index is applied. In contrast to the original Metric Index, $MI_{\text{mod}}(m,l)$ is built upon the inverted file library and does not rely on the mechanism for mapping feature objects to one dimensional values for storing them in a B^+ -tree. $MI_{\text{mod}}(m,l)$ can on the lowest cluster level theoretically maintain up to m^l clusters. Each cluster on level l is thus mapped to a single posting list, and

² We believe that this is an optimistic estimate which leads to an underestimation of the true M-tree index size. As an indicator, the M-tree library offers statistics which show that the average number of leaf node entries is 4.94 for PM(0,0) and 3.24 for PM'(0,0).

thus an object reference is only stored in a single posting list. In the posting lists, the $d(c_i, o)$ values used for pivot filtering are stored as before.

The parameters of MI_{mod} are set to $m = 40$ and $l = 3$. Experiments in [32] based on the CoPhIR dataset show that a static structure with a fix value of $l = 3$ already achieves good retrieval performance and even a dynamic tree structure evolving by limiting the number of feature objects per cluster cell can only slightly improve retrieval results according to the number of necessary distance computations. In order to compare $MI_{\text{mod}}(40,3)$ with IF4MI both approaches use the same set of pivots for pivot filtering and thus $n' = m = 40$. The value of n is set to 347—which is the maximum possible value so that the memory size of $IF(n,40)$ remains below the corresponding memory requirements of $MI_{\text{mod}}(40,3)$. Results according to the number of necessary distance computations are similar (cf. Tab. 5) which was expected since the same set of reference objects is used for pivot filtering and the same pruning constraints are applied. Retrieval performance of the static Metric Index is thus possible for IF4MI and IF4MI can be extended to follow the space partitioning of the Metric Index.

Table 5 Comparison of IF4MI and Metric Index based space partitioning

	distance computations	memory requirements
$MI_{\text{mod}}(40,3)$	58,828.3	66.50 MB
$IF(347,40)$	58,822.0	66.49 MB

To show that the similar performance of $IF(347,40)$ and $MI_{\text{mod}}(40,3)$ according to the number of necessary distance computations is not completely due to the pruning power of pivot filtering, Fig. 12 visualizes the number of database objects which are pruned by cluster pruning. It can be observed that $MI_{\text{mod}}(40,3)$ offers a slightly better pruning power according to cluster pruning. However, $IF(347,40)$ can compensate this by excluding more database objects through pivot filtering. The remaining numbers of necessary distance computations in Fig. 12 correspond to the numbers displayed in Tab. 5. The height of the $MI_{\text{mod}}(40,3)$ bar in Fig. 12 is $347 - 40 = 307$ units smaller than the height of the $IF(347,40)$ bar, since for every query $IF(347,40)$ requires 347 distance computations to compute the list L (cf. algorithm in Fig. 5) whereas $MI_{\text{mod}}(40,3)$ needs only 40 distance computations to determine the cluster where the query lies in.

The memory requirements of $MI_{\text{mod}}(m, l)$ can also be determined with the help of formula 2 on page 25. Since $n' = m$ and since the number of clusters resulting from the l -permutations is m^l , the memory requirements of $MI_{\text{mod}}(m, l)$ are:

$$\mathcal{S}_{MI_{\text{mod}}(m,l)} = \alpha \cdot (m \cdot |O| + 2 \cdot m^l) + \omega \cdot |O| \quad (3)$$

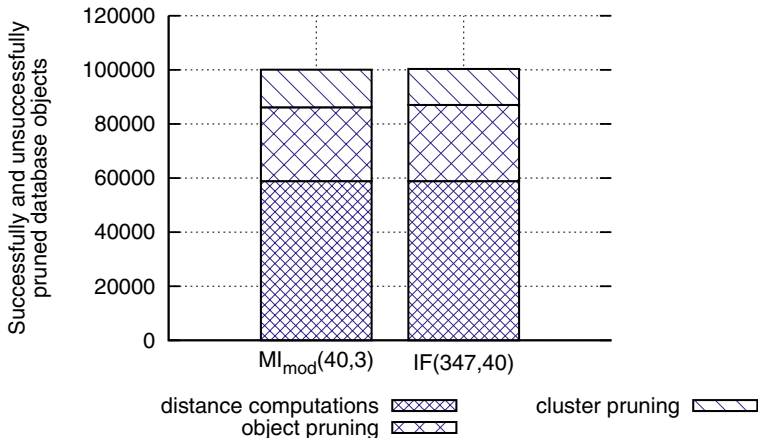


Fig. 12 Successfully and unsuccessfully pruned database objects for IF(347,40) and MI_{mod}(40,3)

4.5 Improvements to Pivot Filtering

Pivot filtering (cf. formula 1 on page 15) is applied by different MAMs such as the Metric Index, the PM-tree, and IF4MI. There are analyses of how to best choose pivots c_i (cf. e.g. [7, 9, 49]). In this section, it is analyzed how to best evaluate the condition outlined in formula 1. This is of special importance to IF4MI since the number of reference objects used for pivot filtering might be large. In this section, four different strategies are analyzed for determining the order in which the centers c_i should be processed to break as soon as the object-pivot distance constraint used in formula 1 is fulfilled for the first time.

Random: In this case no specialized strategy for determining the best order of how to evaluate different centers c_i is applied. Thus, the centers c_i are processed from $i = 1, 2, \dots, n$ which corresponds to a random ordering since centers are initially chosen at random. This random ordering represents the baseline technique against which the three alternative approaches presented in the following are compared.

L_q Reverse and L_q Order: When performing pivot filtering, lower bound distances $\check{d}_i = |d(c_i, q) - d(c_i, o)|$ of the true distance $d(q, o)$ are computed with the help of reference objects c_i in order to determine if the possibly expensive computation of $d(q, o)$ can be avoided. The consideration of the cluster centers c_i in the above formula can be stopped as soon as $\check{d}_i > r$ is fulfilled for the first time with r denoting the current search radius (cf. formula 1). Thus, it is desirable to first check centers c_i with large lower bound distances \check{d}_i . An analysis of the formula $\check{d}_i = |d(c_i, q) - d(c_i, o)|$ shows that the resulting

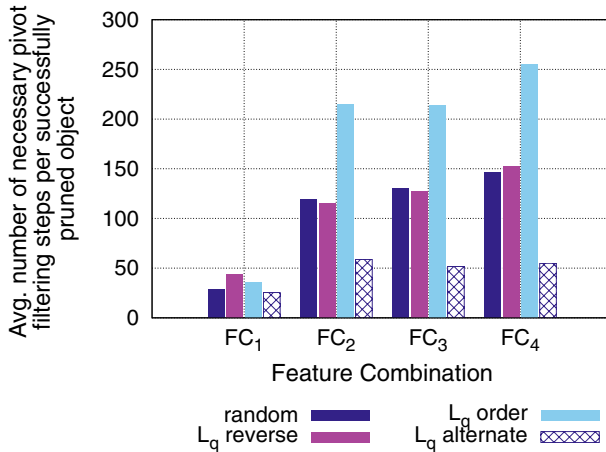


Fig. 13 Necessary pivot filtering steps per successfully pruned database object for different pivot filtering heuristics in case of IF(1024,1024)

absolute value can be high either because $d(c_i, q)$ is high and $d(c_i, o)$ is low or because $d(c_i, q)$ is low and $d(c_i, o)$ is high. As a consequence, especially medium values for $d(c_i, q)$ are candidates with a limited potential for selective lower bounds. Since the list L_q (denoted as L in the algorithm presented in Fig. 5 on page 16) is anyway computed at the beginning of the query process, L_q can directly be applied exploiting this observation. L_q corresponds to an ordering of the centers c_i by decreasing distance $d(c_i, q)$ —in opposition to the increasing ordering denoted as L_q order typical for many scenarios (cf. e.g. [16] or the algorithm in Fig. 5). Results in Fig. 13 indicate that both strategies cannot outperform the processing of the reference objects in random order. Thus, a hybrid combination is investigated, too.

L_q Alternate: Here, the two approaches L_q reverse and L_q order are combined. Therefore, the centers are chosen from L_q which is anyway computed during query processing in an alternate fashion. First, the center at position $n - 1$ of the list L_q is selected, then it is proceeded with the center at position 0, center at position $n - 2$, center at position 1, and so on. Consequences for query processing can be observed from Fig. 13. This approach clearly outperforms the other techniques. In case of FC₄, if a database object can be successfully pruned from query processing by applying pivot filtering, approximately 55 lower bound distances are computed on average compared to 146 without optimization. Since on average 42,198.1 database objects are pruned for this particular setting (cf. object pruning for FC₄ and IF(1024,1024) in Fig. 6), the number of lower bound distance computations can—on average—be reduced per query from approx. 6.2 to 2.3 million.

Instead of using L_q and thus $d(c_i, q)$, the decision of choosing reference objects for pivot filtering could also be based on L_o , i.e. an ordering of all $c_i \in C$ according to $d(c_i, o)$. L_o is computed during the insertion of a database object o but not explicitly present in the current index structure at query time. Cluster indexes i could be stored in addition to now sorted $d(c_i, o)$ values (either in L_o order or reversely) in order to be able to compute the lower bound distances without having to recompute L_o at query time. Alternatively, with the initial design of an individual posting as displayed in Fig. 4, the list of $d(c_i, o)$ values could be sorted at query time in order to identify the largest value, the smallest, the second largest, the second smallest, etc. While the first alternative would increase the size of the index, the second would negatively influence runtime performance. Furthermore, initial measurements indicate that an alternate strategy based on L_o cannot outperform L_q *alternate*.

4.6 Processing Filter Queries

Recent studies in CBIR such as [33] indicate that query processing purely based on content-based techniques is not sufficient for effective retrieval. Textual information provides an important additional search criterion which is frequently applied. Not only in the domain of CBIR, it is necessary to integrate various search criteria and filter searches. The same applies for web search engines where for example file type, language, or date are recognized as important filter criteria.

In this section, it is shown how IF4MI can benefit from the extensive knowledge in the field of inverted files. The skip pointers are used in order to make query processing more efficient since they prevent the unnecessary reading of posting list entries. Textual filter conditions are performed in combination with content-based k -NN queries. Therefore, also the tags of the CoPhIR collection are indexed. Of course, instead of applying tags as filters, various other aspects (as depicted in Sect. 1) are also possible—especially in the context of quality-aware IR.

We analyze a random run with 200 queries. The queries are constructed as follows: We use randomly chosen images as described in Sect. 4 and define the tag with the highest document frequency among all tags associated with the image as our filtering query term. Among the 200 randomly chosen query images only 138 are tagged at all, leaving 138 test queries.

IF4MI based query processing considers the inverted list representing image content properties only, if the document frequency of the filter term is higher than the number of clusters n . Otherwise, the number of distance calculations $d(c_i, q)$ in order to compute the list L (cf. the algorithm outlined in Fig. 5) already exceeds the number of necessary distance computations when calculating the object-query distances $d(q, o)$ directly for all objects fulfilling the textual filter criterion. As shown in the gray shaded row of Tab. 6, only for 26 query images the document frequency of the most frequent tag is higher

than 1024, only for 70 query images the document frequency of the most frequent tag is higher than 512, and only for 88 query images the document frequency of the most frequent tag is higher than 256.

Table 6 (white background) shows for FC_4 how many postings can be skipped when performing 20-NN queries and additionally applying a textual query filter. The default parameter values of Lucene such as a skip interval of 16 are directly applied without optimization. We can see that with decreasing n the number of skipped postings increases. This is especially beneficial in case of a main-memory index when not disk access but the reading and decoding of posting lists becomes the dominating cost factor [3]. It should be noted that e.g. for $x = 512$ an average value of approx. 25,000 skips does not mean that approx. 75,000 database objects are accessed since cluster pruning is applied and posting lists are excluded from query processing whenever possible. In order to benefit from skipping—under the assumption that distance computations are the dominant cost factor—one might also think of a simultaneous maintenance of multiple indexes with different values of n such as 1024, 512, 256, etc.

Table 6 Skipped postings for $IF(n,*)$ in case of textual filter queries

	$n=1024$	$n=512$	$n=256$
#queries	26	70	88
average	9,765.9	24,855.2	35,670.0
minimum	2,494	5,017	9,779
maximum	19,906	67,655	72,680
median	7,631	20,046	32,775
1st quartile	6,212.0	13,514.3	23,623.0
3rd quartile	15,401	28,877.5	43,896

5 Conclusion and Outlook

In this chapter $IF4MI$ has been presented—an exact MAM based on inverted files. To our best knowledge it is the first metric access method which brings the accuracy of existing MAMs to inverted files allowing for exact k -NN and range queries. In addition, it has been shown that $IF4MI$ can outperform an M-tree and a PM-tree implementation under typical parameter settings according to the number of necessary distance computations. Moreover, $IF4MI$ can compete with the Metric Index without relying on an underlying B^+ -tree or a similar data structure. $IF4MI$ can benefit from the extensive knowledge in the field of inverted files and e.g. inherently provides a multi-feature MAM which opens doors for a wide range of quality-aware IR applications.

It is our goal to build a flexible, scalable retrieval framework based on $IF4MI$ and thus inverted files supporting multiple search criteria (text, media content, date & time, geographic data, quality aspects), different data types (quality profiles, multimedia objects, plain and structured text), and various query modes (similarity search, faceted search, etc.). In this regard,

it will be interesting to analyze how IF4MI can integrate with powerful software libraries such as Elasticsearch (<http://www.elasticsearch.org/>) and Katta (<http://katta.sourceforge.net/>) according to distributed large-scale query processing, Apache Solr (<http://lucene.apache.org/solr/>) according to faceted search capabilities, etc.

References

1. Bayer, R., McCreight, E.: Organization and maintenance of large ordered indexes. *Acta Informatica* 1(3), 173–189 (1972)
2. Bille, P.: A survey on tree edit distance and related problems. *Theor. Comput. Sci.* 337(1-3), 217–239 (2005)
3. Boldi, P., Vigna, S.: Compressed perfect embedded skip lists for quick inverted-index lookups. In: Consens, M.P., Navarro, G. (eds.) *SPIRE 2005*. LNCS, vol. 3772, pp. 25–28. Springer, Heidelberg (2005)
4. Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Piccioli, T., Rabbitti, F.: CoPhIR: a Test Collection for Content-Based Image Retrieval. *CoRR*, abs/0905.4627v2 (2009), <http://arxiv.org/abs/0905.4627v2> (last visit: September 12, 2011)
5. Brin, S.: Near Neighbor Search in Large Metric Spaces. In: *Proc. of the 21st Intl. Conf. on Very Large Data Bases*, pp. 574–584. Morgan Kaufmann, Zurich (1995)
6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: *Proceedings of the 7th Intl. World Wide Web Conf.*, pp. 107–117. Elsevier Science Publishers, Amsterdam (1998)
7. Brisaboa, N., Pedreira, O., Seco, D., Solar, R., Uribe, R.: Clustering-Based Similarity Search in Metric Spaces with Sparse Spatial Centers. In: Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., Bieliková, M. (eds.) *SOFSEM 2008*. LNCS, vol. 4910, pp. 186–197. Springer, Heidelberg (2008)
8. Bustos, B., Keim, D., Saupe, D., Schreck, T.: Content-based 3d object retrieval. *IEEE Comput. Graph. Appl.* 27(4), 22–27 (2007)
9. Bustos, B., Navarro, G., Chávez, E.: Pivot selection techniques for proximity searching in metric spaces. *Pattern Recogn. Lett.* 24, 2357–2366 (2003)
10. Chandrasekaran, K., Gauch, S., Lakkaraju, P., Luong, H.P.: Concept-based document recommendations for citeseer authors. In: Nejdil, W., Kay, J., Pu, P., Herder, E. (eds.) *AH 2008*. LNCS, vol. 5149, pp. 83–92. Springer, Heidelberg (2008)
11. Chávez, E., Navarro, G., Baeza-Yates, R.A., Marroquín, J.L.: Searching in Metric Spaces. *ACM Comput. Surv.* 33(3), 273–321 (2001)
12. Ciaccia, P., Patella, M., Zezula, P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In: *Proc. of the 23rd Intl. Conf. on Very Large Data Bases*, pp. 426–435. Morgan Kaufmann, Athens (1997)
13. Connor, R., Simeoni, F., Iakovos, M., Moss, R.: A bounded distance metric for comparing tree structure. *Inf. Syst.* 36(4), 748–764 (2011)
14. Croft, W.B., Metzler, D., Strohman, T.: *Search Engines – Information Retrieval in Practice*. Pearson, Upper Saddle River (2010)
15. Datta, R., Li, J., Wang, J.Z.: Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In: *Proc. of the Intl. Conf. on Image Processing*, pp. 105–108. IEEE, San Diego (2008)

16. Eisenhardt, M., Müller, W., Henrich, A., Blank, D., El Allali, S.: Clustering-based source selection for efficient image retrieval in peer-to-peer networks. In: Proc. of the 8th Intl. Symp. on Multimedia, pp. 823–830. IEEE, San Diego (2006)
17. Esuli, A.: MiPai: Using the PP-Index to Build an Efficient and Scalable Similarity Search System. In: Proc. of the 2nd Intl. Workshop on Similarity Search and Applications, pp. 146–148. IEEE, Washington, DC (2009)
18. Figueroa, K., Chavez, E., Navarro, G., Paredes, R.: Speeding up spatial approximation search in metric spaces. *J. Exp. Algorithmics* 14, 6:3.6–6:3.21 (2010)
19. Gennaro, C., Amato, G., Bolettieri, P., Savino, P.: An Approach to Content-Based Image Retrieval Based on the Lucene Search Engine Library. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 55–66. Springer, Heidelberg (2010)
20. Hetland, M.L.: The Basic Principles of Metric Indexing. In: Coello, C.A.C., Dehuri, S., Ghosh, S. (eds.) *Swarm Intelligence for Multi-objective Problems in Data Mining*. SCI, vol. 242, pp. 199–232. Springer, Heidelberg (2009)
21. Hu, X., Chiueh, T.C., Shin, K.G.: Large-scale malware indexing using function-call graphs. In: Proc. of the 16th ACM Conf. on Computer and Communications Security, pp. 611–620. ACM, New York (2009)
22. Jing, Y., Baluja, S.: Pagerank for product image search. In: Proc. of the 17th Intl. Conf. on World Wide Web, pp. 307–316. ACM, New York (2008)
23. Kunze, M., Weske, M.: Metric trees for efficient similarity search in large process model repositories. In: Muehlen, M.z., Su, J. (eds.) *BPM 2010 Workshops*. LNBP, vol. 66, pp. 535–546. Springer, Heidelberg (2011)
24. Lalmas, M.: XML retrieval. *Synthesis Lectures on Information Concepts, Retrieval and Services*. Morgan & Claypool Publishers (2009), <http://www.morganclaypool.com/doi/abs/10.2200/S00203ED1V01Y200907ICR007>
25. Lee, J.: A graph-based approach for modeling and indexing video data. In: Proc. of the 8th IEEE Intl. Symp. on Multimedia, Washington, DC, USA, pp. 348–355 (2006)
26. Mamou, J., Mass, Y., Shmueli-Scheuer, M., Sznajder, B.: A Unified Inverted Index for an Efficient Image and Text Retrieval. In: Proc. of the 32nd Intl. Conf. on Research and Development in Information Retrieval, pp. 814–815. ACM, New York
27. Manjunath, B.S., Salembier, P., Sikora, T. (eds.): *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley & Sons (2002)
28. Micó, M.L., Oncina, J., Vidal, E.: A new version of the Nearest-Neighbour Approximating and Eliminating Search Algorithm (AESAs) with linear preprocessing time and memory requirements. *Pattern Recogn. Lett.* 15, 9–17 (1994)
29. Müller, H., Squire, D.M., Müller, W., Pun, T.: Efficient access methods for content-based image retrieval with inverted files. In: *Multimedia Storage and Archiving Systems IV*, pp. 461–472 (1999)
30. Myrvold, W., Ruskey, F.: Ranking and unranking permutations in linear time. *Inf. Process. Lett.* 79(6), 281–284 (2001)
31. Novak, D., Batko, M.: Metric Index: An Efficient and Scalable Solution for Similarity Search. In: Proc. of the 2nd Intl. Workshop on Similarity Search and Applications, pp. 65–73. IEEE, Washington, DC (2009)
32. Novak, D., Batko, M., Zezula, P.: Metric index: An efficient and scalable solution for precise and approximate similarity search. *Inf. Syst.* 36, 721–733 (2011)

33. Lestari Paramita, M., Sanderson, M., Clough, P.: Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsirikia, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 45–59. Springer, Heidelberg (2010)
34. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)
35. Samet, H.: Foundations of Multidimensional and Metric Data Structures. Morgan Kaufmann Publishers Inc., San Francisco (2006)
36. Schindler, U., Diepenbroek, M.: Generic xml-based framework for metadata portals. *Comput. Geosci.* 34(12), 1947–1955 (2008)
37. Skopal, T.: Where are you heading, metric access methods?: a provocative survey. In: Proc. of the 3rd Intl. Conf. on Similarity Search and Applications, pp. 13–21. ACM, New York (2010)
38. Skopal, T., Bustos, B.: On nonmetric similarity search problems in complex domains. *ACM Computing Surveys* 43(4), 34:1–34:50 (2011)
39. Skopal, T., Pokorný, J., Krátký, M., Snášel, V.: Revisiting M-tree building principles. In: Kalinichenko, L.A., Manthey, R., Thalheim, B., Wloka, U. (eds.) ADBIS 2003. LNCS, vol. 2798, pp. 148–162. Springer, Heidelberg (2003)
40. Skopal, T., Pokorný, J., Snášel, V.: Nearest Neighbours Search Using the PM-Tree. In: Zhou, L.-Z., Ooi, B.-C., Meng, X. (eds.) DASFAA 2005. LNCS, vol. 3453, pp. 803–815. Springer, Heidelberg (2005)
41. Socorro, R., Micó, L., Oncina, J.: A fast pivot-based indexing algorithm for metric spaces. *Pattern Recogn. Lett.* 32(11), 1511–1516 (2011)
42. Sznajder, B., Mamou, J., Mass, Y., Shmueli-Scheuer, M.: Metric inverted-an efficient inverted indexing method for metric spaces. In: Proc. of the Efficiency Issues in Information Retrieval Workshop (2008), <http://irlab.dc.fi.udc.es/ecir/sznajder.pdf> (last visit: March 7, 2011)
43. Tellez, E.S., Chávez, E.: On Locality Sensitive Hashing in Metric Spaces. In: Proc. of the 3rd Intl. Conf. on Similarity Search and Applications, pp. 67–74. ACM, New York (2010)
44. Traina Jr., C., Traina, A.J.M., Seeger, B., Faloutsos, C.: Slim-trees: High performance metric trees minimizing overlap between nodes. In: Zaniolo, C., Grust, T., Scholl, M.H., Lockemann, P.C. (eds.) EDBT 2000. LNCS, vol. 1777, pp. 51–65. Springer, Heidelberg (2000)
45. Vidal, E.: New formulation and improvements of the nearest-neighbour approximating and eliminating search algorithm (AESAs). *Pattern Recogn. Lett.* 15, 1–7 (1994)
46. Vidal, R.: An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recogn. Lett.* 4, 145–157 (1986)
47. Wojna, A.: Center-Based Indexing in Vector and Metric Spaces. *Fundam. Inf.* 56, 285–310 (2002)
48. Yao, A.C.C.: On random 2-3 trees. *Acta Inf.* 9, 159–170 (1978)
49. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach. Springer New York, Inc., Secaucus (2005)
50. Zobel, J., Moffat, A.: Inverted files for text search engines. *ACM Comput. Surv.* 38(2) article 6 (2006)

Chapter 3

Performance Prediction for Quality Recommendations

Josephine Griffith, Colm O’Riordan, and Humphrey Sorensen

Abstract. Work in the area of collaborative filtering continues to predominantly focus on prediction accuracy as a measure of the quality of the systems. Other measures of quality of these systems have been explored but not to the same extent. The work described in this chapter considers quality from the perspective of performance prediction. Per user, the performance of a collaborative filtering system is predicted based on rules learned by a machine learning approach. The experiments outlined aim, using three different datasets, to firstly learn the rules for performance prediction and to secondly test the accuracy of the rules produced. Results show good performance prediction accuracy can be found for all three datasets. The work does not step too far from the idea of prediction accuracy as a measure of quality but it does consider prediction accuracy from a different perspective, that of predicting the performance of a collaborative filtering system, per user, in advance of recommendation.

1 Introduction

The original foundations of collaborative filtering came from the idea of “automating the word of mouth process” that commonly occurs within social networks, where people will seek recommendations from people with whom they share similar preferences [29].

Given a set of users, a set of items, and a set of ratings, collaborative filtering (CF) systems attempt to recommend items to users based on user ratings. Collaborative filtering systems traditionally make use of one type of information, that is,

Josephine Griffith · Colm O’Riordan

College of Engineering and Informatics, National University of Ireland, Galway, Ireland
e-mail: {josephine.griffith, colm.oriordan}@nuigalway.ie

Humphrey Sorensen

Dept. of Computer Science, University College Cork, Ireland
e-mail: h.sorensen@cs.ucc.ie

prior ratings, or implicit indications of ratings, that users have given to items. However, additional information, such as content and explicit social information, has also been used. To date, application domains have predominantly been concerned with recommending items for sale (e.g. music, movies, books, restaurants) and with small amounts of text such as Usenet articles, email messages and bookmarks to websites. The datasets within these domains have their own characteristics, but they can be predominantly distinguished by the fact that they are both large and sparse, i.e., in a typical domain, there are many users and many items but any given user rates only a small percentage of all the items in the dataset.

Similar to software quality in general, and information retrieval system quality specifically, measures of collaborative filtering quality typically are linked to a system’s *superiority* or *non inferiority* in meeting a set of requirements with respect to one or more metrics. These measures have important consequences in that the relative merits of systems and algorithms can be compared empirically and assurances of quality with respect to these metrics can be associated with systems and algorithms.

Whilst viewing quality in terms of predictive accuracy is the focus of the majority of collaborative filtering studies, a user-centric view of quality has also been considered where measures such as perceived accuracy, novelty, transparency, and trustworthiness are considered [25].

The work described in this chapter also aims to view quality from a perspective different to that of pure predictive accuracy. The motivation of the work comes from the fact that recommender systems are increasingly becoming more prevalent to the extent that, rather than being a tool useful in dealing with information overload, they are becoming another source of information overload. Users are often presented with too many recommendations, too frequently, and often without being given any control over the quality of the recommendations. We believe that good recommender systems can distinguish themselves by only supplying recommendations when there is an associated level of confidence that the recommendations are of a sufficient quality for a particular user.

We believe that collaborative filtering recommender systems have information at their disposal which can allow them to predict whether a user is likely to receive good recommendations or not. In the scenario proposed here, a collaborative filtering system could, prior to recommendation, predict the performance it is likely to achieve per user. If the predicted performance for a user is high, the system could produce recommendations. However, if the predicted performance is low, the system should, unless requested otherwise by a user, desist from making recommendations to that user. The idea is that a user is likely to have more confidence in a system which has indicated that its recommendations are poor, and the user may be motivated to return to the system and enter more information to the system in the hope of receiving better recommendations. This is in contrast to a system which will present recommendations to a user irrespective of how good these recommendations are and may cause users to lose confidence in the system when they are recommended items that they know they do not like.

Performance prediction has become an established field in the area of IR where techniques are used to predict the accuracy of the results that are likely to be achieved given some input query [35], [37]. This can be informative for users allowing them to put more or less credence in the results returned by a system. The information can also be used to modify the query in advance of retrieval so as to improve performance if possible. In the collaborative filtering domain, Bellogín et al. directly map the idea of a “clarity score” from IR to the collaborative filtering domain with the aim of improving the performance of the system [4].

The focus of the work described in this chapter is also performance prediction in a collaborative filtering domain. The aim is to predict, per user, the performance of a collaborative filtering system so that this information could be presented to a user in advance of recommendation. The performance prediction approach involves deriving statistical measures from the user rating information in the dataset and using a machine learning approach to learn general rules about the predictive accuracy of these derived measures.

The outline of the chapter is as follows: Section 2 presents an overview of the quality measures used in information retrieval and collaborative filtering in addition to an overview of the area of performance prediction in collaborative filtering and information retrieval. An overview of the three datasets used in the work described here is given in Section 3. Section 4 presents details of the performance prediction approach proposed and Section 5 outlines how the performance prediction approach is evaluated. Results are presented in Section 6 and conclusions are presented in Section 7.

2 Previous Work

The quality of the results returned from information retrieval and recommender systems is predominantly measured in terms of these system’s *superiority* or *non inferiority* in satisfying a set of requirements. This evaluation approach allows experiments to be repeated and allows the rigorous and scientific comparison of systems and their underlying algorithms.

Systems have information at their disposal which can be used to offer feedback to users on the predicted, or likely, quality (or accuracy) of the results that will be returned to them. Recent work in IR in the area of performance prediction does this by aiming to predict the accuracy of the results that are likely to be achieved given some input query [35], [37] and these ideas from IR have also been adapted to the collaborative filtering domain [4].

This section briefly overviews the main themes in information retrieval and collaborative filtering evaluation. Work in the area of performance prediction in both information retrieval and collaborative filtering is also presented.

2.1 Evaluating Information Retrieval Quality

In information retrieval, given a user query, a document collection and an IR system, the quality of the retrieval system is usually defined in terms of retrieving documents relevant to the query and not retrieving documents which are not relevant to the query. Systems are generally evaluated using metrics such as precision and recall [28].

In addition, some systems also consider the relative quality of the information returned, ranking results based on both predictive accuracy metrics and information quality measures. Information quality measures have included link analysis of hyperlinked documents, finding authorities and hubs and incorporating user behaviour [38], [18], [1].

2.2 Evaluating Collaborative Filtering Quality: Predictive Accuracy Focus

A number of studies have been carried out in order to compare, from a predictive accuracy perspective, the quality of the recommendations produced by various collaborative filtering algorithms. Intermediate steps within a particular approach can also be evaluated. In most work, metrics are used to measure the ability of the system to provide a recommendation on a given item (coverage) and to measure the correctness of the recommendations generated for a given item by the system (accuracy). Herlocker et al. provide a comprehensive review of many suitable evaluation metrics [15]. The most commonly used metrics include mean absolute error (MAE), normalised MAE (NMAE), and root mean square error (RMSE) which compare the exact rating value given to an item by a user to the exact value predicted for the item by the system. The closer the two values are, the lower the error. The MAE score is defined as:

$$\frac{\sum_{i=1}^N |(p_i - r_i)|}{N} \quad (1)$$

where for N test items, on which the system returns predictions, p_i is the predicted rating for item i from the collaborative filtering system and r_i is the actual rating given by a user to item i .

Other evaluation metrics which have been used include Pearson correlation, Spearman rank correlation, area underneath an ROC-4 and ROC-5 curve, half life utility metric and the mean average precision at the $top - N$ documents returned.

The difference between the systems evaluated by these metrics lies in the underlying algorithms and approaches used. One large body of collaborative filtering work investigates the difference in prediction accuracy when different weighting schemes per user are used. The weighting schemes typically try to model some underlying bias or feature of the dataset with many approaches weighting the rating values. For example, items that are rated frequently by many users are penalised by giving the items a lower weight [6], [36]; items with a high rating variance are

weighted higher and items with a low rating variance are weighted lower [14], [36]; using the idea of a *tf-idf* weighting scheme from information retrieval for row normalisation of the dataset [17] or as part of a probabilistic framework [32]; learning the optional weights to assign to items [7], [16], [27]; giving a higher weighting to user neighbours who have provided good recommendations in the past [23]; and giving a higher weight to items which are recommended more frequently [11].

More recent work in collaborative filtering has abandoned these approaches in favour of latent factor models, such as matrix factorization techniques. These techniques have shown better accuracy than previous techniques [20].

2.3 Evaluating Collaborative Filtering Quality: User-Centric Focus

McNee et al. claim that the focus on predictive accuracy as a measure of recommender system quality has been detrimental to the field [22]. They claim that the most accurate recommendations are not necessarily the most useful from a user perspective. Despite the very strong focus on algorithms and predictive accuracy, some work does focus on quality from a user's perspective. The metrics used are more subjective, relying as they do on a user's opinion of the system in addition to the user's opinion of the actual recommendations.

Swearingen and Sinha evaluated the quality of six recommender systems from a user perspective but the user study was too small (with only 19 users) to draw any conclusive results [30]. A larger study with 210 users and seven recommender systems was carried out by Cremonesi et al. [8]. Each recommender system differed in the algorithm they used but each had the same user interface and the same dataset in an effort to reduce the number of factors that might influence results. The focus of the work was to compare the user's perceived quality of the results (in terms of accuracy, novelty and overall satisfaction) against the evaluated quality of the results using standard accuracy metrics. Results show that the user's perceived accuracy of the recommendations did not correlate with the evaluated accuracy of the recommendations.

Other user-centric studies have focused on including explanation interfaces [31]; generating a more diverse list of recommendations [39]; motivating users to rate items [3] and motivating users to state preferences [21].

Pu and Chen develop an overall model to allow for user evaluation of recommender system quality so that different studies can use a common set of measures. The model can thus facilitate meaningful comparisons between studies [25]. The model contains eight different categories of measures: perceived quality of recommended items; interaction adequacy; interface adequacy; perceived ease of use; perceived usefulness; control/transparency; attitude; and behavioural intentions. The perceived quality of the recommended items category consists of measures such as perceived accuracy, familiarity, novelty, attractiveness, enjoyableness, diversity and content compatibility. Each measure is represented by a questionnaire statement.

Users in a study can indicate their answers to each of the questions using a 1-5 Likert scale (1 indicating “strongly disagree” and 5 indicating “strongly agree”).

2.4 Performance Prediction in Information Retrieval

Query performance prediction in information retrieval aims to predict the effectiveness of a given query with respect to a retrieval system and a document collection [35], [37]. If query performance can be predicted in advance of, or during, retrieval, then retrieval results may be improved for specific types of queries. Two categories of performance prediction algorithms are studied: pre-retrieval and post-retrieval. The aim of pre-retrieval performance prediction is to estimate the performance of a query before any documents are retrieved. In comparison, post-retrieval performance prediction uses the ranked list of documents, or performance scores, returned from the search system to predict performance.

Examples of post-retrieval approaches are those proposed by Cronen-Townsend et al. using a *clarity score* [9], those proposed by Amati et al. using a measure of *query difficulty* to predict query performance [2] and approaches using the distributions of scores in the ranked list of documents returned by the system [24], [10].

He and Ounis present a pre-retrieval approach using a list of statistical values which can be derived from the query prior to retrieval [13]. Pre-retrieval approaches are generally less computationally expensive than post-retrieval approaches but as they are not using the returned information from the search system they are also generally less accurate than post-retrieval approaches.

2.5 Performance Prediction in Collaborative Filtering

In earlier, related work, the idea proposed is that a collaborative filtering system has the information available to provide evidence as to whether the recommendations produced by the system are likely to be weakly or strongly supported [12]. A user can thus be provided with the information with which to judge the quality of the recommendations which have been produced. The information used to obtain this evidence is already available in the collaborative filtering dataset and some of the information is calculated as part of the recommendation process. In that work, six features were extracted from the collaborative filtering dataset and analysed with respect to their effect on recommendation accuracy [12]. Each feature value above a set threshold was used to provide one piece of “evidence” with respect to the likely accuracy of the system in producing recommendations. Thresholds were chosen based on the analysis of the effect of the features on prediction accuracy. The more positive or negative pieces of evidence that exist for a given user, the more likely that the recommendation results will be accurate (for positive evidence) or inaccurate (for negative evidence). Results showed that a large percentage of users were correctly identified as having weak or strong evidence. The results suggested that it was worthwhile to investigate these and other features (and their effects) in more detail.

Bellogín et al. apply techniques from the area of performance prediction in information retrieval to the collaborative filtering domain [4]. They modify the clarity score used in performance prediction in IR to define performance predictors for recommender systems. The clarity score can be viewed as the difference between a user model and the background model. The idea is that a user model which is close to the background model (i.e. similar to the background model) is a sign of ambiguity as it is too similar to the background model to be distinguished from it.

In further experiments, Bellogín et al. use the clarity score to dynamically weight neighbour’s contributions based on the prediction of the neighbour’s performance using the clarity score [5]. A CF system using this dynamic weighting is compared to a standard CF approach without dynamic neighbour weighting. Results showed improved accuracy when using small neighbourhood sizes and comparable accuracy with larger neighbourhood sizes.

3 Datasets

Three datasets are used to test the proposed performance prediction approach: MovieLens [14], Bookcrossing [39] and Lastfm [19]. These datasets were chosen as they vary sufficiently from each other in terms of basic characteristics such as number of users, number of items and sparsity. The Lastfm dataset differs from the other two datasets in that the number of times a user listens to a music track (the playcount) is stored rather than a discrete value for an item which would explicitly indicate preference. A brief summary of the datasets is outlined in Table 1.

Table 1 Comparison of datasets

	MovieLens	Bookcrossing	Lastfm
Domain	Movies	Books	Music
Num. Users	943	77805	3080
Num. Items	1682	185968	30520
%Sparsity	87.66%	99.9%	99.1%
Value Range	1-5	1-10	1 to 7939

The distribution of the ratings for MovieLens, the playcounts for Lastfm, and the ratings for Bookcrossing are shown in Figures 1, 2 and 3 respectively. The MovieLens and Bookcrossing dataset distributions show a similar trend with a large proportion of the ratings associated with higher (“liked”) values. This indicates that users were more likely to rate positively than negatively. In comparison, the Lastfm dataset has a very long tail distribution due to the playcount being stored rather than a discrete rating value being stored. The majority of playcounts are in the range [1 – 100], specifically, approximately 94% of playcounts are in the range [1 – 90] and approximately 86% of playcounts are in the range [1 – 50].

Our intuition was that it did not make sense to work with the raw playcount data given this long tail distribution. In a collaborative filtering scenario it would mean

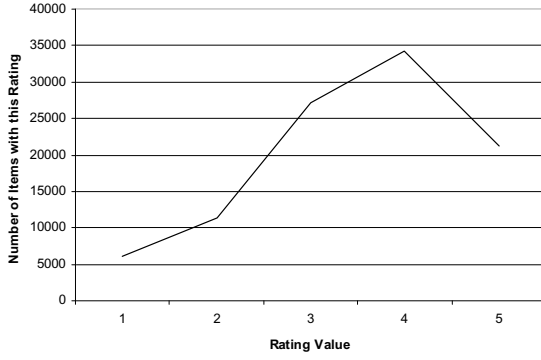


Fig. 1 Distribution of MovieLens ratings

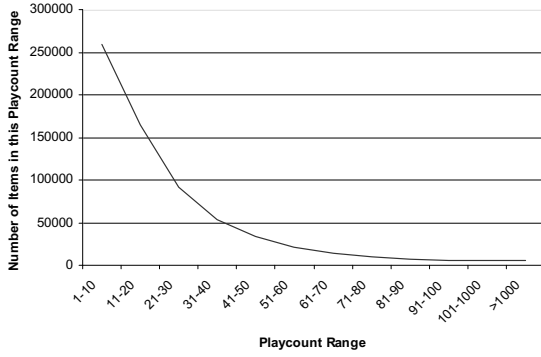


Fig. 2 Distribution of Lastfm playcounts

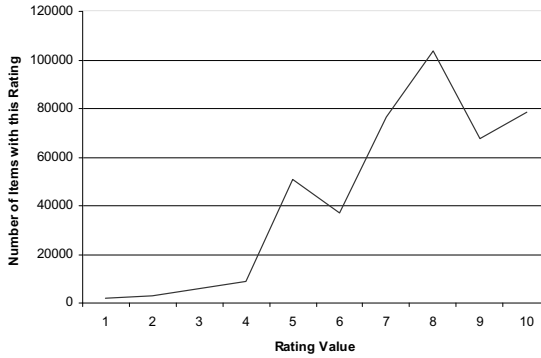


Fig. 3 Distribution of Bookcrossing ratings

that the relatively few ratings in the long tail would bias the results for the majority of users. Various normalisations were considered for the Lastfm dataset with the motivation to map the playcount values to ratings in a set range where the majority of playcounts would be distinguished more clearly from each other. After some experimentation, the log normalisation shown in Equation 2 was used.

$$ratingvalue = \log_{10}(1 + playcount) \quad (2)$$

With this normalisation, values are in the range $[0.3 - 3.89]$. These real values are then mapped to discrete values in the range $[1 - 12]$ based on 12 “buckets”, where the first 11 buckets contain values in steps of 0.3, e.g. playcounts in the range ≥ 0.3 and < 0.6 are mapped to 1, playcounts in the range ≥ 0.6 and < 0.9 are mapped to 2, etc. The final bucket (with value 12) contains playcount values ≥ 3.6 . Figure 4 shows the new distribution of the dataset using this normalisation and mapping. The long tail of the distribution still exists but, up to when the long tail begins at values 5 and 6, the ratings show a distribution more similar to the Movielens and Bookcrossing distributions.

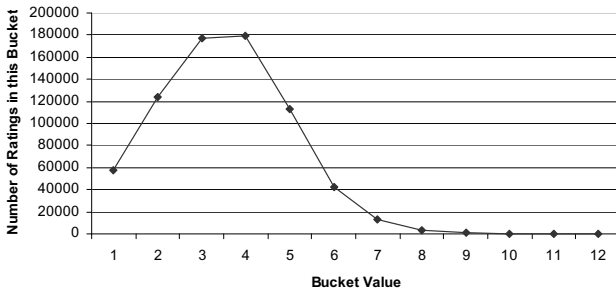


Fig. 4 Distribution of log normalised Lastfm playcounts

Table 2 Average MAEs for each dataset

	Movielens	Lastfm	Bookcrossing
Avg. MAE	0.728	0.629	5.89

Table 2 lists the average MAEs for each dataset. A standard collaborative filtering test approach is used to find these MAE values where a portion of test users and a portion of their items are removed. A Pearson correlation nearest neighbourhood approach is used to find neighbours and a weighted average of the neighbour’s ratings are used to generate recommendations for the removed test items. The predicted values are compared to the actual values in the removed test portion to find the MAE values. MAE results are averaged over 10 runs. Bookcrossing has a very high MAE in comparison to the MAEs for the other two datasets. This is a result of the high sparsity in the Bookcrossing dataset and has been shown in other studies also [33].

4 Performance Prediction Approach

The aim of the performance prediction work described in this chapter attempts to test the validity of a collaborative filtering performance prediction approach. It does not attempt to improve the prediction result by incorporating the performance prediction information in the collaborative filtering algorithm, as Bellogín et al. do [4]. In addition, the approach used is different to the approach used by Bellogín et al. The work described here does not directly map concepts such as the *clarity score* from IR to collaborative filtering. Instead statistical measures of the user rating information are derived from the dataset and a machine learning approach is used to learn general rules about the predictive accuracy of the derived statistical values.

Figure 5 outlines a general overview of the approach where, per user, derived measures of the user’s rating information, along with the rating information of other users, is used in a simple rule which returns a prediction on how well the system can produce recommendations for the user. The three different datasets outlined in Section 3 are tested with this scenario.

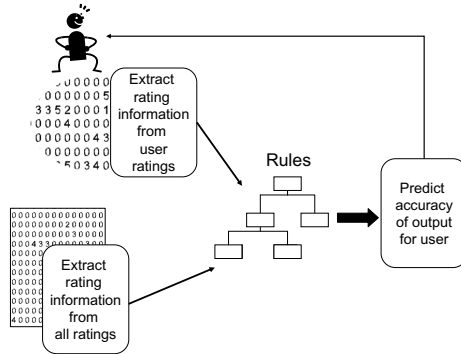


Fig. 5 Performance prediction scenario in a collaborative filtering domain

4.1 Learning the Performance Prediction Rules

Figure 6 gives an overview of the steps involved in learning, per dataset, some simple rules which can be used to predict, per user, the performance of the system.

Initially a holdout set of test users (up to 10% depending on the dataset) are removed to be used to evaluate the rules learned (as will be described in Section 5). The remainder of the dataset comprises the training data.

4.1.1 Extract Rating Information

Based on previous work, a number of aspects of the user rating information, called *features*, are extracted from the collaborative filtering datasets [12]. The motivation is to choose aspects of the user rating history that would seem likely to affect a user’s

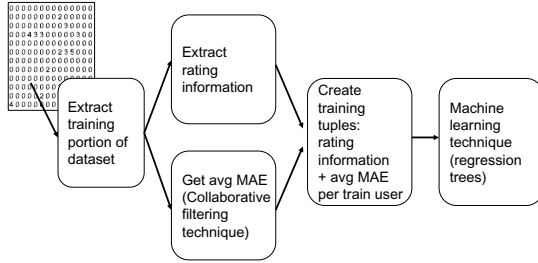


Fig. 6 Steps to learn the performance prediction rules

prediction accuracy. The data extracted range from simple calculations (such as the average user rating) to values derived using some item and neighbour information. The first set of features (features 1 to 6 and feature 10) are standard collaborative filtering measures, i.e. find similar users based on the ratings received by items. The remaining set of features tie together the relationship between users and the items they have rated: that is, to what degree a user “stands out” in the dataset in terms of the items they have rated.

A list of the 11 features follows, with details and formulas where required:

1. Number of ratings a user has given.
2. Average rating a user has given.
3. Standard deviation of the average rating a user has given.
4. Number of neighbours a user has. This is calculated by first using a Pearson correlation similarity measure to find the similarity between users. Any user with similarity to the current user above a set threshold (in this case 0.1) is counted as a neighbour.
5. Average similarity to the top closest 30 neighbours (using the Pearson correlation similarity values from feature 4 and, having ordered by similarity, picking the top 30 users).
6. The clustering coefficient of a user’s group, i.e., the level of connectedness between a user’s neighbours, defined for some user a as:

$$\frac{2 * actual + n}{n^2 - n} \tag{3}$$

where n is the number of neighbours of user a (as found in feature 4) and $actual$ is the number of user a ’s neighbours who are neighbours with each other (using the same similarity as in feature 4). This measures the cohesiveness of a neighbourhood.

7. The popularity of the items the user has rated, which is based on the number of ratings each item has received (and not considering the actual rating value). The formula per user a is:

$$\frac{\sum_{i=1}^M numratings_i}{M} \tag{4}$$

for M items rated by the user a and $numratings_i$ is the number of ratings item i has received from all users in the dataset.

8. How well-liked by all users are the items rated by the current user. This measure is calculated using the actual rating value given to items. The formula used per user a is:

$$\frac{\sum_{i=1}^M avgval_i}{M} \quad (5)$$

for M items rated by the user a and $avgval_i$ is the average rating value item i has received from all users in the dataset who have rated item i .

9. The importance, or influence, of a user in a dataset. This is based on the idea of term frequency and inverse document frequency from information retrieval and is the proportion of items the user has rated multiplied by how frequently-rated those items are in the dataset. Frequently-rated items get low values (similar to the IDF component in IR where frequently occurring terms across all documents receive lower scores). The formula used is:

$$\frac{numratings_a}{numitems} \times \sum_{i=1}^M \left(\log \frac{numusers}{numratings_i} \right) \quad (6)$$

for M items rated by the user a , where $numitems$ is the number of items in the dataset and $numratings_a$ is the number of ratings user a has given, i.e. this is the ratio of the number of ratings the user actually gave over the number of ratings the user could have given (all items); $numusers$ is the number of users in the dataset and $numratings_i$ is the number of ratings item i received, i.e., this is the ratio of the number of ratings an item could have received (a rating from all users in the dataset) over the number of ratings it actually received.

10. The average Jaccard co-efficient per user with all other users with whom they have co-rated items. The Jaccard coefficient is a measure of the similarity of two sets. In this case, for two users the sets, u_1 and u_2 , are the items rated by the two users, and the Jaccard coefficient is calculated by:

$$\frac{|u_1 \cap u_2|}{|u_1 \cup u_2|} \quad (7)$$

that is, the number of items co-rated by both users divided by the number of items rated by both users. For example, if two sample users have each rated the same items, and no other items, their Jaccard coefficient is 1.

11. The average item entropy of the items a user has rated. The formula used to calculate the entropy of an item x is:

$$- \sum_{i=1}^n \Pr(x_i) \log_2 \Pr(x_i) \quad (8)$$

where $\Pr(x_i)$ is the probability of x_i , for $i = 1 \dots n$ for the n possible values of the item, x .

The feature values are all normalised by min/max normalisation such that each value is in the range [0.0-1.0].

4.1.2 Collaborative Filtering Technique

For each user, in each of the three datasets, the feature values outlined in Section 4.1.1 are extracted. In addition, some score which represents how well a collaborative filtering system can predict items for these users is required for learning. The machine learning approach will learn over this score, ideally associating some feature values with low scores and other features values with high scores, and thus finding the predictive power of the features in terms of the accuracy score.

This experiment requires a measurement which is comparable across the three datasets, and which can be suitably averaged so that it can be used as a score over which the machine learning approach will learn. Initial experiments performed using the MAE (see Equation 1) found it was suitable as it is reasonably strict while being widely-used and well-understood.

In order to obtain an MAE score per training user, a collaborative filtering system is required to produce recommendations for a portion of ratings removed from the training user's ratings. Any standard collaborative filtering technique can be used. For this experiment, a nearest neighbour collaborative filtering technique was used using Pearson correlation to find similar neighbours and using a weighted average of the neighbour's ratings of test items to produce recommendations.

4.1.3 Create Training Tuples

For this experiment the percentage of test to training users for the collaborative filtering approach was kept constant at up to 10% test users per run. Using only this 10% would result in a small training file for the machine learning approach. The solution was to repeatedly re-run the collaborative filtering approach on an additional 10% of test users until most of the users had been picked from the dataset (the original holdout set of test users is not used). For each test user, 10% of their rated items is chosen as the test items and the remaining portion of the dataset is used to predict values for the removed ratings of the test items. The average MAE over these test items (comparing actual with predicted scores) is calculated for each test user. For any given user, their user ID along with their average MAE value and the 11 aforementioned features (from Section 4.1.1) comprise the user tuples in the training dataset.

4.1.4 Machine Learning Technique

All the data in this experiment is numeric. The target variable (the MAE) is known for each training tuple and therefore a supervised machine learning approach is suited to the problem. Often a neural network approach would be used in the classification scenario where labeled numeric data exists. However, we wish to understand the underlying patterns and correlations between the feature values and precision

scores. We therefore require a technique which will produce one or more rules. The technique used is regression trees which are similar to ordinary decision trees except they can be used with numeric data [34]. The regression tree used is the model tree inducer $M5'$ [26]. The machine learning package WEKA is used which has an implementation of $M5'$ [34].

The results of choosing attribute selection prior to running the $M5'$ approach was also tested. Attribute selection reduces the complexity of the rules, that is, the number of features used in the rules. As a result of attribute selection, the most predictive features with respect to the class (MAE) are chosen. This results in simpler rules. As the rules are to be used either prior to, or in conjunction with, producing recommendations, the quicker a performance prediction can be generated the better. Thus, simpler rules are better in this case.

5 Evaluation: Testing the Rules

To test whether the rules produced by the machine learning technique are predictive of system performance the holdout set of test users are used. Figure 7 gives an overview of the steps involved in the comparison of the *actual* and *predicted* performance for the holdout set of test users.

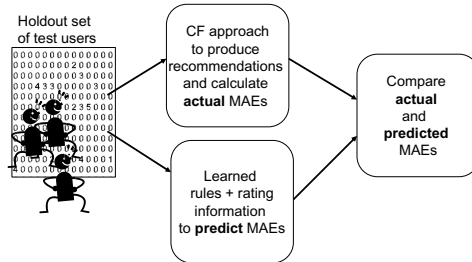


Fig. 7 Steps to test the performance prediction rules

Firstly a predicted MAE based on the feature values and the learned rules are produced for each user in the holdout set. These are the *predicted* MAE values with one prediction value per user. A collaborative filtering system is used to produce predictions for a set of items for the same users. An MAE score is calculated based on the ratings the user has given the items versus the ratings the collaborative filtering approach produced. The collaborative filtering run is repeated 10 times per user where for each run, up to 10% of the user’s items are randomly chosen as the test items. Finally accuracy scores per user are averaged over the 10 runs. This is the *actual* MAEs and again there is one (averaged) MAE value per user. Thus there are two lists per dataset where each list contains a user ID and an associated MAE score: the *actual* accuracy list, where the MAE scores are based on the average of 10 runs of the collaborative filtering system; and the *predicted* accuracy list, where

the MAE scores are based on the learned rules and feature values. Both lists, actual MAEs and predicted MAEs, are sorted by user ID and are compared to ascertain how accurate the performance prediction approach is. This comparison is done by:

1. finding the Pearson correlation between the two lists.
2. finding the MAE between the two lists.
3. dividing the actual MAE list into two sets, based on the average MAE ($avgmae$) for the dataset (as shown in Table 2). All values below this threshold (i.e., $< avgmae$) are considered as cases where the system returned *good* results. All values at or above the threshold are considered as cases where the system returned *poor* results. The idea is then to compare the level of agreement of the actual scores in both sets with the corresponding predicted scores for the same users. This approach does not penalise small variations in MAE (apart from small variations at the threshold value) and thus gives a less strict comparison of the actual and predicted lists in terms of the percentage of *good* results predicted correctly and the percentage of *poor* results predicted correctly.

6 Results

6.1 Movielens

For the Movielens dataset, using the training data and with no attribute selection chosen prior to learning (i.e. all 11 feature values are used) the rule found is:

$$mae = 1.15 \times stdev - 0.49 \times sim30neighs - 0.19 \times cluscoeff + 0.78 \quad (9)$$

where $stdev$ is the standard deviation of the user's ratings and $sim30neighs$ is the average similarity to the top closest 30 neighbours and $cluscoeff$ is the clustering coefficient of a user's group. The mean absolute error of the rule on the training data is 0.1528.

When choosing to first select the best features with attribute selection, the features found are $stdev$ and $sim30neighs$ and the rule found (with mean absolute error of 0.15) is:

$$mae = 1.17 * stdev - 0.47 * sim30neighs + 0.62 \quad (10)$$

There are 94 test users in the holdout test set for the Movielens dataset. Using these users, actual MAEs are found by producing recommendations for the user's test items using the collaborative filtering testing scenario. Predicted MAEs are found using the rule with two attributes in Equation 10. The two lists of predicted and actual MAEs are compared. The results from the three evaluation scenarios are:

- The Pearson correlation of the two lists is 0.819, thus showing a high correlation between the actual and predicted accuracy scores.
- The MAE of the two lists is 0.075 which shows a very low error between the actual and predicted accuracy scores.

- Taking the threshold between *poor* and *good* recommendations to be the best average found with the dataset, 0.728, the percentage of accuracy scores predicted correctly as *good* (with MAE scores < 0.728) is 85.9%. The percentage of accuracy scores predicted correctly as *poor* (≥ 0.728) is 94.5%. This indicates that for a high percentage of the test users, performance was predicted correctly as *good* or *poor*.

6.2 Lastfm

For the Lastfm dataset, with no attribute selection, the rule learned is long and unusable for the task of quickly producing a performance prediction score. When choosing attribute selection before learning the rules, 3 attributes are chosen: 1) standard deviation, 2) the number of ratings, and 3) how well-liked by all users are the items rated by the current user. The rule found consists of 3 sub-rules (with a mean absolute error of 0.105). There are 308 test users in the holdout test set for the Lastfm dataset. Using these withheld test users, actual MAEs are found using the collaborative filtering testing scenario and predicted MAEs are found using the rule with three attributes. The results of comparing the actual and predicted MAEs using the three evaluation scenarios are:

- The Pearson correlation of the two lists is 0.807, thus showing a high correlation between the actual and predicted accuracy scores.
- The MAE of the two lists is 0.067 which shows a very low error between the actual and predicted accuracy scores.
- Taking the threshold between *poor* and *good* recommendations to be the best average found with the dataset, 0.629, the percentage of accuracy scores correctly predicted as *good* is 88.7%. The percentage of accuracy scores correctly predicted as *poor* is 85.7%. Again for this dataset, performance can be predicted correctly as *good* or *poor* for a high percentage of the test users.

6.3 Bookcrossing

For the Bookcrossing dataset, with no attribute selection, the rule learned is again too long to be useful for the task. When choosing attribute selection before learning, four attributes are chosen: 1) the number of ratings, 2) the number of neighbours, 3) the popularity of items the user has rated, and 4) the average Jaccard coefficient of the items the user has rated. The rule found with these four attributes has quite a high error, with an MAE of 1.66. As a result we were not expecting the results to be as good as they were with the other two datasets. Using the withheld 139 Bookcrossing test users, and finding the actual and predicted MAE scores as before, the results from comparing the lists are:

- The Pearson correlation between the two lists of predicted and actual MAE scores is 0.768 which is a reasonably high correlation.

- The MAE of the two lists is 0.807 which is relatively high error, and thus shows inaccuracies between the two lists.
- Using the best average MAE found for the dataset (see Table 2), the percentage correctly identified as *good* (< 5.89) is 70% and the percentage correctly identified as *poor* is 79.7%.

6.4 Performance Prediction Scenario

Given the results outlined in the previous sections we can be confident of good performance prediction results in at least two of the datasets (Movielens and Lastfm). Although not tested in this work, the good results for the two datasets suggest that the following user scenario would be a viable approach to performance prediction:

1. Perform pre-processing steps per dataset to learn rules to extract the user feature values, to find the average MAE per train user and per dataset, and to create the training tuples.
2. Per user, use the learned rule, and feature values for that user, to produce a performance prediction score.
3. The performance prediction score can be returned to the user with an explanation of what it means (e.g., a lower MAE score is better) or a prediction of *good* or *poor* performance can be given. This can be produced by comparing the predicted MAE value to the average MAE value for the dataset. As in the evaluation scenario, if the predicted MAE is lower than the average MAE then a prediction of *good* can be returned. If the predicted MAE is equal to or higher than the average MAE then a prediction of *poor* can be returned.

7 Conclusions

The experiment described in this paper views an aspect of collaborative filtering quality in terms of a performance prediction approach.

The approach outlined extracts user rating information (feature values) that describe the user in the dataset. The user feature values are used, in conjunction with an MAE score, to learn rules. These rules, using some feature values, can predict the performance of the system per user. Three datasets were investigated: Movielens, Lastfm and Bookcrossing.

Given the differences in the dataset characteristics it is not surprising that there is no full agreement in terms of the features selected and the rules found for each dataset. However there are some very similar trends in the features selected and some of the same features are selected across datasets. Despite 11 features, some of which are relatively complex, it is some of the simpler features which are selected. For example, the number of ratings feature is chosen for the Lastfm and Bookcrossing datasets; the standard deviation feature is chosen for the Movielens and Lastfm datasets; a feature based on neighbours, *sim30neighs* and the number of neighbours, is chosen for the Movielens and Bookcrossing datasets; and a feature based on how

popular and well-liked the items a user has rated are, is chosen in the Lastfm and Bookcrossing datasets.

Results for both the Movielens and Lastfm datasets are very encouraging showing good performance across all evaluation methods. For the two strictest evaluations, high correlations of 0.819 and 0.807 respectively and low MAEs of 0.075 and 0.067 respectively were found. As expected, given the error associated with the Bookcrossing rule, the results for Bookcrossing were not as good as this.

Future work will consider additional datasets and a quicker performance prediction approach. Although it is acceptable to do some initial computationally expensive work per dataset, ideally the performance prediction per user must be very quick to warrant its use. Future work will also consider how stable the learned rules are when new ratings are added to the dataset.

References

1. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR, pp. 19–26 (2006)
2. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 127–137. Springer, Heidelberg (2004)
3. Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., Kraut, R.: Using social psychology to motivate contributions to online communities. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, pp. 212–221 (2004)
4. Bellogín, A., Castells, P.: Predicting neighbor goodness in collaborative filtering. In: Andreasen, T., Yager, R.R., Bulskov, H., Christiansen, H., Larsen, H.L. (eds.) FQAS 2009. LNCS, vol. 5822, pp. 605–616. Springer, Heidelberg (2009)
5. Bellogín, A., Castells, P.: A performance prediction approach to enhance collaborative filtering performance. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 382–393. Springer, Heidelberg (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 (1998)
7. Cheung, K., Tian, L.: Learning user similarity and rating style for collaborative recommendation. *Information Retrieval* 7, 395–410 (2004)
8. Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A., Turrin, R.: Comparative evaluation of recommender system quality. In: Proceedings of the 2011 International Conference on Human Factors in Computing Systems: Extended Abstracts, CHI EA 2011, pp. 1927–1932 (2011)
9. Cronen-Townsend, S., Zhou, Y., Croft, W.: Predicting query performance. In: 25th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR (2002)
10. Cummins, R., Jose, J., O’Riordan, C.: Improved query performance prediction using standard deviation. In: Proceedings of the 34th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR (2011)

11. De Bruyn, A., Lee Giles, C., Pennock, D.M.: Offering collaborative-like recommendations when data is sparse: The case of attraction-weighted information filtering. In: De Bra, P.M.E., Nejd, W. (eds.) AH 2004. LNCS, vol. 3137, pp. 393–396. Springer, Heidelberg (2004)
12. Griffith, J., O’Riordan, C., Sorensen, H.: Using user model information to support collaborative filtering recommendations. In: Proceedings of the 18th Irish Conference on Artificial Intelligence and Cognitive Science, pp. 71–80 (2007)
13. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) SPIRE 2004. LNCS, vol. 3246, pp. 43–54. Springer, Heidelberg (2004)
14. Herlocker, J., Konstan, J., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR, pp. 230–237 (1999)
15. Herlocker, J., Konstan, J., Terveen, L., Riedl, J.: Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22(1), 5–53 (2004)
16. Jin, R., Chai, J., Si, L.: An automatic weighting scheme for collaborative filtering. In: Proceedings of the 27th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR, pp. 337–344 (2004)
17. Karypis, G.: Evaluation of item-based top-n recommendation algorithms. In: Proceedings of the 10th International Conference on Information and Knowledge Management (2001)
18. Kleinberg, J.: Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5), 604–632 (1999)
19. Konstan, I., Stathopoulos, V., Jose, J.: On social networks and collaborative recommendation. In: Proceedings of the 32nd Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR, pp. 195–202 (2009)
20. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *IEEE Computer* 42(8), 30–37 (2009)
21. McNee, S., Lam, S., Konstan, J., Riedl, J.: Interfaces for eliciting new user preferences in recommender systems. In: Brusilovsky, P., Corbett, A.T., de Rosi, F. (eds.) UM 2003. LNCS, vol. 2702, pp. 178–187. Springer, Heidelberg (2003)
22. McNee, S., Riedl, J., Konstan, J.: Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: CHI 2006 Extended Abstracts on Human Factors in Computing Systems, pp. 1097–1101 (2006)
23. O’Donovan, J., Smyth, B.: Trust in recommender systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, pp. 167–174 (2005)
24. Pérez-Iglesias, J., Araujo, L.: Standard deviation as a query hardness estimator. In: Chavez, E., Lonardi, S. (eds.) SPIRE 2010. LNCS, vol. 6393, pp. 207–212. Springer, Heidelberg (2010)
25. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys 2011, pp. 157–164 (2011)
26. Quinlan, J.: Learning with continuous classes. In: *Artificial Intelligence*, pp. 343–348. World Scientific (1992)
27. Rashid, A., Karypis, G., Riedl, J.: Influence in ratings-based recommender systems: An algorithm-independent approach. In: SIAM International Conference on Data Mining (2005)
28. Robertson, S.: Evaluation in information retrieval. In: Agosti, M., Crestani, F., Pasi, G. (eds.) ESSIR 2000. LNCS, vol. 1980, pp. 81–92. Springer, Heidelberg (2001)

29. Shardanand, U., Maes, P.: Social information filtering: Algorithms for automating word of mouth. In: Proceedings of the Annual ACM SIGCHI on Human Factors in Computing Systems (CHI 1995), pp. 210–217 (1995)
30. Swearingen, K., Sinha, R.: Beyond algorithms: An HCI perspective on recommender systems. In: SIGIR 2001 Workshop on Recommender Systems (2001)
31. Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: IEEE 23rd International Conference on Data Engineering: Workshop, pp. 801–810 (2007)
32. Wang, J., de Vries, A., Reinders, M.: Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: Proceedings of the 29th International ACM Conference on Research and Development in Information Retrieval, SIGIR, pp. 501–508 (2006)
33. Weng, L.-T., Xu, Y., Li, Y., Nayak, R.: Improve recommendation quality with item taxonomic information. In: Filipe, J., Cordeiro, J. (eds.) ICEIS 2008. LNBIP, vol. 19, pp. 265–279. Springer, Heidelberg (2009)
34. Witten, I., Frank, E., Hall, M.: Data Mining Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann (2011)
35. Yom-Tov, E., Fine, S., Carmel, D., Darlow, A.: Learning to estimate query difficulty. In: 28th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR (2005)
36. Yu, K., Xu, X., Tao, J., Kri, M., Kriegel, H.: Feature weighting and instance selection for collaborative filtering: An information-theoretic approach. Knowledge and Information Systems 5(2) (2003)
37. Zhou, Y., Croft, B.: Ranking robustness: a novel framework to predict query performance. In: 15th ACM conference on Information and Knowledge Management, CIKM (2006)
38. Zhu, X., Gauch, S.: Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In: Proceedings of the 23rd Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR, pp. 288–295 (2000)
39. Ziegler, C., McNee, S., Konstan, J., Lausen, G.: Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web, WWW 2005, pp. 22–32 (2005)

Chapter 4

Automated Cleansing of POI Databases

Guy De Tré, Daan Van Britsom, Tom Matthé, and Antoon Bronselaer

Abstract. In the context of geographic information systems (GIS), points of interest (POIs) are descriptions that denote geographical locations which might be of interest for some user purposes. Examples are public transport facilities, historical buildings, hotels and restaurants, recreation areas, hospitals etc. Because information gathering with respect to POIs is usually resource consuming, the user community is often involved in this task. In general, POI data originate from different sources (or users) and are therefore vulnerable to imperfections which might have a negative impact on data quality. Different POIs referring to, or describing the same physical geographical location might exist. Such POIs are said to be coreferent POIs. Coreferent POIs must be avoided as they could harm the data(base) quality and integrity. In this chapter, a novel soft computing technique for the (semi-)automated cleansing of POI databases is proposed. The proposed technique consists of two consecutive main steps: the detection of collections of coreferent POIs and the fusion, for each collection, of all coreferent POIs into a single consistent POI that represents all the POIs in the collection. The technique is based on fuzzy set theory, whereas possibility theory is used to cope with the uncertainties in the data. It can be used as a component of (semi-)automated data quality improvement strategies for databases and other information sources.

1 Introduction

Geographic information systems are characterized by a tremendous amount of data, which must be collected, processed and represented in an efficient, user-friendly way. Moreover, some of these data must regularly be actualised as geographic objects like roads, buildings or borderlines often change. A specific kind of

Guy De Tré · Daan Van Britsom · Tom Matthé · Antoon Bronselaer
Ghent University, Dept. of Telecommunications and Information Processing,
St.-Pietersnieuwstraat 41, B9000 Ghent (Belgium)
e-mail: {guy.detre, tom.matthe, daan.vanbritsom,
antoon.bronselaer}@UGent.be

information concerns the description of geographic locations or entities at geographic locations. In general, such information is modelled by objects which are called *points of interest* (POIs). Examples of POIs are objects that describe historical buildings, public services, hotels, restaurants and bars, panoramic views, interesting places to visit, etc. Usually, POIs contain information about location (coordinates) and a short textual description, but also other information such as the category the POI belongs to, multimedia like pictures and video and meta-data like the creator's name, the timestamp of creation, the file size, etc. can be provided.

In practice and due to their specific content, POI databases often contain data that are obtained from different heterogeneous sources, of which some might be maintained by user communities. User communities are often involved in data collection processes in cases where detailed, not commonly known data have to be inserted and maintained. When POIs originate from different sources or are entered by a user community, taking care of data consistency and correctness needs special attention. Indeed, such data are extremely vulnerable to errors, which might among others be due to uncertainty, imprecision, vagueness or missing information.

A problem that seriously harms the overall quality of a geographical information system (GIS) occurs when different POIs, denoting the same geographic entity, are inserted in the system. Such POIs are called *coreferent POIs*: they differ from each other, but all describe the same geographic location or object at a geographic location. Coreferent POIs can introduce uncertainty and inconsistency in the data, result in a storage and data processing overhead and moreover can cause low quality or, even worse, incorrect information retrieval results [26].

It is therefore important and relevant to develop techniques to *detect* coreferent POIs. Once detected, the problem of coreference has to be solved. Two basic approaches can be identified. In the first approach, the existence of coreferent POIs is *prevented* with techniques that, e.g., inform users about POIs that are detected to be in the neighbourhood of a new POI. As such, it is up to the user to check and verify whether the insertion makes sense. In the second approach, which is handled in more detail in this chapter, the responsibility for the correctness of the database is to a considerable extent shifted to the database management system. Coreferent POIs have to be *merged* (or *fused*) by the database management system into one single, consistent POI and the duplicates have to be removed. Perhaps, the simplest merging strategy is to keep one of the coreferent POIs and then remove all the others. As this simple merging strategy often introduces an information loss, more advanced merging techniques are required.

The research described in this chapter contributes to as well automatic detection, as automatic merging of coreferent POIs. The automatic detection of coreferent POIs has been approached as an uncertain Boolean problem. This means that two POIs are either coreferent or not (i.e., a boolean matter), but uncertainty about this decision must be dealt with. In order to determine this uncertainty, the POI structure is decomposed into elementary attributes (i.e., atomic sub objects). In this chapter it has been explicitly assumed that all POIs share the same structure. Thus, the issue of POI schema matching is not taken into account here. For each elementary attribute, an elementary *evaluator* is allocated. Such an evaluator determines the uncertainty

about the coreference of two values of the attribute's domain (which is the set of all allowed values for the attribute). The returned uncertainty is modelled by a possibilistic truth value [35, 18]. Because the POI's coordinates are among the main characteristics of a POI, special attention is paid to the detection of co-location of POIs, i.e., the definition of an appropriate evaluator for geographic coordinates. To obtain the overall uncertainty on the coreference of two POIs, the elementary evaluators are applied and their resulting possibilistic truth values are aggregated. For this aggregation, a variant of the Sugeno integral, as presented in [8] is used. The proposed merging approach for coreferent POIs uses the possibilistic truth values that are returned from the elementary evaluations and the aggregation, to determine how and which parts of two coreferent POIs should be merged to obtain a single deduplicated POI. Different strategies are described in the chapter.

The presented work also contributes to research on data quality issues in information retrieval by studying techniques that allow to automatically improve the data quality of information sources. Although applied in the context of geographic POI databases, the presented techniques can also be used and further extended for coreference detection and handling in other data sources and web sources. An improved data quality on its turn will automatically lead to better database query and information retrieval results. In cases where the data sources are read-only and hence can not be updated, coreference handling can be postponed until the data querying or information retrieval results are retrieved. Coreferent results can then be automatically filtered out and adequately handled before presenting them to the users.

The remainder of the paper is structured as follows. In Section 2, a brief overview of related work is given. Next, in Section 3, some preliminary definitions and notations with respect to objects and POIs are presented. Then, in Section 4 the problem of determining the uncertainty about the coreference of two POIs is dealt with. Herewith special attention is respectively paid to the definition of evaluators for atomic objects (in Subsection 4.1), the determination of the uncertainty about the co-location of two POIs in a two-dimensional space (in Subsection 4.2), and the computation of the overall uncertainty about the coreference of the overall POIs, i.e., the definition of aggregators for complex objects (in Subsection 4.3). Section 5, discusses the problem of merging two coreferent objects. Some general merge functions are described. These general functions allow one to develop a specific merge technique for POIs. The presented techniques for the detection and merging of coreferent POIs are illustrated in Section 6. Finally, in Section 7, some conclusions and indications for further work are given.

2 Related Work

Both the topics of coreference detection and of the merging of coreferent data have already been studied from different perspectives. In the next subsections we briefly give an overview of related work in these areas.

2.1 Coreference Detection

Coreference detection is already being studied since the late '60s, at which time it was commonly described as *record linkage*. A basic work on record linkage is [25]. Both traditional and fuzzy approaches exist.

In traditional approaches, coreference detection is typically done by means of a clustering method. An example is the DBSCAN algorithm [24]. When applying the DBSCAN algorithm to a POI database, clusters of coreferent POIs are expanded by adding similar POIs. Similarity between POIs is often determined by means of some multidimensional similarity measure, which is a weighted linear combination of spatial, linguistic and semantic measures. Spatial similarity is usually measured by calculating the distance between two POIs [34] and map this to inverse values in the interval $[0, 1]$, where 1 denotes an identical location and 0 represents the maximal distance. Linguistic similarity is usually measured by applying the Jaro-Winkler or another string comparison metric [29, 43] and semantic similarity can be computed by comparing the relative positions of the concepts under consideration in a taxonomic ontology structure [37].

In fuzzy approaches, the problem of detecting coreferent POIs is usually addressed by considering that duplicates are due to uncertainty and by explicitly handling this uncertainty by means of fuzzy set theory [47] and its related possibility theory [48, 21] (see, e.g., [40, 20]). Fuzzy ranges are then used to model spatial uncertainty about the co-location of two POIs. In [40], rectangular ranges are used, whereas in [20] context dependent circular ranges are proposed that are based on the scales of the maps in which the POIs are entered. In the remainder of this chapter, fuzzy set theory is used to further enhance spatial similarity measures so that these better cope with imperfections in the descriptions (of the locations) of the POIs. The problem of detecting co-location and merging of co-located data is also somewhat related to issues of conflation in GIS (see, e.g., [27]). Conflation is the complex process of combining information from two digital maps to produce a third map which is better than either of its component sources. In [36] the software agent technology paradigm has been applied as a conflation solution. Agent system techniques are hereby combined with expert system techniques to provide a feasible system architecture for distributed conflation.

2.2 Merging of Coreferent Data

The scientific foundations of POI merging lay in the research on information fusion, which deals with the combination of information provided by independent sources into one piece of information. The challenge hereby is to resolve inconsistencies between the different sources. An interesting aspect of information fusion is its applicability in many different contexts.

In a mathematical context, information fusion has led to the development of numerous aggregation operators such as generalized means [45, 23], t-norms and t-conorms [22] and uninorms [46]. Aggregation operators fuse information that is represented as an element of a complete lattice (L, \leq) . The information typically

expresses *facts*, for example the opinion or score of an agent. A flexible spatial data fusion approach based on a generalized ordered weighted averaging operator reflecting the concept of a fuzzy majority is presented in [16, 5]. Next to aggregation operators, a significant body of research deals with the case where deductive knowledge, such as inference rules and (integrity) constraints is used to combine information from different sources. Hereby, each source is considered to be a *propositional belief base* modelled as a first-order theory (see, e.g., [4, 1, 2, 31, 32, 30]). A typical difference between propositional belief bases and aggregation operators, is the presence of non-factual knowledge, such as inference rules and integrity constraints. As a consequence, the interest here is to *combine* all information in a maximal first-order theory. Such a setting occurs, amongst others, in heterogeneous databases [7]. A third type of information fusion deals with the case where each source provides knowledge by means of a possibility distribution (see, e.g., [38, 19]). In this case, it is assumed that the different sources have to cope with imprecision and/or incomplete knowledge and the key question is how uncertainty can be processed when dealing with different sources, that can provide conflicting information. Other approaches include heterogeneous data source fusion based on semantic rules (e.g., [33]) or ontologies (e.g., [6]).

Despite these related research areas, surprisingly the problem of merging coreferent data has not been as deeply investigated as the problem of coreference detection. An interesting overview of information combination operators for data fusion is given in [3]. In [12] the properties of object merging functions are investigated and a general framework for the merging of coreferent objects is proposed. In this paper we investigate and illustrate how this general framework can be applied in the context of POI merging.

3 Some Preliminaries

In this section we give some basic definitions and properties of objects and points of interest (POIs). These definitions form the formal basis for the techniques presented in the remainder of the chapter.

3.1 Basic Concepts on Objects

A fundamental concept in this chapter is that of an object. An *object* is axiomatically defined as a piece of data that describes an entity. A distinction is made between atomic and complex objects. Atomic objects are objects of which the universe is non compound, while complex objects belong to a universe O that is composed of non compound universes, i.e., $O = U_1 \times \dots \times U_n$. The appropriate universe of entities is denoted by \mathcal{E} and the link between objects and entities is formalised by a surjective function $\rho : O \rightarrow \mathcal{E}$. Objects that refer to the same entity in \mathcal{E} through ρ are said to be coreferent. Formally:

$$\forall(o_1, o_1) \in O^2 : (o_1 \leftrightarrow o_2) \Leftrightarrow (\rho(o_1) = \rho(o_2)). \quad (1)$$

The universe of an object is always equipped with a label function $l : O \rightarrow \mathcal{L}$, where \mathcal{L} represents the appropriate set of labels. The label of a universe represents the class of entities that objects in the universe are describing. For example, consider $l(\mathbb{R}) = \text{'latitude'}$, then we know that objects in \mathbb{R} are describing entities of the class 'latitude', i.e., describe the geographic latitude coordinate of a location on the earth's surface.

In addition, complex objects are equipped with a tree structure in the sense that there exist logical groups of labels that belong together. For example, in objects that describe geographic entities, the universes with label 'street', 'house number', and 'postal code' form a logical group, i.e., the address. Formally, for a complex universe O and with the understanding that $\mathcal{P}(U)$ denotes the power set of U (i.e., the set of all subsets of U , including the empty set and U itself), there exists a function:

$$\lambda : \mathcal{P}(\{l(U_i)\}_{i=1\dots n}) \rightarrow \{0, 1\}. \quad (2)$$

such that λ indicates for each group of labels, whether these labels form a logical group or not. As the structure that corresponds to λ must be a tree structure, some constraints must be satisfied. The labels themselves must represent leaf nodes and the root node is given by the set of all labels, which means that:

$$\forall i \in \{1, \dots, n\} : \lambda(\{l(U_i)\}) = 1 \quad (3)$$

$$\lambda(\{l(U_1), \dots, l(U_n)\}) = 1. \quad (4)$$

Also, the parent child relation must be respected. In terms of λ , this means that for two arbitrary sets of labels, the following constraint must be satisfied:

$$(\lambda(A) = \lambda(B) = 1) \Rightarrow (A \subseteq B \vee B \subseteq A \vee A \cap B = \emptyset) \quad (5)$$

which states that two logical groups A and B are either connected through the ancestor relation or are disjoint.

3.2 Basic Concepts on POIs

Reconsider the universe of entities \mathcal{E} . A *point of interest* (or POI) is axiomatically understood as a piece of data that describes a geographic entity in the real world that is modelled by \mathcal{E} . A POI is hence a special kind of complex object which is commonly used to describe an interesting location (or an entity at an interesting location).

By applying the function ρ that has been introduced in the previous subsection we obtain that two POIs POI_1 and POI_2 are coreferent, i.e., $POI_1 \leftrightarrow POI_2$ iff

$$(POI_1 \leftrightarrow POI_2) \Leftrightarrow (\rho(POI_1) = \rho(POI_2)). \quad (6)$$

Note that with the previous assumptions, we aim to keep the automated cleansing approach as general as possible and thus applicable to any data(base) model. The only requirements are that the data(base) model should support the modelling of

complex objects which belong to a compound universe $O = U_1 \times \dots \times U_n$ and for which there exists a label function l . The universe O is moreover equipped with a tree structure that is modelled by a function λ , which specifies logical groups of labels that belong together.

Example 1. An example of a compound universe that can be used to model POIs is

$$O_{POI} = U_1 \times U_2 \times U_3 \times U_4 \times U_5 \times U_6$$

where

$$\begin{aligned} U_1 &= S \\ U_2 &= S \\ U_3 &= [-90, 90] \\ U_5 &= [-180, 180] \\ U_5 &= S \\ U_6 &= C. \end{aligned}$$

Herewith, S is the set of all character strings and C is an enumerated list of allowed POI types. The label function l is specified as follows

$$\begin{aligned} \forall u \in U_1 : l(u) &= \text{identifier} \\ \forall u \in U_2 : l(u) &= \text{name} \\ \forall u \in U_3 : l(u) &= \text{latitude} \\ \forall u \in U_4 : l(u) &= \text{longitude} \\ \forall u \in U_5 : l(u) &= \text{description} \\ \forall u \in U_6 : l(u) &= \text{type} \\ \forall o \in O_{POI} : l(o) &= \text{POI}. \end{aligned}$$

The tree structure that is specified on O_{POI} is given by the function λ which is specified as follows (all subsets of labels that are not explicitly mentioned map to 0):

$$\begin{aligned} \lambda(\{l(U_1)\}) &= \lambda(\{\text{identifier}\}) = 1 \\ \lambda(\{l(U_2)\}) &= \lambda(\{\text{name}\}) = 1 \\ \lambda(\{l(U_3)\}) &= \lambda(\{\text{latitude}\}) = 1 \\ \lambda(\{l(U_4)\}) &= \lambda(\{\text{longitude}\}) = 1 \\ \lambda(\{l(U_5)\}) &= \lambda(\{\text{description}\}) = 1 \\ \lambda(\{l(U_6)\}) &= \lambda(\{\text{type}\}) = 1 \\ \lambda(\{\text{identifier}, \text{name}, \text{latitude}, \text{longitude}, \text{description}, \text{type}\}) &= 1 \\ \lambda(\{l(U_3), l(U_4)\}) &= \lambda(\{\text{latitude}, \text{longitude}\}) = 1. \end{aligned}$$

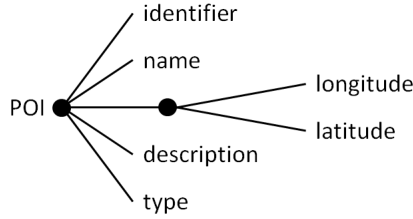


Fig. 1 Tree structure on labels corresponding to the mappings of Example 1

The tree structure corresponding to these mappings is presented in Figure 1. The semantics of the POIs under consideration can then be described as follows. The element of U_1 is the unique identifier of the POI, the element of U_2 is the name of the POI. The elements of U_3 and U_4 are connected to each other and together represent the location of the POI, which is given by a latitude and a longitude. Both latitudes and longitudes are expressed in decimal degrees (where 0.000001 degrees corresponds to 0.111 metre). The element of U_5 is a free description, provided by the user and modelled by full text. Finally, the element of U_6 is the type (or category) of the POI. It is assumed that this type is chosen from a given list. \square

Because each POI is an element of a universe $O = U_1 \times \dots \times U_n$, it can be denoted by a n -tuple (u_1, \dots, u_n) , where $u_i \in U_i, i = 1, \dots, n$.

Example 2. Reconsider the POI structure as introduced in Example 1. The following four 6-tuples are illustrations of POIs.

- $(POI_1, \text{'Friday market'}, 51.056934, 3.727112, \text{'Friday Market, Ghent'}, \text{'Market'})$
- $(POI_2, \text{'St-Bavo'}, 51.053036, 3.727015, \text{'St-Bavo's Cathedral, Ghent'}, \text{'Church'})$
- $(POI_3, \text{'Ghent cathedral'}, 51.053177, 3.726382, \text{'St-Bavo Cathedral'}, \text{'Cathedral'})$
- $(POI_4, \text{'St-Bavo'}, 51.033333, 3.700000, \text{'St-Bavo – Ghent'}, \text{'Cathedral'})$.

POI_2, POI_3 and POI_4 are examples of coreferent POIs. All four POIs have a different location. \square

4 Detection of Coreferent POIs

In this section, the problem of determining the uncertainty about the coreference of two POIs is dealt with. A possibilistic solution for finding coreferent objects consists of finding functions that express the uncertainty of coreference by means of *possibilistic truth values* [35, 42, 17, 18], which are possibility distributions over the Boolean domain $\mathbb{B} = \{T, F\}$. Thus, for a given Boolean proposition p , the possibilistic truth value (or PTV) \tilde{p} :

$$\tilde{p} = \{(T, \mu_{\tilde{p}}(T)), (F, \mu_{\tilde{p}}(F))\} \quad (7)$$

expresses the possibility that p is true (T) and the possibility that p is false (F). The domain of all possibilistic truth values is denoted $\mathcal{F}(\mathbb{B})$, i.e., the power set of normalised fuzzy sets over \mathbb{B} . In what follows, we shall adopt the couple shorthand notation for possibilistic truth values, i.e., $\tilde{p} = (\mu_{\tilde{p}}(T), \mu_{\tilde{p}}(F))$. Let us define the order relation \geq on the set $\mathcal{F}(\mathbb{B})$ as follows:

$$\tilde{p} \geq \tilde{q} \Leftrightarrow \begin{cases} \mu_{\tilde{p}}(F) \leq \mu_{\tilde{q}}(F), & \text{if } \mu_{\tilde{p}}(T) = \mu_{\tilde{q}}(T) = 1 \\ \mu_{\tilde{q}}(T) \leq \mu_{\tilde{p}}(T), & \text{else} \end{cases} \quad (8)$$

An evaluator is a function that estimates a possibilistic truth value in order to express uncertainty about coreference [9].

Given a universe of objects O , an *evaluator* over O is defined as a function E_O :

$$E_O : O^2 \rightarrow \mathcal{F}(\mathbb{B}) \quad (9)$$

An evaluator compares two objects and yields a possibilistic truth value that expresses both the possibility that the objects are coreferent and the possibility that the objects are not coreferent. An evaluator is

- *Reflexive* if and only if:

$$\forall (o_1, o_2) \in O^2 : (o_1 = o_2) \Rightarrow (E_O(o_1, o_2) = (1, 0)) \quad (10)$$

- *Strong reflexive* if and only if:

$$\forall (o_1, o_2) \in O^2 : (o_1 = o_2) \Leftrightarrow (E_O(o_1, o_2) = (1, 0)) \quad (11)$$

- *Commutative* if and only if:

$$\forall (o_1, o_2) \in O^2 : E_O(o_1, o_2) = E_O(o_2, o_1) \quad (12)$$

In what follows, evaluators are always assumed to be commutative and at least reflexive. Finally, an evaluator is called *transitive* if and only if, for every triplet $(o_1, o_2, o_3) \in O^3$:

$$\begin{aligned} 1 - \mu_{E_O(o_1, o_3)}(F) &\geq \min(1 - \mu_{E_O(o_1, o_2)}(F), 1 - \mu_{E_O(o_2, o_3)}(F)) \\ 1 - \mu_{E_O(o_1, o_3)}(T) &\geq \min(1 - \mu_{E_O(o_1, o_2)}(F), 1 - \mu_{E_O(o_2, o_3)}(T)) \\ 1 - \mu_{E_O(o_1, o_3)}(T) &\geq \min(1 - \mu_{E_O(o_1, o_2)}(T), 1 - \mu_{E_O(o_2, o_3)}(F)). \end{aligned}$$

In the next subsections we successively describe evaluators for atomic objects, evaluators for determining co-location and evaluators for complex objects.

4.1 Elementary Evaluators for Atomic Objects

When it comes to the evaluation of atomic objects (i.e., objects with a non compound universe), some existing approaches are useful in the detection of coreferent

POIs. More specifically, the comparison of character strings and numerical data has already been studied extensively and is the basis for the development of general purpose evaluators for character strings and numerical data. Such evaluators are briefly introduced in the next subsections.

4.1.1 Evaluators for Character Strings

First, *syntactical* evaluators have been proposed. These evaluators allow for the comparison of two character strings, taking into account the occurrence of spelling errors, abbreviations, ... [10, 13]. Hereby, strings are decomposed into a multiset of substrings. These multisets are then compared such that similarities between elements are taken into account [9]. The evaluators are called ‘syntactical’, because they decide upon coreference of two objects by comparing the syntactical construction of objects. Syntactical evaluators for strings are for example well suited for comparison of POI names and descriptions.

Secondly, *semantical* evaluators have been proposed [11]. As opposed to syntactical evaluators, semantical evaluators reject the idea that a decision of coreference must be based on a syntactical similarity between two objects. Instead, it accepts the fact that the existence of some (semantical) relationship between two objects can be sufficient to decide that these objects are coreferent. Examples of such relationships are the synonym relationship, the specification/generalization relationship, ... In [11], an approach is proposed for the dynamical discovery of (semantical) relationships between objects. In the case of POIs, semantical evaluators are well suited for the comparison of POI types.

4.1.2 Evaluators for Numerical Data

Evaluators for character strings can also be used for coreference detection of numerical data too. Indeed, coreferent numerical values refer to the same number, but can differ from each other due to typing errors or uncertainty. A typical example are telephone numbers or bank account numbers. In such cases, depending on the context in which the numbers are used, either syntactical and/or semantical evaluators can be applied for coreference detection.

Coreferency of numerical data can also be due to imprecision. In such a case the difference between two numbers can be used as the basis for evaluation. If two numbers a and b are close enough, i.e., if $|a - b| \leq \varepsilon$, then a and b can be considered as being coreferent, else they are not considered to be coreferent. Hereby, ε acts as a threshold value and depends on the application under consideration. An example on how the value of ε can be determined is given in the next subsection.

4.2 Evaluators for Co-location

Next to these general purpose evaluators described in the previous subsection, the case of POIs requires some case-specific evaluators for the comparison of locations. More details on these evaluators are discussed below.

Perhaps the most important aspect of a POI is its registered geographic location. POI's are considered to be zero dimensional objects, whereas geographic entities in the real world are generally two or three dimensional objects and hence can be denoted by multiple locations. Consider for example all locations of the surface of a bridge, park or lake or all locations in a building. To construct the POI, one of these locations has to be chosen as the representative location (or point). The location of a POI is hence, due to its nature, already very vulnerable to imprecision what is one of the main causes for coreferency. Beside of this inherent imprecision, coreferent POIs can also be assigned to different locations due to uncertainty or a lack of information.

In the remainder of this subsection a soft technique for estimating the uncertainty about the co-location of two POIs is presented. First, a basic technique commonly used in fuzzy geographic applications is presented. Secondly, this basic technique is further enhanced in order to explicitly cope with the scale at which the POI is entered by the user.

4.2.1 Basic Technique

The geographic location of a POI is usually modelled in a two-dimensional space by means of a latitude lat and longitude lon , as has been illustrated in Example 1. Consider two POIs POI_1 and POI_2 with locations (lat_1, lon_1) and (lat_2, lon_2) respectively. In geographic applications, the distance (in metres) between the two locations is usually approximately computed by

$$d(POI_1, POI_2) = 2R \arcsin(h) \quad (13)$$

where $R = 6367000$ is the radius of the earth in metres and

$$h = \min \left(1, \sqrt{\sin^2 \left(\frac{lat_2^r - lat_1^r}{2} \right) + \cos(lat_1^r) \cos(lat_2^r) \sin^2 \left(\frac{lon_2^r - lon_1^r}{2} \right)} \right)$$

with $lat_j^r = \frac{\pi}{180} lat_j$ and $lon_j^r = \frac{\pi}{180} lon_j$, for $j = 1, 2$, being the conversions in radians of lat_j and lon_j [39]. The higher the precision of the measurement of the latitude and longitude, the higher the precision of this distance.

From a theoretical point of view, POIs are considered to be geographic locations. Hence, two POIs are considered to be co-located if their distance equals zero. In practice however, one has to deal with imperfect positioning specifications of locations. Therefore, it is more realistic to consider two POIs as being co-located if they refer to the same area and are thus *close enough*. In traditional approaches '*close enough*' is usually modelled by a threshold $\varepsilon > 0$, such that two POIs POI_1 and POI_2 are ε -close if and only if

$$d(POI_1, POI_2) \leq \varepsilon. \quad (14)$$

The problem with such a single threshold is that it puts a hard constraint on the distance, which implies an ‘all or nothing’ approach: depending on the choice for ε , two POIs will be considered as being co-located or not. If an inadequate threshold value is chosen, this will yield in a bad decision. A single threshold neither offers the flexibility to use different criteria in different contexts.

Fuzzy sets [47] have been used to soften the aforementioned hard constraint. In general, a fuzzy set with a membership function $\mu_{\varepsilon-close}$, as presented in Figure 2, is used to model ‘close enough’. This membership function is defined by

$$\mu_{\varepsilon-close} : [0, +\infty) \rightarrow [0, 1]$$

$$d \mapsto \begin{cases} 1 & , \text{ if } d \leq \varepsilon \\ \frac{\delta - d}{\delta - \varepsilon} & , \text{ if } \varepsilon < d \leq \delta \\ 0 & , \text{ if } d > \delta. \end{cases} \quad (15)$$

The extent to which two POIs POI_1 and POI_2 are considered to be co-located is then given by $\mu_{\varepsilon-close}(d(POI_1, POI_2))$. Hence, for distances below ε , $\mu_{\varepsilon-close}$ denotes co-location, for distances larger than δ no co-location is assumed, whereas for distances between ε and δ , there is a gradual transition from co-location to no co-location. Other membership function definitions can be used.

4.2.2 Enhanced Technique

A practical problem with fuzzy approaches as described above, is that the membership function has to reflect reality as adequate as possible. This implies that adequate values for ε and δ must be chosen. Values that are too stringent (too small) will result in false negatives, i.e., some POIs will falsely be identified as not being co-located, whereas values that are too soft (too large) will result in false positives, i.e., some POIs will falsely be identified as being co-located. In this subsection, it is considered that different POIs can originate from different sources or users. Such a situation often occurs in practical cases where data of different origins have to be collected and combined. Under this consideration, it makes sense to study how

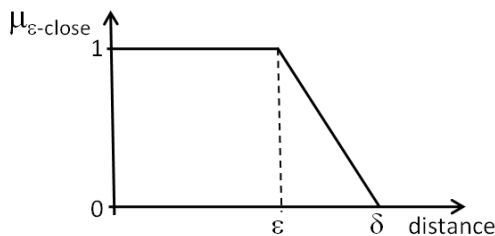


Fig. 2 Fuzzy set with membership function $\mu_{\varepsilon-close}$ for representing ‘close enough’

the parameters ε and δ are influenced by the *context* in which the POI has been originally registered. Eq. (15) can then be further enhanced in order to better reflect the imperfection and the context of the placement of the POI.

In practice, the exact coordinates of the location of a POI will not always be known. In such a case, the location of the POI has to be approximated. When user communities are involved in the construction and maintenance of a POI database, users might be asked to denote the position of the POI on a map. User communities are often involved when the content of the database changes regularly, which is for example the case with locations of speed control devices, locations that denote dangerous road conditions, and locations that denote interesting points to visit during walking or cycling activities.

If POI locations are entered via geographic maps the quality of the data will to some extent depend on the context the user is working in. Next, we focus on two aspects of this work context, namely scale and precision, and show how explicitly coping with these can help to improve Eq. (15).

If users work with maps on computer screens or screens of mobile devices when entering or maintaining (locations of) POIs, they work with a representation of (a part of) the real world that is drawn at a specific *scale* $(1 : s)$, which means, e.g., that 1 cm on the scale corresponds to s cm in reality. For example, a map of Europe on a computer screen can be drawn at scale $(1 : 15000000)$, a map of Belgium at scale $(1 : 1000000)$ and a map of Ghent at scale $(1 : 125000)$. It is clear that the precision with which a user can place a POI on a map depends on the scale of the map. Denoting a POI that represents the Eiffel tower on a map of Europe will be less precise than on a map of France, which on its turn will be less precise than on a map of Paris. On the other hand, depending on his or her knowledge about the location of the new POI the user can zoom-in or zoom-out on the map to enter the POI at the map with the most appropriate detail for the user. Considering the different scales used in the different sources or used by different users, a scale $(1 : s_{min})$ corresponding to the most detailed level and a scale $(1 : s_{max})$ corresponding to the least detailed level can be determined. Hence, all occurring scales $(1 : s)$ will be within the range $(1 : s_{min}) \leq (1 : s) \leq (1 : s_{max})$.

Another aspect to take into account is the *precision* with which the user can denote the location of a POI on the screen. Usually, when working at an appropriate scale $(1 : s)$, the user will be able to place a point on the screen with a precision of a couple of centimetres, i.e., the exact location of the point will be within a circle with the denoted point as centre and radius d_s . This radius can be considered to be a parameter that depends on the scale $(1 : s)$ and the user's abilities for accurately denoting the POI on the screen. Therefore, in practical applications, d_s could be adjustable by the user or by a user feedback mechanism.

The scales $(1 : s)$, $s_{min} \leq s \leq s_{max}$, and corresponding radii d_s can now be used to further enhance the definition of the membership function $\mu_{\varepsilon-close}$ that is used in Eq. (15).

Estimating the Value of ε

In order to better approach reality, ε should reflect the maximum distance for which two POIs are indistinguishable and hence must be considered as being co-located.

If no further information about the geographical area of the POI is available, then the POI is positioned at the location that is entered by the user and modelled by its latitude and longitude. Two POIs are then indistinguishable if they are mapped to the same latitude and longitude. The maximum precision can be approximated by the dot pitch of the screen and be used to estimate the value of ε . The dot pitch d_p of a screen is defined as the diagonal distance between two pixels on the screen and usually has a standard value of 0.28mm. Considering the minimum scale ($1 : s_{min}$), the value of ε can then be approximated by

$$\varepsilon = d_p s_{min}. \quad (16)$$

If information about the geographical area of the POI is given, then the length l of the diagonal of the minimum bounding rectangle that surrounds this area can be used to approximate ε . Indeed, all POIs that are placed in the rectangle can reasonably be considered as being co-located. If the POI location of POI_1 and POI_2 is respectively entered at a scale ($1 : s_1$) and ($1 : s_2$), the value of ε can be approximated by

$$\varepsilon = \max\left(\frac{l}{2}s_1, \frac{l}{2}s_2\right) \quad (17)$$

where the maximum operator is used to take the roughest, largest approximation (which is due to the least precise scale) in cases where both POIs were entered at a different scale.

Estimating the Value of δ

Taking into account the scale ($1 : s_1$) and precision d_{s_1} with which a user entered POI_1 and the scale ($1 : s_2$) and precision d_{s_2} with which POI_2 was entered, the value of δ can be defined by

$$\delta = \varepsilon + \max(s_1 d_{s_1}, s_2 d_{s_2}) \quad (18)$$

where the maximum operator is again used to take the roughest approximation in cases where both POIs were entered at a different scale. With this definition the precisions d_{s_1} and d_{s_2} are handled in a pessimistic way. Alternative definitions for δ are possible.

4.2.3 Evaluator for Co-location

The membership function $\mu_{\varepsilon-close}$ can now be used to define an evaluator E_{loc} for the determination of co-location. Such an evaluator should satisfy Eq. (9) and hence result in a PTV, expressing the uncertainty about the colocation of two locations of POIs.

A proposal for a simple definition for E_{loc} is

$$E_{loc} : ([-90, 90] \times [-180, 180])^2 \rightarrow \mathcal{F}(\mathbb{B})$$

$$((lat_1, lon_1), (lat_2, lon_2)) \mapsto (\mu_{\bar{p}}(T), \mu_{\bar{p}}(F)) \quad (19)$$

where the membership grades $\mu_{\bar{p}}(T)$ and $\mu_{\bar{p}}(F)$ are defined by

$$\mu_{\bar{p}}(T) = \frac{\mu_{\varepsilon-close}(d)}{\max(\mu_{\varepsilon-close}(d), 1 - \mu_{\varepsilon-close}(d))} \quad (20)$$

$$\mu_{\bar{p}}(F) = \frac{1 - \mu_{\varepsilon-close}(d)}{\max(\mu_{\varepsilon-close}(d), 1 - \mu_{\varepsilon-close}(d))}. \quad (21)$$

where the distance $d = d((lat_1, lon_1), (lat_2, lon_2))$ is computed using Eq. (13) and the membership function $\mu_{\varepsilon-close}$ is defined by Eq. (15) with the parameter values ε and δ being estimated as described above. An example of the use of the evaluator E_{loc} is given in Section 6.

The evaluator E_{loc} can be used as a component of a technique to determine whether two POIs are coreferent or not. The resulting PTVs as obtained by Eq. (20) and (21), then denote a measure for the uncertainty about the co-location or spatial similarity of the POIs.

4.3 Evaluators for Complex Objects

Once atomic objects have been compared, a comparison of complex objects can be performed by aggregating the results of atomic comparisons. For that purpose, an extension of the Sugeno integral to the domain of PTVs has been proposed [8].

This integral uses two fuzzy measures (γ^T and γ^F). The measure γ^T (resp. γ^F) provides the conditional necessity that two complex objects are (not) coreferent, given that some set of attributes are (not) coreferent. In the case of POIs, $\gamma^T(\{\text{'name'}, \text{'type'}\})$ is a number in the unit interval that represents the necessity that two POIs are coreferent, provided that their names and types are coreferent. Similarly, $\gamma^F(\{\text{'name'}, \text{'type'}\})$ is a number in the unit interval that represents the necessity that two POIs are not coreferent, provided that their names and types are not coreferent. As required by the definition of fuzzy measures, γ^T and γ^F are normalised between \emptyset and \mathcal{L} and are monotonic.

It is noted that the fuzzy measures can be used to take structural information of objects into account. For example, it can be reflected in γ^T and γ^F that street, zip code and city together constitute an address by introducing dependencies between these atomic objects. This can be easily automated by usage of the function λ as introduced by Eq. (2).

The Sugeno integral introduced in [8] combines conditional necessity (γ^T and γ^F) with marginal necessity (the PTVs obtained from atomic comparison) into one PTV that reflects the uncertainty about the fact that two complex objects are coreferent. The inference used for this combination is purely possibilistic in nature and

is therefore a valid and well suited aggregation method for PTVs in the case of coreference.

With the understanding that \tilde{P} denotes a finite set of PTVs $\tilde{P} = \{\tilde{p}_1, \dots, \tilde{p}_n\}$, the Sugeno integral of \tilde{P} with respect to γ^T and γ^F is defined by

$$S_{\gamma^{T,F}}(\tilde{P}) : \mathcal{F}(\mathbb{B})^n \rightarrow \mathcal{F}(\mathbb{B}) : \tilde{P} \mapsto \tilde{p} \quad (22)$$

so that

$$\begin{aligned} \mu_{\tilde{p}}(T) &= Pos_{\tilde{p}}(T) \\ &= 1 - Nec_{\tilde{p}}(F) \\ &= 1 - \bigvee_{i=1}^n Nec\left(\tilde{P}_{(i)F} = F\right) \wedge \gamma^F\left(\tilde{P}_{(i)F}\right) \\ &= 1 - \bigvee_{i=1}^n \left(\min_{\tilde{p} \in \tilde{P}_{(i)F}} (1 - \mu_{\tilde{p}}(T)) \right) \wedge \gamma^F\left(\tilde{P}_{(i)F}\right) \end{aligned}$$

and

$$\begin{aligned} \mu_{\tilde{p}}(F) &= Pos_{\tilde{p}}(F) \\ &= 1 - Nec_{\tilde{p}}(T) \\ &= 1 - \bigvee_{i=1}^n Nec\left(\tilde{P}_{(i)T} = T\right) \wedge \gamma^T\left(\tilde{P}_{(i)T}\right) \\ &= 1 - \bigvee_{i=1}^n \left(\min_{\tilde{p} \in \tilde{P}_{(i)T}} (1 - \mu_{\tilde{p}}(F)) \right) \wedge \gamma^T\left(\tilde{P}_{(i)T}\right) \end{aligned}$$

where $\cdot_{()T}$ and $\cdot_{()F}$ are permutations on the elements of \tilde{P} . With the understanding that $\tilde{p}_{(i)T}$ (resp. $\tilde{p}_{(i)F}$) denotes the i th element of the permutation $\cdot_{()T}$ (resp. $\cdot_{()F}$) and that \leq is the order relation for PTVs as defined by Eq. 8, the permutations $\cdot_{()T}$ and $\cdot_{()F}$ are defined as follows:

$$\forall i \in \{1, \dots, n-1\} : \tilde{p}_{(i+1)T} \leq \tilde{p}_{(i)T}. \quad (23)$$

In other words $\cdot_{()T}$ is a permutation that orders the elements of \tilde{P} according to largest PTV first. Furthermore the permutation $\cdot_{()F}$ on the elements of \tilde{P} is defined by

$$\forall i \in \{1, \dots, n-1\} : \tilde{p}_{(i)F} \leq \tilde{p}_{(i+1)F}. \quad (24)$$

This is the reciproque permutation of $\cdot_{()T}$.

More details about (the use of) the Sugeno integral can be found in [8].

Because a POI is considered to be a special kind of a complex object, the evaluators for complex objects can be used to determine the PTV expressing the overall uncertainty that two POIs are coreferent or not. This will be illustrated in Section 6.

5 Merging of Coreferent POIs

Once coreferent POIs are detected, their duplicate information should be removed and their non-duplicate information should be merged. The challenges hereby are to avoid information loss and to resolve the inconsistencies that might exist among the different coreferent data.

A general *merge function* for coreferent objects of a universe O has been formally defined by

$$\varpi_O : \mathcal{M}(O) \rightarrow O \quad (25)$$

where $\mathcal{M}(O)$ denotes the set of all multisets drawn from the universe O [13, 15]. The merge function thus takes a multiset of objects and produces one single object as a result. As proposed by Yager [44], a multiset M over O is hereby characterized by a counting function $\omega_M : O \rightarrow \mathbb{N}$. For $v \in O$, $\omega_M(v)$ then represents the number of times that v occurs in M .

In the next Subsections 5.1 and 5.2, specific merge functions for atomic and complex objects will be defined. These functions will then be further fine-tuned for the purpose of POI merging in Subsection 5.3. More information on the properties of the proposed functions is given in [13].

5.1 Merge Functions for Atomic Objects

Let us first introduce merge functions ϖ_U where U is a non compound universe. Recall that the context in which ϖ_U is to be used, is that of coreference. As such, we can assume that upon merge time, an evaluator E_U is available. Let M be a multiset of coreferent objects that are identified by a coreference detection framework. Then, for each object $u \in M$, $|M| = \sum_{u \in U} \omega_M(u)$ PTVs can be calculated by comparing u with all objects in M . Due to reflexivity of E_U , the PTV $(1, 0)$ occurs at least $\omega_M(u)$ times. As such, for each object $u \in M$ a collection of PTVs is obtained where each \tilde{p} indicates the uncertainty about the proposition that two objects are coreferent. In [28], a method is proposed to construct a possibility distribution $\pi_{\mathbb{N}}$ (a fuzzy integer) from a collection of PTVs. Hereby, $\pi_{\mathbb{N}}(k)$ indicates the possibility that exactly k propositions are true. Hence, for each element $u \in M$, a possibility distribution $\pi_{\mathbb{N}}^u$ can be constructed, where $\pi_{\mathbb{N}}^u(n)$ represents the possibility that ‘exactly n values in M are coreferent with u ’.

The method described in [28] has been used for the construction of a confidence based merge function as it allows to express the uncertainty about the number of coreferent objects according to a given evaluator E_U . It works as follows. Let P be a set of independent Boolean propositions and let \tilde{P} be the multiset of corresponding PTVs which results from the evaluation of the proposition in P . Then, the quantity of true propositions in P is modelled by the possibility distribution $\pi_{\mathbb{N}}$ such that:

$$\pi_{\mathbb{N}}(k) = \min \left(\sup \{ \alpha \in [0, 1] \mid |\{p \in P \mid \mu_{\tilde{p}}(T) \geq \alpha\}| \geq k \}, \sup \{ \alpha \in [0, 1] \mid |\{p \in P \mid \mu_{\tilde{p}}(F) < \alpha\}| \geq k \} \right). \quad (26)$$

Eq. (26) states that the possibility $\pi_{\mathbb{N}}(k)$ is the minimum of the possibility that at least k propositions are true and the possibility that at most $|P| - k$ propositions are false. The possibility $\pi_{\mathbb{N}}(k)$ can be efficiently calculated by adopting the following notations. For a multiset \tilde{P} , let $\tilde{p}_{(i)}$ denote the i^{th} largest PTV with respect to the order relation defined in Eq. (8). The following then holds:

$$\pi_{\mathbb{N}}(k) = \begin{cases} \mu_{\tilde{p}_{(k)}}(F) & , \text{ if } k = 0 \\ \mu_{\tilde{p}_{(k)}}(T) & , \text{ if } k = |M| \\ \min(\mu_{\tilde{p}_{(k)}}(T), \mu_{\tilde{p}_{(k+1)}}(F)) & , \text{ else} \end{cases} \quad (27)$$

Figure 3 shows two example multisets, each consisting of five PTVs $(\mu_{\tilde{p}}(T), \mu_{\tilde{p}}(F))$, where \circ denotes the possibility $\mu_{\tilde{p}}(T)$ of T and \times denotes the possibility $\mu_{\tilde{p}}(F)$ of F . The derived possibility distributions $\pi_{\mathbb{N}}$, computed using Eq. (27), are shown below the PTVs. Note that the membership functions of the derived fuzzy integers $\pi_{\mathbb{N}}$ are always *convex*.

Applying this method allows us to express the number of coreferent objects, *according to* the evaluator E_U . Hence, although we already know that objects in M are coreferent, the distributions $\pi_{\mathbb{N}}$ express the uncertainty about this statement, at least, according to the evaluator E_U . Based on these observations, a merge function can be defined, considering that the result of the merging should be the object which has the highest number of coreferent objects according to E_U . We then obtain a merging technique where the uncertainty model of E_U is used to choose the best representative.

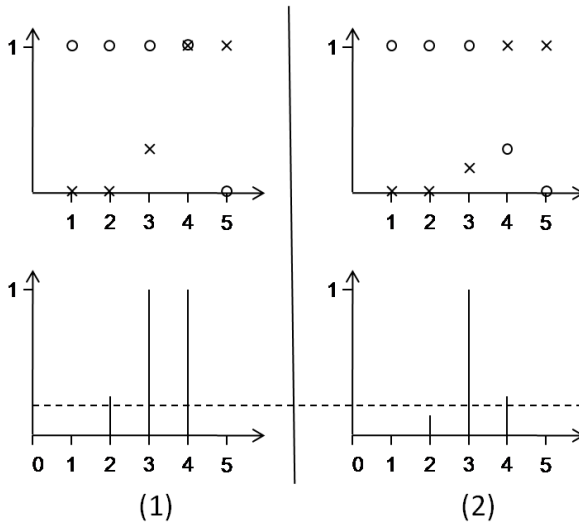


Fig. 3 Two example sets of five PTVs $(\mu_{\tilde{p}}(T), \mu_{\tilde{p}}(F))$ where \circ and \times respectively denote $\mu_{\tilde{p}}(T)$ and $\mu_{\tilde{p}}(F)$ (top) and their corresponding derived fuzzy integer (bottom)

For this purpose, a method for comparing fuzzy integers is required. Many methods have been proposed. The most common technique is to *defuzzify* the fuzzy integer, for example by means of the center of gravity [22]. Fuzzy integers are then compared by comparing the results of defuzzification. The method that we shall adopt here, is not based on defuzzification, but is rather possibilistic in nature. We propose two order relations for fuzzy integers, one constructed from the viewpoint of possibility and one constructed from the viewpoint of necessity.

For two fuzzy integers, \tilde{n} and \tilde{m} , the sup-order relation \prec_{sup} is defined by

$$\tilde{n} \prec_{\text{sup}} \tilde{m} \Leftrightarrow \sup \tilde{n}_\alpha < \sup \tilde{m}_\alpha. \quad (28)$$

Hereby, \tilde{n}_α is the α -cut of \tilde{n} where α is chosen such that:

$$\alpha = \sup\{x \mid \sup \tilde{n}_x \neq \sup \tilde{m}_x\}.$$

Also, for two fuzzy integers, \tilde{n} and \tilde{m} , the inf-order relation \prec_{inf} is defined by

$$\tilde{n} \prec_{\text{inf}} \tilde{m} \Leftrightarrow \inf \tilde{n}_\alpha < \inf \tilde{m}_\alpha. \quad (29)$$

Hereby, \tilde{n}_α is the α -cut of \tilde{n} where α is chosen such that:

$$\alpha = \sup\{x \mid \inf \tilde{n}_x \neq \inf \tilde{m}_x\}.$$

The sup-order of fuzzy integers searches for the highest α , such that the α -cuts have a different supremum and then chooses the fuzzy number for which the α -cut has the higher supremum. It can be seen that this method is equivalent to first searching the fuzzy integers that have the maximal k , say k_{max} , for which $\pi_{\mathbb{N}}(k_{\text{max}}) = 1$. If multiple such fuzzy integers exist, the decision is obtained by applying the leximax-operator on the sequence $\pi_{\mathbb{N}}(k_{\text{max}} + 1), \dots, \pi_{\mathbb{N}}(|M|)$. The dual is true for \prec_{inf} . Note that both \prec_{sup} and \prec_{inf} are partial orders. If multiple fuzzy numbers are equivalent, a random choice is made. Note that two non-equal convex fuzzy integers are always comparable by either \prec_{inf} or \prec_{sup} .

Consider the fuzzy integers shown in Figure 3. The order relation \prec_{sup} denotes the leftmost fuzzy integer as the largest, because the 1-cut of the leftmost fuzzy integer has a higher supremum (4) than the rightmost (3). However, the order relation \prec_{inf} denotes the rightmost fuzzy integer as the largest, because the 0.2-cut (denoted by the dashed line) of the leftmost fuzzy number has a lower infimum (2) than the 0.2-cut of the rightmost fuzzy integer (3).

Using the order of fuzzy integers, it is possible to define a merge function $\overline{\omega}_U$, which is driven by an evaluator E_U for atomic universes U . For example, using the order relation \prec_{sup} , the confidence-based merge function $\overline{\omega}_U$ for coreferent objects of an atomic universe U has been defined by

$$\overline{\omega}_U(M) = \arg \max_{u \in M} \pi_{\mathbb{N}}^u \quad (30)$$

where $\pi_{\mathbb{N}}^u$ is a possibility distribution, representing a fuzzy integer, that is obtained from the multiset \tilde{P}_u of PTVs for which

$$\forall u' \in M : \omega_{\tilde{P}_u}(E_U(u, u')) = \omega_M(u').$$

As such, $\overline{\omega}_U(M)$ selects the object $u \in M$ that has the largest corresponding fuzzy number $\pi_{\mathbb{N}}^u$ according to the order relation \prec_{sup} . Selecting the largest fuzzy number hereby reflects that the object with the largest confidence has been chosen as the result of the merging. Illustrations of such merge functions for atomic objects are given in Subsection 6.2.

5.2 Merge Functions for Complex Objects

In order to merge coreferent objects of a complex, composite universum O , a *composite merge function* is used. A possible strategy in doing so is to consider an evaluator E_O and to construct merge functions for complex universes as explained in the previous subsection.

Another way of defining composite merge functions is to combine the *projection* operator on the compound universe O with merge functions for the atomic universes. Doing so, yields the following definition.

Consider a complex universe $O = U_1 \times \dots \times U_n$. A *composite merge function* $\overline{\omega}_O$ over O is defined by

$$\overline{\omega}_O : \mathcal{M}(O) \rightarrow O \quad (31)$$

where

$$\overline{\omega}_O(M) = (\overline{\omega}_{U_1}(\text{Proj}_1(M)), \dots, \overline{\omega}_{U_n}(\text{Proj}_n(M)))$$

with $\text{Proj}_i(M) \in \mathcal{M}(U_i)$ such that

$$\omega_{\text{Proj}_i(M)}(u) = \sum_{o \in M \wedge o_i = u} \omega_M(o).$$

5.3 Merging of Coreferent POIs

The general merge strategies presented in the previous subsections can be used to develop a merge technique for coreferent POIs. Because POIs are complex objects (as specified in Subsection 3.2), a composite merge function as defined by Eq. (31) can be used to merge coreferent POIs.

This approach is motivated by the fact that we prefer to keep only the best (partial) information from each coreferent POI in the resulting merged POI. Hence, we do not prefer to select and preserve one of the existing POIs as the result of the merging operation. With this strategy, we explicitly opt to cleanse the (regular) POI database without introducing uncertain data in it. Indeed, alternatively one might also choose to work with a ‘fuzzy’ POI database in which uncertainty about the possible values of the POI attributes is explicitly stored. As such the informative richness of the many sources provided by the user communities can be better

maintained. However, such an approach will result in POIs that are more difficult to interpret and to process. For that reason this approach is not further considered within the scope of the work presented in this chapter.

In order to specify the composite merge function in accordance with Eq. (31), merge functions for atomic objects, handling locational, descriptive and categorical data must be provided. The use and selection of such merge functions is discussed in the next subsections.

5.3.1 Merging of Locational Data

In POIs, locational data is usually specified by means of a latitude and a longitude value, each of them being modelled by an atomic object, respectively taken from the atomic universa $[-90, 90]$ and $[-180, 180]$ (cf. Example 1). Because POIs are often descriptions of geographic areas (buildings, parks, lakes, etc.) latitude and longitude data are often imprecise. A good merge function for latitude and longitude data has to reduce this imprecision as good as possible. Hence, an aggregation function like arithmetic mean could be a good candidate. Two situations are distinguished:

- If we have no information about the scale $1 : s$ of the map on which the POIs are entered by the user (or if no map is used to enter POIs), then the latitude value lat (resp. longitude value lon) of the merged POI, resulting from the merging of the coreferent POIs POI_1, \dots, POI_n is obtained by taking the arithmetic mean of the latitudes $lat_i, i = 1, \dots, n$ (resp. longitudes lon_i) of all coreferent POIs, i.e.,

$$lat = \frac{\sum_{i=1,n} lat_i}{n}, \quad lon = \frac{\sum_{i=1,n} lon_i}{n}. \quad (32)$$

Alternatively, to eliminate the impact of outliers, the median of the latitudes (resp. longitudes) can be taken as merge result.

- If the POI locations have been inserted by users using maps, then we have scale information and only the latitudes (resp. longitudes) of the POIs at the most detailed scale are considered in the computation of the arithmetic means, i.e.,

$$lat = \frac{\sum_{\substack{i=1,n \\ s_i=s_{min}}} lat_i}{\sum_{\substack{i=1,n \\ s_i=s_{min}}} 1}, \quad lon = \frac{\sum_{\substack{i=1,n \\ s_i=s_{min}}} lon_i}{\sum_{\substack{i=1,n \\ s_i=s_{min}}} 1} \quad (33)$$

where $1 : s_i$ is the scale at which POI_i is entered and $s_{min} = \min\{s_i | i = 1, 2, \dots, n\}$.

This approach guarantees that only POIs that are entered at the scale with the highest precision among the scales that are used for the coreferent POIs under consideration are involved in the merge operation.

5.3.2 Merging of Descriptive Data

For descriptive, atomic POI components, a confidence-based merge function, as defined by Eq. (30) can be used. This is motivated by the assumption that the description for which the possible quantity of coreferent descriptions is maximised, is a

good candidate for the merge result. Indeed, because for this description we have the highest confidence that it is coreferent with most of the other descriptions.

On the one hand, by selecting only one description from the descriptions of the coreferent POIs, the risk for an inconsistent description in the merged POI is minimised as one could assume that users most likely provide consistent descriptions. However, on the other hand, by neglecting the descriptions of the non-selected POIs, information not present in the description of the selected POI might be lost. A solution for this is to apply a multi-document summarising technique to the descriptions of all coreferent POIs. Such summarising techniques have been described in [41, 15].

5.3.3 Merging of Categorical Data

For atomic POI components that contain categorical data, a confidence-based merge function, as defined by Eq. (30) can also be used. This is motivated by the assumption that the category for which the possible quantity of coreferent descriptions is maximised, is a good candidate for the merge result.

The underlying assumption at this point is that if different category labels are used in the coreferent POIs, these are most likely the result of user mistakes. Hence, keeping only the label for which the confidence is the highest might be a good merge strategy.

Alternatively, if the category labels are organised in a hierarchical structure, reflecting category-subcategory relationships, then the most common ancestor of the category labels in the coreferent POIs, might be taken as the merge result. In such a case, there is less chance for mistakes, but specific category label information might get lost.

6 An Illustrative Example

To illustrate the coreference detection and merging of POIs as described in the previous sections, the POIs of Example 2 are reconsidered. First we deal with coreference detection in Subsection 6.1, next in Subsection 6.2 the merging is illustrated.

6.1 Illustration of Coreference Detection

As illustrated in Example 1, the POIs under consideration are objects of a complex universe $O_{POI} = U_1 \times U_2 \times U_3 \times U_4 \times U_5 \times U_6$ that consists of six non compound universa U_1, \dots, U_6 of which only the five universa U_2, \dots, U_6 are relevant with respect to coreference detection. Indeed, the universe U_1 is used to model the identifier of a POI which by definition should be unique and which is either provided by the user or generated by the system. Hence it is assumed that the semantics of the identifier do not contribute to the coreference detection process. In the next example, we illustrate POI coreference detection on the basis of the universa U_2, \dots, U_6 .

Example 3. Consider the four POIs of Example 2 and assume that all of them have been entered by users using a map interface. POI_1 , POI_2 and POI_3 are entered at scale 1 : 10000 which corresponds to a street map of Ghent, whereas POI_4 is entered at scale 1 : 1000000 which corresponds to a map of Belgium. The latitude, longitude, scale, radius of screen precision, and parameter value for ε of these POIs (cf. Subsection 4.2.2) are summarised in Table 1. The minimum scale supported is assumed to be 1 : 10000. For all POIs, the same precision $d_s = 0.01m$ is used. This precision is assumed to be provided by the user (or could alternatively be set by default in the system).

Table 1 Information about the POIs used in Example 3

POI	lat	lon	1 : s	d_s	$\varepsilon = d_p s_{min}$
POI_1	51.056934	3.727112	1:10000	0.01m	2.8m
POI_2	51.053036	3.727015	1:10000	0.01m	2.8m
POI_3	51.053177	3.726382	1:10000	0.01m	2.8m
POI_4	51.033333	3.700000	1:1000000	0.01m	2.8m

We now present the calculation of the uncertainty of coreference for objects of each of the constituting relevant universa U_2, \dots, U_6 .

- **Coreference detection for objects of the universum U_2 .** This universum is used to model the name of the POI. As explained before, the uncertainty of coreference for names is preferably determined by means of a syntactical evaluator E_{name} . By using the evaluators described in [10, 12], the PTVs ($\mu_{\bar{p}}(T), \mu_{\bar{p}}(F)$) in Table 2 are obtained. These PTVs express the uncertainty about the coreference of the names of the POIs under consideration, i.e., POI_1 , POI_2 , POI_3 and POI_4 . From these results it can be seen that the names of POI_2 and POI_4 are certainly coreferent because reflexivity of the evaluator requires that equal object value are certain to be coreferent. In addition, other POI names are certainly not coreferent, due to a lack of sufficient syntactical similarities between names.

Table 2 Uncertainty about the coreference of the names of POI_x and POI_y

POI_x	POI_y	$E_{name}(POI_x, POI_y)$
POI_1	POI_2	(0,1)
POI_1	POI_3	(0,1)
POI_1	POI_4	(0,1)
POI_2	POI_3	(0,1)
POI_2	POI_4	(1,0)
POI_3	POI_4	(0,1)

- **Coreference detection for objects of the universa U_3 and U_4 .** Universa U_3 and U_4 are respectively used to model the latitude and longitude of a POI. In order to apply the techniques presented in Subsection 4.2 both the latitude and longitude of a POI have to be considered together. An evaluator $e_{location}$, which uses Eq. 19, is applied to compute the PTV that reflects the (un)certainty about the co-location of two POIs. Table 3 gives an overview of the results obtained from the application of the evaluator $E_{location}$ for the POIs under consideration. The third column gives the distances between the POIs as computed by using Eq. (13). The fourth column contains the values for the parameter δ as computed by using Eq. (18). Whereas the last column represents the resulting PTVs ($\mu_{\bar{p}}(T), \mu_{\bar{p}}(F)$) denoting the (un)certainty about the co-location of the POIs as obtained by applying Eq. 19.

Table 3 Uncertainty about the co-location of POI_x and POI_y

POI_x	POI_y	$d(POI_x, POI_y)$	$\delta = \varepsilon + \max(s_1 d_{s_1}, s_2 d_{s_2})$	$E_{location}(POI_x, POI_y)$
POI_1	POI_2	433.2m	102.8m	(0,1)
POI_1	POI_3	420.6m	102.8m	(0,1)
POI_1	POI_4	3235.2m	10002.8m	(1,0.48)
POI_2	POI_3	46.9m	102.8m	(1,0.79)
POI_2	POI_4	2890.8m	10002.8m	(1,0.41)
POI_3	POI_4	2874.1m	10002.8m	(1,0.40)

These results reflect that POI_1 is not co-located with POI_2 and POI_3 , which is reflected by the PTV (0, 1). Remind that it has been assumed in the example that POI_4 is entered at scale 1 : 1000000, which is less precise than scale 1 : 10000. This makes that there is no certainty about the co-location of POI_4 with POI_1 , POI_2 and POI_3 what is respectively reflected in the PTVs (1, 0.48), (1, 0.41) and (1, 0.40). Due to their possibilistic interpretation each of these PTVs expresses that it is either completely possible that there is co-location ($\mu_{\bar{p}}(T) = 1$) or that it is either to a lower extent possible that there is no co-location ($\mu_{\bar{p}}(F)$ resp. being equal to 0.48, 0.41 and 0.40). Likewise, the PTV (1, 0.79) expresses that it is either completely possible ($\mu_{\bar{p}}(T) = 1$) that POI_2 and POI_3 are co-located, or that it is either possible to a lower extent $\mu_{\bar{p}}(F) = 0.79$ that these are not co-located. This rather high value of 0.79 is due to the pessimistic assumption of ε being only 2.8m, where Saint-Bavo cathedral has a diagonal of about 110m. Alternatively, using Eq. (17), we obtain that $\varepsilon = 55m$ and applying Eq. (18) yields $\delta = 155m$. So, using this alternative approach, the resulting PTV becomes $\{(T, 1)\}$, what corresponds to true and illustrates the efficiency of Eq. (17).

- **Coreference detection for objects of the universum U_5 .** Universum U_5 is used to model the description of the POI. Similarly as for the name, the coreference

detection for the description of a POI is preferably done using a syntactical evaluator E_{descr} . The PTVs in Table 4 show the uncertainty about coreference of the descriptions of the POIs under consideration and are obtained by applying the evaluators that have been described in [10, 12]. From these results it follows that

Table 4 Uncertainty about the coreference of the descriptions of POI_x and POI_y

POI_x	POI_y	$E_{descr}(POI_x, POI_y)$
POI_1	POI_2	(0.5,1)
POI_1	POI_3	(0,1)
POI_1	POI_4	(0.3,1)
POI_2	POI_3	(1,0.1)
POI_2	POI_4	(1,0.1)
POI_3	POI_4	(1,0.1)

the POI description of POI_1 is certainly not coreferent with that of POI_3 (PTV (0, 1)). There is also higher confidence that this description is not coreferent with that of POI_2 ($\mu_{\bar{p}}(F) = 1$ in PTV (0.5, 1)) and POI_4 ($\mu_{\bar{p}}(F) = 1$ in PTV (0.3, 1)) than there is confidence that the description of POI_1 is coreferent with the description of POI_2 ($\mu_{\bar{p}}(T) = 0.5$ in PTV (0.5, 1)) and the description of POI_4 ($\mu_{\bar{p}}(T) = 0.3$ in PTV (0.3, 1)). Furthermore, there is higher confidence that the descriptions of POI_2 , POI_3 and POI_4 are coreferent (PTVs (1, 0.1)) than there is confidence that these descriptions are not coreferent.

- **Coreference detection for objects of the universum U_6 .** Universum U_6 is used to model the category class of the POI. As opposed to the POI name and description, the type of the POIs is compared in a semantical manner. Therefore, a binary relation R between POI types is constructed dynamically as described in [11]. Then, based on this binary relation, uncertainty about category values can be inferred using the semantic evaluator $E_{category}$. Table 5 presents the results of these computations.

As can be seen, because of the PTV (0,1) the category value of POI_1 ('Market') is not coreferent with the category values of POI_2 ('Church'), POI_3 ('Cathedral') and POI_4 ('Cathedral'). The category values of POI_3 and POI_4 are the same ('Cathedral') and therefore coreferent, what is reflected by the PTV (1, 0). Furthermore, the category value of POI_3 and POI_4 ('Cathedral') is connected through an 'is-a' relation with the category value of POI_2 ('Church'). This connection is reflected in the binary relation R (not shown in the chapter) and resulted in a PTV (1, 0.5) describing that there is higher confidence that the value 'Church' is related to the value 'Cathedral' ($\mu_{\bar{p}}(T) = 1$ in PTV (1, 0.5)) than there is confidence that both values are not coreferent ($\mu_{\bar{p}}(F) = 0.5$ in PTV (1, 0.5)).

Table 5 Uncertainty about the coreference of the category values of POI_x and POI_y

POI_x	POI_y	$E_{category}(POI_x, POI_y)$
POI_1	POI_2	(0,1)
POI_1	POI_3	(0,1)
POI_1	POI_4	(0,1)
POI_2	POI_3	(1,0.5)
POI_2	POI_4	(1,0.5)
POI_3	POI_4	(1,0)

Finally, given the above uncertainties about the coreference for all the objects of the universa U_2, \dots, U_6 (i.e., marginal possibilities), the uncertainty about the coreference of POIs can be calculated using a complex evaluator E_{POI} . For that purpose, an aggregation technique based on the Sugeno integral is used. As has been proposed in [8], such an approach requires two necessity measures γ^T and γ^F . These fuzzy measure γ^T (resp. γ^F) evaluates subsets of POI attributes and expresses the necessity that coreference of the values of the attributes in the set implies coreference (resp. does not imply coreference) of the POIs containing those values. The necessity measures used in this example are given as shown in Table 6. The given measures reflect that marginal knowledge about less than three attributes is considered to provide us with no necessity at all about the coreference of the POIs. However, marginal knowledge of three or more attributes allows us to infer necessity about (non) coreference. Note that the fuzzy measures satisfy the normalisation constraint:

$$\forall L \in \mathcal{L} : \min(\gamma^T(L), \gamma^F(\bar{L})) = 0. \tag{34}$$

Combining the conditional necessity as given in Table 6 with the marginal necessities that can be derived from the marginal PTVs from Tables 2, 3, 4 and 5 is then done using the Sugeno integral for PTVs, which is defined by Eq. (22). Applying the Sugeno integral with the PTVs from Tables 2, 3, 4 and 5 leads to the aggregated PTVs shown in Table 7. More details about (the use of) the Sugeno integral are given in [8]. □

6.2 Illustration of Merging

Reconsider the POIs of Example 2. Based on the coreference detection results presented in Table 7, we can safely conclude that POI_2 , POI_3 and POI_4 are coreferent (to some extent). In the next example we illustrate the merging of these three coreferent POIs.

Example 4. By applying the techniques described in Section 5, the merging of coreferent POIs is done in two steps. In the first step, merge functions for the relevant universa U_2, \dots, U_6 are specified and applied.

Table 6 The given necessity measures γ^T and γ^F used in the Sugeno integral in order to reflect how conditional knowledge about the values of POI attribute subsets leads to knowledge about the coreference of the POIs

$L \subseteq \mathcal{L}$	$\gamma^T(L)$	$\gamma^F(L)$
\emptyset	0	0
{name}	0	0
{location}	0	0
{description}	0	0
{type}	0	0
{name, location}	0	0
{name, description}	0	0
{name, type}	0	0
{location, description}	0	0
{location, type}	0	0
{description, type}	0	0
{name, location, description}	0.9	1
{name, location, type}	0.6	1
{name, description, type}	0	1
{location, description, type}	0.8	1
{name, location, description, type}	1	1

Table 7 Overall uncertainty about the coreference of POI_x and POI_y

POI_x	POI_y	$E_{POI}(POI_x, POI_y)$
POI_1	POI_2	(0,1)
POI_1	POI_3	(0,1)
POI_1	POI_4	(0.3,1)
POI_2	POI_3	(1,0.79)
POI_2	POI_4	(1,0.41)
POI_3	POI_4	(1,0.40)

- **Merging of objects of the universum U_2 .** Objects of the universe U_2 represent POI names. The names of the coreferent POIs POI_2 , POI_3 and POI_4 are respectively, ‘St-Bavo’, ‘Ghent cathedral’ and ‘St-Bavo’. Reconsider the PTVs obtained from the coreference detection of POI names given in Table 2. By applying Eq. (30), it is obtained that the name with the largest possible quantity of coreferent names is ‘St-Bavo’.

Indeed, for each coreferent POI POI_i , the corresponding fuzzy number $\pi_{\mathbb{N}}^{POI_i}$ is obtained as follows:

– For POI_2 :

$$\begin{aligned} E_{name}(POI_2, POI_2) &= (1, 0) \\ E_{name}(POI_2, POI_3) &= (0, 1) \\ E_{name}(POI_2, POI_4) &= (1, 0). \end{aligned}$$

This allows us to construct the multiset $\tilde{P} = \{(1, 0), (1, 0), (0, 1), (1, 0)\}$ where the first POI $(1, 0)$ is added to obtain a correct modelling for $\pi_{\mathbb{N}}^{POI_i}(0)$. Applying Eq. (8) yields the ordered list of POIs

$$[(1, 0), (1, 0), (1, 0), (0, 1)].$$

Applying Eq. (27) then yields

$$\begin{aligned} \pi_{\mathbb{N}}^{POI_2}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_2}(1) &= \min(1, 0) = 0 \\ \pi_{\mathbb{N}}^{POI_2}(2) &= \min(1, 1) = 1 \\ \pi_{\mathbb{N}}^{POI_2}(3) &= 0. \end{aligned}$$

– For POI_3 :

$$\begin{aligned} E_{name}(POI_3, POI_3) &= (1, 0) \\ E_{name}(POI_3, POI_2) &= (0, 1) \\ E_{name}(POI_3, POI_4) &= (0, 1). \end{aligned}$$

This allows us to construct the extended multiset $\tilde{P} = \{(1, 0), (1, 0), (0, 1), (0, 1)\}$ and the ordered list of POIs

$$[(1, 0), (1, 0), (0, 1), (0, 1)].$$

Applying Eq. (27) then yields

$$\begin{aligned} \pi_{\mathbb{N}}^{POI_3}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_3}(1) &= \min(1, 1) = 1 \\ \pi_{\mathbb{N}}^{POI_3}(2) &= \min(0, 1) = 0 \\ \pi_{\mathbb{N}}^{POI_3}(3) &= 0. \end{aligned}$$

– For POI_4 :

$$\begin{aligned} E_{name}(POI_4, POI_4) &= (1, 0) \\ E_{name}(POI_4, POI_2) &= (1, 0) \\ E_{name}(POI_4, POI_3) &= (0, 1). \end{aligned}$$

This yields the extended multiset $\tilde{P} = \{(1, 0), (1, 0), (1, 0), (0, 1)\}$ and the ordered list of POIs

$$[(1, 0), (1, 0), (1, 0), (0, 1)].$$

Applying Eq. (27) then yields

$$\begin{aligned}\pi_{\mathbb{N}}^{POI_4}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_4}(1) &= \min(1, 0) = 0 \\ \pi_{\mathbb{N}}^{POI_4}(2) &= \min(1, 1) = 1 \\ \pi_{\mathbb{N}}^{POI_4}(3) &= 0.\end{aligned}$$

Applying Eq. (28) results in

$$\pi_{\mathbb{N}}^{POI_3} \prec_{\text{sup}} \pi_{\mathbb{N}}^{POI_2} \text{ and } \pi_{\mathbb{N}}^{POI_3} \prec_{\text{sup}} \pi_{\mathbb{N}}^{POI_4}$$

Such that using Eq. (30) returns

$$\bar{\omega}_{\text{name}}(\{POI_2, POI_3, POI_4\}) = \arg \max_{u \in \{POI_2, POI_3, POI_4\}} \pi_{\mathbb{N}}^u = POI_2 \text{ or } POI_4.$$

Hence the name with the largest possible quantity of coreferent names is the name of POI_2 or POI_4 , which is in both cases ‘St-Bavo’. So, the merged value for the POI name is ‘St-Bavo’. As a side effect of this merge technique the (less specific) information ‘Ghent cathedral’ is lost.

- **Merging of objects of the universa U_3 and U_4 .** Universa U_3 and U_4 together model the location of a POI. These universa were handled together in the coreference detection process. Recall from Table 1 that POI_2 and POI_3 have been entered at scale 1 : 10000, whereas POI_4 has been entered at a less detailed map scale 1 : 1000000. Because we have scale information and not all coreferent POIs have been entered at the same scale, Eq. (33) can be used to compute the latitude and longitude value of the merged POI. Hereby, only the information related to the most detailed scale, i.e., the data from POIs POI_2 and POI_3 , are considered. Using the data given in Table 1, this yields

$$lat = \frac{51.053036 + 51.053177}{2} = 51.053106$$

and

$$lon = \frac{3.727015 + 3.726382}{2} = 3.726699.$$

The differences between the latitude and longitude of POI_2 and POI_3 are inherent to the fact that both POIs are representing (the geographical area of) St.-Bavo cathedral, which has a diagonal of about 110m, at a scale with a precision of 0.01m.

- **Merging of objects of the universum U_5 .** Objects of the universe U_5 represent POI descriptions. The descriptions of the coreferent POIs POI_2 , POI_3 and POI_4 are respectively, ‘St-Bavo’s Cathedral, Ghent’, ‘St-Bavo Cathedral’ and ‘St-Bavo – Ghent’. The same technique as previously used for POI names can be applied. Hence, Eq. (30) can now be applied with the PTVs obtained from the coreference detection of POI descriptions given in Table 4.

For each coreferent POI POI_i , the corresponding fuzzy number $\pi_{\mathbb{N}}^{POI_i}$ is obtained as follows:

- For POI_2 :

$$\begin{aligned} E_{descr}(POI_2, POI_2) &= (1, 0) \\ E_{descr}(POI_2, POI_3) &= (1, 0.1) \\ E_{descr}(POI_2, POI_4) &= (1, 0.1). \end{aligned}$$

This allows us to construct the extended multiset $\tilde{P} = \{(1, 0), (1, 0), (1, 0.1), (1, 0.1)\}$ and the ordered list of POIs

$$[(1, 0), (1, 0), (1, 0.1), (1, 0.1)].$$

Applying Eq. (27) then yields

$$\begin{aligned} \pi_{\mathbb{N}}^{POI_2}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_2}(1) &= \min(1, 0.1) = 0.1 \\ \pi_{\mathbb{N}}^{POI_2}(2) &= \min(1, 0.1) = 0.1 \\ \pi_{\mathbb{N}}^{POI_2}(3) &= 1. \end{aligned}$$

- For POI_3 :

$$\begin{aligned} E_{descr}(POI_3, POI_3) &= (1, 0) \\ E_{descr}(POI_3, POI_2) &= (1, 0.1) \\ E_{descr}(POI_3, POI_4) &= (1, 0.1). \end{aligned}$$

This yields the extended multiset $\tilde{P} = \{(1, 0), (1, 0), (1, 0.1), (1, 0.1)\}$ and the ordered list of POIs

$$[(1, 0), (1, 0), (1, 0.1), (1, 0.1)].$$

Applying Eq. (27) then yields

$$\begin{aligned} \pi_{\mathbb{N}}^{POI_3}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_3}(1) &= \min(1, 0.1) = 0.1 \\ \pi_{\mathbb{N}}^{POI_3}(2) &= \min(1, 0.1) = 0.1 \\ \pi_{\mathbb{N}}^{POI_3}(3) &= 1. \end{aligned}$$

– For POI_4 :

$$\begin{aligned} E_{descr}(POI_4, POI_4) &= (1, 0) \\ E_{descr}(POI_4, POI_2) &= (1, 0.1) \\ E_{descr}(POI_4, POI_3) &= (1, 0.1). \end{aligned}$$

This yields the extended multiset $\tilde{P} = \{(1, 0), (1, 0), (1, 0.1), (1, 0.1)\}$ and the ordered list of POIs

$$[(1, 0), (1, 0), (1, 0.1), (1, 0.1)].$$

Applying Eq. (27) then yields

$$\begin{aligned} \pi_{\mathbb{N}}^{POI_4}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_4}(1) &= \min(1, 0.1) = 0.1 \\ \pi_{\mathbb{N}}^{POI_4}(2) &= \min(1, 0.1) = 0.1 \\ \pi_{\mathbb{N}}^{POI_4}(3) &= 1. \end{aligned}$$

Using Eq. (30) returns

$$\bar{\omega}_{descr}(\{POI_2, POI_3, POI_4\}) = \arg \max_{u \in \{POI_2, POI_3, POI_4\}} \pi_{\mathbb{N}}^u = POI_2 \text{ or } POI_3 \text{ or } POI_4.$$

Hence, all three descriptions qualify as the description with the largest possible quantity of coreferent descriptions. A choice has to be made. Considering the fact that we want to minimise information loss, the description which consists of most characters will be chosen in such a case. So, the merged value for description becomes ‘St-Bavo’s Cathedral, Ghent’.

- **Merging of objects of the universum** U_6 . Objects of the universe U_6 represent the categorical data about the POI. For categorical data, the same confidence-based merge technique as used before is applied. The POI categories in the coreferent POIs POI_2 , POI_3 and POI_4 are respectively, ‘Church’, ‘Cathedral’ and ‘Cathedral’. Using the PTVs obtained from the coreference detection of POI (category) types given in Table 5 yields that the type with the largest possible quantity of coreferent types is ‘Cathedral’. This follows from the following computations.

For each coreferent POI POI_i , the corresponding fuzzy number $\pi_{\mathbb{N}}^{POI_i}$ is obtained as follows:

– For POI_2 :

$$\begin{aligned} E_{category}(POI_2, POI_2) &= (1, 0) \\ E_{category}(POI_2, POI_3) &= (1, 0.5) \\ E_{category}(POI_2, POI_4) &= (1, 0.5). \end{aligned}$$

This allows us to construct the extended multiset $\tilde{P} = \{(1,0), (1,0), (1,0.5), (1,0.5)\}$ and the ordered list of POIs

$$[(1,0), (1,0), (1,0.5), (1,0.5)].$$

Applying Eq. (27) then yields

$$\begin{aligned}\pi_{\mathbb{N}}^{POI_2}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_2}(1) &= \min(1, 0.5) = 0.5 \\ \pi_{\mathbb{N}}^{POI_2}(2) &= \min(1, 0.5) = 0.5 \\ \pi_{\mathbb{N}}^{POI_2}(3) &= 1.\end{aligned}$$

– For POI_3 :

$$\begin{aligned}E_{category}(POI_3, POI_3) &= (1, 0) \\ E_{category}(POI_3, POI_2) &= (1, 0.5) \\ E_{category}(POI_3, POI_4) &= (1, 0).\end{aligned}$$

This allows us to construct the extended multiset $\tilde{P} = \{(1,0), (1,0), (1,0.5), (1,0)\}$ and the ordered list of POIs

$$[(1,0), (1,0), (1,0), (1,0.5)].$$

Applying Eq. (27) then yields

$$\begin{aligned}\pi_{\mathbb{N}}^{POI_3}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_3}(1) &= \min(1, 0) = 0 \\ \pi_{\mathbb{N}}^{POI_3}(2) &= \min(1, 0.5) = 0.5 \\ \pi_{\mathbb{N}}^{POI_3}(3) &= 1.\end{aligned}$$

– For POI_4 :

$$\begin{aligned}E_{category}(POI_4, POI_4) &= (1, 0) \\ E_{category}(POI_4, POI_2) &= (1, 0.5) \\ E_{category}(POI_4, POI_3) &= (1, 0).\end{aligned}$$

This allows us to construct the extended multiset $\tilde{P} = \{(1,0), (1,0), (1,0.5), (1,0)\}$ and the ordered list of POIs

$$[(1,0), (1,0), (1,0), (1,0.5)].$$

Applying Eq. (27) then yields

$$\begin{aligned}\pi_{\mathbb{N}}^{POI_4}(0) &= 0 \\ \pi_{\mathbb{N}}^{POI_4}(1) &= \min(1, 0) = 0 \\ \pi_{\mathbb{N}}^{POI_4}(2) &= \min(1, 0.5) = 0.5 \\ \pi_{\mathbb{N}}^{POI_4}(3) &= 1.\end{aligned}$$

Such that using Eq. (30) returns

$$\bar{\omega}_{category}(\{POI_2, POI_3, POI_4\}) = \arg \max_{u \in \{POI_2, POI_3, POI_4\}} \pi_{\mathbb{N}}^u = POI_3 \text{ or } POI_4.$$

Thus, the type with the largest possible quantity of coreferent types is the type of POI_3 or POI_4 , which is in both cases ‘Cathedral’. Hence, the incorrect category value ‘Church’ is neglected by the merge strategy.

In the second step, the results of the previous merge operations are combined using the composite merge function given by Eq. (31). The resulting merged POI then finally becomes

$$(POI_{merge}, \text{‘St-Bavo’}, 51.053106, 3.726699, \text{‘St-Bavo’s Cathedral, Ghent’}, \text{‘Cathedral’}).$$

This POI gives a consistent description of St-Bavo’s cathedral. \square

The case study presented above is limited, though chosen for exemplifying the presented coreference detection and merging mechanisms. Other, more specific and statistically relevant tests, covering more extended data sets, have been performed and published in [14]. These tests proof the efficiency of the presented methods in terms of precision and recall as compared to the other methods presented in the literature.

7 Conclusions and Further Work

7.1 Contribution

In this chapter, a novel soft computing approach to cleanse POI databases is described. In essence, this approach consists of two parts. In the first part, the uncertainty about the potential coreference of two POIs is estimated and subsets of potentially coreferent POIs are identified (two POIs are considered to be coreferent if they describe the same geographical location or object at a geographical location). In the second part, coreferent POIs are merged into a new POI which acts as a representation of all information present in the coreferent POIs.

At the basis of the approach is the concept of evaluators for coreference detection. Such an evaluator takes two objects as input and returns an estimation of the (un)certainly that these objects are coreferent, expressed by means of a possibilistic

truth value (PTV). Evaluators have been proposed for atomic objects, co-location detection and complex objects.

The specific evaluators for co-location detection are especially suited for cases where latitude and longitude coordinates of POIs are entered by users using a map interface, which is often the case with POI databases that are maintained by a user community. The evaluators allow to explicitly cope with the context (scale and precision) with which the locational data have been entered. Fuzzy ranges are used to determine in a flexible way whether two POI locations can be considered to be close enough to conclude that they are co-located.

Coreferent POIs are merged using merge functions. A merge function takes a finite number of objects as input and returns a (new) object that acts as a representation of the input objects. Merge functions have been proposed for atomic objects and complex objects. The presented merge functions for atomic objects are based on an evaluator. Complex objects are merged using a composite merge function. Typical for composite merge functions is that they do not preserve any of the coreferent POIs, but combine the best (most confident) parts of each of them to construct a novel, merged POI.

7.2 Context

The presented work contributes to research on data quality issues in information retrieval. On the one hand it offers automatic data cleansing techniques which could be developed further and generalised in order to improve data quality in information sources. Information retrieval processes could benefit from an improved data quality and provide better results as the data quality will be propagated in data processing results.

On the other hand such data cleansing techniques can also be applied to cleanse the results of information retrieval operations that run on unclean data. Coreference detection techniques can be used to detect coreferent results, which in their turn eventually can be merged using merging techniques.

Moreover, the computed uncertainty measures obtained from the coreference detection can be communicated to the users as an indication of the quality of the retrieval results.

The presented approach is based on soft computing techniques and allows to reflect human reasoning with respect to coreference detection and object merging in an adequate way. This leads to more justifiable results as compared to those obtained by using existing approaches. This is the main advantage of the proposed approach. Statistically relevant experiments on different data sets reported in [14] reveal that the proposed techniques for coreference detection overall perform better in terms of precision and recall than the related techniques that were mentioned in Section 2. More extended tests to validate the performance of the proposed merge techniques are required and are currently under development. Note that such tests are more difficult to implement as the ground truth for object merging is much more difficult to obtain.

7.3 Further Work

Further research is required and planned. The techniques presented in this chapter have been specifically developed for the cleansing of POI databases. An important aspect that will be further investigated is the generalisation of the approach so that it will become applicable for the cleansing of other, more general databases. For that purpose, among others, it is worth investigating whether other aggregation techniques like, e.g., the technique used in logic scoring of preference (LSP) which is based on the generalized conjunction/disjunction (GCD) function [23], offer better aggregation facilities for coreference detection than the approach based on the Sugeno integral. Furthermore, the desired mathematical properties of merge functions should be better understood and new families of merging functions able to model different kinds of desired behavior should be developed. For example, in some cases it might be preferable to keep as much information as possible in the resulting merged object. In such cases, rather than selecting the most confident part, the merging function should concatenate, summarise or combine all available data in an intelligent way.

Another aspect to investigate further concerns the optimization of the object comparison technique. Optimization is possible as not all pairs in a set of objects must necessarily be checked to detect all coreferent objects. Moreover, not all components of a complex object must necessarily in all cases be evaluated to come to a conclusion regarding coreference.

References

1. Baral, C., Kraus, S., Minker, J.: Combining multiple knowledge bases. *IEEE Transactions on Knowledge and Data Engineering* 3(2), 208–220 (1991)
2. Baral, C., Kraus, S., Minker, J., Subrahmanian, V.: Combining knowledge bases consisting of first-order theories. *Computational Intelligence* 8(1), 45–71 (1992)
3. Bloch, I.: Information Combination Operators for Data Fusion: A Comparative Review with Classification. *IEEE Transactions on Systems, Man and Cybernetics, Part A* 26(1), 52–67 (1996)
4. Borgida, A., Imielinski, T.: Decision making in committees – a framework for dealing with inconsistency and non-monotonicity. In: *Proc. of the Workshop of Nonmonotonic Reasoning*, pp. 21–32 (1984)
5. Bordogna, G., Pagani, M., Pasi, G.: Imperfect Multisource Spatial Data Fusion Based on a Local Consensual Dynamics. In: Kacprzyk, J., Petry, F.E., Yazici, A. (eds.) *Uncertainty Approaches for Spatial Data Modeling and Processing: A Decision Support Perspective*. *SCI*, vol. 271, pp. 79–94. Springer, Heidelberg (2010)
6. Boury-Brisset, A.-C.: Ontology-based approach for information fusion. In: *Proc. of the 6th International Conference on Information Fusion*, pp. 522–529 (2003)
7. Bright, M., Hurson, A., Pakzad, S.: A taxonomy and current issues in multidatabase systems. *Computer* 25(3), 50–59 (1992)
8. Bronselaer, A., Hallez, A., De Tré, G.: Extensions of Fuzzy Measures and Sugeno Integral for Possibilistic Truth Values. *International Journal of Intelligent Systems* 24(2), 97–117 (2009)

9. Bronselaer, A., Hallez, A., De Tré, G.: A possibilistic view on set and multiset comparison. *Control and Cybernetics* 38(2), 341–366 (2009)
10. Bronselaer, A., De Tré, G.: A possibilistic approach to string comparison. *IEEE Transactions on Fuzzy systems* 17(1), 208–223 (2009)
11. Bronselaer, A., De Tré, G.: Semantical evaluators. In: *Proc. of the IFSA 2009 International Conference*, Lisbon, Portugal, pp. 663–668 (2009)
12. Bronselaer, A., De Tré, G.: Properties of possibilistic string comparison. *IEEE Transactions on Fuzzy systems* 18(2), 312–325 (2010)
13. Bronselaer, A., De Tré, G.: Aspects of object merging. In: *Proc. of the NAFIPS 2010 International Conference*, Toronto, Canada, pp. 27–32 (2010)
14. Bronselaer, A. (2010) Coreferency of atomic and complex objects. PhD thesis, Ghent University, Ghent, Belgium (December 2010) (in Dutch)
15. Bronselaer, A., Van Britsom, D., De Tré, G.: A framework for multiset merging. *Fuzzy Sets and Systems* 191, 1–20 (2012)
16. Carrara, P., Bordogna, G., Boschetti, M., Brivio, P.A., Nelson, A., Stroppiana, D.: A flexible multi-source spatial-data fusion system for environmental status assessment at continental scale. *International Journal of Geographical Information Science* 22(7), 781–799 (2008)
17. De Cooman, G.: *Evaluatieverzamelingen en - afbeeldingen. Een ordetheoretische benadering van vaagheid en onzekerheid*. Ph.D. dissertation, Ghent University, Belgium (1993) (in Dutch)
18. De Cooman, G.: Towards a possibilistic logic. In: Ruan, D. (ed.) *Fuzzy Set Theory and Advanced Mathematical Applications*, pp. 89–133. Kluwer Academic, Boston (1995)
19. Destercke, S., Dubois, D., Chojnacki, E.: Possibilistic information fusion using maximal coherent subsets. *IEEE Transactions on Fuzzy Systems* 17(1), 79–92 (2009)
20. De Tré, G., Bronselaer, A., Matthé, T., Van de Weghe, N., De Maeyer, P.: Consistently Handling Geographical User Data: Context-Dependent Detection of Co-located POIs. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) *IPMU 2010, Part II. CCIS*, vol. 81, pp. 85–94. Springer, Heidelberg (2010)
21. Dubois, D., Prade, H.: *Possibility Theory*. Plenum Press, New York (1988)
22. Dubois, D., Prade, H. (eds.): *Fundamentals of Fuzzy Sets*. Kluwer Academic Publishers, Boston (2000)
23. Dujmović, J., Larsen, H.L.: Generalized conjunction/disjunction. *International Journal of Approximate Reasoning* 46(3), 423–446 (2007)
24. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226–231. AAAI Press (1996)
25. Fellegi, I., Sunter, A.: A Theory for Record Linkage. *American Statistical Association Journal* 64(328), 1183–1210 (1969)
26. Federal Geographic Data Committee, Content standard for digital geospatial metadata. FGDC-STD-001-1998, Washington D.C., USA (1998)
27. Foley, H., Petry, F.: Fuzzy Knowledge-Based System for Performing Conflation in Geographical Information Systems. In: Loganathara, R., Palm, G., Ali, M. (eds.) *IEA/AIE 2000. LNCS (LNAI)*, vol. 1821, pp. 260–269. Springer, Heidelberg (2000)
28. Hallez, A., De Tré, G., Verstraete, J., Matthé, T.: Application of fuzzy quantifiers on possibilistic truth values. In: *Proc. of the Eurofuse Workshop on Data and Knowledge Engineering*, Warsaw, Poland, pp. 252–254 (2004)
29. Jaro, M.: *Unimatch: A record linkage system: User's manual*. US Bureau of the Census, Technical Report (1976)

30. Konieczny, S., Pérez, R.: Merging information under constraints: a logical framework. *Journal of Logic and Computation* 12(1), 111–120 (2002)
31. Lin, J., Mendelzon, A.: Knowledge base merging by majority. In: Pareshi, R., Fronhöfer, B. (eds.) *Dynamic Worlds: From the Frame Problem to Knowledge Management*, pp. 195–218. Kluwer Academic, Boston (1994)
32. Lin, J., Mendelzon, A.: Merging databases under constraints. *International Journal of Cooperative Information Systems* 7(1), 55–76 (1998)
33. Nachouki, G., Ouafafou, M.: Multi-data source fusion. *Information Fusion* 9(4), 523–537 (2008)
34. National Geospatial Intelligence Agency (NGA), Geodetic System 1984: Its Definitions and Relationships with Local Geodetic Systems. NIMA Technical Report 8350.2 (2004)
35. Prade, G.: Possibility sets, fuzzy sets and their relation to Lukasiewicz logic. In: *Proc. of the International Symposium on Multiple-Valued Logic*, pp. 223–227 (1982)
36. Rahimi, S., Cobb, M., Ali, D., Paprzycki, M., Petry, F.E.: A Knowledge-Based Multi-Agent System for Geospatial Data Conflation. *Journal of Geographic Information and Decision Analysis* 6(2), 67–81 (2002)
37. Rodríguez, M.A., Egenhofer, M.J.: Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science* 18, 229–256 (2004)
38. Sandri, S., Dubois, D., Kalfsbeek, H.: Elicitation, assessment and pooling of expert judgements using possibility theory. *IEEE Transactions on Fuzzy Systems* 3(3), 313–335 (1995)
39. Sinnott, R.W.: Virtues of the Haversine. *Sky and Telescope* 68(2), 159 (1984)
40. Torres, R., Keller, G.R., Kreinovich, V., Longpré, L., Starks, S.A.: Eliminating Duplicates under Interval and Fuzzy Uncertainty: An Asymptotically Optimal Algorithm and Its Geospatial Applications. *Reliable Computing* 10(5), 401–422 (2004)
41. Van Britsom, D., Bronselaer, A., De Tré, G.: Automatically Generating Multi-Document Summarizations. In: *Proc. of the 11th International Conference on Intelligent Systems Design and Applications*, Cordoba, Spain, pp. 142–147 (2011)
42. Van Schooten, A.: *Ontwerp en implementatie van een model voor de representatie en manipulatie van onzekerheid en imprecisie in databanken en expert systemen*. Ph.D. dissertation, Ghent University, Belgium (1988) (in Dutch)
43. Winkler, W.E.: *The State of Record Linkage and Current Research Problems*. R99/04, Statistics of Income Division, U.S. Census Bureau (1999)
44. Yager, R.: On the theory of bags. *International Journal of General Systems* 13(1), 23–27 (1986)
45. Yager, R.: On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man and Cybernetics* 18(1), 183–190 (1988)
46. Yager, R., Rybalov, A.: Uninorm aggregation operators. *Fuzzy Sets and Systems* 80(1), 111–120 (1996)
47. Zadeh, L.A.: Fuzzy Sets. *Information and Control* 8(3), 338–353 (1965)
48. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* 1, 3–28 (1978)

Chapter 5

A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web

Andrea Ballatore, David C. Wilson, and Michela Bertolotto

Abstract. Over the past decade, rapid advances in web technologies, coupled with innovative models of spatial data collection and consumption, have generated a robust growth in geo-referenced information, resulting in spatial information overload. Increasing ‘geographic intelligence’ in traditional text-based information retrieval has become a prominent approach to respond to this issue and to fulfill users’ spatial information needs. Numerous efforts in the Semantic Geospatial Web, Volunteered Geographic Information (VGI), and the Linking Open Data initiative have converged in a constellation of open knowledge bases, freely available online. In this article, we survey these open knowledge bases, focusing on their geospatial dimension. Particular attention is devoted to the crucial issue of the quality of geo-knowledge bases, as well as of crowdsourced data. A new knowledge base, the OpenStreetMap Semantic Network, is outlined as our contribution to this area. Research directions in information integration and Geographic Information Retrieval (GIR) are then reviewed, with a critical discussion of their current limitations and future prospects.

1 Introduction

In 1998, U.S. Vice President Al Gore delivered a speech at the California Science Center about what he named Digital Earth, a “multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of

Andrea Ballatore · Michela Bertolotto
School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin 4, Ireland
e-mail: {andrea.ballatore, michela.bertolotto}@ucd.ie

David C. Wilson
Department of Software and Information Systems, University of North Carolina,
9201 University City Boulevard, Charlotte, NC 28223-0001, USA
e-mail: davils@uncc.edu

geo-referenced data” [47, p. 89]. Much of the unprecedented amount of information produced and released on the Internet is about a specific place on the Earth. However, Gore pointed out, most of this informational wealth is generated and left untapped. Among the key aspects that would enable a more efficient exploitation of geo-data, interoperability and metadata were considered of particular importance. Multiple data sources should be combined together using a common framework, and metadata should describe online resources in a clear, standardised way [47].

Over the past 14 years, several geospatial initiatives have been undertaken, oriented towards the implementation of the Digital Earth [45]. The Open Geospatial Consortium (OGC)¹ has defined and promoted several standards to distribute geographic data, while the Global Spatial Data Infrastructure (GSDI) association aims at fostering “spatial data infrastructures that support sustainable social, economic, and environmental systems integrated from local to global scales” [1, p. 1]. Despite these efforts, standard formats are often ignored in favour of application-specific formats. As Fonseca et al. put it, heterogeneity emerges spontaneously in a free market of ideas and products, and standards cannot reduce it by decree [33].

The spectacular growth of unstructured information online has affected all domains, prompting Tim Berners-Lee to envisage the advent of the so-called Semantic Web [10]. The Semantic Web project aims to develop a standard semantic format to describe online data, originating a network of machine-readable, semantically clear documents. Data semantics is expressed in predicate logic-based languages such as RDF,² in large collections of statements about real world entities. This vision was further formalised through the Linked Data initiative, which promotes the release of datasets in an inter-connected web of semantic data [12].

The information explosion in geographic data has not only been quantitative, but also qualitative [113]. With the rise of Web 2.0, Internet users have become active producers of geo-referenced information, utilising collaborative web tools in large projects [85]. Several collaborative efforts emerged to create and maintain large datasets, resulting in crowdsourcing, impacting initially on non-spatial information and subsequently also on the geographic domain [61]. In the geospatial context, the term ‘neogeography’ has been used in order to refer to this rapid and complex nexus of technological and social practices [106]. Goodchild termed the crowdsourcing of geographic information as Volunteered Geographic Information (VGI), emphasising its production through voluntary labour [44]. Haklay et al. have surveyed VGI projects [53], while Coleman et al. have discussed the practices and motivations of ‘producers’, users/producers of geographic data [25]. In addition, Sui used the term ‘wikification’ to describe the practice of crowdsourcing of non-textual data, emulating the Wikipedia model in the geographic domain [104].

The impact of neogeography is not restricted to non-profit, academic organisations. Private institutions such as Google, Microsoft, and Yahoo! are progressively offering facilities for sharing geo-data, expanding their services beyond the

¹ <http://www.opengeospatial.org> (acc. June 5, 2012).

² <http://www.w3.org/RDF> (acc. June 5, 2012).

routing systems that dominated the first phase of web-based Geographic Information Systems (GIS).³ In this sense, geo-wikification is identifiable in the growth of web services allowing users, with some degree of freedom, to create or edit spatial data. As Priedhorsky notes, however, most interactive geo-services are essentially ‘digital graffiti,’ i.e. annotations on a static geographic image [90]. Beyond the specificities of each case, it can be argued that all neogeographic and VGI phenomena share the characteristics of being volunteered, crowdsourced, wikified, and web-based.

Even though the popular claim that 80% of information is geo-referenced has been questioned [51], it can be stated safely that, over the past decade, geo-information has experienced a remarkable growth [71]. As happened in other fields subject to an information explosion and subsequently to information overload, the issue of semantics of geo-data – or lack thereof – has become critical. The deluge of semantically ambiguous geo-data caused Egenhofer to advocate the emergence of a Semantic Geospatial Web, a spatial extension of the Semantic Web [29]. In Egenhofer’s view, this new framework for geospatial information retrieval should rely on the semantics of spatial and terminological ontologies. Thanks to inter-operable semantic representations of the data, the Semantic Geospatial Web will increase the relevance and quality of results in geographic retrieval systems.

As a result of the synergy between crowdsourcing, VGI, and the Semantic Geospatial Web, several large-scale collaborative projects have emerged. While Wikipedia⁴ is without doubt the most visible text-based crowdsourcing project, OpenStreetMap (OSM) has applied the wiki model to create an open world vector map [54]. Several geo-knowledge bases have then been created by structuring existing datasets into Semantic Web formats: the projects LinkedGeoData, GeoNames, and GeoWordNet are salient examples [6, 41]. Research efforts have been undertaken on the development, maintenance, and merging of open geo-knowledge bases, to enhance the geographic intelligence of information retrieval systems, beyond the traditional text-based techniques [33, 113, 6, 31].

Moreover, GIR has attempted to increase the geographic awareness of text-based information retrieval systems. On top of traditional flat gazetteers (dictionaries of toponyms and geo-coordinates), GIR has started exploiting geo-knowledge bases to reduce the ambiguity of geographic terms and enable spatial reasoning [67, 86]. Despite these efforts, the knowledge contained in such computational artifacts is left largely untapped. We believe that these open geo-knowledge bases have potential in addressing the challenges of GIR, and deserve particular attention. For this reason, we provide a survey of currently active knowledge bases with particular emphasis on their geospatial content, and we review the state of the art in information integration and GIR, including our contribution to these areas.

The remainder of this chapter is organised as follows: Section 2 surveys the constellation of online open knowledge bases containing geographic knowledge. Applications of geo-knowledge bases are discussed in Section 3, from recent efforts in

³ See <http://maps.google.com>, <http://www.bing.com/maps>, <http://maps.yahoo.com>

⁴ <http://www.wikipedia.org> (acc. June 5, 2012).

ontology alignment and merging (Section 3.1), to ontology-powered GIR (Section 3.2). Section 4 presents our work in this area, describing the OSM Semantic Network⁵ and the semantic expansion of the OSM dataset, connecting it to DBpedia, in order to enrich spatial data with a richer knowledge base [8]. The issue of quality of geo-knowledge bases is discussed in Section 5, and Section 6 offers a review of current limitations of these computational artifacts that need particular consideration in their usage. Finally, Section 7 discusses the challenges lying ahead in this field and the further research required to identify solutions to these challenges.

2 Survey of Open Linked Geo-Knowledge Bases

This section provides a survey of open, collaborative geo-knowledge bases, which constitute an important part of semantic technologies. To avoid terminological confusion, it is beneficial to provide a definition of the related and sometimes overlapping terms used in knowledge representation. A ‘knowledge base’ is a collection of facts about a domain of interest, typically organised to perform automatic inferences [50]. A knowledge base contains a terminological conceptualisation (typically called ‘ontology’) and a set of individuals. Widely used both in philosophy and in computer science, the meaning of the term ‘ontology’ is particularly difficult to define [99]. Among the many definitions, “an explicit specification of a conceptualization” and “shared understanding of some domain of interest” are of particular relevance, as they stress the presence of an explicit formalisation, and the general aim of being understood within a given domain [50, p. 587]. Winter notes that ontologies became part of Geographic Information Science (GIScience) towards the end of the 20th century [112].

A ‘thesaurus’ is a list of words grouped together according to similarity of their meaning [92], whilst a digital ‘gazetteer’ is specifically geographic, and contains toponyms, categories, and spatial footprints of geographic features [59]. In the Web 2.0 jargon, a ‘folksonomy’ is a crowdsourced classification of online objects, based on an open tagging process [109]. Finally, a ‘semantic network,’ a term which originated in psychology, is a graph whose vertices represent concepts, and whose edges represent semantic relations between concepts [91].⁶ We define a ‘geo-knowledge base’ as a knowledge base containing some geographic information.

In the context of geographic information, a knowledge base is generally made up of an ontology, defining classes and their relationships (abstract geographic concepts such as ‘lake’), and then populated with instances of these classes, generally referring to individual entities (e.g. Lake Victoria and Lake Balaton). In this survey, we restrict the scope to projects having global coverage, discussing their

⁵ <http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork> (acc. June 5, 2012).

⁶ Unlike ontologies, semantic networks focus on psycho-linguistic aspects of the terms. However, some knowledge bases, such as WordNet, defy this distinction by showing aspects of both ontologies and semantic networks.

spatial content. These knowledge bases are the result of combined efforts in crowdsourcing, VGI, and the Semantic Geospatial Web, and offer useful resources for GIR, and other areas of geo-science.

The Semantic Web and the Linked Data initiatives promote the adoption of semantic formats, which can be used to add an open, machine readable semantic structure to online data [12, 46]. In this context, several collaborative projects have emerged, resulting in a growing number of freely available geo-knowledge bases. Among these numerous resources, we focus on eleven datasets that have a global scope (as opposed to local projects), are mostly generated through crowdsourcing, released under Creative Commons/Open Database licences,⁷ and which are available as fully downloadable dumps in popular Semantic Web formats such as OWL and RDF. Some of the selected projects are focused specifically on geographic data (e.g. GeoNames and OpenStreetMap), while others are more general-purpose but contain valuable geographic knowledge (e.g. DBpedia and Freebase). These knowledge bases provide open datasets, and are inter-connected with one another. Our own contribution to this area of research, the OSM Semantic Network, is described in Section 4. Relevant characteristics of each knowledge base are summarised in Table 1.

CONCEPTNET. This semantic network is focused on natural language processing and understanding [58]. ConceptNet is a large semantic network, whose nodes represent concepts in the form of words or short phrases of natural language. The graph edges represent labelled relationships. Each statement in ConceptNet has justifications pointing to it, explaining where it comes from and how reliable the information seems to be. The ontology includes 1.6 million assertions gathered from Wikipedia, Wiktionary, WordNet, and the 700,000 sentences from the Open Mind Common Sense project [97]. Efforts to encode ConceptNet in RDF are being undertaken [48].

DBPEDIA. One of the leading projects of the Semantic Web, DBpedia is a Semantic Web version of Wikipedia [5]. The knowledge base currently contains 3.6 million entities, encoded in a billion RDF triples, including 526,000 places. As DBpedia is strongly interconnected with other knowledge bases (e.g. WordNet W3C, GeoNames, LinkedGeoData), it is considered the central hub of Linked Data.

FREEBASE. Designed as an open repository of structured data, Freebase allows web communities to build data-driven applications [13]. The knowledge base is structured around terms (classes), and unique entities (instances), where an entity can be a specific person, a place, or a thing, and is described by facts. It currently contains 22 million entities, of which 1 million are locations. As entities are described by facts corresponding to a directed graph, it can be easily converted into RDF.

⁷ See <http://creativecommons.org> and <http://opendatacommons.org> (acc. June 5, 2012).

Table 1 A survey of open ontologies. All of these projects are currently active, release open data, have global scope, and are interconnected with other projects. *Beginning of the project.

<i>Project Name</i>	<i>Year*</i>	<i>Type & Content</i>	<i>Data sources</i>	<i>Formats</i>
CONCEPTNET	2000	Ontology, semantic network; 1.6 million assertions, 700,000 natural language sentences	Wikipedia, WordNet, and others	JSON
DBPEDIA	2007	Ontology, semantic network; 320 classes, 740K Wikipedia types, 3.6M entities, 1 billion triples	Wikipedia	OWL/RDF
FREEBASE	2007	Ontology, knowledge base; 22M+ entities, 1M locations	Crowdsourced	Tab separated text
GEO NAMES	2006	Gazetteer; 650 classes, 10M+ toponyms	Gazetteers, Wikipedia, crowdsourced	OWL/RDF
GEO WORDNET	2010	Semantic network, thesaurus, gazetteer; 330 classes, 3.6M entities	WordNet, GeoNames, MultiWordNet	RDF
LINKED GEO DATA	2009	Gazetteer; 1K classes, 380M geographic entities	OpenStreetMap	RDF
OPEN CYC	1984	Ontology, semantic network; 50K classes, 300K facts	Expert-authored Cyc knowledge base	OWL/RDF
OPEN STREET MAP	2004	Vector map, gazetteer; User-defined tags, 1.2B nodes, 114M ways	Crowdsourced, free GIS datasets	XML
WIKIPEDIA	2001	Semantic network, dictionary, thesaurus; Semi-structured (infoboxes), 3.9M articles in English	Crowdsourced	XML
WORDNET	1985	Semantic network, dictionary, thesaurus; 117K synsets	Expert-authored knowledge base	OWL/RDF
YAGO	2006	Ontology, semantic network; 10M+ entities, 460M facts	Wikipedia, GeoNames, WordNet	RDF

GEO NAMES. Combining multiple data sources, GeoNames aims at offering a large, volunteered gazetteer.⁸ The knowledge base contains over 10 million toponyms, structured in 650 classes. GeoNames integrates geographical data such as names of places in various languages, elevation, and population. The data is collected from traditional gazetteers such as National Geospatial-Intelligence Agency's (NGA) and the U.S. Geological Survey Geographic Names Information System (GNIS), and crowdsourced online.

⁸ <http://www.geonames.org> (acc. June 5, 2012).

GEOWORDNET. GeoWordNet is the result of the integration of WordNet, GeoNames and the Italian part of MultiWordNet [41]. It is a hybrid project, combining a semantic network, a dictionary, a thesaurus, and a gazetteer. It was developed in response to the limited WordNet coverage of geospatial information and lack of concept grounding with spatial coordinates. The knowledge base contains 3.6 million entities, 9.1 million relations between entities, 334 geographic concepts, and 13,000 (English and Italian) alternative entity names, for a total of 53 million RDF triples.

LINKEDGEODATA (LGD). Since OpenStreetMap has gathered a large collection of geographic data, LinkedGeoData is an effort to republish it in the Semantic Web context [6]. The OSM vector dataset is expressed in RDF according to the Linked Open Data principles, resulting in a large spatial knowledge base. The knowledge base currently contains 350 million nodes, 30 million ways (polygons and polylines in the OSM terminology), resulting in 2 billion RDF triples. Some entities are linked with the corresponding ones in DBpedia.

OPENCYC. This is the open source version of Cyc, a long running artificial intelligence project, aimed at providing a general knowledge base and common sense reasoning engine.⁹ Even though OpenCyc covers a limited number of geographic instances, it contains a rich representation of specialised geographic classes, such as *salt lake* and *monsoon forest*. The OpenCyc classes are interlinked with DBpedia nodes and Wikipedia articles.

OPENSTREETMAP (OSM). The OSM project aims at constructing a world vector map [54]. The leading VGI initiative, the dataset represents the entire planet, gathering data from existing datasets, GPS traces, and crowdsourced knowledge. To date, the vector dataset contains 1.2 billion nodes (points), and 115 million ways (polygons and polylines).

WIKIPEDIA. A collaborative writing project, Wikipedia is a multilingual, universal encyclopedia, and has become the most visible crowdsourcing phenomenon.¹⁰ The English version currently contains 3.9 million articles, resulting in a 2 billion-word corpus. Because of high connectivity between its articles, Wikipedia is sometimes used as a semantic network [102]. This vast repository of general knowledge has been used for different purposes, including semantic similarity and ontology extraction [110, 84]. The project has also attracted interest in the area of GIScience [3].

WORDNET. Initially conceived as a lexical database for machine translation, WordNet has become a widely used resource in various branches of computer science, where it is used as a semantic network and as an ontology [32]. Currently it contains 117,000 ‘synsets’, groups of synonyms corresponding to a concept, connected to other concepts through several semantic relations. The dataset has been encoded and released in RDF, becoming a highly linked knowledge base in the web of Linked Open Data.¹¹ Even though the spatial content of WordNet is

⁹ <http://www.cyc.com/opencyc>, <http://sw.opencyc.org> (acc. June 5, 2012).

¹⁰ <http://www.wikipedia.org> (acc. June 5, 2012).

¹¹ <http://www.w3.org/TR/wordnet-rdf> (acc. June 5, 2012).

limited, the ontology holds a high quality, expert-authored conceptualisation of geographic concepts.

YAGO. Yet Another Great Ontology (YAGO) is a large knowledge base extracted from Wikipedia and Wordnet [103]. Recently YAGO has been extended with data from GeoNames, with particular emphasis on the spatial and temporal dimensions [60]. The current version of the knowledge base contains 10 million entities, encoded in 460 million facts. YAGO is inter-linked with DBpedia and Freebase.

Figure 1 presents the constellation of geo-knowledge bases, showing a schematised data path from the data producers to the knowledge bases. Bearing in mind the complexity of these collaborative processes, the main actors in this constellation, involved in the production of information and the generation of open linked knowledge bases, can be grouped as follows:

1. **Data providers.** Traditionally, geographic data was collected exclusively by experts and professionals in large public and private institutions. As Web 2.0 and VGI have emerged, a new category of non-expert users/producer ('producers') has entered the production process [25]. Crowdsourced primary sources include contributions from a wide variety of information producers, ranging from experts operating within public and private institutions to non-expert, unpaid, pro-active users.
2. **Primary sources.** Projects such as Wikipedia and OSM collect a large amount of information about the world through crowdsourced efforts. On the other hand, primary sources such as WordNet are expert-authored, while other projects combine both crowdsourcing and expert control. Most knowledge bases rely heavily on these primary sources, often aligning and merging them into larger knowledge bases. Inconsistencies and contradictions in primary sources can be propagated onto the derived knowledge bases. For example, an incorrect piece of information in a Wikipedia article will be also found in DBpedia and YAGO. For this reason, assessing the quality of these primary sources bears particular importance (see Section 5).
3. **Geo-knowledge bases.** Typically, open knowledge bases consist of structured and aggregated versions of existing semi-structured or unstructured primary sources. However, some datasets lie at the boundary between primary sources and knowledge bases, as they are both interlinked with existing knowledge bases and produce new data through crowdsourcing and expert contributions (e.g. Freebase and OpenCyc). Several knowledge bases encode the same primary data into different formalisms, such as DBpedia and YAGO.

These three actors are part of an open system, in which more or less structured data flows in complex patterns that determine the nature, quality and limitations of the resulting projects. Investigations on such collaborative open processes have been carried out, both in the area of general crowdsourcing and VGI [25]. The next section covers applications where these knowledge bases play an important role, in particular in relation to ontology alignment, and GIR.

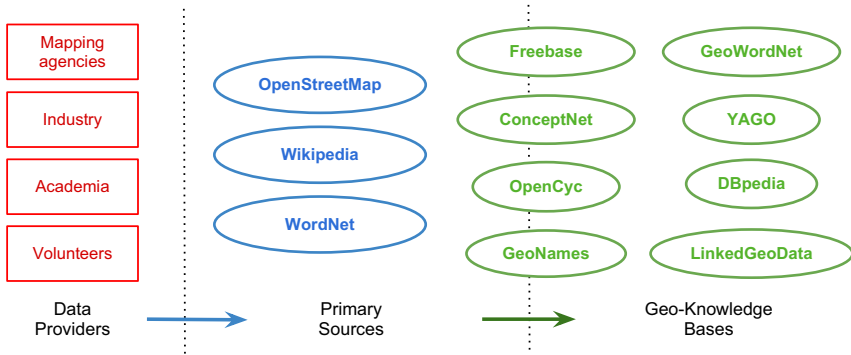


Fig. 1 The constellation of open geo-knowledge bases. The data path is schematised from the data providers to semi-structured primary sources, and finally structured into knowledge bases. Some projects defy classification by producing new knowledge in structured knowledge bases, and extracting knowledge from primary sources.

3 Open Geo-Knowledge Bases in Action

Since the late 1990s, geospatial knowledge bases have been of fundamental importance in many geographic applications [112], including semantic geographic information systems [2], GIR [3], and toponym disambiguation [87]. In general, geo-knowledge bases are used to achieve semantic interoperability between local geographic datasets modelled on incompatible ontologies. Geo-knowledge bases can also be useful in cases where advanced geographic knowledge is necessary to interpret unclear, fuzzy spatial information queries and needs [57, 41]. We focus initially on ontology alignment and merging (Section 3.1), and we subsequently discuss the usage of ontologies in GIR (Section 3.2). Our own contributions to this area of research are presented in Section 4.

3.1 Mapping, Aligning, and Merging Geo-Knowledge Bases

Online geo-data is stored in many different formats, leading to a radical heterogeneity of data formats, ontologies, and semantic models [33]. Standards by the Open Geospatial Consortium (OGC) and International Organization for Standardization (ISO) are used in geospatial modelling, while other standards are developed and promoted by the World Wide Web Consortium (W3C) in the context of web technologies. To date, despite some initial efforts, there is no clear sign of convergence towards broad adoption of joint standards between those two communities [95]. From an information retrieval perspective, the issue of coverage is critical, as GISs want to access geo-data from as many sources as possible in a consistent way. For these reasons, the field of integration of geo-knowledge bases has received a lot of attention, and is currently an active research area [73, 88].

The integration of heterogeneous data sources relies on the semantic matching of geo-knowledge bases, which present similar conceptualisations of geographic entities using different syntax, structure, and semantics [107]. According to Giunchiglia et al., considering ontologies as graphs, their alignment consists of the production of a “set of correspondences between the nodes of the graphs that correspond semantically to each other” [42, p. 1]. Combining geo-knowledge bases poses unresolved challenges, mostly due to the dynamic nature of online open datasets, semantic ambiguity, and inconsistent use of the same vocabulary, which is defined by Vaccari et al. as the ‘semantic heterogeneity problem’ [107].

Such an integration can operate at the class level (e.g. identify and inter-link the concept ‘lake’ in all the data sources), or at the instance level (e.g. find the entities representing Lake Ontario in all of the ontologies). Moreover, the semantic mapping can be applied to data, such as geo-ontologies, vector datasets, and gazetteers, or to services, in particular web services, focusing on the public interface signatures [33, 66]. Choi et al., surveying this area of research, have defined three categories of ontology mapping. The first category includes mapping between local ontologies and a higher-level ontology. The second category of mapping is performed between local ontologies. The third category of mapping is part of ontology merging, in which existing ontologies are combined in a bigger ontology [22].

In order to identify semantically close classes and instances in different ontologies, semantic similarity measures are particularly useful. Semantic similarity measures specific for geographic classes have been surveyed by Schwering [96]. She classifies the existing measures into geometric, feature, network, alignment, and transformational models. Janowicz et al. have proposed a formal framework for geo-semantic similarity [65]. This new framework responds to the ambiguity and lack of clear theoretical grounds that characterise the area of semantic similarity measurement.

Some form of ontology alignment and merging, either partly or fully automatic, has been utilised to generate most of the geo-knowledge bases surveyed in Section 2. The process leading to the creation GeoWordNet, for example, relies on the alignment between GeoNames and WordNet at the class level. Some of the concepts modelled by GeoNames were not defined in WordNet, prompting the creation of new synsets. After the ontologies were aligned at the class level, it was possible to align them at the instance level, resulting in the new, integrated ontology [41].

Similarly, LinkedGeoData has mapped some of its instances to corresponding entities in DBpedia, by aligning the ontologies along feature type, spatial distance, and name similarity [6]. The ontology YAGO is assembled by aligning WordNet synsets with the less structured Wikipedia articles [103]. Along similar lines, Buscaldi et al. have linked existing gazetteers with WordNet, and Wikipedia [21]. Their system extracts place names from the freely available Geonet Names Server (GNS) and the Geographic Names Information System (GNIS). Relevant place names are then filtered and enriched using semantic knowledge in Wikipedia and WordNet. The particular challenge that this system addresses is the combination of semantically flat place names with nodes in complex semantic networks.

The system GeoMergeP uses a layered architecture to combine and merge local ontologies, through the ISO 19100 standard [17]. Surveying recent ontology

integration works, Bucella et al. identify three main techniques, which at times are used in isolation, and at times in combination: (1) top-level ontology, (2) logical inferences, and (3) matching/similarity functions. GeoMergeP combines all of the three approaches to overcome their limitations.

While most approaches are top-down, the integration can be a bottom-up process. A bottom-up ontology alignment, focused on geographic linked data, has been carried out by Parundekar et al. [88]. This work adopts the approach of common extension comparison, i.e. two classes in different ontologies are considered similar if they are linked by similar instances, to align DBpedia, GeoNames, and Linked-GeoData (see Section 2). The mapping is done through an alignment hypothesis, built bottom-up, starting from instance pairs, up to the most general classes in the ontology.

Ontology alignment is also used by Smart et al. to combine multiple gazetteers through a common, high-level ontology [98]. Their Geo-Feature Integration module combines toponyms from OSM, GeoNames, Wikipedia, and other sources into a unified gazetteer. The module relies on spatial and textual similarity to match places across the selected data sources. In addition to traditional text similarity measures, this system uses the SoundEx algorithm to match phonetically similar sounding terms to detect alternative – and wrong – place name spellings.

Once the geo-knowledge bases have been integrated, they can be used to support various spatial tasks. In particular, GIR has emerged as a prominent area that can benefit from geo-knowledge bases [68]. The next Section surveys recent work in the area of geo-knowledge bases applied to GIR.

3.2 Ontology-Powered Geographic Information Retrieval (GIR)

Information retrieval (IR) is a vast and rapidly evolving area of computer science. Manning et al. define it as “finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections” [78, p. 1]. Users’ information needs often include some spatial information, such as a geo-location, a street name, and so on. In the context of ever-growing online information, the geographic dimension of information has become a promising way to increase the chances of meeting information needs.

Traditionally, search engines have treated spatial-related terms as any other textual information. Over the past decade, however, the area of GIR has emerged to develop techniques to give geographic information a special treatment, increasing the system’s geographic intelligence [67]. Geographic information is often implicit in the documents: broad geographic entities are omitted when they are assumed to be known to the readers, e.g. Ireland is not mentioned when referring to Dublin in the Irish media. Toponyms (place names) also have a high degree of semantic ambiguity, as there are many terms to indicate the same geographical entity in different cultural and social groups, and several places have the same name (e.g. more than 40 North-American towns are called Greenville). Moreover, toponyms pose particular challenges across different natural languages, where historical and spelling variations are very common [86].

Geo-knowledge bases have been identified as a promising support tool to develop more sophisticated GIR systems [68]. While describing their project YAGO, Weikum et al. advocate the usage of knowledge bases to go beyond the limitations of current keyword-based search engines [111, 103]. As they put it, the main challenge is “how to extract the important facts from the Web and organize them into an explicit knowledge base that captures entities and semantic relationships among them” [111, p. 61]. To illustrate YAGO’s knowledge representation, the entity representing Max Planck is displayed, including geographic knowledge about the town where the German physicist was born. The underlying intuition is that geographic knowledge is generally not provided explicitly, therefore knowledge bases can be used to discover implicit connections between entities.

Ontologies have been used in information retrieval to increase the system intelligence. In GIR, Lutz and Klien described an ontology-based system [74]. A shared vocabulary is used to translate queries across multiple ontologies, without defining a full global ontology. Their GIR system allows for user-friendly queries, translating generic queries to specific, local geo-knowledge bases. This is accomplished in a transparent way using Description Logics (DL), a family of knowledge representation languages based on first-order logic that has gained popularity in Semantic Web applications [7]. Fouad et al. have devised a location-based service to retrieve semantic information based on the user’s location [34]. Their application performs keyword-based queries on DBpedia, LinkedGeoData, and GeoNames, with the aim of displaying semantically enriched web maps. Furthermore, in the area of location-based services, DBpedia Mobile demonstrates the possibility of obtaining rich semantic information about the user’s surroundings [9].

Among others, we regard the following areas to be particularly promising as application domains for geo-knowledge bases:

Named Entity Recognition and Classification (NERC). Several systems rely on NERC techniques to identify location, people and organisations names in raw text. Nadeau and Sekine surveyed NERC techniques, from the field inception in 1991 to 2006, and discussed the main strategies to evaluate them [83]. While most NERC approaches are at least partly supervised, Cimiano and Völker have developed an ontology-based, unsupervised NERC procedure [23]. Overell has developed an approach to recognise and classify geo-referenced entities in Wikipedia articles [86].

Toponym Disambiguation. As many geographic locations share the same name, resolving the correct referent in a given context is far from being trivial. Toponym disambiguation is a specific case of proper name disambiguation, where the proper names refer to explicit or implicit spatial relationships. Knowledge-based techniques exploit geo-knowledge bases [20]. For example, Overell and Rürger have utilised Wikipedia as a knowledge base to perform place name disambiguation [87]. A co-occurrence model is extracted from Wikipedia to provide not only a list of synonyms for each location, but also the context in which each synonym is used. Toponym disambiguation is tightly connected to the issue of toponym resolution.

Toponym Resolution. By definition, a toponym refers to a geographic location. For this purpose geo-coding (toponyms to locations) and reverse geo-coding (locations to toponyms) services have been built using open geo-knowledge bases, extending traditional gazetteers with richer semantic structures. Smart et al., for example, have integrated several toponym sources into an ontological geo-coding and reverse geo-coding service [98]. In order to associate geographic locations to entities, Odon et al. have extracted textual evidence from Wikipedia articles [3]. Using Wikipedia as a semantic network, the importance of the entities is assessed by their connectivity with other entities. In this way, a representation of the geographic content of Wikipedia articles can be obtained. A related task is that of ‘spatial co-reference resolution,’ i.e. determining whether two digital representations refer to the same real-world entity. De Tré et al. developed approaches to detect co-referent features based on possibility theory, and applied it to the issue of duplicate detection [16, 26].

Spatial Footprints. Text or multimedia documents can be associated with a spatial footprint, which can be a simple geo-coordinate, a minimum bounding rectangle, or a complex polygon. Suitable spatial footprints can be computed and indexed, allowing for efficient retrieval and combination with pure text-based indexing. Fu et al. have devised an ontology and footprint-based query expansion mechanism [36]. Spatial entities are identified in a geographic ontology, and the spatial footprint of terms is computed and used in the retrieval process. Similarly, Vaid et al. have described different indexing approaches for text documents, showing that spatial indexing can enrich pure textual indexing to search large collections of text documents [108]. In the same area, Martins et al. discuss the ‘geo-scope’ of a text document, which is essentially a spatial footprint to be matched against the query footprint [79].

Spatial Reasoning. In information retrieval, queries are often expressed in natural language. Words such as ‘in’ and ‘near’ can convey important spatial semantics, which should be taken into account to meet users’ information needs. GIR systems can utilise ontologies to carry out inferences on such geospatial hints. For example, the fuzzy query ‘lakes near Dublin’ could be translated into a ‘within’ spatial query with a radius appropriate to the user context. Fu et al. have conducted work in this area in the frame of the project SPIRIT (Spatially-Aware Information Retrieval on the Internet) [36]. SPIRIT relies on dedicated ontologies to interpret spatial relationships between query terms, in the format *(what, relationship, where)*. Valid relationships include, among others, ‘near’, ‘north’, and ‘outside of’. More recently, the user context emerged as an important element that should be handled by GIR systems.

User Context. Any information need, whether containing a geographic dimension or not, is relative to a user context. For example, a user might want to retrieve all the Italian restaurants in London to conduct a socio-economic analysis, or simply to go out for dinner. The user context contains diverse information about the user, such as interests, current location, habits, language, computational device, etc, and can be exploited to refine semantic similarity measures and GIR [69]. Keßler et al. have devised a semantic language to enrich OWL with context

sensitivity [70]. The *Geo-Finder* system extracts fuzzy spatial footprints from text documents, determining the scope of the search based on the user location and speed [14]. The spatial context is taken into account by Mobile Geotumba, a GIR system optimised for handheld devices, to retrieve local information [35].

All of the aforementioned areas of research are active, and open geo-knowledge bases reviewed in Section 2 can provide useful tools to explore novel approaches. Despite the promising results obtained in several works, such ontologies have clear limitations with respect to quality that should not be ignored by researchers and general users alike. The quality and limitations of geo-knowledge bases will be discussed in Sections 5 and 6.

The next section describes our contribution to the area of geo-knowledge bases, in particular presenting the OSM Semantic Network and a semantic enhancement technique to link OSM entities to DBpedia.

4 The OSM Semantic Network

In this section we describe our own contribution to the area of geo-knowledge bases and information integration. The OSM Semantic Network is a resource that we have extracted from OpenStreetMap (OSM) data to provide a semantic support tool. OSM is the leading project of VGI, and its vector dataset has been discussed, evaluated, and utilised in various contexts [54, 52]. The OSM Semantic Network is extracted through a dedicated web crawler we have developed, and provides a detailed representation of the conceptualisation underlying OSM.

In OSM, the semantics of map entities is described through tags, fragments of text with a key and a value (e.g. *amenity=park, name='Central Park'*). Such tags are proposed, defined, and discussed on a wiki website, which hosts detailed definitions and usage guidelines for the project.¹² In the wiki pages, users often link the OSM tags to similar concepts in Wikipedia. Overall, the tagging process is deliberately informal and open to revision. Contributors are encouraged to stick to well-known tags, but the creation of new tags is not discouraged, resulting in highly dynamic – and often inconsistent – semantics.

In the context of the Linked Open Data, LinkedGeoData (LGD) has converted and published the OSM vector dataset in RDF, linking it to a formally defined ontology [6]. However, the LinkedGeoData ontology is a simple, shallow tree representing tags. To the best of our knowledge, the rich semantic information on the OSM wiki website has not been included. In order to fill this knowledge gap, we have developed an open source tool, the OSM Wiki Crawler, which extracts an RDF graph from the OSM Wiki website. The crawler extracts a semantic network in RDF, whose vertices represent tags, and edges relationships between tags. Tags are linked to Wikipedia pages, and to existing LinkedGeoData classes. The edge labels specify a number of different relationships between vertices, ranging from a generic

¹² http://wiki.openstreetmap.org/wiki/Map_Features
(acc. June 5, 2012).

Table 2 The OSM Semantic Network (extracted on June 10, 2011). Vertices marked with * are leaf vertices, i.e. have only incoming edges. ‘osmwiki:’ stands for <http://wiki.openstreetmap.org/wiki/>

<i>RDF Prefix</i>	<i>Vertex Type</i>	<i>Instances</i>
osmwiki:Key:	OSM Key.	884
osmwiki:Tag:	OSM Tag.	2,047
osmwiki:Proposed_features	OSM Proposed Tag.	340
other	LGD and Wikipedia nodes.*	1,398
<i>RDF Prefix</i>	<i>Edge Type</i>	<i>Instances</i>
osmwiki:link	Internal link.	11,982
osmwiki:valueLabel	A value of a OSM tag.	2,926
osmwiki:keyLabel	OSM key.	2,251
rdf:rdf-schema#comment	OSM Tag description.	1,892
osmwiki:key	Link to OSM key page.	1,891
osmwiki:combinedWith	Tag is combined with target tag.	1,257
osmwiki:link	A link to a Wikipedia page.	1,118
osmwiki:redirect	Redirect to a OSM wiki page.	478
osmwiki:implies	Tag implies target tag.	97

internal link (`link`) to a logical implication (`implies`). The detailed content of the current RDF graph is summarised in Table 2. In addition to the OSM Wiki Crawler, pre-extracted RDF graphs are available online.¹³ Among other applications, this ontology can be used as a support to compute semantic similarity between tags [65], as well as aligning OSM and LinkedGeoData to other geo-knowledge bases [41].

In the context of ontology alignment, we have developed an integration technique between LinkedGeoData and DBpedia, matching geographic features across the datasets. As discussed in Section 2, some LinkedGeoData instances are linked to corresponding nodes in DBpedia, in particular cities, airports, lakes, and other well-defined entities. This alignment was performed in the context of pessimistic assumptions, favouring precision over coverage. As a result, only a small subset of OSM objects is linked to DBpedia. Thus, to obtain a wider coverage, we have adopted more flexible heuristics, based on geographic proximity and a tag matching mechanism based on key words. A web application was built to allow users to visually explore the OSM dataset, and extract DBpedia nodes and concepts related to the geographic entities displayed in the current web map. A preliminary evaluation, published in [8], suggests a promising performance of this ontology-based system, but further work is needed to explore its strengths and weaknesses.

¹³ <http://wiki.openstreetmap.org/wiki/OSMSemanticNetwork> (acc. June 5, 2012).

5 The Quality of Crowdsourced Geo-Knowledge Bases

In order to utilise open knowledge bases successfully, it is crucial to assess their quality with respect to the user's requirements. For example, a geo-knowledge base might have sufficient quality to enrich the semantics of a web-based GIR system, but is likely to fail to meet the standards needed by the transport industry. Assessing the quality of knowledge bases can benefit project owners, contributors, and users, indicating criteria to select the best available resource for a given task and highlighting limitations and design flaws.

A crucial trade-off in the geo-knowledge bases discussed in this survey is between coverage and precision. Wikipedia-based ontologies such as DBpedia and YAGO cannot aim at pristine perfection, but can still obtain a reasonable precision [103]. On the other hand, expert-authored resources such as WordNet have very high precision, but are unable to compete with the coverage of crowdsourced projects. A similar trade-off applies to the geospatial dimension: geo-knowledge bases can either reach high, expert-validated spatial quality, or can be updated very frequently by a large number of volunteers, but it is difficult for these two elements to co-exist.

In recent years, several quantitative approaches to assess the quality of an ontology have been discussed [43, 15, 101, 37, 100, 49]. In our view, the approaches to evaluate the quality of geo-knowledge bases can be classified in four families:

1. **Manual evaluation:** Domain experts and intended users analyse manually the knowledge base, highlighting issues and giving qualitative judgements on the mapping between the knowledge base and the real world domain that the knowledge base is supposed to capture [40]. Although human subjects can easily detect design flaws in the schema, the labour cost of human experts can make it impractical. Moreover, even in the presence of considerable resources, large knowledge bases cannot be fully evaluated manually, and automatic methods are needed.
2. **Within-knowledge-base evaluation:** Particular properties of a knowledge base are observed without comparison with external sources. These approaches are based on relationship patterns, distributional patterns, and logical inconsistencies [19, 105]. Although this approach is inexpensive, and can be adopted for any knowledge base, its effectiveness is largely context-dependent. The average connectivity between objects, for instance, can vary across different domains, without being a reliable indicator of quality.
3. **Between-knowledge-base evaluation:** Two knowledge bases covering the same domain can be compared, cross-checking their quality. If one of the two knowledge bases has comparatively high quality, it can be used as a 'ground truth' or 'gold standard' to validate the other. For example, datasets collected and validated by national mapping agencies tend to obtain higher spatial accuracy than equivalent crowdsourced data [52]. Clearly, this approach cannot be used when comparable knowledge bases are unavailable, a rather common situation.
4. **Application-based evaluation:** Ultimately, knowledge bases are designed and populated to provide support for real-world applications. Hence, an approach consists of observing the performance of a task with and without a given

knowledge base, and measuring the differential as an indicator of quality. In this framework, different knowledge bases can be compared indirectly, bearing in mind that knowledge bases can obtain varying performances on different tasks. Strasunskas and Tomassen have proposed a scheme to evaluate the ‘ontology fitness’ with respect to search tasks [101].

In practice, quality assessment strategies can combine these four approaches in different ways, along multiple dimensions. The quality of a knowledge base can be measured at the class level and at the instance level, looking at the statistical properties of the knowledge base. For example, it is possible to have a solid, well-designed schema but noisy, insufficient instances, or vice-versa. A combination of these two aspects can offer a comprehensive picture of the ontology quality.

Specific approaches to ontology evaluation focus on a set of dimensions. Tartir et al. [105], for example, outlined a within-ontology approach based on a triangular model, in which three dimensions of quality can be observed: between the real world and the schema, between the real world and the knowledge base, and between the schema and knowledge base. In their formulation, metrics for schema quality include ‘relationship richness,’ ‘attribute richness,’ and ‘inheritance richness,’ while instance metrics capture ‘class importance,’ ‘cohesion,’ ‘connectivity,’ and ‘readability.’ Logical inconsistencies in the knowledge base can also be detected and used to measure quality [4]. For example, a knowledge base can contain the conflicting statements ‘Canada *southOf* USA’ and ‘USA *southOf* Canada.’

Moreover, Burton-Jones et al. addressed the issue of ontology quality from a semiotic viewpoint, proposing a within-ontology evaluation framework [19]. The quality is observed from four perspectives: ‘syntactic quality’ (richness of lexicon and correctness), ‘semantic quality’ (interpretability, consistency and clarity), ‘pragmatic quality’ (comprehensiveness, accuracy and relevance), and ‘social quality’ (authority and history). An overall indicator of quality is obtained with a linear combination of these four dimensions.

In the context of the open geo-knowledge bases that we have described in Section 2, the quality of primary sources such as Wikipedia and OSM has a great impact of the derived ontologies. The reliability of Wikipedia has fostered a major academic and intellectual debate, without reaching a monolithic verdict [75]. A typical way of assessing the quality of Wikipedia is based on a between-knowledge-base comparison of a random sample of articles against a well-established, expert-authored encyclopedia [40]. The results indicate that Wikipedia has excellent coverage, but the quality of its articles can vary from poor to excellent. Hu et al. have proposed within-ontology quality measures for Wikipedia articles, based on the authoritative-ness of the contributors [62].

Although false information, hoaxes and spam are generally corrected in a timely manner, Wikipedia articles at a given time can always have errors being introduced and removed. Therefore, when a snapshot of the Wikipedia website is stored and analysed, any particular article might happen to be captured right after being vandalised or after a thorough revision by a domain expert. To date, no easy solution to this issue exists.

Assessing the quality of geographic data is a well known area of GIScience, traditionally developed in the framework of cartography [18, 63]. Several dimensions can be observed to assess the quality of spatial information, including positional accuracy (how accurate the object location is with respect to the real world), completeness (how many objects are represented in the map versus all the existing objects), and logical consistency (duplicate objects, inconsistent topological relationships, etc.). Moreover, semantic aspects are particularly important for GIR, such as attribute and semantic accuracy, which focus on the quality of the metadata. The temporal quality, i.e. the rate and accuracy of updates, bears particular importance in several geospatial applications [52].

Indeed, the advent of VGI introduces additional challenges. The quality and reliability of OSM has been debated since its inception, and is now considered a critical research area for VGI [11, 82]. Like other crowdsourced projects, OSM has experienced recurring and extensive vandalism, urging the project founder to call for action [24]. Allegations have been put forward that some vandalism might be carried out by corporate competitors [38]. This sort of ‘spatial vandalism’ in open data poses peculiar challenges for project administrators, and has not yet been studied on a systematic basis.

Analogously to Wikipedia, precision and coverage of the OSM spatial data can vary greatly. An approach to quantify quality consists of adopting a map from a trusted source (e.g. a national mapping agency), and comparing it with OSM. Thus, Haklay have compared a sample from the OSM vector dataset against the corresponding data from the British Ordnance Survey [52]. OSM obtains a positional accuracy of 70%, with drops to 20%, a range that Haklay considers to be “not dissimilar to commercial datasets” [52, p. 700]. Along similar lines, Mooney et al. have conducted a quality analysis on a European subset of OSM [82]. Their study confirms the high variability in the data quality, identifying several geographical divides: rural and low-income areas tend to have lower coverage than wealthy, urban areas; natural features tend to be less covered than man-made features.

To date, the lack of standardised ‘fitness’ metrics to indicate the quality of open geo-knowledge bases makes their adoption problematic, particularly in areas in which the requirements are strict, e.g. logistics and transport. However, mainly for economic reasons, a number of online services are moving from commercial Web maps to VGI data sources – the popular social network FourSquare being the most prominent case – indicating a rising trust in crowdsourced geographic data [56].

6 Current Limitations of Geo-Knowledge Bases and GIR

Given the promise of geo-knowledge bases, GIR and the Semantic Web in general, it is important to be aware of the current limitations and drawbacks of such technologies. The Semantic Web is a broad and ambitious project that has made undeniable progress, but many of its issues are largely unresolved [81]. Polleres et al. have identified critical problems affecting the web of Linked Open Data, ranging from cases where there is too little data, poor quality data, or too much data [89].

We identify issues affecting the usability of linked geo-knowledge bases, restricting the discussion to aspects relevant to the ontologies reviewed in Section 2.

Ambiguity. Because of the wide variety of data linked by ontologies, the same vocabulary can have very different usages depending on the context. A paradigmatic case is the `owl:sameAs` predicate, which has become ambiguous in real datasets [55]. The difficulties in specifying geographic information share a common root in the complexity of the concept of place in natural languages. The conceptualisation of place is a cultural and language-dependent process, is intrinsically vague, refers to ever-shifting cultural borders, depends on other complex concepts, and is influenced by the context of usage [94]. Moreover, the web of open data lacks a meta-ontology framework to describe ontologies in a unified way [64].

Coverage. In some cases there is too little data, and missing entities or links prevent queries from retrieving results. When an entity has not been published in RDF and loaded in a public triple repository, it is simply unreachable. RDF adoption online is sparse, and most RDF triples are coming from mass imports from unstructured or semi-structured datasets [89]. When using open ontologies, the coverage/quality dilemma has to be taken into account: increasing coverage normally entails a drop in quality, and vice-versa. Projects aiming at global coverage often stumble upon the difficulty of keeping large knowledge bases in the same coherent semantic framework. Coverage also varies depending on fine-grained, project-specific aspects. In OSM, for example, man-made features are generally better covered than natural features [82]. The coverage of the interlinking between ontologies, can also show high variability, leaving vast areas of ontologies unlinked [89].

Quality. Most geo-knowledge bases contain a vast amount of data imported from crowdsourced projects. As discussed in Section 5, while crowdsourcing has clear advantages in terms of coverage and cost, precision is inevitably neglected. Moreover, when inconsistent, incomplete or inaccurate information is entered in Wikipedia or OSM, it will be propagated into DBpedia, YAGO, LinkedGeoData, and many other derived ontologies. The quality of VGI and crowdsourced data in general is hotly debated, and high variability has to be expected (see Section 5). The difficulties related to creating, maintaining and interpreting metadata were bluntly but persuasively described in the ‘Metacrap’ article by Doctorow [27]. However, several ontology quality metrics have been devised. Stransunkas and Tomassen, while presenting a framework for ontology evaluation for information retrieval, survey existing ontology metrics [101]. Beside formal quality metrics based on structure, coherence, and other aspects, an open geo-ontology can be evaluated indirectly on the basis of results obtained in real-world tasks.

Expressivity. Modelling geographic knowledge into an ontology poses specific challenges. RDF triplification, however simple it might be in most cases, can be very complex and counter-intuitive for certain facts [89]. This issue is due to well-known representational limitations of semantic networks. Additionally, OWL expressivity for spatial data is very limited. As Abdelmoty et al. pointed out, OWL does not support spatial types, and common spatial operations such as

distance are not available [2]. For spatial reasoning, OWL has to be used in conjunction with spatial databases, preventing a seamless integration with existing infrastructures [28].

Complexity. Spatial reasoning has been often identified as a fundamental instrument to increase intelligence in GIS [30]. However, applying complex spatial reasoning over large geo-knowledge bases poses remarkable challenges. Even in an ideal situation – data without noise and logical contradictions – reasoning in OWL Full is undecidable, and OWL DL is not designed to reason over massive, distributed datasets [89]. Further research is needed to enable efficient spatial reasoning in noisy, large, distributed knowledge bases.

Several efforts are being undertaken to tackle these issues in the context of the Linked Data initiatives [89]. However, it is reasonable to assume that the presence of noise, varying quality, and limited expressivity can be reduced but never fully resolved. Therefore, when developing applications relying on open geo-knowledge bases, caution is needed in order to deal with unexpected contradictions, inconsistencies, ambiguity, and a varying amount of noise in the data.

A prominent application area for geo-knowledge bases, GIR is a relatively young discipline, and its achievements are particularly difficult to assess [80]. Most of the works in the area present a preliminary evaluation, leaving the effectiveness of the approaches to be verified empirically in real world applications. To date, the most important large-scale evaluation is represented by the four GeoCLEF challenges, run from 2005 to 2008 [39, 77]. Focus was placed on open data in GikiCLEF 2009, an evaluation contest conceived to explore cultural and linguistic issues in Wikipedia-based GIR [93].

The driving intuition behind such initiatives is that adding geographic knowledge to an IR system would improve its performance when dealing with information needs with a spatial component. However, as Mandl noted, complex GIR systems have not consistently obtained better results than geographically naive systems [76]. According to Leveling, the contradictory results of GeoCLEF show possible areas of research that might improve the overall results of GIR, strengthening the usage of natural language processing with semantic indexing, handling metonyms, and topological relations beyond simple inclusion [72].

7 Conclusions and Future Work

In this chapter we presented a survey on recent advances in open geo-knowledge bases and GIR. In Section 1, we framed these areas of research in the combined visions about the Digital Earth, the Semantic Geospatial Web, and the emergence of Volunteered Geographic Information, which have changed the face of geographic information over the past decade [29, 44]. The linked open geo-resources available online that we discuss in this chapter are realising, at least in part, the vision of the ‘collaboratory,’ a collaborative geo-laboratory envisaged by Al Gore in 1998 to promote the development of geographic digital technologies [47].

A survey of free, open geo-knowledge bases with global coverage was presented in Section 2, including GeoNames, DBpedia, YAGO, GeoWordNet, ConceptNet, and others. Those knowledge bases are created by extracting knowledge from Wikipedia, OSM and traditional GIS data sources, merging different knowledge bases. Section 3.1 provides an overview of recent work in the area of geo-ontology alignment and merging. In order to cope with the growing amount of online geographic information, GIR has emerged. Section 3.2 surveys recent work in the usage of geo-knowledge bases to increase the geographic intelligence of GIR systems. Our own contribution to the area of open geo-knowledge bases, the OSM Semantic Network and an OSM/DBpedia alignment approach, is subsequently outlined in Section 4.

Section 5 surveyed the existing strategies to assess the quality of geo-knowledge bases, with particular emphasis on the quality of crowdsourced data sources. Despite undeniable advances towards the Semantic Geospatial Web and the increased coverage and quality of open geo-knowledge bases, it is important to recognise its current limitations. Section 6 highlights current issues which researchers using open geo-knowledge bases frequently encounter, identifying the core issues in coverage, quality, expressivity, and complexity of geo-knowledge bases. Similarly, current GIR systems have not met the expected increase in performance over traditional information retrieval, indicating that geographic intelligence needs refinement to become effective in its applications [76].

These issues notwithstanding, promising applications of open geo-knowledge bases are to be found in GIR, ontology alignment, toponym resolution, and related areas. In this respect, it can be argued that the most effective way to counter scepticism lies not only in formal, academic evaluations such as GeoCLEF, but in the production and dissemination of usable web applications for Internet users. For this purpose, more collaboration with the human computer interaction community might help devise appropriate interfaces to interact with open geo-data, exploiting these knowledge bases in convincing ways [76]. Work on open geo-knowledge bases should never lose contact with the ultimate stakeholders in information retrieval systems, the human users with their diversified and often unexpected information needs.

Acknowledgements. The research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

References

1. Strategic Plan 2009–2013. Tech. rep., Global Spatial Data Infrastructure Association (2009), http://portal.gsdi.org/files/?artifact_id=544
2. Abdelmoty, A., Smart, P., Jones, C.: Building place ontologies for the semantic web: issues and approaches. In: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, pp. 7–12. ACM (2007)

3. Odon de Alencar, R., Davis Jr., C., Gonçalves, M.: Geographical classification of documents using evidence from Wikipedia. In: Proceedings of the 6th ACM Workshop on Geographic Information Retrieval, GIR 2010, pp. 12:1–12:8. ACM (2010)
4. Arpinar, I., Giriloganathan, K., Aleman-Meza, B.: Ontology quality by detection of conflicts in metadata. In: Proceedings of the 4th International Workshop on Evaluation of Ontologies for the Web, EON 2006, pp. 1–7. IW3C2 (2006)
5. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
6. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (2009)
7. Baader, F., Horrocks, I., Sattler, U.: Description Logics as Ontology Languages for the Semantic Web. In: Hutter, D., Stephan, W. (eds.) Mechanizing Mathematical Reasoning. LNCS (LNAI), vol. 2605, pp. 228–248. Springer, Heidelberg (2005)
8. Ballatore, A., Bertolotto, M.: Semantically Enriching VGI in Support of Implicit Feedback Analysis. In: Kim, K.-S. (ed.) W2GIS 2011. LNCS, vol. 6574, pp. 78–93. Springer, Heidelberg (2010)
9. Becker, C., Bizer, C.: DBpedia mobile: A location-enabled linked data browser. In: Proceedings of the WWW 2008 Workshop on Linked Data on the Web, LDOW 2008. CEUR Workshop Proceedings, vol. 369 (2008)
10. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* 284(5), 28–37 (2001)
11. Bishr, M., Kuhn, W.: Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In: The European Information Society: Leading the Way with Geo-information, Proceedings of the 10th AGILE Conference, LNGC, pp. 365–387. Springer (2007)
12. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems* 5(3), 1–22 (2009)
13. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250. ACM (2008)
14. Bordogna, G., Ghisalberti, G., Psaila, G.: Geographic information retrieval: Modeling uncertainty of user’s context. *Fuzzy Sets and Systems* 196, 105–124 (2011)
15. Brank, J., Grobelnik, M., Mladenić, D.: A survey of ontology evaluation techniques. In: Proceedings of the Conference on Data Mining and Data Warehouses, SiKDD 2005, pp. 166–169. Information Society (2005)
16. Bronselaer, A., De Tré, G.: Semantical evaluators. In: Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, pp. 663–668. EUSFLAT (2009)
17. Buccella, A., Cechich, A., Gendarmi, D., Lanubile, F., Semeraro, G., Colagrossi, A.: GeoMergeP: Geographic information integration through enriched ontology matching. *New Generation Computing* 28(1), 41–71 (2010)
18. Burrough, P., Rijn, R., Rikken, M.: Spatial Data Quality and Error Analysis: GIS Functions and Environmental Modeling. In: GIS and Environmental Modelling: Progress and Research Issues, pp. 29–34. John Wiley & Sons, Hoboken (1996)

19. Burton-Jones, A., Storey, V., Sugumaran, V., Ahluwalia, P.: A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering* 55(1), 84–102 (2005)
20. Buscaldi, D.: Approaches to disambiguating toponyms. *SIGSPATIAL Special* 3(2), 16–19 (2011)
21. Buscaldi, D., Rosso, P., Peris, P.: Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval. In: 3rd Workshop on Geographic Information Retrieval, SIGIR. ACM (2006)
22. Choi, N., Song, I., Han, H.: A survey on ontology mapping. *ACM Sigmod Record* 35(3), 34–41 (2006)
23. Cimiano, P., Völker, J.: Towards large-scale, open-domain and ontology-based named entity classification. In: Recent Advances in Natural Language Processing, RANLP 2005, pp. 166–172. ACL (2005)
24. Coast, S.: Enough is enough: disinfecting OSM from poisonous people. OpenGeoData (August 10, 2010), <http://opengeodata.org/enough-is-enough-disinfecting-osm-from-poison>
25. Coleman, D., Georgiadou, Y., Labonte, J.: Volunteered Geographic Information: the nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research* 4(2009), 332–358 (2009)
26. De Tré, G., Bronselaer, A., Matthé, T., Van de Weghe, N., De Maeyer, P.: Consistently Handling Geographical User Data: Context-Dependent Detection of Co-located POIs. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010, Part II. CCIS, vol. 81, pp. 85–94. Springer, Heidelberg (2010)
27. Doctorow, C.: Metacrap: Putting the torch to seven straw-men of the meta-utopia (2001), <http://www.well.com/~doctorow/metacrap.htm>
28. Dolbear, C., Hart, G., Goodwin, J.: What OWL has done for geography and why we don't need it to map read. In: Proceedings of the OWLED 2006 Workshop on OWL: Experiences and Directions. CEUR Workshop Proceedings, vol. 216, pp. 10–11 (2006)
29. Egenhofer, M.: Toward the Semantic Geospatial Web. In: Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, pp. 1–4. ACM (2002)
30. Egenhofer, M., Mark, D.: Naive geography. In: Kuhn, W., Frank, A.U. (eds.) COSIT 1995. LNCS, vol. 988, pp. 1–15. Springer, Heidelberg (1995)
31. Euzenat, J., Ferrara, A., Meilicke, C., Nikolov, A., Pane, J., Scharffe, F., Shvaiko, P., et al.: Results of the Ontology Alignment Evaluation Initiative 2010. Tech. rep., Ontology Alignment Evaluation Initiative (2010)
32. Fellbaum, C. (ed.): WordNet: An electronic lexical database. MIT Press, Cambridge (1998)
33. Fonseca, F., Egenhofer, M., Agouris, P., Câmara, G.: Using ontologies for integrated geographic information systems. *Transactions in GIS* 6(3), 231–257 (2002)
34. Fouad, R., Badr, N., Talha, H., Hashem, M.: On Location-Centric Semantic Information Retrieval in Ubiquitous Computing Environments. *International Journal of Electrical & Computer Sciences* 10(6), 1–7 (2010)
35. Freitas, S., Afonso, A., Silva, M.: Mobile Geotumba: Geographic information retrieval system for mobile devices. In: Proceedings of the 4th MiNEMA Workshop in Sintra, pp. 83–87. Department of Informatics, University of Lisbon (2006)

36. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-based spatial query expansion in information retrieval. In: Meersman, R., et al. (eds.) CoopIS, DOA, and ODBASE 2005. LNCS, vol. 3761, pp. 1466–1482. Springer, Heidelberg (2005)
37. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 140–154. Springer, Heidelberg (2006)
38. Garling, C.: Google Workers Caught ‘Vandalizing’ Open Source Maps. *Wired* (January 17, 2012), <http://www.wired.com/wiredenterprise/2012/01/osm-google-accusation>
39. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track overview. In: Peters, C., et al. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 908–919. Springer, Heidelberg (2006)
40. Giles, J.: Internet encyclopaedias go head to head. *Nature* 438(7070), 900–901 (2005)
41. Giunchiglia, F., Maltese, V., Farazi, F., Dutta, B.: GeoWordNet: A Resource for Geospatial Applications. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 121–136. Springer, Heidelberg (2010)
42. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic matching: Algorithms and implementation. In: Spaccapietra, S., et al. (eds.) *Journal on Data Semantics IX*. LNCS, vol. 4601, pp. 1–38. Springer, Heidelberg (2007)
43. Gómez-Pérez, A.: Evaluation of ontologies. *International Journal of Intelligent Systems* 16(3), 391–409 (2001)
44. Goodchild, M.: Citizens as Sensors: the world of volunteered geography. *GeoJournal* 69(4), 211–221 (2007)
45. Goodchild, M.: NeoGeography and the nature of geographic expertise. *Journal of Location Based Services* 3(2), 82–96 (2009)
46. Goodwin, J., Dolbear, C., Hart, G.: Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. *Transactions in GIS* 12, 19–30 (2008)
47. Gore, A.: The Digital Earth: Understanding our planet in the 21st century. *The Australian Surveyor* 43(2), 89–91 (1998)
48. Grassi, M., Piazza, F.: Towards an RDF encoding of ConceptNet. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) *ISNN 2011, Part III*. LNCS, vol. 6677, pp. 558–565. Springer, Heidelberg (2011)
49. Guarino, N.: Toward a formal evaluation of ontology quality. *IEEE Intelligent Systems* 19(4), 78–79 (2004)
50. Guarino, N., Giarretta, P.: Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: Mars, N. (ed.) *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pp. 25–32. IOS Press, Amsterdam (1995)
51. Hahmann, S., Burghardt, D., Weber, B.: “80% of All Information is Geospatially Referenced”??? Towards a Research Framework: Using the Semantic Web for (In)Validating this Famous Geo Assertion. In: *Proceedings of the 14th AGILE International Conference on Geographic Information Science*, vol. 158, pp. 1–9. Association Geographic Information Laboratories, Europe (2011)
52. Haklay, M.: How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37(4), 682–703 (2010)

53. Haklay, M., Singleton, A., Parker, C.: Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass* 2(6), 2011–2039 (2008)
54. Haklay, M., Weber, P.: OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7(4), 12–18 (2008)
55. Halpin, H., Hayes, P.: When owl:sameAs isn't the same: An analysis of identity links on the Semantic Web. In: *Linked Data on the Web WWW 2010 Workshop, LDOW 2010. CEUR Workshop Proceedings*, vol. 628, pp. 1–5 (2010)
56. Hardy, Q.: Facing Fees, Some Sites Are Bypassing Google Maps. *New York Times* (March 20, 2012), <http://www.nytimes.com/2012/03/20/technology/may-sites-chart-a-new-course-as-google-expands-fees.html>
57. Harvey, F., Kuhn, W., Pundt, H., Bishr, Y., Riedemann, C.: Semantic interoperability: A central issue for sharing geographic information. *The Annals of Regional Science* 33(2), 213–232 (1999)
58. Havasi, C., Speer, R., Alonso, J.: ConceptNet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: *Recent Advances in Natural Language Processing, RANLP 2007*, pp. 27–29. *ACL* (2007)
59. Hill, L.L.: Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: Borbinha, J.L., Baker, T. (eds.) *ECDL 2000. LNCS*, vol. 1923, pp. 280–290. Springer, Heidelberg (2000)
60. Hoffart, J., Suchanek, F., Berberich, K., Lewis-Kelham, E., De Melo, G., Weikum, G.: YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In: *Proceedings of the 20th International Conference Companion on World Wide Web*, pp. 229–232. *ACM* (2011)
61. Howe, J.: The Rise of Crowdsourcing. *Wired Magazine* 14(6), 1–4 (2006)
62. Hu, M., Lim, E., Sun, A., Lauw, H., Vuong, B.: Measuring Article Quality in Wikipedia: Models and Evaluation. In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pp. 243–252. *ACM* (2007)
63. Hunter, G., Bregt, A., Heuvelink, G., Bruin, S., Virrantaus, K.: Spatial data quality: problems and prospects. In: *Research Trends in Geographic Information Science. LNGC*, pp. 101–121. Springer (2009)
64. Jain, P., Hitzler, P., Yeh, P., Verma, K., Sheth, A.: Linked data is merely more data. In: *AAAI Spring Symposium on Linked Data Meets Artificial Intelligence*, pp. 82–86. *AAAI* (2010)
65. Janowicz, K., Raubal, M., Kuhn, W.: The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science* 2(1), 29–57 (2011)
66. Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P., Stasch, C.: Semantic enablement for spatial data infrastructures. *Transactions in GIS* 14(2), 111–129 (2010)
67. Jones, C.B., Alani, H., Tudhope, D.: Geographical Information Retrieval with Ontologies of Place. In: Montello, D.R. (ed.) *COSIT 2001. LNCS*, vol. 2205, pp. 322–335. Springer, Heidelberg (2001)
68. Jones, C., Purves, R.: GIR 2005 ACM workshop on Geographical Information Retrieval. *ACM SIGIR Forum* 40, 34–37 (2006)
69. Keßler, C.: Similarity measurement in context. In: Kokinov, B., Richardson, D.C., Roth-Berghofer, T.R., Vieu, L. (eds.) *CONTEXT 2007. LNCS (LNAI)*, vol. 4635, pp. 277–290. Springer, Heidelberg (2007)

70. Keßler, C., Raubal, M., Wosniok, C.: Semantic Rules for Context-Aware Geographical Information Retrieval. In: Barnaghi, P., Moessner, K., Presser, M., Meissner, S. (eds.) EuroSSC 2009. LNCS, vol. 5741, pp. 77–92. Springer, Heidelberg (2009)
71. Kitsuregawa, M., Nishida, T.: Special issue on information explosion. *New Generation Computing* 28(3), 207–215 (2010)
72. Leveling, J.: Challenges for Indexing in GIR. *SIGSPATIAL Special* 3(2), 29–32 (2011)
73. Lopez-Pellicer, F.J., Silva, M.J., Chaves, M., Javier Zarazaga-Soria, F., Muro-Medrano, P.R.: Geo Linked Data. In: Bringas, P.G., Hameurlain, A., Quirchmayr, G. (eds.) DEXA 2010, Part I. LNCS, vol. 6261, pp. 495–502. Springer, Heidelberg (2010)
74. Lutz, M., Klien, E.: Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science* 20(3), 233–260 (2006)
75. Magnus, P.: On trusting Wikipedia. *Episteme* 6(1), 74–90 (2009)
76. Mandl, T.: Evaluating GIR: geography-oriented or user-oriented? *SIGSPATIAL Special* 3(2), 42–45 (2011)
77. Mandl, T., et al.: GeoCLEF 2007: The CLEF 2007 cross-language geographic information retrieval track overview. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 745–772. Springer, Heidelberg (2008)
78. Manning, C., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)
79. Martins, B., Silva, M., Andrade, L.: Indexing and ranking in Geo-IR systems. In: Proceedings of the 2005 Workshop On Geographic Information Retrieval, GIR 2005, pp. 31–34. ACM (2005)
80. Martins, B., Silva, M., Chaves, M.: Challenges and resources for evaluating geographical IR. In: Proceedings of the 2005 Workshop on Geographic Information Retrieval, GIR 2005, pp. 65–69. ACM (2005)
81. Millard, I., Glaser, H., Salvadores, M., Shadbolt, N.: Consuming multiple linked data sources: Challenges and Experiences. In: First International Workshop on Consuming Linked Data, COLD 2010. CEUR Workshop Proceedings, vol. 665, pp. 1–12 (2010)
82. Mooney, P., Corcoran, P., Winstanley, A.: Towards quality metrics for OpenStreetMap. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 514–517. ACM (2010)
83. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1), 3–26 (2007)
84. Nakayama, K., Hara, T., Nishio, S.: Wikipedia Link Structure and Text Mining for Semantic Relation Extraction. Towards a Huge Scale Global Web Ontology. In: Proceedings of the Workshop on Semantic Search (SemSearch 2008), 5th European Semantic Web Conference (ESWC 2008). CEUR Workshop Proceedings, vol. 334, pp. 59–73 (2008)
85. O’Reilly, T.: *What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software* (2005), <http://oreilly.com/web2/archive/what-is-web-2.0.html>
86. Overell, S.: *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. Ph.D. thesis, Imperial College London, Department of Computing (2009)
87. Overell, S., Rüger, S.: Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science* 22(3), 265–287 (2008)

88. Parundekar, R., Knoblock, C., Ambite, J.: Aligning Ontologies of Geospatial Linked Data. In: Workshop on Linked Spatiotemporal Data at the 6th International Conference on Geographic Information Science, GIScience 2010. CEUR Workshop Proceedings, vol. 691 (2010)
89. Polleres, A., Hogan, A., Harth, A., Decker, S.: Can we ever catch up with the Web? *Semantic Web Journal* 1(1), 45–52 (2010)
90. Priedhorsky, R., Terveen, L.: The Computational Geowiki: What, Why, and How. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW 2008, pp. 267–276. ACM (2008)
91. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics* 19(1), 17–30 (1989)
92. Roget, P., Kirkpatrick, B.: *Roget's Thesaurus*. Penguin, London (1998)
93. Santos, D., Cardoso, N., Cabral, L.: How geographic was GikiCLEF?: a GIR-critical review. In: Proceedings of the 6th ACM Workshop on Geographic Information Retrieval, GIR 2010, pp. 1–2. ACM (2010)
94. Santos, D., Chaves, M.: The place of place in geographical IR. In: 3rd Workshop on Geographic Information Retrieval, SIGIR, pp. 5–8. ACM (2006)
95. Schade, S., Maué, P.: Standardizing the Geospatial Semantic Web? In: Semantic Web meets Geospatial Applications Workshop, 11th International Conference on Geographic Information Science, AGILE 2008, pp. 1–5. Association Geographic Information Laboratories, Europe (2008)
96. Schwering, A.: Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS* 12(1), 5–29 (2008)
97. Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., Li Zhu, W.: Open Mind Common Sense: Knowledge acquisition from the general public. In: Meersman, R., Tari, Z. (eds.) CoopIS 2002, DOA 2002, and ODBASE 2002. LNCS, vol. 2519, pp. 1223–1237. Springer, Heidelberg (2002)
98. Smart, P.D., Jones, C.B., Twaroch, F.A.: Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. In: Fabrikant, S.I., Reichenbacher, T., van Kreveld, M., Schlieder, C. (eds.) GIScience 2010. LNCS, vol. 6292, pp. 234–248. Springer, Heidelberg (2010)
99. Smith, B.: Ontology. In: Floridi, L. (ed.) *The Blackwell Guide to the Philosophy of Computing and Information*, pp. 153–166. Blackwell, Oxford (2003)
100. Staab, S., Gómez-Pérez, A., Daelemana, W., Reinberger, M., Noy, N.: Why evaluate ontology technologies? Because it works! *IEEE Intelligent Systems* 19(4), 74–81 (2004)
101. Strasunskas, D., Tomassen, S.L.: Empirical insights on a value of ontology quality in ontology-driven web search. In: Meersman, R., Tari, Z. (eds.) OTM 2008, Part II. LNCS, vol. 5332, pp. 1319–1337. Springer, Heidelberg (2008)
102. Strube, M., Ponzetto, S.: WikiRelate! Computing semantic relatedness using Wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI 2006, vol. 2, pp. 1419–1424. AAAI (2006)
103. Suchanek, F., Kasneci, G., Weikum, G.: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706. ACM (2007)
104. Sui, D.: The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems* 32(1), 1–5 (2008)

105. Tartir, S., Arpinar, I., Moore, M., Sheth, A., Aleman-Meza, B.: OntoQA: Metric-based ontology quality analysis. In: IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, at the 5th IEEE International Conference on Data Mining 2005, ICDM 2005, pp. 1–9. IEEE (2005)
106. Turner, A.: Introduction to Neogeography. O'Reilly Media, Sebastopol (2006)
107. Vaccari, L., Shvaiko, P., Marchese, M.: A geo-service semantic integration in spatial data infrastructures. *International Journal of Spatial Data Infrastructures Research* 4, 24–51 (2009)
108. Vaid, S., Jones, C.B., Joho, H., Sanderson, M.: Spatio-textual indexing for geographical search on the web. In: Medeiros, C.B., Egenhofer, M., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 218–235. Springer, Heidelberg (2005)
109. Vander Wal, T.: Folksonomy (2007), <http://vanderwal.net/folksonomy.html>
110. Völkel, M., Krötzsch, M., Vrandečić, D., Haller, H., Studer, R.: Semantic Wikipedia. In: Proceedings of the 15th International Conference on World Wide Web, pp. 585–594. ACM (2006)
111. Weikum, G., Kasneci, G., Ramanath, M., Suchanek, F.: Database and information-retrieval methods for knowledge discovery. *Communications of the ACM* 52(4), 56–64 (2009)
112. Winter, S.: Ontology: buzzword or paradigm shift in GI science? *International Journal of Geographical Information Science* 15(7), 587–590 (2001)
113. Xiao, R., Liu, W., Du, Y., He, Y.: An intelligent bay geo-information retrieval approach based on Geo-ontology. In: Proceedings of the 17th International Conference on Geoinformatics 2009, pp. 1–6. IEEE (2009)

Chapter 6

Fact Based Search Engine: News Fact Finder Utilizing Naive Bayes Classification

Ricardo Salmon, Cristina Ribeiro, and Swathi Amarala

Abstract. There are a number of quality news sources available on the Internet. Searching through all these sources for facts related to a certain subject would be exhaustive for a user. We developed a niche sentence level search engine called News Fact Finder in order to provide users with factual information relevant to the query. Sentence level search is based on the intuition that if all the query words are within the same sentence, that result is more relevant than a result containing the query words in remote parts of the text. We therefore use suffix arrays which excel at exact substring matching to index our database. Our framework uses a Naive Bayes classifier for classification of sentences as facts and opinions. Ranking was performed at the document level, such that a document with many related facts would be ranked higher. News Fact Finder performs competitively on a large collection of news documents in providing relevant fact-based results to users. This is a novel approach to perform quality-based searching, ranking, indexing and categorization of news information.

1 Introduction

The proliferation of reputable news sources on the Internet has made it difficult for a user to search through these sources for facts related to a certain subject. An example would be looking for facts related to a new health treatment that has been hyped up in the media. We would want to know the facts including the negative side effects, beneficial effects, and cost. Whereas, we are not interested in emotional suggestions or prejudice opinions of treatment or medical advice. A typical news article presents a mixture of these facts and opinions. Also, a single news article might not mention the whole range of facts. Therefore, it is advantageous to develop an automated system to extract the facts from the multiple news sources.

Ricardo Salmon · Cristina Ribeiro · Swathi Amarala
Computer Science, University of Waterloo, 200 University Avenue, Waterloo, Canada
e-mail: {rsalmon, cribeiro, samarala}@uwaterloo.ca

Most of the top search engines perform document level search. Hence, when long queries are typed in by a user, these engines will return a result that contains most of the specified keywords spread out over the document. The document as a whole may not be relevant to the intended search. Taking this into consideration can enhance the precision of the results set without affecting sensitivity. Generally it is hard to increase the specificity of a search without decreasing the sensitivity and thus losing a number of relevant results.

Existing document level search engines claim that they provide results with exact substring matching of long queries within quotes. This is only true when the quotes contain well-known text, such as that of a Shakespeare play. Otherwise, it will return, “No results found”, even though such text exists on the web. Furthermore, the results it suggests without quotes are usually not relevant. Where document level search engines fail, sentence level search would provide relevant results to queries requiring exact substring matching. When performing document level search the unit used is a document, where as sentence level search it is a sentence.

Queries that require exact substring matching in order to obtain very relevant results are not well established. We compare the various data structures available for indexing sentences. Specifically, we explore the potential of suffix arrays, which excel at exact substring matching. However, there is an inherent difficulty in implementing sentence level search at a large scale due to the fact that there is no fast and easy way to index and search sentences in contrast to an inverted index. The scalability of this solution for large web-sized corpuses is explained in section 6.

We developed a niche sentence level search engine, News Fact Finder using Naive Bayes to filter out opinions and to provide the users with factual information relevant to the topic of the query. The dataset for our experiments is obtained by crawling on a number of reliable news sources to extract the web pages covering various topics. Our search engine is based on suffix arrays, which is convenient for substring matching, a vital component for sentence level search. The indexed sentences were classified as facts by training a scalable classifier, Naive Bayes, on a set of fact and opinion sentences for quality-based sentence categorization. Ranking was performed at the document level, such that a document with many related facts would be ranked higher. In addition, each of the fact-based sentences were ranked by relevance to the query.

A number of existing news analysis systems typically use keyword matching and topic classification techniques to group related news into various clusters [1]. However, these clusters contain both factual and opinionated information. In such situations, the user ends up performing an exhaustive search for facts in the resulting clusters. Our goal is to provide the user with just the facts relevant to the given query. The system developed by [2] uses a subjective word list to classify sentences as facts and opinions. Its limitation was the comprehensive subjective word list used, which included weakly subjective words. Hence, we propose a different approach, namely Naive Bayes classification, to reliably classify facts from opinions and improve performance.

NFF Naive Bayes, performs competitively on a large collection of news documents with an accuracy of 82% and an F_2 measure of 0.96, in providing relevant

fact based results. There has been minimal research conducted on the plausibility of sentence level search engines. Therefore, this approach is a novel contribution to perform sentence-level fact-based information retrieval.

The outline of this chapter is as follows. In section 2 we review research being conducted to test the feasibility and scalability of sentence level search engines. In section 3, we present the architecture of our search engine in detail. In section 4, we outline a method of extracting facts from opinions using a classifier. Section 5 presents the experimental results and detailed analysis. Finally, the conclusion is presented in section 6.

2 Sentence Search Engine

Multi-word queries are the majority of search queries issued on the Internet. It is intuitive that the presence of all query words within each document is a necessary condition for retrieving relevant articles. However, this is not a sufficient condition to obtain the most relevant articles, as the relationship between the query words plays an important role in determining the search engines recall. Experiments suggest that if two words occur within an article, the probability that they are related is higher when the words occur within adjacent sentences rather than in remote sentences [3]. We therefore developed a search engine based on sentence level techniques.

To the best of our knowledge there are very few sentence level search engine in the literature because it is a new research area. We survey the existing sentence search engines Relemed, PubMed, AnswerFinder and AnswerBus. Finally, present different data structures and explore the most optimal ones for use in sentence level search.

2.1 Sentence Search

Relemed [3] is a sentence level search engine for MEDLINE, which is a database of over 15 million biomedical articles. Because of the large size of the corpus, it is typical for document level search techniques to return extraneous results. Extraneous results are defined as documents in which all the query words are contained but are not relevant to the user's query. There are over thirty different search engines for MEDLINE, PubMed being one of the most popular, but they are based on document level search. Relemed is the first to explore sentence level searching.

Relemed is implemented using a database with two tables: one containing the text of each sentence along with its document and sentence IDs, and the other containing all of the document IDs linked to the corresponding citations. The sentences are broken down using the delimiters '.', '?', and '!' but include special cases to handle instances where '.' does not mark the end of a sentence (such as abbreviations or decimal numbers). Queries can consist of query words separated by spaces, or can be written in PubMeds query language. Like many other search engines, Relemed uses the Unified Medical Language System to automatically map terms to keywords in order to improve the sensitivity of the search.

Relemed uses a relevance metric in conjunction with sentence level search to rank results. There are eight relevance levels. The first level contains results in which all of the query words are in the title, the same sentence of the abstract, and the keywords. The other levels require some permutations of these, until the last level, where the only requirement is that the query words need to appear anywhere in the text. This is equivalent to PubMeds search, and is document level rather than sentence level. Relemed can sort these results to the end of the results list, since they appear to be least relevant and it is more likely they are false positives. Two case studies were conducted to compare Relemed and PubMeds results. These preliminary studies found that Relemed and PubMed returned the same number of results, and that Relemed displayed a high number of relevant results in the first few pages, with over 98% precision [3]. Relemed also was shown to rank false positives lower. Since PubMed posts results chronologically, the systems cannot be compared on this level. However, Relemed was successful at accomplishing its goals. These results show the promise of sentence level search to improve precision and ranking of search queries.

Similar sentence level search engines include specialty search engines that deal only with questions as queries, such as AnswerFinder and AnswerBus. AnswerFinder and AnswerBus' performance was evaluated using the TREC 2002 question set. AnswerFinder answered approximately 26% of the questions correctly [4]. AnswerBus is an open-domain natural language question answering system based on sentence level information retrieval. Its accuracy was 70.5% and the average time taken to respond to a question is 7 seconds [5]. Their average response time is better than many other sentence level search engines.

2.2 Search Engine Data Structure

The data structure used in the implementation of a search engine is very important. It dictates the search time, the relevance of the results retrieved, and the space required to hold the data. We compare different data structures to determine the ideal one for use in a sentence level search engine. We consider the following data structures: inverted index, suffix trees, string B trees, and suffix arrays.

2.2.1 Inverted Index

The inverted index is the most popular data structure used in document retrieval systems. An inverted index provides a mapping between terms and their location of occurrence in the documents [6]. It is simple to implement, scales well to large corpuses and is extremely efficient in performing AND queries. However, it's performance in phrase and proximity matching is very poor.

2.2.2 String B-Tree

A string B-tree is an external memory data structure that efficiently implements operations such as prefix search, range query, substring search and string insertions

and deletions on a collection of arbitrary long strings [7]. String B-trees are a combination of Btrees and Patricia tries for internal node indices that are made more effective by adding extra pointers to speed up search and update operations.

String B-trees have the same worst case complexity as B-trees but they manage unbound-length strings and perform much more powerful search operations such as the ones supported by suffix trees. Since we are performing a substring search in RAM, we are not considering string B-trees as a possible data structure in our implementation.

2.2.3 Suffix Trees

The suffix tree is a data structure known for its characteristics that enable it to perform string sequence matching. This makes it a perfect candidate for sentence level search engines. McCreight [8] came up with a path compression process, where the individual edges in the tree represent sequences of text instead of individual characters. As a result, suffix trees have a linear construction in time and space, $O(n)$. The algorithm has a few disadvantages. In particular, the tree has to be built in reverse order, so that the characters are added from the end of the input. It is due to this impediment that suffix trees are not a suitable data structure for search engines. Later Ukkonen [9] modified McCreight's algorithm to work from left to right.

2.2.4 Suffix Arrays

A suffix array is an array of integers, which represents the starting positions of suffixes of a string and is organized in lexicographical order. They have good space and time complexity. The linear construction time is a one time cost, which is done offline, thereby not adding to the search time. Suffix arrays are better at substring searching than inverted index.

In [10], it is proposed that suffix arrays are a space saving alternative to suffix trees. Their space complexity is proportional to the size of the document and searching for a pattern is proportional to the length of the search query, this is independent of the size of the document. This is even more apparent when they are compressed as discussed in [11, 12, 13, 14, 15]. The use of self-indexing structures can search and return results without storing the original text [11, 13, 16, 17, 18] and requires approximately only 30% of the space required to store the original text, which could also act as a compression scheme [19].

Even with the significant amount of research over the years and effort in attempting to minimize the size of suffix arrays, there still exists a scalability issue [19]. Compressed suffix arrays are still not scalable, requiring $O(m+\log n)$ seeks [13]. This limitation allows us realize the benefits of quicker searching. Another limitation is the time it takes for the construction of the suffix array. However, for the application of a search engine, its construction is required to be performed only once offline. Therefore this does not affect the length of time required to search and return the top relevant results.

3 News Fact Finder Naive Bayes

The News Fact Finder Naive Bayes (NFF-NB) is a search engine designed to retrieve facts relevant to the user’s query. NFF-NB is optimized for sentence level search queries. It returns exact substring matches from the users parsed query. The exact substring matches are then classified as facts or opinions, and the opinions are removed. For the experiment, we consider a sentence to be an opinion if one or more judges classified the sentence as an opinion. In cases where all of the judges classify the sentence as a fact, the sentence is considered a fact. This method of classification was used in experiments conducted in [20]. We expect that our system will return a relevant result set to the user based on our systems heuristics. In cases where there is doubt it is preferred that more information be provided to the user to interpret whether the sentence is a fact or an opinion. We next present the description of the NFF-NB algorithm.

3.1 Algorithm

The implementation of the NFF-NB system is comprised of the following components: crawler, parser, index, classifier, query, and ranking. The architecture is shown in Figure 1.

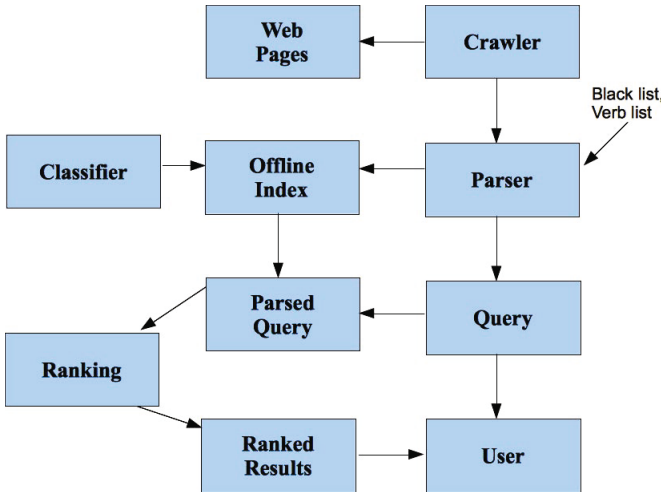


Fig. 1 News Fact Finder Naive Bayes architecture

The crawler is used to extract the relevant webpages from news websites. The details of the dataset are given in Section 5. The parser then parses each document into sentences. The text is split into sentences on the occurrence of a sentence delimiter, such as ‘.’, ‘!’ and ‘?’, including special cases for ambiguous delimiters. For

example, sentences are not broken down on the occurrence of abbreviations such as U.S., R.S.V.P, etc., or on the occurrence of initials such as Dan F. McCormick.

The second stage of the parser reads each sentence from the parsed documents, splits the sentence into a list of words. It then eliminates all the words from the black list, which are stopwords and converts all the verbs to their root forms. The details of the black list and verb list are given in Section 4. The parser then joins all the remaining words back to a sentence and writes the sentence to a text file. The query is also parsed in a similar way.

The index builder reads in one document at a time and turns each sentence into a suffix array, and also keeps track of its document ID and sentence ID. Opinion sentences are replaced by blank lines in the final parsed document. However, suffix arrays are not created for them, but are given sentence IDs. This enables tracing of these opinion sentences back to the parsed documents and exclude them from being returned in the results as neighbouring facts. The suffix arrays are stored in one list that is iterated through when searching. Each of the following units: document ID, sentence ID, and suffix array are stored as a tuple in the list.

Ranking is performed at the document level, i.e., documents with higher number of keyword matches to the query are ranked higher compared to documents with fewer keyword matches. This is based on the intuition that matched sentences are a good indicator of document relevance to the user. We then return the matched sentences along with k sentences in close proximity to the original matching sentence. We typically use $k = 2$, meaning the preceding and the succeeding two sentences are returned along with the original matched sentence. This is based on the locality principle; we assume related information is closer in proximity on a page. We consider k small since we only wish to return relevant results that the user will have time to read.

3.2 Data Structure

The data structure we use for indexing is the suffix array. It contains all pointers to the text suffixes, which are sorted in lexicographical order [4, 20, 21, 22, 23]. Each suffix is a string starting at a particular position in the original text and ending at the end of the text, as seen in Figure 2.

The suffix array class stores two attributes: 1) the sentence (which is actually a list of words in the sentence); and 2) the suffix array. It contains functions to build a suffix array given a sentence, and search the suffix array for exact matches, given a string of one or more words. It also has a function that returns a string of the sentence and the suffix array, separated by the delimiter '|', which is how the suffix array is saved to disk.

4 Fact Extraction

Finn et. al [21] develop a system which automatically classifies news articles according to whether they present opinions or facts. They first started with a classifier

Suffix										Index	
a	b	r	a	c	a	d	a	b	r	a	0
	b	r	a	c	a	d	a	b	r	a	1
		r	a	c	a	d	a	b	r	a	2
			a	c	a	d	a	b	r	a	3
				c	a	d	a	b	r	a	4
					a	d	a	b	r	a	5
						d	a	b	r	a	6
							a	b	r	a	7
								b	r	a	8
									r	a	9
										a	10

Fig. 2 Suffix Array

built based on the occurrence of words in the text as features in conjunction with a Naive Bayes classifier. When the classifier was trained and tested within the sports and politics domain, the system achieved an accuracy of 88%. It was found that this basic classifier couldn't be generalized to new domains.

The authors came up with a second approach that examines the type of language in the document. The idea behind this approach is that the kind of language used in opinion documents is different than that used in factual articles. The documents are first processed using Brill's Part-Of-Speech (POS) tagger [22] and then the documents are represented as the fraction of words for each POS. This enabled the system to be generalized for a wider variety of news articles, but it did not improve the accuracy.

4.1 *Opinion Sentence Search*

A sentence based opinion search engine in open domain topics was developed by Furuse et. al [20]. Their search engine involves extracting opinion sentences from Japanese blog pages that are relevant to a users query. As the size of the web grows, a significant proportion of content is user generated, in the form of blog pages, emails, and social networks. The motivation behind these types of search engines would be to infer opinions on various products, beliefs regarding different topics. This knowledge can also be informative for decision-making tasks.

Opinion sentences are classified based on the particular type of opinion, i.e., sentiments, neutral opinions, requests, advice, or thoughts. Each opinion sentence is identified if an opinion clue is explicitly found. For example "I am glad" from the sentence "I am glad to see you" or "extremely" from "They played extremely well". Sentences containing the exclamation mark and conditionally subjective phrases are also identified as opinions. These clues are encoded as features in a Support Vector Machine to perform the classifying if a sentence is opinionated. Opinion clues are also augmented with semantic categories to compensate relations between

co-occurring phrases in sentences. The semantic categories encode a hierarchical relationship between words. An example would be:

- (1) X is beautiful
- (2) X is pretty

The adjectives beautiful and pretty can be classified into a more general category: appearance. Now, if (1) is determined to be an opinion sentence, based off the word beautiful, (2) should also be considered as an opinion since they are of the same form.

The architecture of the system works as follows: a crawler collects pages online that are later filtered for opinions, and stores only sentences which are considered opinions by the system. When a user queries the system, a second module is called which extracts only opinion sentences that are relevant to the users query from the previous index, built offline.

The evaluation consists of collecting pages from the web and having three judges manually labeling sentences they judge to be opinions. To illustrate the subjective nature of the task they found that all three judges only agreed that a sentence is an opinion 22% of the time [20]. In terms of performance, the system was tested against (a) the baseline method, (b) a proposed model with expression clues for opinion extraction, and, (c) the effects of adding semantic categories. The results show that the opinion sentence search engine outperformed the baseline method on the test criteria of precision, recall and, accuracy.

4.2 Fact Finder

In [2], we used a subjective word list that contains words that suggests an opinion or a point of view. However, by using this approach we sometimes found it too restrictive and it required manual tuning. Additionally, we adopt a more autonomous approach by using a probabilistic classifier Naive Bayes. We model each word in our vocabulary as a binary value and represented a sentence as a feature vector for classification.

In particular, a sentence of length n can be written as $s = (w_1, w_2, \dots, w_n)$ where each $w_i, i = 1, 2, \dots, n$ is a binary value that represents the presence or absence of the word i . A sentence label is either a fact ($y = 1$) or an opinion ($y = 0$). The set of labeled sentence pairs are of the form (s_j, y_j) . That allows us to efficiently calculate $p(y|w_1, w_2, \dots, w_n)$ as proportional to $p(w_1|y)p(w_2|y) \dots p(w_n|y)p(y)$. The estimated parameters are stored for future classification.

Afterwards, whenever we retrieve a new sentence, s_* , from the web, we have to determine if it should be included in the index based on its label generated by the classifier. The classifier finds the class of y_* that maximizes $p(y_*|s_*)$. If it is the case that $y_* = 0$ is most likely, then the sentence is discarded, else it is added to the index.

After the parsing stage, various transformations are applied to the documents. In order to ensure that our system retrieves the most relevant information and only

fact based results, we created a set of word lists to aid in the task. We developed three lists: a verb word list, a black word list, and a subjective word list. Each list is used in various components of the architecture. We present further details in the following sections.

4.2.1 Verb Word List

Various forms of verbs in the sentences are converted into the base verb form using a verb list. The verb list is from WordNet, a large English lexical database [24]. We modified the list by adding missing verbs and their corresponding root forms. The verb list format is shown in Table 1.

Table 1 Verb list for document uniformity

Non Base Form Verbs	Verb Base Form
Believed	Believe
Believes	Believe
Believing	Believe

The reason we convert various verb forms to the root form is to have uniformity of verbs throughout the document and hence retrieve all the relevant sentences to a users query.

4.2.2 Black Word List

The black word list is used by the parser to eliminate common words found in the English language that are not relevant to the topic, such as the word "the". This includes definite and indefinite articles, prepositions, some verbs, and conjunctions. The black list used in the News Fact Finder consists of 131 words.

This is much more efficient than the alternative of using a white list that contains only valid words we would allow into the database. The language being used is constantly changing and we would not want to severely restrict the type of queries answerable by our system. The black list consist of a small constant set of high frequency words.

4.2.3 Subjective Word List

The subjective word list contains words used in the English language to describe a topic that suggests an opinion or a point of view. The word list was created using a subjectivity clues database, which contains subjective words and classifies them as either weakly subjective or strongly subjective [25, 26].

Ribeiro et al. [2] created a list containing only the strong subjective words, to ensure that the system is not too restrictive by classifying the sentences as opinions. The weak subjective terms do not necessarily apply in all cases and contexts.

Even with natural language processing it is difficult to determine when the weak subjective words represents an opinion or a fact.

Despite the fact that only strong subjective words are included, the exhaustive list contains 4,745 words. By using this list in the system, [2] are testing a rule based approach to classifying sentences as facts or opinions.

5 Experimentation

The sentences are then classified as either a fact or an opinion. We used Naive Bayes (NB) classifier for sentence classification. NB is a probabilistic classifier with the strong assumption that each word is sampled independently of all other words. The advantage of a NB classifier is its speed in particular for large datasets. The library used was Biopython-1.55 [27]. We used 780 manually classified sentences that consisted of 322 facts and 458 opinions to train the classifier on a total of 2,867 unique words. Based on studies by [28] 56% of news articles were judged to be objective. Next, during the creation of the index we discard all sentences that were classified as an opinion.

We tested the ability of NFF-NB to retrieve relevant facts and to filter sentences it classifies as opinions. The experiments conducted consisted of testing the NFF-NB on 27 queries covering various articles that pertain to different topic areas found in the corpus. Some of the topics include politics, health, entertainment, sports, and opinions. The corpus was obtained by crawling the CNN website and extracting news articles. Our intention was to gather articles from a niche domain instead of highly unstructured pages on the World Wide Web. The date of the articles ranges from November 6th to December 3rd, 2009, is over 120 MB in size and consists of 4,756 documents. This is a reasonable data set, as it indexes only reputable news sites and not the entire Internet.

Four judges were assigned to read the selected articles and classify the sentences as either fact or opinion. There were many ambiguous scenarios which were difficult to classify, which are described below.

Sentences which contain both facts and opinions: News articles tend to contain lengthy and complex sentences in which one part is a fact, and another part is an opinion.

Sentences that contain a personal statement from someone being interviewed: These sentences are formulated as X said Y, which in itself is a fact, regardless if Y is a fact or an opinion.

Sentences containing informed opinions: In some cases the reporters obtain opinions from experts. There was disagreement between the judges whether these types of sentences should be classified as facts or as opinions.

Sentences that contain subjective truths: In cases where an eyewitness states an account of their experience, this can be construed as fact, as no one else can argue that it did not occur to that individual. Yet it is still their opinion of what they experienced.

It is difficult to obtain a high inter-human agreement on classifying sentences into facts and opinions. Experiments conducted by [20] show that only 22% of the judges agree that a sentence is an opinion given that at least one of the judges marked it as an opinion. We used the same approach as [20] in our experiments. We assume a sentence is an opinion if one or more judges classified the sentence as an opinion. In cases where all the judges classify the sentence as a fact, the sentence is considered a fact. There were many ambiguous cases, which were difficult to classify. For example: the user may consider certain opinions from experts as facts based on their credibility. In such cases, the NFF-NB leaves the decision up to the users discretion by including those sentences in the results set.

We expect that our system will return a relevant results set to the user based on our systems heuristics. We are careful to avoid accidentally excluding sentences where we cannot agree whether they are facts or opinions. It is preferred that more information be provided to the user, rather than less. If the system consistently did not return enough information it would render the system useless. As a result, in the cases where doubt exists, we leave it up to the users to interpret whether the sentence is a fact or an opinion.

5.1 *Experimental Results*

There is a large body of knowledge and research being conducted on filtering opinions and only a few in the area of fact filtering. Although these two problems are similar in that we are classifying sentences, they use different semantics. As a result, the only direct comparison we have is the NFF-WL previously built [2]. The NFF-WL is a sentence level search engine that uses subjective word lists to extract fact based sentences. In this section we compare the NFF-NB results with NFF-WL.

Each of the query results was compared to the judges classification. If all of the sentences classified as facts were displayed in the results set, and none of the sentences classified as opinions were displayed to the user, this would be considered a successful result. Of the 27 queries tested, the NFF-NB returned 21 successful results and achieved a accuracy of 82%. Each query was processed in 0.2 seconds. This is a considerably better processing time compared to the other sentence level based search engines. AnswerBus performs better than its competitors, with an average of 7 seconds query-processing rate.

The queries classified as opinions included sentences containing strong subjective words, for example, 'I believe'. We anticipated that the NFF-NB would not return any results matching this query. However, it returned results that the classifier failed to filter as opinions. This is due to our training data not containing many sentences that include direct opinions such as the query I believe. The NFF-NB method incorrectly classified these sentences as facts. Successful results returned only factual sentences relevant to the query.

There are two possible error categories. The first error type was false positive and the second was false negative. The false positive results included opinion sentences appearing in the results set, as though they were facts. These sentences were

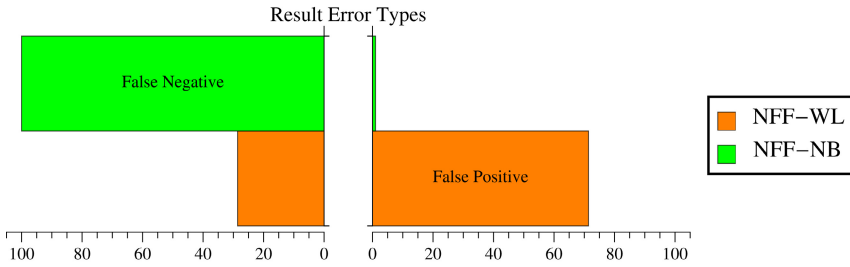


Fig. 3 Result Error Types

classified as opinions by the judges. The false negative results case is comprised of fact-based sentences that were classified as opinions. The Naive Bayes classifier had six queries, in which, the results returned were incorrect. All of these results were false positive. There were no false negative results. Again, this may be due to the skewed data used to train the classifier, being more fact based rather than opinion based material. Figure 3 shows the comparison of error types and frequency between NFF-WL and NFF-NB.

In most cases the NFF-NB method outperformed the NFF-WL method previously used. Eight of the successful results produced significantly more relevant facts from articles in the corpus that the previous method had missed. In total 39 additional facts were returned, which on average resulted in 2.05 more pertinent facts per query.

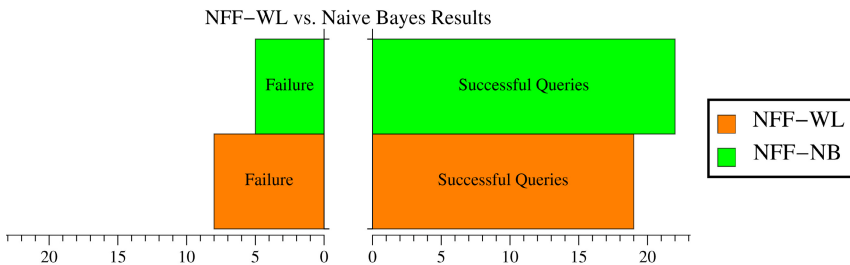


Fig. 4 NFF-WL vs. Naive Bayes Results

The sentence based question answering systems, Answer Finder and Answer bus have an accuracy of 26% and 70% respectively [5, 4]. Whereas the NFF-WL [2] method has an accuracy of 73%, in contrast the NFF-NB achieved an 82% accuracy along with significantly better results for each of the successful queries. A comparison is shown below in Figure 4.

NFF-NB obtained an F-measure of 0.89, which is a weighted average of precision and recall. The users of a search engine place significantly more emphasis on recall over precision. A user who wants a higher recall would be required to manually review an inconceivable number of sources and extremely large data sets that would

be subjected to human error. This is why the F_2 -measure is a better effectiveness measure of our search engine. NFF-NB produced an F_2 -measure of 0.96.

We tested both these systems on the same set of 27 queries. We present the different scenarios in which the results are different, and listed below. The queries are numbered, in order to easily reference them throughout the following section.

<i>Query 1:</i>	a christmas carol
<i>NFF-WL Results Returned:</i>	<p data-bbox="448 389 947 419">‘New Moon’ wins another box office weekend</p> <ol data-bbox="448 437 1026 749" style="list-style-type: none"> <li data-bbox="448 437 1026 560">1. With a \$24.1 million total so far, Old Dogs hasnt captured the same men-of-a-certain-age crowds that drove Travoltas Wild Hogs to a \$39.7 million debut back in 2007. <li data-bbox="448 566 1026 684">2. Meanwhile, Disneys A Christmas Carol (fifth place, \$16 million) got a holiday bump, jumping 30 per cent over last weekend to a total of \$105.3 million. <li data-bbox="448 689 1026 749">3. In its first wide-release weekend, and The Road grossing \$1.5 million at 111 theaters. <p data-bbox="448 784 794 814">‘New Moon’ banks at box office</p> <ol data-bbox="448 832 1026 1049" style="list-style-type: none"> <li data-bbox="448 832 1026 892">1. Vampires and werewolves weren’t the only champions at the box office, either. <li data-bbox="448 897 1026 989">2. At fifth, Disney’s A Christmas Carol continued to hold on strong, dropping 45 percent for \$12.2 million and \$79.8 million total. <li data-bbox="448 994 1026 1049">3. Two other limited release debuts had varying success.
<i>Success:</i>	Results were returned relating to the movie A Christmas Carol, including previous success, current success, ranking among movies, and a comparison to other debuts in the past. These are excellent results that a user could use to determine whether or not to watch the movie.

<p><i>NFF-NB Results Returned:</i></p>	<p>‘New Moon’ wins another box office weekend</p> <ol style="list-style-type: none"> 1. With a \$24.1 million total so far, Old Dogs hasn’t captured the same men-of-a-certain-age crowds that drove Travolta’s Wild Hogs to a \$39.7 million debut back in 2007. 2. Meanwhile, Disney’s A Christmas Carol (fifth place, \$16 million) got a holiday bump, jumping 30 per cent over last weekend to a total of \$105.3 million. 3. Specialty pics found modest success, with the animated Fantastic Mr. Fox earning \$7 million Friday-Sunday. <p>‘New Moon’ banks at box office</p> <ol style="list-style-type: none"> 1. The weekend’s real casualty was the animated sci-fi comedy Planet 51, which managed to open at just \$12.6 million for fourth place. 2. At fifth, Disney’s A Christmas Carol continued to hold on strong, dropping 45 percent for \$12.2 million and \$79.8 million total. 3. And Precious: Based on the Novel ‘Push’ by Sapphire pulled in \$11 million for sixth place on just 629 screens; after only three weeks of a limited, platform release, the Oscar favourite has grossed a stunning \$21.4 million. <p>8 to watch: This holiday season’s movies wrapped up</p> <ol style="list-style-type: none"> 1. Some of the most anticipated movies of 2009 will be released this winter, as studios aim to take advantage of the holiday crowds and slip in Oscar hopefuls ahead of awards season. 2. Recent releases like Disney’s modern interpretation of the traditional Charles Dickens tale A Christmas Carol have already started to build up the holiday mood. 3. Other upcoming highlights include big-budget 3D epic Avatar, which is due to be released December 18 worldwide, and Peter Jackson’s film adaptation of Alice Sebold’s heart-wrenching novel The Lovely Bones.
--	---

<i>Success:</i>	Several additional facts were returned including all of the NFF-WL facts. The word list used for NFF-WL was comprehensive in that it included weak subjective words. Its strictness resulted in the incorrect classification.
<i>Query 2:</i>	Fantastic Mr. Fox
<i>NFF-WL Results Returned:</i>	None
<i>Success:</i>	One of the judges classified the sentence containing this sub-string as an opinion. Therefore NFF-NB should not display this sentence in the results.
<i>Explanation:</i>	The words fantastic and modest are contained in the subjective word list. As a result, the sentence containing this sub-string was classified as an opinion by system and removed. This query is a movie title. If the sentence did not contain modest, it would have still been filtered out, due to the word fantastic.
<i>NFF-NB Results Returned:</i>	None
<i>Success:</i>	Same result achieved as the NFF-WL
<i>Query 3:</i>	I think
<i>NFF-WL Results Returned:</i>	None
<i>Success:</i>	All of the judges classified the sentence containing this sub-string as an opinion. Therefore NFF-WL should not display this result.
<i>NFF-NB Results Returned:</i>	None
<i>Failure:</i>	These sentences were not classified as opinions by NFF-NB.
<i>Explanation:</i>	We attribute this failure to the fact that the training data included very few instances of phrases such as I think. Hence the classifier couldn't identify these phrases as opinions.

Query 4:	health problem
NFF-WL Results Returned:	<p>Alcohol takes its toll on Russians health</p> <ol style="list-style-type: none"> 1. Alcohol takes its toll on Russians health 2. The biggest health problem facing Russia is the very high level of mortality among working aged men, says Martin McKee, an expert in Russian public health at the London School of Hygiene and Tropical Medicine. 3. A new dynamism appears to be taking hold of Russia as it aims to raise its prominence on the world stage. <p>Vitamin D: Hyped or true wonder?</p> <ol style="list-style-type: none"> 1. If you eat a balanced diet, are you likely to develop vitamin D deficiency? 2. Back in the 1930s, rickets was a major public health problem. 3. As a result, the United States started a milk fortification program.
Failure:	The sentence containing the sentence expected from the articles reviewed was not returned in the results set.
Explanation:	The word extremely is in subjective word list. As a result this sentence was classified as an opinion by the NFF-NB and not indexed. It should have been classified as a fact and displayed.
Implementation:	With the use of Natural Language Processing, used in conjunction with the subjective word list, these cases may be avoided.

NFF-NB Results Returned:

Task force opposes routine mammograms for women age 40-49

1. Watch Dr. Gupta explain the controversy With its new recommendations, the [task force] is essentially telling women that mammography at age 40 to 49 saves lives; just not enough of them, Dr. Otis Brawley, chief medical officer for the American Cancer Society.
2. Breast cancer is a serious health problem facing adult women, and mammography is part of our solution beginning at age 40 for average-risk women, it says.
3. It recommends annual exams beginning at that age.

10 surprising facts about cholesterol

1. For example, 33 percent of people ages 20 to 74 had high cholesterol (defined as above 240 mg/dL) in the early 1960s, and the average was 222 mg/dL; in 2003 to 2006, about 16 percent of people in that age group had high cholesterol and the average was 200 mg/dL.
2. Elevated cholesterol, which was unrecognized as a serious health problem 50 years ago, is dropping mainly because of more awareness of its dangers, which has resulted in healthier diets , more cholesterol screening, and the widespread use of statin medications.
3. Health.com: Are you cholesterol smart?

Alcohol takes its toll on Russians' health

1. Life expectancy for Russian men is well below that of western European countries like Germany, where men have an average life span of 77 years, according to World Health Organization figures.
2. The biggest health problem facing Russia is the very high level of mortality among working aged men, says Martin McKee, an expert in Russian public health at the London School of Hygiene and Tropical Medicine.
3. Life expectancy for men has stagnated for quite some time, and a major culprit has been high levels of alcohol consumption.

	<p>Sex drive vs.</p> <ol style="list-style-type: none"> 1. Long banned in Olympic sports from bobsled to diving, beta blockers also reduce anxiety and muscle tremor and can sharpen focus and precision. 2. Barron has another health problem: extremely low testosterone. 3. Since 2005, he has been getting monthly shots of synthetic testosterone, which is also banned from most professional sports. <p>Vitamin D: Hyped or true wonder?</p> <ol style="list-style-type: none"> 1. The National Cancer Institute and the Centers for Disease Control and Prevention concluded vitamin D levels in the blood were not related to overall cancer mortality. 2. Back in the 1930s, rickets was a major public health problem. 3. Rickets is a skeletal disease that weakens the bones, especially in children.
<i>Success:</i>	<p>Several additional facts were returned including all of the NFF-WL facts. The improved performance is due to the same reason as listed in Query 1.</p> <p>Consider the query, 'Obama announcing new plan for Afghanistan', the NFF-NB classification method ranked very relevant facts at the top. This is an improvement over the NFF-WL method.</p>
Query 5:	Obama was announcing his new plan for Afghanistan
<i>NFF-WL Results Returned:</i>	<p>McChrystal on Afghanistan</p> <ol style="list-style-type: none"> 1. The whole world is listening. 2. In a statement issued as Obama was announcing his new plan for Afghanistan, McChrystal said the president had provided him with a clear mission and sufficient resources. 3. Some question whether we can do it, or if we should do it, McChrystal told the forces
<i>Failure:</i>	This is a false positive result.

<i>Explanation:</i>	The second sentence was classified as an opinion by all the judges, and therefore should not be returned, although it was classified as such by taking into account the context of the article. Having read the preceding sentences, it is clear that this sentence is not a fact. However, analysis of this sentence based solely on its sentence structure, out of context of the article, would classify it as a fact.
<i>NFF-NB Results Returned:</i>	McChrystal on Afghanistan <ol style="list-style-type: none"> 1. The whole world is listening. 2. In a statement issued as Obama was announcing his new plan for Afghanistan, McChrystal said the president had provided him with a clear mission and sufficient resources. 3. Some question whether we can do it, or if we should do it, McChrystal told the forces <p>U.S. commander: 'I have exceptional confidence right now'</p> <ol style="list-style-type: none"> 1. We must make sure we are of one mind. 2. In a statement issued as Obama was announcing his new plan for Afghanistan, McChrystal said the president had provided him with a clear mission and sufficient resources. 3. The general said the situation had improved with the commitment of additional troops, giving the mission better clarity, capacity, commitment and confidence.
<i>Success:</i>	Several additional facts were returned including all of the NFF-WL facts.
<i>Query 6:</i>	European Union Summit
<i>NFF-WL Results Returned:</i>	None
<i>Failure:</i>	Should have returned the sentence containing European Union (EU) Summit.
<i>NFF-NB Results Returned:</i>	None
<i>Failure:</i>	Same result achieved as the NFF-WL

<i>Explanation:</i>	A good example of where sub-string matching in our system fails to provide the user with relevant results. The abbreviation (EU) would not be in the users query, and as a result this fact was not returned. However, if our corpus included a larger variety of topics (more breadth rather than depth), we would expect that this would not be the only fact existing with the sub-string European Union Summit. As a result, our system would still return valuable results to the user.
---------------------	--

6 Conclusion

In this chapter we proposed the use of sentence level quality-based indexing as a replacement for document level searching that is employed by all the major search engines including Google and Yahoo!. However, our framework is not meant to index the entire web but is ideal for a subset of the web. The focus of our method is on semi-formal text that is plentiful on news related websites. This provides a bounded niche where more sophisticated data-structures can be used but scalability remains practical.

We developed a sentence level search engine, News Fact Finder, which conducts sentence level searching through reputable online news media sites. The system removes opinions and extracts fact-based sentences by preprocessing the data using various word lists and a Naive Bayes for sentence classification.

Suffix arrays were used to store each sentence across documents and news sites. In building the initial index our method scales linearly with the number of sentences, n , each of which are stored as a suffix array, where the index can be constructed in time $\Theta(n \cdot |s|)$ and space usage of $O(n \cdot |s|)$ based on improvements by [29, 28] for a string, s , of length $|s|$. The news websites are frequently updated with breaking news. Therefore search engines are continually indexing new pages. However, rebuilding the index to insert additional sentences is not required. There is an approach by [30] that allows the index to be updated dynamically without rebuilding.

We used manually annotated news articles as the training data set for the Naive Bayes classifier. This data, due to its nature contains fewer opinions and may have hindered the classifiers ability to correctly classify opinions and is a difficult obstacle to overcome in the context of our system. In dealing only with reputable news media web sites, the style of writing must also be taken into account.

The classifier was very successful in classifying facts, such that there were fewer false positives than NFF-WL. The overall accuracy was 82% and achieved an F_2 measure of 0.96, which is significantly better than NFF-WL. In addition, the results included more facts per query, where the other method misclassified those facts as opinions.

Acknowledgements. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) grant funded by the Canadian government.

References

1. An Introduction to the Europe Media Monitor Family of Applications. In: Information Access in a Multilingual World - Proceedings of the SIGIR, Workshop (SIGIR- CLIR 2009), Boston, USA (2009)
2. Sentence Level Fact Based Search Engine: News Fact Finder, Las Vegas, NV, USA (2010)
3. Siadat, M.S., Shu, J., Knaus, W.A.: Relemed: sentence-level search engine with relevance score for the medline database of biomedical articles. *BMC Medical Informatics and Decision Making* 7(1) (2007)
4. Proceedings of the 7th Annual Research Colloquium of the UK, Special Interest Group for Computational Linguistics, Birmingham, UK (2004)
5. AnswerBus question answering system. In: Proceedings of HLT 2002, Second International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., San Francisco (2002)
6. Buttcher, S., Clarke, C., Cormack, G.: *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press (2010)
7. Ferragina, P., Grossi, R.: The string b-tree: a new data structure for string search in external memory and its applications. *Journal of the ACM (JACM)* 46(2), 236–280 (1999)
8. McCreight, E.M.: A space economical suffix tree construction algorithm. *Journal of the ACM (JACM)* 23(2), 262–272 (1976)
9. Ukkonen, E.: On-line construction of suffix trees. *Algorithmica* 14(3), 249–260 (1995)
10. Manber, U., Myers, G.: Suffix arrays: A new method for on-line string searches, pp. 319–327. Society for Industrial and Applied Mathematics, Philadelphia
11. Ferragina, P., Manzini, G.: Opportunistic data structures with applications, pp. 390–398
12. Sadakane, K.: Succinct representations of lcp information and improvements in the compressed suffix arrays, pp. 225–232. Society for Industrial and Applied Mathematics, Philadelphia
13. When indexing equals compression: Experiments with compressing suffix arrays and applications. In: 15th Annual ACM-SIAM Symposium on Discrete Algorithms (2004)
14. The performance of linear time suffix sorting algorithms
15. Mäkinen, V., Navarro, G.: Succinct suffix arrays based on run-length encoding. In: Apostolico, A., Crochemore, M., Park, K. (eds.) *CPM 2005*. LNCS, vol. 3537, pp. 45–56. Springer, Heidelberg (2005)
16. Mäkinen, V., Navarro, G.: Compressed compact suffix arrays. In: Sahinalp, S.C., Muthukrishnan, S.M., Dogrusoz, U. (eds.) *CPM 2004*. LNCS, vol. 3109, pp. 420–433. Springer, Heidelberg (2004)
17. Ferragina, P., Manzini, G.: An experimental study of an opportunistic index, pp. 269–278. Society for Industrial and Applied Mathematics, Philadelphia
18. Navarro, G.: Indexing text using the ziv-lempel trie. *Journal of Discrete Algorithms* 2(1), 87–114 (2004)
19. Puglisi, S.J., Smyth, W.F., Turpin, A.: Suffix arrays: what are they good for? p. 18. Australian Computer Society, Inc.
20. Opinion sentence search engine on open domain blog. In: *Artificial Intelligence (IJCAI 2007)*, Hyderabad, India (2007)

21. Fact or fiction: Content classification for digital libraries. In: The Proceedings of Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin, Ireland (2001)
22. Brill, E.: Some advances in transformation-based parts of speech tagging. In: AAAI (1994)
23. Identifying collocations for recognizing opinions. In: Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL), Toulouse, France (2001)
24. Fellbaum, C.: Wordnet: An electronic lexical database (1998)
25. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 165–210 (2005)
26. Wilson, T.: Fine-grained subjectivity analysis (2008)
27. Biopython, v1.55 (August 31, 2010)
28. Kärkkäinen, J., Sanders, P.: Simple linear work suffix array construction. In: Baeten, J.C.M., Lenstra, J.K., Parrow, J., Woeginger, G.J. (eds.) ICALP 2003. LNCS, vol. 2719, pp. 943–955. Springer, Heidelberg (2003)
29. Ko, P., Aluru, S.: Space efficient linear time construction of suffix arrays. In: Baeza-Yates, R., Chávez, E., Crochemore, M. (eds.) CPM 2003. LNCS, vol. 2676, pp. 200–210. Springer, Heidelberg (2003)
30. Salson, M., Leonard, M., Lecroq, T., Mouchard, L.: Dynamic extended suffix arrays. *Journal of Discrete Algorithms* (2009)

Chapter 7

Quality-Based Knowledge Discovery from Medical Text on the Web

Example of Computational Methods in Web Intelligence

Andreas Holzinger, Pinar Yildirim, Michael Geier, and Klaus-Martin Simonic

Abstract. The MEDLINE database (Medical Literature Analysis and Retrieval System Online) contains an enormously increasing volume of biomedical articles. Consequently there is need for techniques which enable the quality-based discovery, the extraction, the integration and the use of hidden knowledge in those articles. Text mining helps to cope with the interpretation of these large volumes of data. Co-occurrence analysis is a technique applied in text mining. Statistical models are used to evaluate the significance of the relationship between entities such as disease names, drug names, and keywords in titles, abstracts or even entire publications. In this paper we present a selection of quality-oriented Web-based tools for analyzing biomedical literature, and specifically discuss PolySearch, FACTA and Kleio. Finally we discuss Pointwise Mutual Information (PMI), which is a measure to discover the strength of a relationship. PMI provides an indication of how more often the query and concept co-occur than expected by chance. The results reveal hidden knowledge in articles regarding rheumatic diseases indexed by MEDLINE, thereby exposing relationships that can provide important additional information for medical experts and researchers for medical decision-making and quality-enhancing.

1 Introduction

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database for the life sciences and includes bibliographic information

Andreas Holzinger · Michael Geier · Klaus-Martin Simonic
Institute for Medical Informatics, Statistics and Documentation, Medical University Graz
e-mail: {andreas.holzinger,klaus.simonic}@medunigraz.at,
m.geier@hci4all.at

Pinar Yildirim
Department of Computer Engineering, Okan University, Istanbul
e-mail: pinar.yildirim@okan.edu.tr

for papers of academic journals covering a broad range of biomedical and health care topics. Moreover, MEDLINE covers much of the literature in biology and biochemistry. Maintained by the United States National Library of Medicine (NLM), MEDLINE is available for free on the Web and searchable via tools such as PubMed [1] and Entrez [2]. The MEDLINE database is the primary resource for biomedical researchers and contains currently 21,763,549 total records [3]. Within this big data, a wealth of scientific information is existing and knowledge on relationships between biomedical concepts including genes, diseases and cellular processes is hidden [4]. All the information contained in the database is stored as text. The rapid growth of these text collections makes it difficult for humans to access the required data in a convenient and effective manner.

Figure 1 shows the vastly increasing number of publications in the MEDLINE database from 1940 until 2011. The number of publications was determined using the PubMed query "*pubyear*"[*Publication Date*], where *pubyear* was replaced by the corresponding years. For the year 2011 986,794 publications are listed, in May 2012 already 420,933 publications are found for the year 2012.

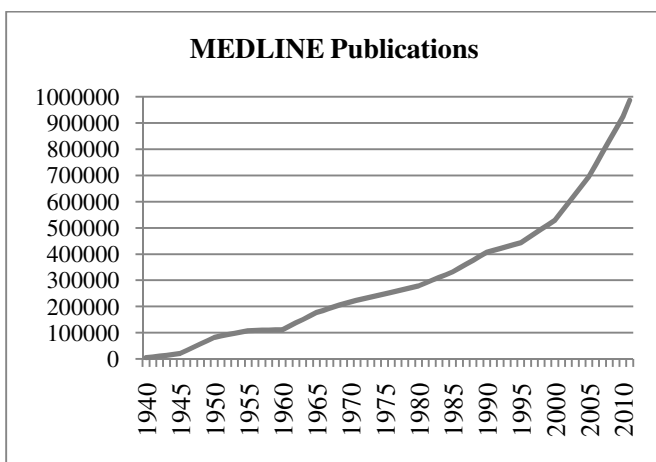


Fig. 1 Yearly number of MEDLINE publications from 1940 to 2011 (queried in steps of five years on 14/05/2012)

In order to make this data accessible, usable and useful, smart information retrieval systems that can operate on these non-standardized text (often called: "free text") are essential [5]. Consequently, there is a strong necessity of developing quality-based methods for the extraction of relevant information (such as keywords related with diseases) from the literature, which is written in natural language.

Data mining on text has been designated at various sources as statistical text processing, knowledge discovery in text, intelligent text analysis, or natural language processing, depending on the application and methodology that is used [6], [7].

Text mining aims at developing technologies helping to cope with the interpretation of these large volumes of publications. A commonly used method to

establish such relationships between biomedical concepts from literature is co-occurrence analysis. Apart from its use in knowledge retrieval, the co-occurrence method is also well suited to discover new, hidden relationships between biomedical concepts. This technique is applied in text mining and the methodologies and statistical models are used to evaluate the significance of relationship between entities such as disease names, drug names, and keywords in titles, abstracts or even entire publications.

Table 1 Feature comparison of various biomedical text mining tools [8]

	Entrez	MedMiner	Alibaba	PolySearch
Type of search supported	Literature, Disease, Gene, Structure, Taxonomy, SNP, Compound, etc.	Gene, Drug, Text Word	Gene, Disease, Drug, Tissues/Organs, Cells, Species	Gene, Disease, Drug, Metabolite, Tissues/Organs, Subcellular Localization, Text Word
Extensive hyperlinking	Most Extensive	Less Extensive	Less Extensive	More Extensive
Text and sentence highlighting	No	Yes	Yes	Yes
Co-occurrence scoring scheme	None	None	Sentence level	Sentence level
Use of keywords for association words	None	Predefined keywords	Predefined keywords	Predefined & custom association words
Sentence pattern recognition	No	No	Yes	Yes
Thesaurus query synonym expansion	Yes, limited	Yes, limited	None	Yes, extensive
Databases	PubMed, OMIM, Gene, MMDB, Taxonomy, dbSNP, PubChem, etc.	PubMed, GeneCards	PubMed	PubMed, OMIM, Swiss-Prot, DrugBank, HMDB, HPRD, GAD, HapMap, dbSNP, CGAP, HGMD

2 Web-Based Tools for Analyzing Biomedical Literature

There are several Web-based tools for the analysis of biomedical literature. Most of the tools provide the analysis of co-occurrence between biomedical entities such as disease, drugs, genes, proteins and organs. Some provide additional measures, such as Pointwise Mutual Information.

Four tools (Entrez, MedMiner, Alibaba, and PolySearch) are compared in Table 1. [9], however, provides a more extensive overview of Web tools for searching biomedical literature. Kleio and FACTA (see later) are not mentioned, PolySearch, however, and more than 25 tools from the following categories are included:

- Ranking PubMed's search results (example: RefMed)
- Clustering and categorizing results into topics (example: McSyBi)
- Extracting and displaying semantics and relations (example: MEDIE)
- Visualization and improving search interface and retrieval experience (example: iPubMed)

PolySearch can produce a list of concepts which are relevant to the user's query by analysing multiple information sources including PubMed, OMIM, Drugbank and Swiss-Prot. It covers many types of biomedical concepts including organs, diseases, genes/proteins, drugs, metabolites, SNPs, pathways and tissues.

The general issue of synonyms and acronyms is handled by PolySearch by optionally automatically expanding the query with synonyms and acronyms. A list of filter words excludes unwanted results. One drawback of PolySearch is the low speed performance of the system.

EBIMed, XplorMed, MedlineR, LitMiner and Anni are commonly used tools and they provide similar functionality with PolySearch [10], [7].

NACTeM (The National Centre of Text Mining) also develops Web-based tools such as FACTA/FACTA+ and Kleio. These are text search engines for MEDLINE abstracts, which are designed particularly to help users browse biomedical concepts (e.g. genes/proteins, diseases, enzymes and chemical compounds) appearing in the documents retrieved by the query. By revealing associations between biomedical concepts, **FACTA** allows to gain new knowledge from the large amount of MEDLINE text data. The distinct advantage of FACTA is that it delivers real-time responses while being able to accept flexible queries [11]. FACTA covers six categories of biomedical concepts: human genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds. The concepts appearing in the documents are recognized by dictionary matching. UMLS (Unified Medical Language System) is used for diseases and symptoms. UMLS constitutes a valuable lexical resource integrating a thesaurus and multilingual vocabulary database of health-related concepts as well as the semantic relationships between them. FACTA receives a query from the user as the input. A query can be a concept name like "Rheumatoid Arthritis", a concept ID or a combination of these. The system then retrieves all the documents that match the query from MEDLINE using word/concept indexes. The concepts contained in the documents are then counted and ranked according to their relevance to the query. For the input query "Rheumatoid Arthritis" with disease as a selected concept, and the system retrieves 94834 documents from MEDLINE. The results are displayed as a table and ranked by their frequencies which indicate how many times selected concept appears in the articles with the query word. For example, "Polyarthritis" which is a kind of Rheumatoid disease appears 4393 times with "Rheumatoid Arthritis" [12].

One issue of FACTA is that synonyms and variations of the spelling of terms are often not considered properly. As shown in Figure 3, it is not distinguished between "weakness" and "Weakness", for example.

FACTA+ Visualizer [13] is an Adobe Flash-based browser application which presents the query results of a FACTA query as tile chart (Figure 2, Figure 3). For supporting data analysis by medical experts, who typically are not aware of the mathematical or technical background of text mining tools, good visualisations of the results are essential.

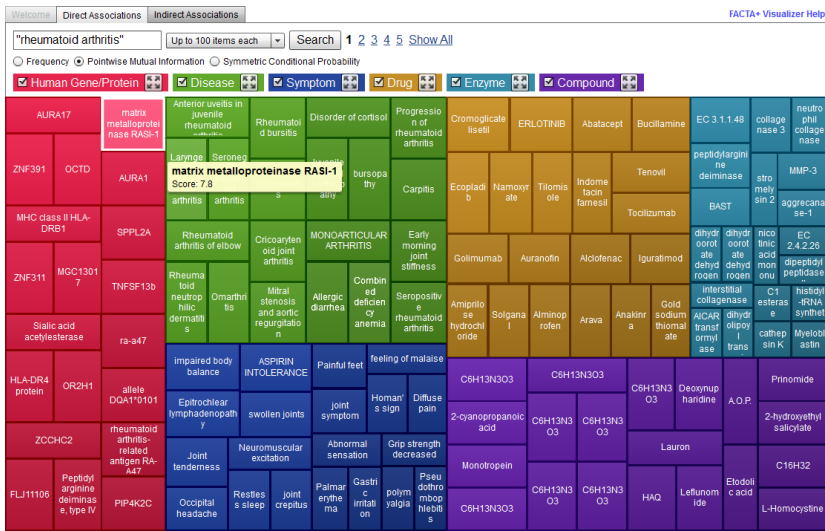


Fig. 2 FACTA+ Visualizer: Pointwise Mutual Information

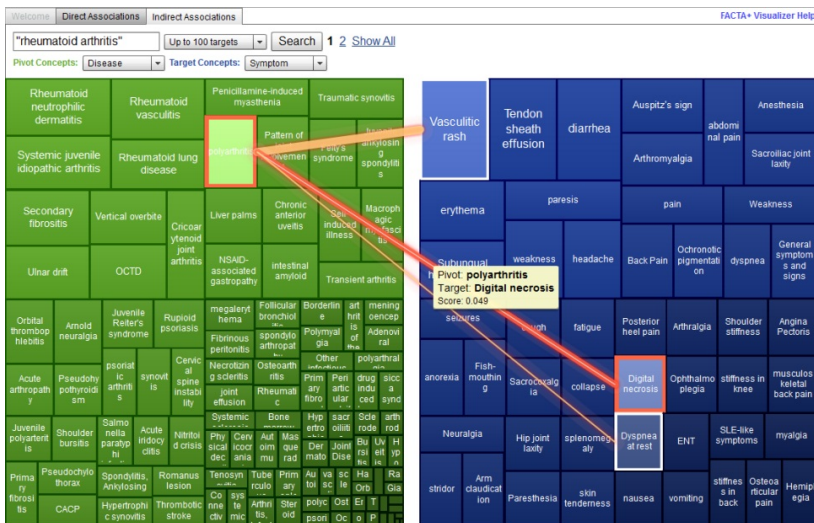


Fig. 3 FACTA+ Visualizer: Indirect associations between pivot concepts and related target concepts

Kleio is an advanced information retrieval (IR) system developed by NaCTeM and offers textual and metadata searches across MEDLINE and provides enhanced searching functionality by leveraging terminology management technologies [14].

Kleio draws upon one of the technologies from the NaCTeM text mining tool kit to enhance automated detection and mark-up of biologically important terms appearing in text, such as gene/protein names. One of these tools is AcroMine, which disambiguates acronyms based upon the context in which they appear. This functionality plays a key role in searching large document collections by allowing users to expand their queries and to include synonymous acronyms without losing the specificity of the original query.

The rich variety of term variants is a stumbling block for information retrieval as these many forms have to be recognized, indexed, linked and mapped from text to existing databases. Typically, most of the currently available information retrieval systems for the biomedical domain fail to deal with the problems of term ambiguity and variability. Kleio addresses this problem for reducing the diversity of term variation. Another key innovation of Kleio is dealing with the variety of names (terms) for denoting the same concept. To map these forms (e.g. IL2, IL-2 and Interlukin-2) to biological databases, machine learning based term normalization techniques which reduce term variation (e.g. il2) is used. An advantage of applying term normalization is to permit efficient look-up and to discover ambiguous and variant terms in the resources [14].

In order to develop a study to discover hidden relationships for biomedical entities such disease-disease relationships, a Web-based text mining tool can be used to find entity names and their co-occurrence frequencies in MEDLINE articles for each entity. Normalisation is another concern for text mining based studies. Biomedical names have some variations such as synonyms. These names need to be normalized to one specific name. For example, Wegener's Granulomatosis and Wegener's Granuloma indicate same diseases and can be mapped to Wegener's Granulomatosis. During the normalisation process, some biomedical resources should be used and interviewing with the biomedical experts can be needed [7].

Statistical techniques also play an important role for text mining studies [15]. There are some measures of co-occurrence analysis. The simplest method to identify relationships is using the co-occurrence assumption: terms that appear in the same texts tend to be related. For example, if a protein is mentioned often in the same abstracts as a disease, it is reasonable to hypothesize that the protein is involved in some aspect of the disease. The degree of co-occurrence can be quantified statistically to rank and eliminate statistically weak co-occurrences.

Web-based tools for discovering such relationships in medical literature may reveal new information and lead to a better understanding of certain concepts and therefore to higher quality of medical treatment.

Table 2 Comparison of three Web-based tools for analyzing biomedical literature


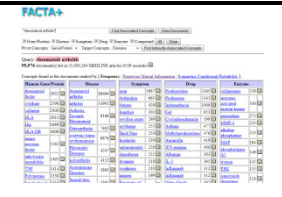

		
<p>Scope</p>		
<p>Finds associated concepts to a given concept.</p>	<p>Finds associated concepts to a given concept.</p>	<p>Search for concepts of certain categories. IR system supported by terminology management technologies</p>
<p>Ranking Algorithm(s)</p>		
<p>Proprietary PolySearch Relevancy Index, PRI [10]</p>	<p>Co-occurrence Frequency, Pointwise Mutual Information, Symmetric Conditional Probability</p>	<p>Date, Score</p>
<p>Data sources</p>		
<p>PubMed, OMIM, DrugBank, Swiss-Prot, HMDB, HPRD, GAD, HapMap, dbSNP, CGAP, HGMD</p>	<p>MEDLINE, UniProt, BioThesaurus, UMLS, KEGG, DrugBank</p>	<p>MEDLINE, BioThesaurus, acronym dictionary (mapping created from MEDLINE)</p>
<p>Strengths</p>		
<p>Use of biomedical thesauruses</p>	<p>Flexible queries. Indexing of concepts (→ quick search results) [11]</p>	<p>Acronym recognition and disambiguation. Normalisation of biology terms. Named entity recognition for gene/protein names. Indexing of terms. Reduction of term variation [14].</p>

Table 3 Comparison of three Web-based tools for analyzing biomedical literature (part two)

PolySearch	FACTA+	Kleio
Limitations		
Slow, Finds associated concepts belonging to only one single category. Novel or newly named terms are not recognized (simple dictionary approach to identify biological or biomedical associations) [10]	Limited synonyms/term variation support.	-
Supported concept categories (in: accepted as input; out: provided as output)		
Disease (in/out), gene/protein (in/out), Drug (in/out), Metabolite (in/out), SNP (RS#) (in/out), Gene sequence (in), Text word (in), Pathway(in/out), Tissue (in/out), Organs (out), Subcellular Localizations (out)	Human Gene/Protein, Disease, Symptom, Drug, Enzyme, Compound	Protein, Gene, Metabolite, Disease, Symptom, Organ, Diagnostic/therapeutic procedure, Medical phenomenon or process, Reagent or diagnostic aid, acronym, author, Publication type

3 Pointwise Mutual Information

A very interesting and useful concept based on information theory is mutual information.

Mutual Information (MI) goes back to Shannon (1948) [16] and is a measure of the mutual dependence between two random variables¹ X and Y . The measure itself and the instantiation for specific outcomes are called Pointwise Mutual Information (PMI). It has been introduced to the text mining community by Church & Hanks (1990) [17] as an alternative measure (association ratio) for measuring word association norms, based on the theoretic concept of mutual information. The association ratio can be scaled up to provide robust estimates of word association norms and has up to date proven to be a very useful association measure in Web-based text mining tasks [4].

Mutual Information can be seen as a measure of the information overlap between X and Y , where the values have probabilities $p(x)$ and $p(y)$. Consequently, the joint probability of $p(x, y)$ is defined as:

¹ Capitalized variable names refer to random variables.

$$I(X; Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Originally, Fano (1961) [18] used the $\log_2(x)$, however, any logarithm can be used, and changing the base of the logarithm changes the unit of measurement of information [19].

The information overlap between X and $Y = 0$, when the two variables are independent, as $p(x)p(y) = p(x,y)$. When X determines Y , $I(X; Y) = H(Y)$, where $H(Y)$ is the entropy of, or lack of information about Y , defined as:

$$H(Y) = - \sum_y p(y) \log p(y)$$

If X and Y are perfectly correlated, i.e. they determine each other, then $I(X; Y)$ reaches a maximum $H(X) = H(Y) = H(X, Y)$, where $H(X, Y)$ is the joint entropy of X and Y see [26], [27] for practical examples.

This leads to the definition of Fano (1961), who stated, that if two points P (information objects, e.g. words), x and y , have probabilities $P(x)$ and $P(y)$, then their Pointwise Mutual Information, $PMI(x, y)$ is defined as:

$$PMI(x, y) = \log \left(\frac{P(x, y)}{P(x)P(y)} \right)$$

In a recent study on disease-disease relationships for rheumatic diseases by Yildirim, Simonic & Holzinger (2012) this measure was used to discover the strength of a relationship and to provide an indication of how more often the query and the concept co-occur. After ranking of the measures and the frequencies together, the results revealed hidden knowledge in articles regarding rheumatic diseases indexed by MEDLINE. Such relationships can provide important additional information for medical experts and researchers for medical decision-making [4]. In its original form, the method is restricted to the analysis of two-way co-occurrences. Problems involving natural language processing, however, need not to be restricted to two-way co-occurrences; often, a particular problem can be more naturally tackled if it is formulated as a multi-way problem; consequently the framework of tensor decomposition, that has recently been introduced analyzes language issues as multi-way co-occurrences [20].

It was shown by [21] that a version of PMI trained on Wikipedia outperformed several publicly available measures of semantic relatedness and might even outperform LSA (latent semantic analysis) when trained with a sufficiently large amount of data. For practical applications this is interesting because similarity judgments are fast and easy to calculate using PMI, even on huge data sets. Furthermore, [22] showed that PMI based topic models coincide well with human perception. It can be summarized that the previously

mentioned sources show the eligibility of PMI for different scenarios in knowledge discovery tasks.

Pointwise Mutual Information is an ideal measure of word association norms based on information theory and we selected this measure to analyze rheumatic diseases. PMI compares the probability of observing two items together with the probabilities of observing two items independently. Therefore, it can be used to estimate whether the two items have a genuine association or are observed at random [12].

Let two words, w_i and w_j , have probabilities $P(w_i)$ and $P(w_j)$. Their mutual information $PMI(w_i, w_j)$ is defined as:

$$PMI(w_i, w_j) = \log \left(\frac{P(w_i, w_j)}{P(w_i) P(w_j)} \right)$$

The other way to discover hidden knowledge between biomedical entities is to use machine learning techniques to analyse the articles. At first, the co-occurrence frequencies of entity-entity can be extracted by using biomedical text mining tool. Most common of them are selected and normalized for each entity to create datasets. For example, relationships between diseases and symptoms can be explored. The frequencies of symptoms for each disease are found by using Web-based biomedical text mining tool. The frequency of the symptom provides the number of times a considered symptom appears in the articles. After normalizing the names, the dataset containing diseases and the frequencies of symptoms can be created. At the last stage, machine learning algorithms can be applied on the dataset to discover similarity between diseases. For instance, cluster analysis can be used to analyse the dataset. Cluster analysis is one area of machine learning of particular interest to data mining.

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster analysis has been also widely used in numerous applications, including pattern recognition, data analysis, image processing and biomedical research. In biomedical text mining studies, cluster analysis can be used to explore similarities between entities such as diseases, gene and organs [23].

4 Symmetric Conditional Probability

FACTA+ not only allows calculating co-occurrence frequencies and PMI, but also the symmetric conditional probabilities (SCP) for identifying associated concepts. [24] proposed SCP as measure for testing the correlation between terms x and y by multiplying the conditional probabilities of x given y and y given x .

$$SCP(x, y) = p(x|y) \cdot p(y|x) = \frac{p(x, y)}{p(y)} \cdot \frac{p(x, y)}{p(x)} = \frac{p(x, y)^2}{p(x) \cdot p(y)}$$

5 FACTAs Scoring Methods: Frequency, PMI, and SCP

In the following paragraphs we will exemplarily compare three scoring methods used by FACTA to rank the associated concepts to a given concept, specified by a textual query.

Table 4 shows the first 27 of 379 results for the query “rheumatoid arthritis” using FACTA+. The related concepts to the search term are listed in descending order, ordered by the score describing the relation to rheumatoid arthritis. Three scoring methods were compared: Frequency of co-occurrence (Frequency), Pointwise Mutual Information (PMI), and Symmetric Conditional Probability (SCP).

In order to get an impression of the “agreement” amongst the three methods, Kendall’s coefficient of concordance (Kendall’s W) was calculated [25]. Kendall’s W describes the agreement amongst raters concerning the ranking of items. In this case the ranking of the retrieved associated concepts is determined by the strength of the relation to rheumatoid arthritis. The “raters” are the tested methods Frequency, PMI, and SCP. A Kendall’s W value of 1 means complete agreement amongst the raters, a value of 0 means no agreement. Kendall’s W for all three methods the overall agreement (the agreement over all 379 result items) is 0.3778. Looking only at Frequency and PMI, the value is 0.5214. For PMI and SCP the value is 0.5577, and for Frequency and SCP it is 0.5210. We can see that, when looking at no more than two methods at the same time, PMI and SCP have a slightly higher agreement than the other combinations of two methods.

However, when looking at the 27 highest rated terms of each method (see Table 4) it can be observed that the methods Frequency and SCP top-rank similar terms while PMI top-ranks different terms.

In practice, when medical professionals use such tools to discover new relations between concepts, especially the highest ranked results are of importance. As mentioned before, in several studies it was shown that PMI has high performance for certain scenarios by not being computationally intensive at the same time. This makes PMI a good candidate for creating high quality Web-based applications.

Table 4 Comparison of FACTAs ranking of related concepts from the category Symptom for the query “rheumatoid arthritis” created by the methods co-occurrence frequency, PMI, and SCP

Frequency		PMI		SCP	
pain	5667	impaired body balance	7,8	swollen joints	0.002
Arthralgia	661	ASPIRIN INTOLERANCE	7,8	pain	0.001
fatigue	429	Epitrochlear lymphadenopathy	7,8	Arthralgia	0.001
diarrhea	301	swollen joints	7,4	fatigue	0.000
swollen joints	299	Joint tenderness	7	erythema	0.000
erythema	255	Occipital headache	6,2	splenomegaly	0.000
Back Pain	254	Neuromuscular excitation	6,2	Back Pain	0.000
headache	239	Restless sleep	5,8	polymyalgia	0.000
splenomegaly	228	joint crepitus	5,7	joint stiffness	0.000
Anesthesia	221	joint symptom	5,5	Joint tenderness	0.000
dyspnea	218	Painful feet	5,5	hip pain	0.000
weakness	210	feeling of malaise	5,5	metatarsalgia	0.000
nausea	199	Homan's sign	5,4	Skin Manifestations	0.000
Recovery of Function	193	Diffuse pain	5,2	neck pain	0.000
low back pain	167	Palmar erythema	5,2	Eye Manifestations	0.000
abdominal pain	141	Abnormal sensation	5,2	low back pain	0.000
cough	126	Gastric irritation	4,8	dyspnea	0.000
analgesia	120	Grip strength decreased	4,8	weakness	0.000
Pain, Postoperative	112	polymyalgia	4,8	Fever of Unknown Origin	0.000
vomiting	106	Pseudothrombophlebitis	4,7	nausea	0.000
neck pain	105	Deep granuloma annulare	4,6	dry eye	0.000
collapse	103	Axillary lymphadenopathy	4,5	diarrhea	0.000
discomfort	101	Calf pain	4,5	Epitrochlear lymphadenopathy	0.000
discomfort	97	gastrointestinal colic	4,3	ASPIRIN INTOLERANCE	0.000
Fever of Unknown Origin	81	Radiating pain	4,3	impaired body balance	0.000
myalgia	79	Musculoskeletal symptoms	4,3	Recovery of Function	0.000
Eye Manifestations	78	Arthralgia	4,2	myalgia	0.000

6 Conclusion

Optimal tools for quality-based text mining and knowledge discovery are of high importance for the MEDLINE database as it is growing extremely fast and will possibly grow even faster in the future. Without such tools many publications will

not be noticed by biomedical professionals, consequently much potentially useful information may be lost. Additionally, yet hidden knowledge can be made visible with knowledge discovery tools.

There is a large amount of Web-based tools available which make it possible to search the MEDLINE database and which allow the discovery of new knowledge, such as hidden relations between concepts. In this work we discussed and compared PolySearch, FACTA, and Kleio, while at the same time had a look on FACTAs ranking algorithms for associated concepts and Pointwise Mutual Information (PMI).

The quality of the results and therefore the applicability and the relevance of the algorithms used for text mining are essential. Moreover, the user interface of Web-based tools must not be neglected to support the accessibility medical professionals in an intuitive and effective way.

7 Future Work

A large number of Web-based tools are available for searching MEDLINE and for supporting knowledge discovery from the MEDLINE data. However, there are still many issues for research in this area: At first, the non-standardized nature of text is still a big issue, and there is much work left for improvement in the area of synonym recognition as, for example, could be seen during our investigation. Second, it can be stated that for efficient performance, the response time of the Web-based tool must be optimized. Therefore, further investigation is necessary in the optimization of existing algorithms as well as in optimal usage of the available server infrastructure in order to deliver results as quickly as possible. Third, research in end-user centred visualisation and visual analytics of the results is urgently needed in order to support efficient and fast sensemaking processes amongst medical professionals.

References

1. <http://www.ncbi.nlm.nih.gov/pubmed>
2. <http://www.ncbi.nlm.nih.gov/Entrez>
3. http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update
4. Holzinger, A., Simonic, K.M., Yildirim, P.: Disease-disease relationships for rheumatic diseases: Web-based biomedical textmining and knowledge discovery to assist medical decision making. In: 36th International Conference on Computer Software and Applications, COMPSAC, pp. 573–580. IEEE, Izmir (2012)
5. Kreuzthaler, M., Bloice, M.D., Faulstich, L., Simonic, K.M., Holzinger, A.: A Comparison of Different Retrieval Strategies Working on Medical Free Texts. *Journal of Universal Computer Science* 17, 1109–1133 (2011)
6. Solka, J.L.: Text data mining: theory and methods. *Statistics Surveys* 2, 94–112 (2008)
7. Yildirim, P., Çeken, Ç., Çeken, K., Tolun, M.R.: Clustering Analysis for Vasculitic Diseases. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) *NDT 2010*. CCIS, vol. 88, pp. 36–45. Springer, Heidelberg (2010)

8. <http://wishart.biology.ualberta.ca/polysearch/cgi-bin/help.cgi#eval1>
9. Lu, Z.: PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011 (2011)
10. Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., Wishart, D.S.: PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research* 36, W399–W405 (2008)
11. Tsuruoka, Y., Tsujii, J., Ananiadou, S.: FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* 24, 2559–2560 (2008)
12. Yildirim, P., Çeken, Ç., Hassanpour, R., Tolun, M.R.: Prediction of similarities among rheumatic diseases. *Journal of Medical Systems*, 1–6 (2010)
13. <http://refine1-nactem.mc.man.ac.uk/facta-visualizer/>
14. Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J., Ananiadou, S.: Kleio: a knowledge-enriched information retrieval system for biology (Year)
15. Schmeier, S., Hakenberg, J., Kowald, A., Klipp, E., Leser, U.: Text mining for systems biology using statistical learning methods, pp. 125–129 (Year)
16. Shannon, C.E.: A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 379–423 (1948)
17. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. *Computational Linguistics* 16, 22–29 (1990)
18. Fano, R.: *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge (1961)
19. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. *From Form to Meaning: Processing Texts Automatically*. In: *Proceedings of the Biennial GSCL Conference*, pp. 31–40. Günter Narr Verlag, Tübingen (2009)
20. Van de Cruys, T.: Two multivariate generalizations of pointwise mutual information. In: *Workshop on Distributional Semantics and Compositionality (DiSCo 2011)*, pp. 16–20. Association for Computational Linguistics (Year)
21. Recchia, G., Jones, M.N.: More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods* 41, 647–656 (2009)
22. Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. In: *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pp. 215–224. ACM, Gold Coast (2010)
23. Takada, T.: Mining local and tail dependence structures based on pointwise mutual information. *Data Min. Knowl. Discov.* 24, 78–102 (2012)
24. Ferreira da Silva, J., Pereira Lopes, G.: A local maxima method and a fair dispersion normalization for extracting multiword units from corpora. In: *Sixth Meeting on Mathematics of Language*, pp. 369–381 (Year)
25. Bar-Ilan, J.: Comparing rankings of search results on the web. *Inf. Process. Manage.* 41, 1511–1519 (2005)
26. Holzinger, A., Stocker, C., Peischl, B., Simonik, K.-M.: On Using Entropy for Enhancing Handwriting Preprocessing. *Entropy* 14, 2324–2350 (2012)
27. Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H., Fred, A.: On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) *AMT 2012. LNCS*, vol. 7669, pp. 646–657. Springer, Heidelberg (2012)

Chapter 8

An Information Reliability Index as a Simple Consumer-Oriented Indication of Quality of Medical Web Sites

Federico Cabitza

Abstract. Since typical healthcare consumers may lack sufficient knowledge to evaluate the reliability of health-related contents published online, recent researches are addressing the usefulness of Web page evaluation tools to help these consumers assess the quality of the indications they retrieve online. This paper contributes in this line by proposing an intentionally simple composite index of information quality, the so called Medical Information Reliability (MIR) index. This index takes the attitudes of potential and actual consumers toward information quality into account, and it is intended to be applied to online sources of medical information as “trust indicator” to provide their potential consumers with a simple percentage score by which to evaluate the reliability of what they are consulting. The main idea underlying this index is to consider information quality a multidimensional aspect of an online resource and relate it to the extent such a resource is compliant with explicit requirements formulated by third-party endorsement bodies. The method to calculate the MIR index on a sample of medical sites is presented in a step-by-step manner, and a user study is discussed that validated its application to the domain of the Alternative and Complementary Medicine.

1 Background and Motivations

Users rely on online resources in regard to their health for a number of reasons: e.g., to see if others complain their same symptoms and see how these had their disorders solved (especially in case of sensitive or socially stigmatized illnesses); to find the actual meaning of unfamiliar terms that had been used by healthcare

Federico Cabitza

Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy

e-mail: cabitza@disco.unimib.it

professionals in previous encounters; to sift opinions on the effectiveness of alternative treatments or on the reliability of a healthcare provider; to find support groups or just someone to exchange ideas and experiences with about their own health [1]. Looking for healthcare-related advice and information on the Web is easy, fast and extremely cheap, especially in comparison to getting access to the often well remunerated opinion of a doctor; this is the reason why this is a common behavior: approximately two thirds of adult population both in the US and in Europe claim to use the Internet for health care purposes [2]; this is also why an increasing number of people claim to ground their healthcare decisions on what they find on the Internet [3]. Our work lies in the research field aimed at providing final consumers and end-users of Web sites with a simple way to assess the reliability of the information they get access to online. In particular, we propose a methodology by which the content published by an online resource can be rated according to a series of standard domain-specific criteria, and be associated with a simple percentage indicator, the Medical Information Reliability (MIR) index; this index is aimed at making consumers aware of the extent online content has been produced according to quality-related guidelines. The main motivation backing this proposal, first presented in [4], lies in the need to address quality assessment from the consumer perspective and to make quality indicators intentionally simple to understand for lay consumers in critical domain like healthcare is.

Thus, our first motivation lies on the fact that for a consumer of health-related information to be able to assess the Information Quality (IQ) of an indication found on the Internet is particularly important [5]; this is especially true in light of two aspects: first, recently a number of research studies that reviewed Web sites providing healthcare-related information have detected that these sites presented several inaccuracies in their content; this finding raises serious concerns about the IQ that health consumers can encounter on the Internet [6–8]. Second: typical consumers of healthcare-related information online have only a limited knowledge of what they are reading (otherwise it is likely that they would not be seeking medical information through that means) and therefore they could be unable to judge its quality and reliability with full confidence by their own [9].

The second motivation for our proposal lies in the fact that the characteristics of individual consumers and their past interactions with a content provider in general are not sufficient to assess the actual reliability of a health-related Web site [10]; therefore there is a need for mechanisms of trust verification that are based on social institutions and intermediaries. To this aim, a number of initiatives have been conceived to help health information consumers seek, find and access high-quality information: these initiatives include gateway sites (portals), evaluation instruments and codes of conduct associated with some “surface markers” [11]. In particular, Chang and Cheung [12] have showed that third-party certifications are the most effective way for a Web site to gain trust from its prospective consumers when its reputation is unknown. The simplest mechanisms of this type are the so called “surface markers” and “trust indicators”; in particular, the mechanism that

is based on a “code of conduct” that is developed by a third party organization and that is associated to a label or logo is called “kite marking” [13]; a kite mark is like a “seal of quality” that a specific Web site can display if it is declaring to abide by the terms of such a code and if it is periodically found compliant with the code guidelines. By means of this kind of trust indicator, visitors of Web sites can get an idea of the IQ of an online resource [14], in the assumption that a certified trustworthy provider, thanks to its internal policies of IQ control, would always publish reliable contents [11]; in this case, then, an indication of trustworthiness is used as a predictor of accuracy [9,15]. Due to their immediacy and easiness to use, a growing number of organizations have recently developed codes of conduct that are associated to an evaluation and certification service that assigns kite marks in the healthcare domain (e.g., the HON code).

In its simplest terms, our proposal is a method to attach a percentage-based “kite mark” to a Web site, the so called MIR index; this is made according to i) a subjective evaluation of the compliance of the Web site to a set of domain specific codes of conduct (by a trained set of raters), and to ii) a consumer-oriented prioritization of the aspects that these codes of conduct regulate on a more or less prescriptive level. In this paper we will present next the original contribution, discussed in light of the relevant related work; then we will present a stepwise method to calculate the MIR index for a specific web resource; finally we will present a user study we undertook to validate its application to a specific healthcare domain, the domain of the Alternative and Complementary Medicine; to this aim, we will present the method adopted, the results and finally, we will discuss them in light of the main objectives outlined in the next section.

2 The Medical Reliability Index Score

The Medical Reliability Index is a *weighted composite index* that we conceived as an evaluation tool and systematic method whose output is a numerical “trust indicator”, to provide the users of an online health-related resource with a simple indication of its “level of reliability” and, hence, of the degree of IQ of the content published therein. In the proposal of the MIR index we have been driven by three main requirements, which we drew from the specialist literature regarding the kind of third-party certification mentioned in the previous section.

1. *Focus on the patients’ (information) needs.* According to [16], the service quality of electronic resources can not prescind from research initiatives that focus on the relative importance assigned by service consumers to different quality dimensions and perceived attributes. To this regard, it has been shown that consumers of health-related information usually develop a personal perception of how accurate they believe a content is by relying on visual elements, like layout, color schemes and icons, which are displayed by the online resource, rather than on content [17]. Moreover, as noted in [18] and [19] in regard to the IQ of the content available online, it is important to distinguish

between the needs of patients and the needs/expectations of healthcare professionals [20]: although for both categories of content consumers there is a need for a rigorous assessment of the quality of health-related websites [21], patients' needs are reported to value "trustworthiness" more than "availability" and "accessibility" [22], which are the main concerns of doctors. As the profile of the typical consumer of health-related content is changing over time, with patients and laymen becoming the main consumers of this kind of offer, it is recognized the increasing need to concentrate more on patients [23]. This calls for visual indicators of IQ that are "simple" and "straightforward" explicitly, since it is also reported that the majority of health information seekers do not check IQ-related indications in a consistent manner, like date of publication and original source of the information [24].

2. *Compliance to the so called "codes of conduct"*. These codes are proposed by a number of endorsement bodies and associations with the aim to guarantee the generic public of health-related Web sites that get the corresponding certification of compliance; these codes may differ with respect to several aspects, like the intended target population, the specific scope, and the declared and actual aims [25]. A recent survey has found that no code of conduct can be considered universally suitable to evaluate the IQ of different health websites [21]. This calls for the requirement that a good indicator must be a *composite* one, that is one that takes existing complementary IQ indicators and aggregates them together in some consistent and systematic way.
3. *Keep it simple, but not simpler*. Although trust indicators and quality seals, e.g., kite marks, are simple to understand even for laymen and straightforward in their meaning (simplistically put, if the indicator is present, then the "site is OK"; otherwise, nothing can be said about its IQ), it is noticed that they may give a false sense of security [26]. This calls for the requirement of avoiding dichotomic measures (e.g., good vs. bad quality mark, pass vs. no pass certification), but rather to adopt a numerical and percentage-oriented approach that could provide consumers with a more precise, yet still qualitative, indication of the extent the resource is compliant with domain-specific guidelines for IQ assurance.

In light of these three requirements, we devised the MIR score as a composite, percentage-like index whose numerical expression is obtained through a systematic eight-step evaluation method to express the extent IQ-related criteria have been met by a specific health-related information provider (online resource, or Web site in what follows) with respect to the ideality, i.e., 100% of the identified criteria being satisfied. More technically speaking, the MIR index is calculated according to Formula 1, where for each of the n IQ dimensions taken in consideration, w_i is the ranked weight for the i -th dimension d_i (e.g., accuracy); and c_k is the k -th IQ criterion associated to d_i (i.e., such that $F(c_k) = d_i$ ¹), and evaluated for a specific Web site, s ; tc_i is the total number of criteria to be met in regard to dimension d_i . See Table 1 for more details on the meaning of the variables involved in Formula 1.

¹ Or, also, the pair $(c_k, d_i) \in M$, that is a set introduced in Table 1.

$$MIR(s) = \frac{\sum_{i=1}^n \left(w_i \bullet \left(\frac{\sum_{k=1}^m c_k(s)}{tc_i} \right) \right)}{\sum_{i=1}^n w_i} \quad (1)$$

The fact of presenting the MIR score as a percentage is obviously aimed at facilitating laymen consumers in understanding “how reliable a content provider is” with respect to a conventional “upper limit” (i.e., all requirements met), which in a traditional “kite mark” approach would be associated to the issuing of a single trust indicator. Beside indicating also a partial compliance of a given online resource with respect to the available best practices (i.e., absolute benchmarking), the fact that the MIR score is numeric allows also for paired-sample comparisons and, consequently, for its adoption as an (internal) audit tool for the continuous improvement of the IQ of health-related content: i.e., on the one hand, it enables the homogeneous comparison of different online resources (i.e., relative benchmarking) and hence their ranking in online directories, gateway portals or search engines; on the other hand, it enables the progressive evaluation of a single resource over time, and hence to detect trends in IQ policies and actual performance.

3 The Evaluation Process Related to the MIR Index

Besides simplicity, another element that is worthy of note in regard to MIR as an evaluation tool is its capability to be “tailored” to different needs, aims and domains, as the case study that we will present on the Complementary and Alternative Medicine will show. This aspect derives from the recognition, mentioned above and reported also in [21], that a truly universal and semi-automatic evaluation tool would be overambitious and probably practically infeasible. Thus, Formula 1 presents two variables, namely the weights by which IQ dimensions are prioritized (i.e., w) and the criteria by which IQ is assessed (i.e., c) that can either be set once and for all; or be object of tailorization according to the actual uses that are intended for the MIR score (e.g., site valorization, benchmarking, trend analysis, continuous IQ improvement, information retrieval). In this light, IQ dimensions (i.e., a third parameter, d , that does not show up in Formula 1) are just a user-centered way to prioritize IQ success criteria, that is a way to take the users’ perceptions and preferences into account [16] (cf. the first requirement mentioned above). This called for the conceptual separation between the evaluation tool (and related score) and the evaluation process; in its turn, the latter one can be further distinguished into a phase of “adaptation” (or inspection [32]) of the tool, where criteria by which a health-related site is considered reliable (or not) and their weights are uniquely identified and set; and a phase of “use” of the tool, where Web sites are manually checked against the above identified criteria and a numerical score is attached to these sites at a given time.

This process can be further articulated in a stepwise manner by identifying eight distinct tasks:

1. Identification of the IQ Dimensions involved
2. Identification of the IQ Criteria involved
3. Criteria Categorization
4. Prioritization of the IQ Dimensions
5. Weight Definition
6. Site review
7. Score calculation
8. Score dissemination

In Table 1, we describe this evaluation process in some details and indicate, for each step listed above, some techniques that can be adopted for its execution, and the intended outputs.

Table 1 The evaluation process toward the definition of a MIR score for a generic online resource

Step No.	Description	Technique(s) involved	Step Outputs
1	IQ Dimension Identification, and definition / characterization of each IQ dimension in simple but unambiguous terms.	User study (survey); Focus group (Delphi method); Literature review; or a combination of these.	A set D , of n IQ dimensions: $D = \{d_1, \dots, d_n\}$. E.g.: d_1 is Completeness, d_2 is Accuracy, d_3 is Timeliness.
2	IQ Criteria Identification and characterization to the original formulation expressed by third-party endorsement providers.	Literature review; Focus group (Delphi method); or a combination of both.	A set C of m success criteria: $C = \{c_1, \dots, c_m\}$, with c_i being a Boolean function that evaluates the i -th criterion, i.e., a single requirement by which to assess the quality of a resource, such as: $c_i : S \rightarrow \{0, 1\}$, with S set of web sites (s) under review. E.g., given a web site s_i , c_1 : "in s_i is the source of information always identified?"; c_2 : "in s_i is the contact information for the site administrator displayed?"; c_3 : "in s_i are medical and other disclaimers posted and easily accessible?".

Table 1 (continued)

3	Categorization of the criteria defined in Step 2 in terms of the dimensions defined at step 1.	Categorization through inspection; Coding through reliable keyword matching (cf. content analysis and inter-coder reliability assessment).	A collection M, of all ordered pairs M: $\{(c_i, F(c_i))\}$ with i from 1 to m with F: $C \rightarrow D$, i.e., a function by which a single success criterion is associated with a single IQ dimension. In other words, we obtain a list of dimensions operationally defined in terms of IQ specifications. E.g., see Table 2.
4	IQ dimension prioritization	User study; Focus groups (Delphi method) or a combination of these.	A total order $\succ \subseteq D \times D$ (cf. D in step 1) by which $d_1 \succ d_2 \succ \dots d_{n-1} \succ d_n$; in other words, we obtain an ordinal ranking of the IQ dimensions found in step 1. E.g.: 1) Accuracy; 2) Completeness; 3) Timeliness.
5	Weight definition and assignment to ordinal ranks.	Literature review; Introspection; Focus groups (Delphi method), or a combination of these.	An ordered set W of weights, $W = \{w_1, \dots, w_n\}$, where w_i is to be associated with a specific d_k , with $i = k$.
6	Review of an online resource (s) and check of its content against the criteria defined at Step 2.	Evaluation by a (pool of) trained expert(s), be it either extensive or upon a random sample of pages from a web site (s).	A collection E, of all ordered pairs E: $(s, C_i(s))$ with C_i defined at step 2. In other words, by applying all the c_i in C to s, we obtain a list of dichotomic evaluation scores (1/0), one for each IQ criterion.
7	MIR score calculation	See Formula 1	The MIR score for s at time T_1 .
8	MIR score dissemination	Site Directory (aka gateway providers); Kite Mark; or both.	

In light of the process outlined in Table 1, two more points are worthy of note. First, although seemingly redundant, it is important to distinguish between Step 6, i.e., the criterion-by-criterion review of an online medical resource, and Step 7, i.e., the to some extent “mere” calculation of the MIR score that follows this review. This is because the evaluation of a Web site with respect to its compliance

with the identified IQ criteria is conceptually, as well as operationally, a different task from associating such a review with a numerical score. This latter task could be repeated for different sets W of weights in order to choose the optimal one that, e.g., makes important differences between, e.g., competing sites more manifest; obviously in this case, there would be no need to replicate the review; or Step 6 could be assigned to a pool of evaluators that, in a similar way to the coding task of Step 3, are collaboratively called to reach a consensus on what criteria are really met in all those cases this is not a trivial task but rather something that requires experience and interpretative skills².

Second: although strictly stepwise and linear in its overall structure, the MIR evaluation process is intrinsically iterative in all of the steps of the adaptation phase, from Step 1 to Step 5 (see Figure 1). All these steps can encompass a collaborative process in which, respectively relevant IQ dimensions, success criteria and weights are defined in a progressive manner through increasing levels of consensus within a group of prospective users, analysts or domain experts. Moreover, the overall process is intended to loop from Step 8 back to Step 6 (the Use phase in Figure 1) to enable continuous IQ improvement and benchmarking, once the IQ dimensions, criteria and weights set in the adaptation phase have been held constant, of course.

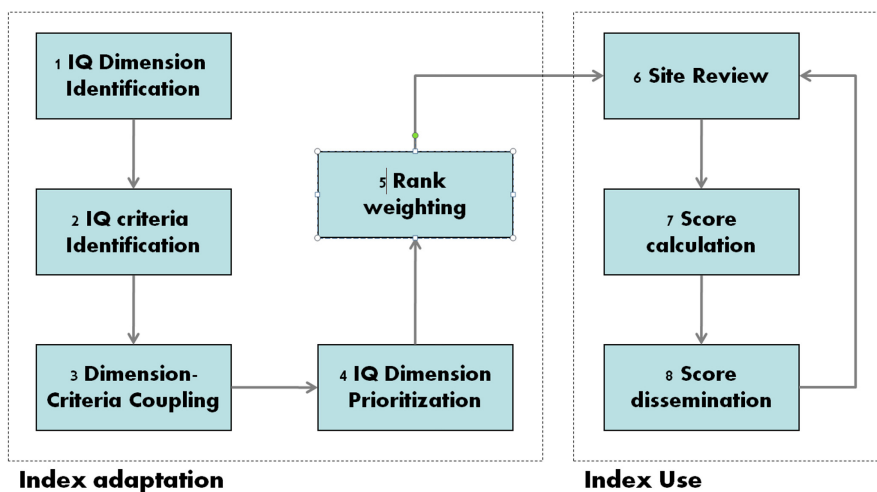


Fig. 1 Graphical representation of the process of the MIR index adaptation and use

4 Validation of the MIR Index in the Medical Domain

In the Introduction we have already made the point of how, due to the intrinsic heterogeneity and extent of the World Wide Web, the quality of health information that a consumer could find online can differ a lot [27]; and how this can be

² In any case, step 7 can be executed only if set W and E (see table 1) are available.

related to legitimate concerns that trusting low quality (or simply not verified) content on health-related matters could have even serious consequences on the consumers' health [28]. This is specially true in those ambits of medicine where there is still a lack of institutional roles acting for consumers as reliable providers of information, like doctors and pharmacists; this is the case of Non Conventional, Complementary and Alternative Medicines, that is the field that is usually referred with the acronym NCM/CAM (in what follows just CAM for simplicity's sake). For this reason, we decided to deploy the MIR evaluation tool in this specific domain and proceed with a preliminary validation that could address the evaluation of the reliability of some Web sites that provide their customers with advices, indications, results from the specialist literature and market news about CAM-related products and remedies in the Italian and English speaking contexts.

The National Center for Complementary and Alternative Medicine (NCCAM) defines CAM as the broad set of healing remedies and resources that are either complementary or alternative to those established within the conventional health-care practice in a particular society [29]. According to the World Health Organization, more than half of the European citizens have used CAM remedies and the expenditure trend on this kind of medicines, therapies and practices is increasing year by year [30].

4.1 Validation Method

To validate the adaptability of the MIR index to the CAM domain and its feasible application to related Web sites, we designed a user study that could follow the step-wise process outlined in Table 1. This study used a mixed methodology with multiple data sources. In particular, in the first step we analyzed the pertinent literature to identify the main dimensions along which IQ in CAM Web resources is usually measured; the same approach was followed in Step 2 to identify a set of criteria that should be met by a Web site to be considered a reliable source of health-related information. Step 3 was conducted by a small panel of coders following the tenets of content analysis [31] and software engineering inspection methods [32]. The prioritization task involved in Step 4 was conducted in virtue of the results of an exploratory empirical user study; in this study, we surveyed a convenience sample of 101 healthcare information consumers on their attitudes and quality expectations for CAM-related information also on the basis of their previous experience with the pursuit and retrieval of such information on the Internet [33]. The participants have been invited to respond to an online questionnaire where they were given the set of IQ dimensions identified in Step 1 and were asked to assess the perceived importance of each dimension with respect to their tasks of seeking and consulting CAM-related information on an ordinal scale ("very important", "important", "moderately Important"; "of little importance") [9]. On the basis of these subjective assessments, we produced a ranking of the dimensions in order of *perceived importance*. Finally, we have tested the applicability of this metrics by applying the MIR index to a convenience selection of Web

sites that publish CAM-related content and hence drew a first indication of their reliability according to our proposal.

4.2 Results

A comprehensive literature survey (including, e.g., [8,10,14,34–39]) allowed us to realize that there is no single list of IQ dimensions or attributes that could be considered the “gold reference” for healthcare website evaluation. For instance, Kim et al. reviewed 29 rating tools presenting explicit criteria to assess health-related Web sites: the most frequently cited criteria regard currency of information, authority of source, ease of use, accessibility/availability, disclosure of authors and content accuracy [37]. Eysenbach, in a systematic review on how health website are evaluated in the specialist literature, found that the most frequently used quality criteria regard accuracy, completeness, readability, design, disclosures, and references provided [8]. More recently, Stvilia et al. [14] analysed thousands of e-mail communication instances in the IPL's Q&A service archives from 2005 to 2007 and identified seven IQ criteria to be relevant to healthcare IQ judgments: accuracy, authority, completeness, currency, objectivity, relevancy, and understandability. In 2009, O'Grady [38] developed an evaluation framework for health Web sites that encompasses the IQ dimensions of: content accuracy, credibility, completeness, understandability, relevance, level of personalization, privacy, security, usability, and accessibility; even more recently (2011), Tao et al. [20], on the basis of a user study focusing on the perspective of healthcare consumers, identified a taxonomy of IQ attributes encompassing understandability, completeness, reputation, adequacy of reference, relevancy, accuracy, site reputation among others. In light of these and other studies, we defined a list of six IQ dimensions that could cover the main quality-related aspects with CAM information published online; for each of these dimensions, we formulated a definition with no ambition of sound comprehensiveness, but rather with the aim to help the coders involved in Step 3 reach a sufficient level of agreement, as well as to provide the participants of the empirical study accomplished in Step 4 with a common ground and shared definition of the terms used in the questionnaire.

The IQ dimensions we adopted for this study are:

- (i) *accuracy*, expressed as ‘the extent a piece of information is true and reliable according to either reality or a “gold standard” reference’ (e.g., a medical dictionary or textbook, a scientific paper);
- (ii) *completeness*, defined as ‘the extent a piece of information is reported in a complete way to inform its consumers according to their needs’;
- (iii) *accessibility*, as ‘the extent a piece of information is easy to be found and consulted’ (cf. availability);
- (iv) *currency*, as ‘the extent a piece of information is up-to-date’;
- (v) *usefulness*, as ‘the extent a piece of information is easy to understand (cf. understandability) and apply to a specific context or need’;
- (vi) *authority*, as ‘the extent the author or source of a given piece of information is known and considered trustworthy’.

For Step 2, a second literature survey was aimed at identifying the main criteria CAM-related Web site must meet to obtain a third-party certification of the quality of its content (e.g.,[40] [10, 11]). We identified three codes of conduct that could fit our aims well:

- 1) The so called “HONcode”, which is issued by the Health On the Net Foundation [12] in terms of a list of eight general requirements that a Web site must satisfy on a yearly basis to get and maintain the related certification.
- 2) The Web Feet Health Criteria for Site Selection³, a detailed collection of criteria that was part of a larger collection of indications to retrieve high quality information on the Internet for school, business and library purposes.
- 3) The checklist issued by the NCCAM, which encompassed ten “questions” addressing as many aspects to consider to judge a Web site a reliable source of CAM-related information (or not).

These evaluation tools were chosen mainly on the basis of the recent and comprehensive literature review reported in [11]: we included the HONcode and the Webfeet collection as these ones were found to cover a superset of the IQ-related aspects addressed by most of the existing other evaluation instruments; and also because they were found to be the most different and hence complementary ones, in terms of rank correlation. We adopted also the NCCAM checklist as this was found to be the only one specifically designed for the CAM domain. In doing so, we were confident to extract from these tools all the main recurring aspects that are covered by the evaluation instruments available online [11], while, at the same time, to also consider the most specific instrument for the domain at hand, and therefore take into account all the relevant aspects or success criteria related to IQ mentioned in the literature. The resulting 42 criteria extracted in Step 2 of our validation study are reported in Table 2.

Table 2 The list of criteria selected from the literature survey. The list is used to evaluate c_k in Formula 1. A requirement is intended to be either satisfied for a site s ($c_k(s) = 1$) or not satisfied ($c_k(s) = 0$), according to an evaluation/judgment task performed by a trained rater.

Evaluation Tool	Criteria Checklist
NCCAM	<ol style="list-style-type: none"> 1. Is who runs this site explicitly reported or otherwise clear? 2. Is Who pays for the site explicitly reported or otherwise clear? 3. Is the purpose of the site explicitly reported or otherwise clear? 4. Is Where the information comes from explicitly reported or otherwise clear? 5. Is What the basis is of the information explicitly reported or otherwise clear? 6. Is How the information is selected explicitly reported or otherwise clear? 7. Is How current the information is explicitly reported or otherwise clear? 8. Is How the site chooses links to other sites explicitly reported or otherwise clear? 9. Is What information about you the site collects (and why) explicitly reported or otherwise clear? 10. Is How the site manages interactions with visitors explicitly reported or otherwise clear?

³ This tool, now apparently discontinued, can be found at the following URL:
<http://www.webcitation.org/5QFjc1Qjk>

Table 2 (continued)

HONcode	<ol style="list-style-type: none"> 1. Authoritative: Are the qualifications of the authors clearly stated? 2. Complementarity: Is it clear that information provided should support, not replace, the doctor-patient relationship? 3. Privacy: Is the privacy and confidentiality of personal data submitted to the site by the visitor fully respected? 4. Attribution: Are the source(s) of published information, date medical and health pages properly cited? 5. Justifiability: Are claims relating to benefits and performance properly backed up? 6. Transparency: Is content presented in an accessible way, and contact information accurate? 7. Financial disclosure: Are all funding sources properly identified and acknowledged? 8. Advertising policy: Is advertising content clearly distinguished from editorial content?
Web Feet Health	<ol style="list-style-type: none"> 1. Is it true that Source of information is identified? 2. Is it true that The contact information for the source or site administrator is displayed? 3. Is it true that The expertise and reputation of the source are considered? 4. Is it true that The expertise and reputation of the site's host are considered? 5. Is it true that The information is not easily available at other sources? 6. Is it true that Reviewers (clinicians, subject-area experts, and researchers) make every effort to ensure that the information is free of errors? 7. Is it true that The information and images are objective, balanced, and unbiased? 8. Is it true that The information has sufficient scope to cover the topic for the intended audience? 9. Is it true that The information is readable and free of spelling and grammatical errors? 10. Is it true that Sponsorship is clearly indicated, and advertising is minimal? 11. Is it true that Medical and other disclaimers are posted? 12. Is it true that Site is updated frequently, typically indicated by a recent "last updated" date? 13. Is it true that Pages list the date of the most recent update and/or the dating of the information is made clear in an accessible area of the site? 14. Is it true that Links work, and they are relevant and appropriate? 15. Is it true that The site loads in a reasonably short time? 16. Is it true that The site is easy to access and navigate? 17. Is it true that Navigation includes clear headings and intuitive icons, menus, and directional symbols that foster independent use? 18. Is it true that Standard multimedia formats such as HTML are used? 19. Is it true that Most information is accessible without special plug-ins such as Adobe Acrobat Reader? 20. Is it true that Logical options are available for printing and downloading all or selected text or graphics? 21. Is it true that The site follows good graphic design principles? 22. Is it true that Information for specific audiences, such as consumer information within a professional site, is easy to locate? 23. Is it true that The site has a text size that is easy to read for the intended audience? 24. Is it true that Product advertising is not intrusive?

In regard to Step 3, we enrolled three coders (including the author) and provided them with the list of the IQ dimensions (and related definitions) detected at Step 1 as the shared “codebook” by which they were called to classify the IQ criteria independently of each other [41]. A score of inter-rater reliability was then calculated by means of the KALPHA macro by Andrew F. Hayes for SPSS v. 17.0, obtaining a Krippendorff’s Alpha score of 0.69. This value is usually associated with a less than optimal reliability, and therefore with exploratory conclusions only (as it is below the conventional threshold for high reliability, i.e., $K \geq 0.8$ [42]); nevertheless, this result made us confident that a representative coding scheme could be eventually found by the coders involved in a subsequent phase, when the resulting pairs criteria-dimension were openly discussed in a Delphi-like manner [43]. In Table 3, we report the result of this collaborative task of characterization of the IQ criteria found in Step 2 in terms of the IQ dimensions identified in Step 1.

Table 3 The definition of each IQ dimension regards what was agreed upon by coders in step 3 as well as the definition given to participants in step 4

IQ Dimension	Definition	Coding of criteria
Accuracy (tc _d : 6 criteria)	This dimension mainly relates to the requirement that the site should not contain commissions, i.e., misleading statements likely to cause physical harm [13].	NCCAM: 5 HONcode: 4 Web Feet Health: 3, 6, 7, 9
Completeness (tc _d : 11 criteria)	This dimension mainly relates to the requirement that the site should not contain omissions, i.e., vital information that should have been mentioned and that All claims should be justified with appropriate references to scientific sources.	NCCAM: 2, 4, 5, 6 HONcode: 5, 7 Web Feet Health: 1, 2, 8, 10, 13
Accessibility (tc _d : 16 criteria)	This dimension mainly relates to how easy it is to find a content that is pertinent to one’s own needs, as well as to retrieve it again over time; thus, this dimension also relates to the persistence of the content itself and to its uniquely identifiability.	NCCAM: 5, 10 HONcode: 8, 4, 5, 6 Web Feet Health: 2, 14, 15, 16, 17, 18, 19, 21, 22, 23
Currency (tc _d : 2 criteria)	This dimension mainly relates to how timely a new content is produced after that a related scientific evidence has been produced and published in the specialist literature, that is the extent the content site is up-to-date with respect to the available knowledge.	NCCAM: 7 HONcode: None Web Feet Health: 12
Usefulness (tc _d : 8 criteria)	This dimension mainly relates to the extent a site presents content that is understandable, interesting, and therefore valuable for the intended consumers; this dimension also relates to the extent an intended consumer can take advantage of the information consulted, that is how easily a piece of information is applicable to either everyday needs or more specific information requirements.	NCCAM: 3 HONcode: 2, 8 Web Feet Health: 8, 11, 16, 20, 22

Table 3 (continued)

Authority (tc _d : 7 criteria)	This dimension mainly relates to the extent the provider is considered reliable, trustworthy and able to satisfy information needs of its customers, and consequently, how easy it is to trace to the author or provider and assess its reliability.	NCCAM: 1, 4 HONcode: 1, 4 Web Feet Health: 1, 4, 5
--	--	--

To perform Step 4, we conducted an online survey that involved a convenience sample of healthcare information consumers that were questioned about their attitudes and quality expectations for CAM-related information.

The survey was conceived as a Computer Assisted Web Interview (CAWI) that was delivered through an online questionnaire platform (Limesurvey⁴). Participants were recruited among acquaintances and colleagues of the author and students of his classes, and were invited to join the study either through a personal email or being forwarded to the survey page through posts published on the main social networks (i.e., Facebook, MySpace e Twitter); word of mouth did the rest. The questionnaire was kept open for 18 days and closed on September 2011, when 101 completed forms had been collected. In Table 4 we report the demographic data extracted from the sample of respondents and the segmentations that were performed for the inference statistical study.

Table 4 Selected demographic information of the respondent sample involved in the empirical study

Characteristic	Options	Responses %
ICT skills	Elementary or basic	39.6
	Advanced or expert	60.4
Interest toward CAM	Very low or low	39.6
	High or very high	60.4
Knowledge about CAM	Very Poor or poor	69.3
	Good or very good	30.7
Frequency for CAM	From Never to Sometimes	68.4
	Frequently or Very frequently	31.7

In Table 5 we report the results coming from the user study. These regard, for each IQ dimension, the average ranking and the ordinal category that better represents the average attitude toward that dimension (i.e., the median of the response distribution). The average rank for each IQ dimension has been calculated counting how many times that dimension was considered the most important among the other ones, how many times the second one, the third one and so on; and then calculating the arithmetic mean of the total score. This is just one of the ways in which large samples of respondents can be challenged about relative rankings without asking them for a ranking directly, so as to minimize acquiescence bias; other techniques can be obviously adopted, as one from those reviewed in [44].

⁴ <http://www.limesurvey.org/>

Table 5 Average Ranking from the empirical study for each IQ dimension

IQ Dimension	Overall Rank⁵	Median Perception⁶
Accuracy	1.56	Very important
Completeness	1.60	Very important
Accessibility	2.07	Very important
Currency	2.22	Very important
Usefulness	2.43	Important
Authority	2.96	Important

Due to the convenience-driven recruitment, the user study employed in Step 4 presents the common limitation of not being based on a sample that is fully representative of the target population, i.e., potential consumers of CAM-related information. To limit non response bias, we stratified the sample into subgroups by age, education, familiarity with ICT (ICT skills in Table 4), knowledge and interest toward CAM, frequency with which information on CAM remedies is either sought or consulted (see Table 4). We associated the subgroups with dichotomic variables (the options column in Table 4) and performed a Mann-Whitney U test (since the assessments were performed on an ordinal scale) for each specific IQ dimension: no significant difference at a 95% confidence level was found among these groups with regard to their assessments of the perceived importance of the IQ dimensions. This fact, as well as the relatively large number of respondents, does not eliminate the bias due to accidental sampling, but makes the results consistent with the requirements of marketing research [45], i.e., suitable to detect attitudes and preferences in potential consumers of CAM-related information.

According to the ranking derived from the user study accomplished in Step 4, we adopted one of the simplest weighting function and assigned each IQ dimension to its corresponding weight, starting from Accuracy ($w_d = 6$, in Formula 1) to Authority ($w_d = 1$), with unitary decrements.

Subsequently, we proceeded in Step 6: in this step, the list of criteria reported in Table 2 is intended to be consulted by a neutral rater⁷ or by anyone specifically instructed to check whether the n -th criterion, associated with the k -th IQ dimension, is actually met by the Web site under evaluation, or not. We then reviewed five Web sites that we selected on a convenience basis among those that were publishing CAM-related information on a daily basis at the time of the validation and we calculated the MIR score for each of these online resources. In Table 6 we report the results from this purposely exemplificatory evaluation.

⁵ Smaller number indicates higher importance. Values are close since we did not forced the respondents to chose a rank for each IQ dimension explicitly (to minimize random bias) but we derived this indication from their single assessments.

⁶ The median values of the distribution of “perceived importance” that are reported in the rightmost column are equal to the modes for that variable, i.e., the value that has been chosen by the majority of the respondents.

⁷ In the case at hand, the author made the review.

Table 6 MIR scores applied reviewing five CAM-related web sites, visited in Summer 2010

Web Site	MIR index score
Italiasalute	42.0%
Viveremeglio	42.0%
Mentalhelp	75.9%
Wiki4cam	51.6%
Altmeds	62.2%

5 Discussion and Concluding Remarks

In this paper we have presented the Medical Information Reliability (MIR) index, a composite weighted score of IQ intended to facilitate healthcare consumers in the task of judging the reliability of an online resource in lack of sufficient knowledge to perform this task without external visual aids.

With respect to other post-hoc IQ evaluation instruments that have been proposed with similar purposes in the healthcare domain [11], the MIR index is novel for its modularity, simplicity and consumer-centredness. First, the MIR index can integrate multiple IQ criteria from instruments that are issued and maintained by various certification bodies. This integration requires only to associate each new dichotomous criterion with the pertinent IQ dimension. Also the ranking weights can be adjusted over time to better fit either local or specific target readerships. Second, the MIR index is purposely conceived as a simple percentage indication of the extent a Web site is compliant with the best practices and guidelines for IQ assurance, where 100% indicates a fully compliant site. As such, it is a tool for health-related information consumers to support them in getting an idea of the reliability of a source of content published in the Web; it is also a tool for gateway providers [11] and, potentially, search engines to refer visitors to better online resources and benchmark them; and it is also a tool for Web sites managers, maintainers and owners, as a means to achieve and guarantee continuous improvement in the eyes of their customers. Lastly, the MIR index is innovative for the idea to include the concept of “requirement prioritization” in a synthetic score: this concept, which is borrowed from the requirement engineering field [32], has inspired the ranking of homogeneous groups of criteria in terms of more understandable meta-level concepts, i.e., the concept of IQ dimension, and suggested to base such a ranking on the actual attitude of potential consumers of health-related information. As part of the further research that is needed to assess the actual value of such a tool, we applied the evaluation process and resulting score to a panel of Web sites publishing consumer-oriented content periodically in the field of the Complementary and Alternative Medicine. This domain was chosen not only because it is receiving strong interest by an increasing population of consumers, but also because the lack of institutional roles and bodies (at least in Europe) that could issue certified indications and proven evidences of effectiveness from the

field makes the development and testing of evaluation tools that could contribute in improving the reliability of online resources a pressing need and an interesting challenge in the agenda of both Academic and professional research.

Acknowledgements. The author would like to gratefully acknowledge the work of Marco Buonvino, who contributed during his Master Degree Thesis in both the definition of the MIR score and in accomplishing the most part of the user study.

References

1. Silience, E., et al.: How do patients evaluate and make use of online health information? *Social Science & Medicine* 64(9), 1853–1862 (2007)
2. Atkinson, N.L., et al.: Using the Internet for Health-Related Activities: Findings From a National Probability Sample. *Journal of Medical Internet Research* 11(1), e4 (2009)
3. Baker, L.: Use of the Internet and E-mail for Health Care Information: Results From a National Survey. *JAMA: The Journal of the American Medical Association* 289(18), 2400–2406 (2003)
4. Cabitza, F.: Introducing a Composite Index of Information Quality for Medical Web Sites. In: *Quality of Life through Quality of Information - Proceedings of the 24th Conference of the European Federation for Medical Informatics, MIE 2012, August 26-29, Pisa, Italy (2012)* (forthcoming)
5. Gustafson, D.: Evaluation of ehealth systems and services. *BMJ* 328(7449), 1150–1150 (2004)
6. Bernstam, E.V., et al.: Commonly cited website quality criteria are not effective at identifying inaccurate online information about breast cancer. *Cancer* 112(6), 1206–1213 (2008)
7. Silberg, W.M., et al.: Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet: Caveant Lector et Viewer—Let the Reader and Viewer Beware. *JAMA* 277(15), 1244–1245 (1997)
8. Eysenbach, G.: Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review. *JAMA* 287(20), 2691–2700 (2002)
9. Xu, Y., et al.: Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology* 57(7), 961–973 (2006)
10. Bailey, B.P., et al.: An examination of trust production in computer-mediated exchange. In: *Proceedings of the 7th Conference on Human Factors and the Web (2001)*
11. Breckons, M., et al.: What Do Evaluation Instruments Tell Us About the Quality of Complementary Medicine Information on the Internet? *JMIR* 10(1), e3 (2008)
12. Chang, M.K., et al.: Online Trust Production: Interactions among Trust Building Mechanisms [Internet], p. 181c. IEEE (cited January 27, 2012)
13. Delamothe, T.: Quality of websites: kite marking the west wind. *British Medical Journal* 7, 843–844 (2000)
14. Stvilia, B., et al.: A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology* 60(9), 1781–1791 (2009)

15. Spink, A., et al.: From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management* 34(5), 599–621 (1998)
16. Zeithaml, V.A., et al.: Service quality delivery through web sites: A critical review of extant knowledge. *Journal of the Academic of Marketing Science* 30, 362–375 (2002)
17. Stanford, J., et al.: Experts vs. online consumers: A comparative credibility study of health and finance web sites. *Consumer Reports WebWatch* (2002)
18. Potts, H.W.W., Wyatt, J.C.: Survey of Doctors' Experience of Patients Using the Internet. *Journal of Medical Internet Research* 4(1), e5 (2002)
19. Sillence, E., et al.: Trust and mistrust of online health sites. In: *Proceedings of CHI 2004*, pp. 663–670 (2004)
20. Tao, D., et al.: Consumer Perspectives on Quality Attributes in Evaluating Health Websites [Internet]. *IEEE* (2012)
21. Hanif, F., et al.: The role of quality tools in assessing reliability of the Internet for health information. *Informatics for Health and Social Care* 34(4), 231–243 (2009)
22. Pletneva, N., et al.: Results of the 10th HON survey on health and medical Internet use [Internet]. *Health on the Net Foundation*, Geneva (2010)
23. Bliemel, M., Hassanein, K.: Consumer satisfaction with online health information retrieval: A model and empirical study. *E-Service Journal* 5(2), 53–84 (2007)
24. Fox, S.: *Online Health Search 2006* [Internet]. *PEW Internet & American Life Project* (2006)
25. Baur, C., Deering, M.J.: Proposed frameworks to improve the quality of health websites: review. *Medscape General Medicine* 26(e35) (2000)
26. Hanif, F., et al.: The quality of information about kidney transplantation on the World Wide Web. *Clinical Transplantation* 21(3), 371–376 (2007)
27. Kalichman, S.C.: Quality of Health Information on the Internet. *JAMA: The Journal of the American Medical Association* 286(17), 2092–2095 (2001)
28. Schmidt, K., Ernst, E.: Assessing websites on complementary and alternative medicine for cancer. *Annals of Oncology* 15(5), 733–742 (2004)
29. Goldstein, M.S.: The growing acceptance of complementary and alternative medicine. In: Bird, C.E., Conrad, P., Fremont, A.M. (eds.) *Handbook of Medical Sociology*, pp. 284–297. *Pearson Education Limited*, Harlow (1999)
30. Roberti di Sarsina, P., Iseppato, I.: Looking for a Person-Centered Medicine: Non Conventional Medicine in the Conventional European and Italian Setting. *Evidence-Based Complementary and Alternative Medicine*, 1–8 (2011)
31. Krippendorff, K.: *Content analysis: An introduction to its methodology*. Sage, Thousand Oaks (2004)
32. Holzinger, A.: Usability Engineering for Software Developers. *Communications of the ACM* 48(1), 71–74 (2005)
33. Holzinger, A., et al.: The effect of previous exposure to technology on acceptance and its importance in usability and accessibility engineering. *Universal Access in the Information Society* 10(3), 245–260 (2010)
34. Eysenbach, G., Koehler, C.: How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 324(7337), 573–577 (2002)
35. Anderson, K.A., et al.: A Systematic Evaluation of Online Resources for Dementia Caregivers. *Journal of Consumer Health on the Internet* 13, 1–13 (2009)
36. Lorence, D., Abraham, J.: A study of undue pain and surfing: using hierarchical criteria to assess website quality. *Health Informatics Journal* 14, 155–173 (2008)

37. Kim, T.R., Deearing, M.J., Maxfield, A.: Published criteria for evaluating health related web sites: Review. *British Medical Journal* 318, 647–649 (1999)
38. O’Grady, L., et al.: Measuring the Impact of a Moving Target: Towards a Dynamic Framework for Evaluating Collaborative Adaptive Interactive Technologies. *Journal of Medical Internet Research* 11, 9 (2009)
39. Charnock, D., et al.: DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology & Community Health* 53(2), 105–111 (1999)
40. Goldschmidt, P.G.: A Report on the Evaluation of Criteria Sets for Assessing Health Web Sites [Internet]. Health Improvement Institute and Consumer Reports WebWatch (2003)
41. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* 22, 249–254 (1996)
42. Krippendorff, K.: Reliability in Content Analysis. Some Common Misconceptions and Recommendations. *Human Communication Research* (3), 411–433 (2004)
43. Okoli, C., Pawlowski, S.D.: The Delphi method as a research tool: an example, design considerations and applications. *Information & Management* 42(1), 15–29 (2004)
44. Valadares Tavares, L.: A model to support the search for consensus with conflicting rankings: Multitrident. *International Transactions in Operational Research* 11(1), 107–115 (2004)
45. Fricker, R.D., Schonlau, M.: Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Methods* 14(4), 347–367 (2002)

Chapter 9

Challenges for Search Engine Retrieval Effectiveness Evaluations: Universal Search, User Intents, and Results Presentation

Dirk Lewandowski

Abstract. This chapter discusses evaluating the quality of Web search engines to effectively retrieve information. It identifies three factors that lead to a need for new evaluation methods: (1) the changed results presentation in Web search engines, called Universal Search, (2) the different query types that represent different user intentions, and (3) the presentation of individual results. It discusses implications for evaluation methodology and provides some suggestions about measures.

Keywords: Web search engines, retrieval effectiveness, evaluation, Universal Search, search engine results page (SERP), user behavior.

1 Introduction

Quality is important in all information retrieval (IR) systems, including Web search engines. The goal of this chapter is to discuss methods for evaluating Web search engines with a focus on the current standards for results presentation and on users' intentions.

The quality of Web search engines is of great importance, as users may choose their preferred search engine based on its perceived quality. The quality of the different search engines is also of great interest to search engine vendors (to improve their system or to benchmark their system with others) and the general

Dirk Lewandowski
Hamburg University of Applied Sciences, Faculty DMI, Department of Information,
Finkenau 35, D – 22081 Hamburg, Germany
e-mail: dirk.lewandowski@haw-hamburg.de

public. Search engines have become a major means to acquire knowledge, and the results they show in the first positions have a great influence on the information that users actually consume.

Evaluation is traditionally an integral part of information retrieval research, and it pays particular attention to a user's examination of the results list presented by the information system from top to bottom. Evaluators also assume that the user's decision to choose one site over another is based on reading the abstract (snippet) presented in the results list.

However, the presentation of search results in Web search engines has changed in recent years, and user behaviour has followed suit. While simple lists were dominant for several years, nowadays, results from different document collections (such as news, images, and video) are presented on the search engine results pages (SERPs) (Höchstötter & Lewandowski, 2009). This type of presentation is called Universal Search, which is the composition of search engine results pages from multiple sources. While in traditional results presentation, results from just one database (the Web index) are presented in sequential order and the presentation of individual results does not differ considerably, in universal search, the presentation of results from the different collections is adjusted to the collections' individual properties.

A search engine results page (SERP) is a complete presentation of search engine results; that is, it presents a certain number of results (determined by the search engine). To obtain more results, a user must select the "further results" button, which leads to another SERP. On a SERP, results from different collection, a.k.a. vertical search engines can be presented. Contrary to the general-purpose search engine, a vertical search engine focuses on a special topic.

The properties of the different results types lead not only to a different presentation of the results pages but also to a different presentation of the individual results. For example, it is clear that SERPs that include image and video results should show preview pictures. Figure 1 shows an example of a Universal Search results page, Fig. 2 provides examples of the presentation of an individual result.

In Figure 1, we can see how results from different sources (i.e., specialized search engine indices or so-called vertical search engines) are injected into the general results list created from the Web index (i.e., the search engine's main index). Additional results in this case come from the image index and from the news index.

In Figure 2, we see a typical results description provided in a results list. It contains a title, a URL, a short description, and, in this case, a social recommendation.

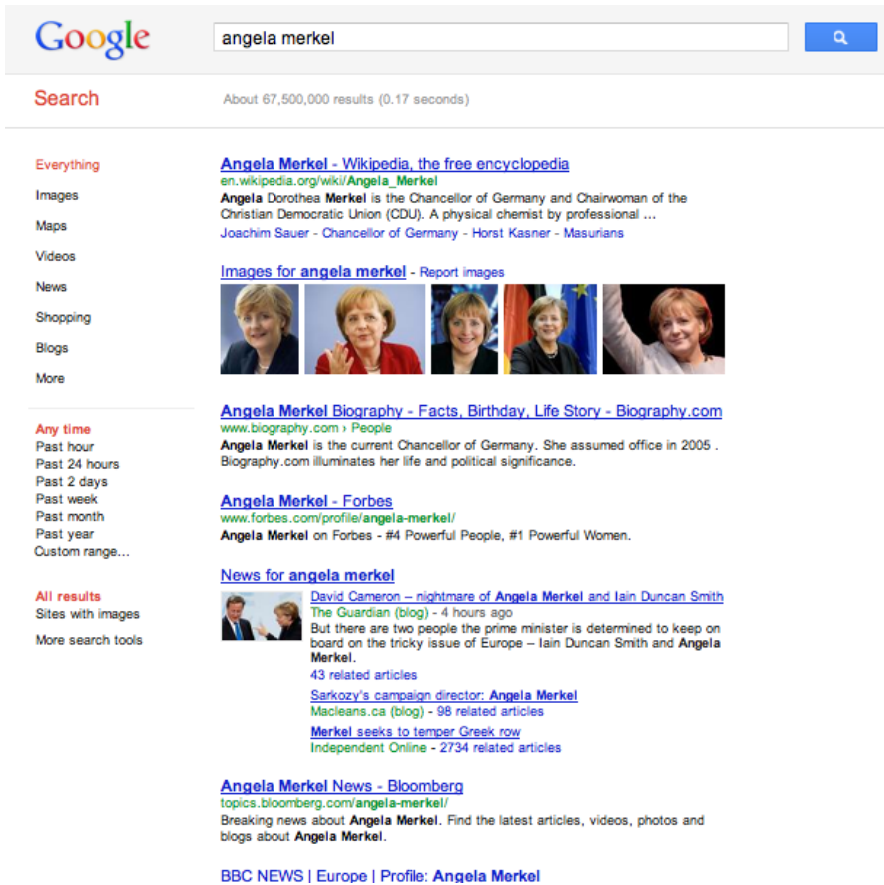


Fig. 1 Search engine results page (example from Google)



Fig. 2 Example of an individual result description (example from Google)

While Universal Search is a concept usually associated with Web search, it may also be applied to such diverse offerings as Web portals, e-commerce websites and Intranets. Therefore, the discussion presented in this chapter may also be applied to search scenarios other than Web searches.

Along with the positioning of the results, the different representations of search results determine the search engine users' viewing and clicking behaviour. Another important factor is the user's goal when entering a query. The classic distinctions between query types posited by Broder (2002) are informational,

navigational, and transactional intentions that are the basis for a further discussion on retrieval effectiveness and success.

To summarize, we will discuss search engine evaluation in the context of

- Results presentation (design of Universal Search results pages)
- Query types
- Results selection

It is obvious that the challenge when measuring the retrieval effectiveness of Web search engines is to develop evaluation methods that consider the three areas mentioned. This chapter provides methods used to evaluate Universal Search results pages and suggestions for designing retrieval effectiveness studies. The structure of this chapter is as follows: First, we will give a short overview of search engine evaluation, then we will discuss users' intentions (as expressed through different types of queries). After that, we will detail Web search engines' results presentations and users' selection behaviour on the search engine results pages. Bringing these three areas together, we will discuss points to consider when evaluating search engines with a Universal Search results presentation. The chapter closes with some conclusions and suggestions for further research.

2 Search Engine Evaluation

When discussing search engine evaluation, it is important to stress that quality measurement goes well beyond retrieval effectiveness, i.e., measuring the quality of the results. Some frameworks for a more complete search engine evaluation have been proposed (e.g., Xie, Wang & Goh, 1998; Mansourian, 2008). Lewandowski and Höchstötter's model (Lewandowski, Höchstötter, 2008) divides Web search engine quality into four major areas:

- **Index Quality:** This area of quality measurement indicates the important role that search engines' databases play in retrieving relevant and comprehensive results. Areas of interest include Web coverage (Gulli, 2005); country bias (Vaughan & Thelwall, 2004; Vaughan & Zhang, 2007), and freshness (Lewandowski, 2008a; Lewandowski, Wahlig, & Meyer-Bautor, 2006).
- **Quality of the results:** Derivates of classic retrieval tests are applied here. However, one needs to consider which measures should be applied and whether or not new measures are needed to satisfy the unique character of a search engines and its users (Lewandowski, 2008b).
- **Quality of search features:** A sufficient set of search features and a sophisticated query language should be offered and should function reliably (Lewandowski, 2004, 2008c).
- **Search engine usability:** The question is whether it is possible for users to interact with search engines in an efficient and effective way.

While all quality factors mentioned are important, results quality is still the major factor in determining which search engine performs best. A good overview of

newer approaches in Web search engine retrieval effectiveness evaluation is provided by Carterette, Kanoulas, and Yilmaz (2012).

In retrieval effectiveness evaluation, two approaches need to be differentiated:

1. “Classic” retrieval effectiveness tests use a sample of queries and jurors to judge the quality of the individual results. These studies use explicit relevance judgements made by the jurors. An overview of search engine retrieval effectiveness studies using explicit relevance judgements is provided by Lewandowski (2008b).
2. Retrieval effectiveness studies analyse click-through data from actual search engine users (e.g., Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G., 2007; Jung, Herlocker & Webster, 2007). As users give their relevance judgements only through their selection behaviour, we speak of implicit relevance judgements here.

Both approaches have their merits. When using click-through data, researchers can rely on large quantities of data and can determine which results are preferred by the actual users of a search engine. The drawback, however, is that these decisions are based on the results descriptions on the SERPs that heavily influence users’ results selections, and users choose only from some of the results presented. For example, a user would not read all the results descriptions and then choose a result from the third results page. On the contrary, he will rely on the first results presented by the search engine and choose from them.

The main advantage of classic retrieval effectiveness tests are that no data from the search engine providers is needed, and jurors can be asked for their opinions, so a researcher can go beyond decisions about whether an individual result is relevant or not. The drawback of such tests, however, is that such studies usually rely on a relatively low number of queries and jurors, and results are seen as independent of one another. This can be illustrated by a user who chooses a completely relevant result and will therefore not need another relevant result that just repeats the information already given.

From this short discussion, one can readily see that a combination of the two approaches described would be the best way to go. However, researchers usually do not have access to click-through data from real search engines, so only the search engine vendors themselves usually do this type of evaluation. However, in recent years, some search engine providers have made datasets including click-through data available to the research community, but there are still many cases where search results need to be evaluated and the researcher does not have access to such data. Furthermore, it is highly unlikely that a researcher will get this data from more than one search engine for benchmarking purposes.

We would like to stress that a basic assumption in this discussion on search engine evaluation is that the researchers do not have access to data owned by the search engines considered. Therefore, we will discuss the use of click-through data merely as an addition to other approaches.

As this short discussion reveals, a challenge in measuring the retrieval effectiveness of Web search engines is to develop evaluation methods that consider

both the results presentation and user behaviour. However, search engine retrieval effectiveness studies to date still lack the integration of explicit user models. Based on the Cranfield paradigm, many evaluation models have been built that provide a robust framework for conducting search engine retrieval effectiveness tests. These evaluations are based on retrieval measures, which largely measure the performance of systems, even though newer approaches try to integrate explicit user models into such evaluations (Carterette et al., 2012). What even integrating users models into the evaluation frameworks may not be enough; furthermore, an explicit *results presentation model* is needed.

3 User Intents

Search engines are used for a multitude of purposes, including navigating to an already known website, simple fact-checking, complex research tasks, and even entertainment purposes. It is clear that, depending on the task, a user will enter different types of queries into the search box. The difference between Web search engines and other information retrieval (IR) systems is usually seen as search engines having no thematic focus and being used by laypersons. But it must be stressed that another important point that differentiates them from other IR systems is that they are used for different user intents, expressed through query types. A simple, yet powerful classification of user intents is Broder's (2002) query type classifications: informational, navigational and transactional.

Navigational queries are used by a person who knows about a Web page or assumes that it exists (for example, the homepage of a company like ebay or people like Angela Merkel). Such queries normally terminate in one correct result. The information need is satisfied when the requested page is found.

In contrast, informational queries require more than one document. The user requires problem-oriented information (see Frants, Shapiro, & Voiskunskii, 1997). The user wishes to become informed about a topic and therefore intends to read several documents. Informational queries aim at static documents to acquire the desired information, which makes further interaction with the Web page unnecessary.

Transactional queries, however, aim at Web pages offering the possibility of a subsequent transaction, such as the purchase of a product, the download of data, or the search of a database.

Broder's query type classification has been refined over the years by researchers like B. J. Jansen, Booth, and Spink (2008), Kang and Kim (2003) and Rose and Levinson (2004). However, all rely on the basic distinction between the three types mentioned. While it is perfectly legitimate to use a more fragmented query type classification, the general query type classification is sufficient for our purposes.

Evaluating Web search engines' performance for the query types mentioned requires adjusting methods and retrieval measures. Many studies on informational queries have been conducted (for an overview, see Lewandowski, 2008b), but

only a few studies have examined navigational queries (Lewandowski, 2011; for an overview of measures, see MacFarlane, 2007). Evaluating search engines' performance on transactional queries is especially difficult, as they require an interaction on the results.

4 Results Presentation in Web Search Engines

Search engine results pages are usually simple lists of search results ("10 blue links"). Each result is presented in the same format, and, apart from a position bias, every result has the same chance of being selected. However, this is not the case, as different result types are now presented on the SERPs, results are given different amounts of space, and some results are highlighted in either through the use of a different background colour or additional graphical elements. This section provides a discussion of the major elements of a SERP. For a more detailed discussion, see Höchstätter and Lewandowski (2009) and Lewandowski and Höchstätter (2009).

As shown in Figure 1, the search engine results pages contain a variety of elements. In the following paragraphs, we will discuss these elements in detail.

4.1 Organic Results

Organic results (also called algorithmic results) are the listings on a search engine results page that are not paid for (Jansen, 2011). The organic results list forms the core of the results presentation, and its ranking is produced by algorithms that aim to determine their relevance. Thereby, all documents in the search engine's index are treated the same.

There is no direct human intervention in the results listings, although in recent years there has been some discussion about whether search engines favour their own offerings (Edelman, 2010). Making an analogy to journalism, Nicholson et al. (2006) spoke of the organic results as "editorial listings."

The basic elements of a description used in an organic result are a heading, a short description, and a URL. Other elements, such as a date and a link to a cache copy, can also be shown. In recent years, search engines have further enriched their results descriptions with ratings (Figure 3a), pictures (Figure 3b), information from social networking sites (Figure 2), and citation information for scholarly articles (Figure 3c). This means that results are no longer equally represented and that more factors than the results position and the contents of the description are influencing users' results selections. Studies show that such "rich snippets" can help users make quicker and more relevant decisions (Li & Shi, 2008), but they can also distract users from relevant results.

The so-called site links are particularly important in organic results. Here, a search engine enriches the result on the first position of the organic list in a special

way. The search engine provides links to popular destinations within the website and describes the homepage. Such a description occupies a lot of space on the SERP.



Fig. 3 Added information shown within results descriptions: (a) ratings from reviews, (b) picture and author information, (c) author and citation information

4.2 *Sponsored Results*

Sponsored results are also called sponsored links or AdWords (named after Google’s sponsored results product). These are text-based advertisements that are presented on the SERPs and are a context response to the query entered by the user. Advertisers are charged for clicks on their ads, and the price per click (PPC) is determined through an auction process (J. Jansen, 2011).

Sponsored results are usually shown on the right side of and on top of the organic results. This means that a user scanning the SERP from top to bottom first sees the sponsored results, which are labelled as advertisements with words such as “ads”, “sponsored”, etc.), and are often highlighted through the use of a different background colour, a technique not used in the organic results. For example, Google uses white as the background colour for the organic results and a light yellow background for the sponsored results.

The basic elements used in a sponsored result are a heading, a short description and a URL. The descriptions of sponsored results can also be enriched by ratings or other graphical information.

One can easily see that the descriptions of sponsored results are built using techniques similar to those used in organic results. As sponsored results are also shown in response to a query and can therefore be relevant to the query (B. J. Jansen, 2007), users might confuse sponsored results and organic results. However, there are only a few studies that have examined this (Bundesverb & Digitale Wirtschaft, 2009; Fallows, 2005), so no clear judgement on this can be given here. However, as we know that users do consider sponsored results (for whatever reason), this results type should be considered in a search engine evaluation.

4.3 *Shortcuts*

Shortcuts provide a direct answer to the query on the SERP itself, and the user does not need to click on a result on the SERP. Shortcuts come from special databases maintained by the search engines, and their inclusion is usually triggered by query words. Search engines have been criticised for injecting results from their

own collections into the SERPs (Edelman & Lockwood, 2011), but others have argued that this practice is completely understandable and does not negatively influence the quality of the results (Sullivan, 2011).

Figure 4 shows an example of a shortcut for a search for the weather in Hamburg. The basic elements of a shortcut cannot be provided as the SERPs vary and they are fitted to the individual topic.



Fig. 4 Example of a shortcut

4.4 Results from Special Collections

Results from special collections (sometimes just called “universal results”) are results that do not come from the general Web index of the search engine but from specialized databases, such as the news index, the video index, or the image index. These results are injected into the list of organic results but are often presented differently. A basic element of results descriptions from special collections is a clickable heading that leads to results obtained only from the vertical index in question. For example, when users click on the heading for the news results, they get a results list containing only results from the news index.

Further elements of the results descriptions are dependent on the document type. While in video results, a title, a still and the source may be shown, results from the database of scientific articles provide very different information. This distinction between the presentations of different document types makes evaluation difficult.

To summarize, there are different results types presented on the search engine results pages. Furthermore, while the presentations of these results are in some ways similar, the presentations differ according to types.

When evaluating Web search results, the amount of space given to an individual result and the graphic presentation of the result should be considered, as both influence users’ results selections. When results presentation changes, so needs evaluation methodology.

5 Users’ Results Selection

While search engines usually return thousands of results, users are not willing to view more than a few (Jansen & Spink, 2006; Keane, O’Brien, & Smyth, 2008; Lorigo et al., 2008; Machill, Neuberger, Schweiger, & Wirth, 2004), so, for all

practical purposes, the first page of a search engine's results is considered most relevant. Even in cases when users are willing to view more than the first few results, the initial ranking heavily influences their perceptions of the results set. That is, when a user does not consider the first few results to be relevant, he or she usually modifies the query or abandons the search.

The first results page must also be used to present results from additional collections, as users usually do not follow links to these additional collections, the so-called "tabs". Search engine expert Danny Sullivan (2003) coined the term "tab blindness" to name this phenomenon. Results beyond organic results also require space on the results page. With the addition of more and more results from special collections (and, in some cases, the addition of more and more ads above the organic results), we have observed a general change in results presentation (Lewandowski, 2008d). Organic results become less important as additional results take their place.

The changed design of the search engine results pages also leads to a larger number of results presented on these pages. While in the classic list-based results presentation, ten results per page (+ads) were displayed, a typical Universal Search results pages can easily contain 30 or even more clickable results links. When considering this more complex results presentation, users' results selection is based on a multitude of factors (see Kammerer & Gerjets, 2012).

An important question behind users' results selection behaviour is whether it is based on an informed decision. By this we mean whether the user actually makes a cognitive judgment about the quality of the individual results and selects the most appropriate result(s) or whether he just clicks on a link shown prominently. In the ongoing discussion on "search neutrality") the position of the search engine vendors is that users indeed make an informed decision (Granka, 2010; Grimmelmann, 2010). If that were not the case, one might question the search engines' revenue model, and if users were not able to distinguish between organic and sponsored results and therefore clicked on ads without knowing that they were doing so, regulation authorities would have to force search engines to label sponsored results more clearly than they are currently doing.

In the following section, we will discuss search results presentation factors influencing users' results selections. We will consider the results position, the results description content, its size and its design.

5.1 Results Position

When considering the probability of a result being clicked, one has to first consider whether the result is included in the visible area of the SERP. The visible area is defined as the part of the SERP that is directly visible to the user, without scrolling down the results list. The size of the visible area depends on the user's screen size, the size of the browser window and the size of the browser's navigation elements, as well as browser toolbars that a user may have installed.

Research shows that users first and foremost consider results within the visible area (also called “above the fold”). However, it is important to understand that navigational queries make up a large proportion of queries, and if the search engine works well, the one desired result will be shown within the visible area, so no scrolling will be necessary. Even when making informational queries, many users desire only one or a few results, and such queries can be satisfied with the results in the visible area.

The second point to consider is the actual ranking of the results. Results presented first get considerably more attention than results presented lower in the ranked lists (Cutrell & Guan, 2007; Hotchkiss, 2006; Joachims, Granka, Pan, Hembrooke, & Gay, 2005; Pan et al., 2007). However, when considering the results positions, one must also decide whether to see the list of sponsored results and the list of organic results as two separate lists or as one continuous list. Eye-tracking research suggests that at least some users consider these two areas as one continuous list (Wirtschaft, 2009).

5.2 Results Description Content

It is evident that users are influenced by the actual contents of the results descriptions. There is a high probability that users will click on results whose descriptions contain the query words and have meaningful titles. However, there is little research on the importance of the results descriptions’ content. Evidence is mainly given by search engine optimizers who emphasize the influence of good description copy for generating clicks on the results (Thurrow, 2007; Thurrow & Musica, 2009) Pirolli (2009) used the concept of information scent to explain the influence of the description content, stating that users follow trails if they are given hints about where to find the desired information.

5.3 Results Description Size

As stated above, the results descriptions occupy different amounts of space on the search engine results page. We can measure this space in pixel size and weigh the probability that a user will click it. We assume that the more space a results description occupies, the higher the probability that a user will click on it.

5.4 Results Description Design

Some results may be presented with additional images or may be emphasized by the use of colour. For example, sponsored links may be highlighted in yellow. Users may perceive these results as being more important and are therefore more likely to click on them. Moreover, given that users know that results highlighted in such a way are advertisements, they may avoid them. Therefore, it is not easy to make a decision about whether to ascribe more or less weight to these results rather than organic results.

While sponsored results are usually highlighted in colour, other results (whether organic or from a special collection) may carry images that draw the users' attention. However, this attention is hard to measure because if the information is displayed on a page cluttered with enriched results, it may get less attention than a result that is the only one highlighted.

In summary, we can see that users' results selections are influenced by a variety of factors. However, the importance of these factors is not easy to measure. Apart from large-scale click-through studies that are usually provided by the search engine vendors themselves and do not allow for comparing search engine, lab-based experiments that track eye movements are used to research selection behaviour (Beiler, 2005; Pan et al., 2007). The advantage of lab-based studies is that the researchers can also ask participants about their reasons for selecting certain results.

6 Approaches to Weighting Results on the SERPs

In a previous study (Lewandowski, 2012), we presented a general framework for measuring the quality of search engines' results. While the influence of results presentation on users' results selections is a part of the model, we intend in this chapter to focus on the weighting of individual results within the results presentation. The factors to be weighted are discussed in the following sections.

6.1 Results to Be Considered

First, one needs to decide whether all results presented on the SERP should be considered or not. The main concern is usually whether one should consider the results in the visible area vs. results in the invisible area and organic vs. sponsored results.

A researcher might decide that results in the invisible area will not be considered as most users do not consider them. However, one should also consider that the size of the visible area varies from user to user. Therefore, one can classify users according to their screen sizes (e.g., "W3 Schools Browser Statistics," 2011).

When deciding whether to include sponsored results that are shown above the organic results in the evaluation, one must decide whether to see them as separate from the organic results or to consider sponsored and organic results as one continuous list. When giving weights to sponsored results, one must decide whether highlighting leads users to this type of results.

6.2 Position

Weighting results according to position is an element included in most retrieval effectiveness measures. However, in the case of evaluating Universal Search

SERPs, one must decide which results types are to be considered as an element of the results list, i.e., those that actually have a position.

6.3 Screen Real Estate

“Screen real estate” is a term introduced by Nielsen & Tahir (n.d.) to measure the amount of space taken up by a Web page that is filled with content in relation to the part that is left blank. The term has been used in different scenarios (e.g., Dewan, Freimer, & Zhang, 2002). Nicholson et al. (2006) used the screen real estate measurement to calculate the space that search engines grant organic results vs. paid results on the first results screen shown to the user. While this is a specific application in the context given, screen real estate can be calculated for every result presented by a search engine. The use of screen real estate may seem obvious when results are presented in different formats; however, it can also be used for list-based results presentation if all results are not equally presented.

Weighting screen real estate can either be applied to the whole SERP or just the visible area. However, one must consider that the size of the visible area is different, depending on the browser window size utilised by individual users.

6.4 Click-Through Rates

As mentioned above, click-through data is usually not available to researchers if they are not affiliated with a search engine vendor. However, when available, the importance of individual results can be weighted according to their CTR. In other words, results with a higher CTR are assigned a higher importance.

As CTRs for the individual queries are often not available, one could also use the average CTR for certain query types. When considering navigational queries, one can see that a large ratio of clicks is aggregated to just one result position. The distribution is wider for informational queries, so one must give more value to more results in the evaluation.

7 Combining Universal Search SERPs and User Intent

When dealing with navigational queries in list-based results pages, one assumes there is one relevant result that should be displayed in the first position of the list, but this raises two problems. First, there is not always a clear distinction between query types. A query can be both navigational and informational. The query type may differ from user to user. Even in the eBay example given above, there may be users who wish to get information about the company and do not wish to navigate to eBay’s website. By using click-through data, however, researchers can see that the overwhelming ratio of users do see this query as navigational.

Considering a more complex example, the query about Angela Merkel depicted in Figure 1 can be navigational or informational. The user may wish to navigate to

her personal website, but he may also want to get information about her life and work. Furthermore, the user may be interested in current news.

This example shows that informational and navigational intentions can be considered in one results list. Furthermore, even when the intent is informational, it may matter considerably whether the user is seeking current information or a general overview of the life and work of Angela Merkel. In such cases, Universal Search results pages can satisfy diverse user intentions.

In evaluating results quality in Web search engines, it is always important to distinguish between query types. It becomes even more important in the context of a Universal Search, as multiple query intentions can be considered within the same results presentation. Therefore, we suggest collecting information on the underlying query intentions for every query one evaluates. If the researcher has access to click-through data from a search engine, navigational queries can be identified reliably. If such data is not available, we suggest asking users about query intentions and using this information to derive weighted query intentions, as Huffman and Hochster (2007) did in their study.

8 Conclusion

In this chapter, we showed that current approaches to Web search engine retrieval effectiveness evaluation have shortcomings and discussed new factors that can be used in evaluations. Often, studies do not differentiate between query types and results that are presented differently so they also ignore the probability of their being selected. We propose an approach that considers all these elements. However, while we were able to show how search engine evaluation is influenced by the decisions of the researcher in this regard, we were not able to provide an empirical evaluation.

Further research needs to examine users' approaches to different results. In particular, the status of sponsored results has been ignored by researchers. While search engine vendors make an overwhelming proportion of their incomes from sponsored results, we still do not know how users actually perceive these results. Some studies suggest that a fairly large ratio of users is not able to differentiate between organic and sponsored results (e.g., Bundesverband Digitale Wirtschaft, 2009), but we need a thorough study on this topic.

In addition, studies that ignore the different results types do not exactly measure what the user gets to see in his searches. As search engine evaluation at least tries to model the user's behaviour, researchers should consider empirical studies of all results types.

Researchers also need to consider browser sizes. While it surely is valid to calculate retrieval effectiveness measures for the whole SERP, the user usually only recognizes a part of it. Users focus on the visible area of the results pages (Höchstötter & Lewandowski, 2009), so results in this area should be considered foremost. Going even further, one could also use eye-tracking research to determine which sections of the results pages are actually seen or considered by the users.

Finally, we must say that there are more open questions in search engine evaluation than answers. We think that in this chapter we have raised more important questions rather than providing answers. Nonetheless, we can see that Web search engine evaluation is not merely a technical issue but also has a societal dimension. When we consider the billions of queries entered into Web search engines each day (ComScore, 2009) and understand that search engines influence searchers' selection behaviours through their results presentations, we recognize that search engine evaluation techniques need to be applied to measuring "search neutrality", i.e., a fair representation of the Web's contents in search engines.

References

- Beiler, M.: Selektionsverhalten in den Ergebnislisten von Suchmaschinen. Modellentwicklung und empirische Überprüfung am Beispiel von Google. In: Machill, M., Schneider, N. (eds.) *Suchmaschinen: Neue Herausforderungen für die Medienpolitik*, vol. 50, pp. 165–189. VISTAS Verl, Berlin (2005)
- Broder, A.: A taxonomy of Web search. *ACM Sigir Forum* 36(2), 3–10 (2002)
- Bundesverband Digitale Wirtschaft, Nutzerverhalten auf Google-Suchergebnisseiten: Eine Eyetracking-Studie im Auftrag des Arbeitskreises Suchmaschinen-Marketing des Bundesverbandes Digitale Wirtschaft (BVDW) e.V. (2009)
- Carterette, B., Kanoulas, E., Yilmaz, E.: Evaluating Web retrieval effectiveness. In: Lewandowski, D. (ed.) *Web Search Engine Research*, pp. 105–137. Emerald, Bingley (2012)
- ComScore. Global search market draws more than 100 billion searches per month (2009), http://www.comscore.com/Press_Events/Press_Releases/2009/8/Global_Search_Market_Draws_More_than_100_Billion_Searches_per_Month (retrieved)
- Cutrell, E., Guan, Z.: Eye tracking in MSN Search: Investigating snippet length, target position and task types. Technical Report, TR-2007-01 2007), <http://research.microsoft.com/pubs/70395/tr-2007-01.pdf> (retrieved)
- Dewan, R., Freimer, M., Zhang, J.: Managing Web sites for profitability: Balancing content and advertising. In: *Proceedings from the 35th Annual Hawaii International Conference on System Sciences, HICSS*, pp. 2340–2347 (2002)
- Edelman, B.: Hard-coding bias in Google "algorithmic" search results (2010), <http://www.benedelman.org/hardcoding/> (retrieved)
- Edelman, B., Lockwood, B.: Measuring bias in an "organic" Web search (2011), <http://www.benedelman.org/searchbias/> (retrieved)
- Fallows, D.: Search engine users: Internet searchers are confident, satisfied and trusting—but they are also unaware and naive. In: *Pew Internet & American Life Project*, pp. 1–36. Pew Internet & American Life Project, Washington, DC (2005)
- Frants, V.I., Shapiro, J., Voiskunskii, V.G.: *Automated information retrieval: Theory and methods*. Academic Press, San Diego (1997)
- Granka, L.: The politics of search: A decade retrospective. *The Information Society* 26(5), 364–374 (2010)

- Grimmelmann, J.: Some skepticism about search neutrality. In: Szoka, B., Marcus, A. (eds.) *The Next Digital Decade: Essays on the Future of the Internet*, pp. 435–460. TechFreedom, Washington, DC (2010)
- Gulli, A.: The indexable Web is more than 11.5 billion pages. In: 14th International Conference on World Wide Web, pp. 902–903. ACM, New York (2005)
- Hotchkiss, G.: Eye tracking report: Google, MSN, and Yahoo! Compared. Enquiro, Kelowna, British Columbia (2006)
- Huffman, S.B., Hochster, M.: How well does result relevance predict session satisfaction? In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 567–574. ACM, New York (2007)
- Höchstötter, N., Lewandowski, D.: What users see – Structures in search engine results pages. *Information Sciences* 179(12), 1796–1812 (2009)
- Jansen, B.J., Booth, D.L., Spink, A.: Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management* 44(3), 1251–1266 (2008)
- Jansen, B.J.: The comparative effectiveness of sponsored and nonsponsored links for Web e-commerce queries. *ACM Transactions on the Web* 1(1), 1–25 (2007)
- Jansen, B.J., Spink, A.: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management* 42(1), 248–263 (2006)
- Jansen, J.: *Understanding sponsored search: Core elements of keyword advertising*. Cambridge University Press, Cambridge (2011)
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting click-through data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154–161. ACM, New York (2005)
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., Gay, G.: Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems* 25(2), article 7 (2007)
- Jung, S., Herlocker, J.L., Webster, J.: Click data as implicit relevance feedback in web search. *Information Processing & Management* 43(3), 791–807 (2007)
- Kammerer, Y., Gerjets, P.: How search engine users evaluate and select Web search results: The impact of the search engine interface on credibility assessments. In: Lewandowski, D. (ed.) *Web Search Engine Research*, pp. 251–279. Emerald, Bingley (2012)
- Kang, I.H., Kim, G.C.: Query type classification for Web document retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71. ACM, New York (2003)
- Keane, M.T., O’Brien, M., Smyth, B.: Are people biased in their use of search engines? *Communications of the ACM* 51(2), 49–52 (2008)
- Lewandowski, D.: Date-restricted queries in Web search engines. *Online Information Review* 28(6), 420–427 (2004)
- Lewandowski, D.: A three-year study on the freshness of Web search engine databases. *Journal of Information Science* 34, 817–831 (2008a)
- Lewandowski, D.: The retrieval effectiveness of Web search engines: Considering results descriptions. *Journal of Documentation* 64(6), 915–937 (2008b)

- Lewandowski, D.: Problems with the use of Web search engines to find results in foreign languages. *Online Information Review* 32(5), 668–672 (2008c)
- Lewandowski, D.: Search engine user behaviour: How can users be guided to quality content? *Information Services & Use* 28, 261–268 (2008d)
- Lewandowski, D.: The retrieval effectiveness of search engines on navigational queries. *ASLIB Proceedings* 61(4), 354–363 (2011)
- Lewandowski, D.: A framework for evaluating the retrieval effectiveness of search engines. In: Jouis, C., Biskri, I., Ganascia, G., Roux, M. (eds.) *Next Generation Search Engines: Advanced Models for Information Retrieval*, pp. 456–479. IGI Global, Hershey (2012)
- Lewandowski, D., Höchstötter, N.: Web searching: A quality measurement perspective. In: Spink, A., Zimmer, M. (eds.) *Web Search: Multidisciplinary Perspectives*, pp. 309–340. Springer, Berlin (2008)
- Lewandowski, D., Höchstötter, N.: Standards der Ergebnispräsentation. In: Lewandowski, D. (ed.) *Handbuch Internet-Suchmaschinen*, pp. 204–219. Akademische Verlagsgesellschaft Aka, Heidelberg (2009)
- Lewandowski, D., Wahlig, H., Meyer-Bautor, G.: The freshness of Web search engine databases. *Journal of Information Science* 32(2), 131–148 (2006)
- Li, Z., Shi, S.: Improving relevance judgment of Web search results with image excerpts. In: *17th International Conference on World Wide Web*, vol. 17, pp. 21–30 (2008)
- Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., et al.: Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology* 59(7), 1041–1052 (2008)
- MacFarlane, A.: Evaluation of web search for the information practitioner. *Aslib Proceedings: New Information Perspectives* 59(4-5), 352–366 (2007)
- Machill, M., Neuberger, C., Schweiger, W., Wirth, W.: Navigating the Internet: A study of German-language search engines. *European Journal of Communication* 19(3), 321–347 (2004)
- Mansourian, Y.: Web search efficacy: definition and implementation. *Aslib Proceedings* 60(4), 349–363 (2008)
- Nicholson, S., Sierra, T., Eseryel, U.Y., Park, J.-H., Barkow, P., Pozo, E.J., Ward, J.: How much of it is real? Analysis of paid placement in Web search engine results. *Journal of the American Society for Information Science and Technology* 57(4), 448–461 (2006)
- Nielsen, J., Tahir, M. (n.d.): *Homepage Usability: 50 websites deconstructed*. New Riders Publishing, Indianapolis
- Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: Google we trust: Users’ decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* 12, 801–823 (2007)
- Pirolli, P.: *Information foraging theory: Adaptive interaction with information*. Oxford University Press, London (2009)
- Rose, D.E., Levinson, D.: Understanding user goals in Web search. In: *Proceedings of the 13th International Conference on World Wide Web*, pp. 13–19 (2004)
- Sullivan, D.: Searching with invisible tabs. *Search Engine Watch* (2003), <http://searchenginewatch.com/showPage.html?page=3115131> (retrieved)
- Sullivan, D.: Study: Google “favors” itself only 19% of the time. *Search Engine Land* (2011), <http://searchengineland.com/survey-google-favors-itself-only-19-of-the-time-61675> (retrieved)

- Thurow, S.: Search engine visibility. New Riders, Berkeley (2007)
- Thurow, S., Musica, N.: When search meets Web usability. New Riders, Berkeley (2009)
- Vaughan, L., Thelwall, M.: Search engine coverage bias: Evidence and possible causes. *Information Processing & Management* 40(4), 693–707 (2004)
- Vaughan, L., Zhang, Y.: Equal representation by search engines? A comparison of websites across countries and domains. *Journal of Computer-Mediated Communication* 12(3), 888–909 (2007)
- W3 Schools Browser Statistics (2011),
http://www.w3schools.com/browsers/browsers_stats.asp
(retrieved)
- Xie, M., Wang, H., Goh, T.N.: Quality dimensions of Internet search engines. *Journal of Information Science* 24(5), 365–372 (1998)

Author Index

- Amarala, Swathi 121
- Ballatore, Andrea 93
- Bertolotto, Michela 93
- Blank, Daniel 5
- Bordogna, Gloria 1
- Bronselaer, Antoon 55
- Cabitza, Federico 159
- De Tré, Guy 55
- Geier, Michael 145
- Griffith, Josephine 35
- Henrich, Andreas 5
- Holzinger, Andreas 145
- Jain, Lakhmi 1
- Lewandowski, Dirk 179
- Matthé, Tom 55
- O’Riordan, Colm 35
- Pasi, Gabriella 1
- Ribeiro, Cristina 121
- Salmon, Ricardo 121
- Simonic, Klaus-Martin 145
- Sorensen, Humphrey 35
- Van Britsom, Daan 55
- Wilson, David C. 93
- Yildirim, Pinar 145