Chapman & Hall/CRC
Social Media and Social Computing Series

MINING USER GENERATED CONTENT

Edited by
Marie-Francine Moens
Juanzi Li
Tat-Seng Chua

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

MINING USER GENERATED CONTENT

Chapman & Hall/CRC
Social Media and Social Computing Series

Series Editor

Irwin King

Published titles

Mining User Generated Content
Marie-Francine Moens, Juanzi Li, and Tat-Seng Chua

Chapman & Hall/CRC
Social Media and Social Computing Series

MINING USER GENERATED CONTENT

Edited by

Marie-Francine Moens
Katholieke Universiteit Leuven
Belgium

Juanzi Li
Tsinghua University
China

Tat-Seng Chua
National University of Singapore
Singapore

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20131114

International Standard Book Number-13: 978-1-4665-5741-3 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Foreword

I am delighted to introduce the first book on multimedia data mining. When I came to know about this book project undertaken by three of the most active researchers in the field, I was pleased that it was coming in the early stages of a field that will need it more than most fields do. In most emerging research fields, a book can play a significant role in bringing some maturity to the field. Research fields advance through research papers. In research papers, however, only a limited perspective can be provided about the field, its application potential, and the techniques required and already developed in the field. A book gives such a chance. I liked the idea that there would be a book that would try to unify the field by bringing in disparate topics already available in several papers, which are not easy to find and understand. I was supportive of this book project even before I had seen any material on it. The project was a brilliant and a bold idea by two active researchers. Now that I have it on my screen, it appears to be even a better idea.

Multimedia started gaining recognition as a field in the 1990s. Processing, storage, communication, and capture and display technologies had advanced enough that researchers and technologists started building approaches to combine information in multiple types of signals such as audio, images, video, and text. Multimedia computing and communication techniques recognize correlated information in multiple sources as well as an insufficiency of information in any individual source. By properly selecting sources to provide complementary information, such systems aspire, much like the human perception system, to create a holistic picture of a situation using only partial information from separate sources.

Data mining is a direct outgrowth of progress in data storage and processing speeds. When it became possible to store a large volume of data and run different statistical computations to explore all possible and even unlikely correlations among data, the field of data mining was born. Data mining allowed people to hypothesize relationships among data entities and explore support. This field has been applied in many diverse domains and continues to experience even more applications. In fact, many new fields are a direct outgrowth of data mining and it is likely to become a powerful computational tool.

Irwin King

This page intentionally left blank

Preface

In recent years, we have witnessed the convergence of social networks, mobile computing, and cloud computing. These trends have encouraged users to carry out most of their social interactions online on social networks and on the move. Through these social networks, users routinely comment on issues, ask questions, provide answers, tweet or blog about their views, and conduct online purchases. Through their mobile devices, they perform spontaneous check-ins to their favorite venues, and readily share their photos and videos of local situations, and so on. The content accumulated has evolved into a huge unstructured source of timely knowledge on the cloud, which forms a rich part of users' social engagements and communication.

Statistics on Facebook[1] and social networking[2] indicate that over 11% of people worldwide now use Facebook (which amounts to 1.15 billion users) with 680 million mobile Facebook users, while 98% of 18 to 24 year olds in the United States are already social network users. Each day, Facebook users share 2.3 billion pieces of content and upload 250 million photos. Outside Facebook, social network users post 190 million tweets on Twitter, and view over 3.1 billion videos on YouTube. In terms of e-commerce, the percentage of retail sales that are made online in United States is 8%, and the number of online users who have made an Internet purchase is 83%.[3] The statistics are even more tilted toward social networking and mobile computing in China.[4] These overwhelming statistics clearly demonstrate the pervasiveness and influence of social media today.

The social media shared by users, along with the associated metadata, are collectively known as user generated content (or UGC). UGC comes from a myriad of sources, including the social networking sites like Facebook and LinkedIn; live microblog sites like Twitter; mobile sharing sites like 4Square and Instagram; information sharing sites like forums and blogs; image and video sharing sites like Flickr and YouTube; and the various community question-answering sites like Wiki-Answers and Yahoo! Answers; as well as their counterparts in China. The content comes in a variety of languages

[1] http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/

[2] Statistic Brain on Social Networking Statistics dated November 2012; http://www.statisticbrain.com/social-networking-statistics/

[3] Statistic Brain on E-Commerce/Online Sale Statistics dated August 2012; http://www.statisticbrain.com/total-online-sales/

[4] http://www.go-globe.com/blog/social-media-china/

such as English and Chinese, and modalities such as text, image, video, and location-based information, and the corresponding metadata. In addition to the multisource, multimodal, and multilingual content, another important element of social media is the users and their relationships. It is noted that the most useful UGC comes mainly from the publicly available data sources that reflect the social interactions of people.

To analyze and fuse these UGCs, we need techniques to deal with the huge amount of real-time multimedia and multilingual data. In addition, we need to tackle the social aspects of these contents, such as user relations and influential users, and so on, with respect to any topics. This offers new challenges that have attracted a lot of active research. Various higher order analytics can be mined and extracted, including structures of UGC with respect to any given topic, live emerging and evolving events/topics; relationships between key users and topics, user communities, and the various events/activities with respect to location, people, and organizations. Key research areas of UGC include: (a) reliable strategies for harvesting representative UGC with respect to any topic; (b) indexing and retrieval of huge media resources arising from these media; (c) organization of unstructured UGC and users on any topic into structured knowledge and user communities; (d) fusion of UGC to generate analytics related to location, people, topic, and organization; and (e) basic research on the analysis and retrieval of text, live discussion streams, images, and videos.

Many large research groups now collect, index, and analyze UGC, with the aim of uncovering social trends and user habits. One example of such an effort is the NExT Research Center jointly hosted at the National University of Singapore and Tsinghua University [141], which focuses on harvesting and mining the huge amount of UGC in real-time and across cultural boundaries. A global effort centering around the idea of the Web Observatory by the Web Science Trust is also taking shape, Web Science Trust: The Web Observatory.[5] The Web Observatory aims to coordinate the common use of social UGC data collected and analytics developed by the various social observatory systems from around the world. Central to the establishment of a Web observatory is the selection of a profile of standards, which each Web observatory node must adopt to facilitate data sharing. This effort is expected to benefit many users and researchers of social media. Given the active range of research and activities on UGC, it is timely to initiate a book that focuses on the mining of UGC and its applications.

This book represents the first concerted effort to compile the state-of-the-art research and future direction on UGC research. The book is divided into four parts. The first part presents the introduction to this new and exciting topic. Part II introduces the mining of UGC of different medium types. Topics discussed include the social annotation of UGC, social network graph construction and community mining, mining of UGC to assist in music

[5] http://Webscience.org/Web-observatory/

retrieval, and the popular but difficult topic of UGC sentiment analysis. Part III then discusses the mining and searching of various types of UGC, including knowledge extraction, search techniques for UGC, and a specific study on the analysis and annotation of Japanese blogs. Finally, Part IV presents the applications, in which the use of UGC to support question-answering, information summarization, and recommendation is discussed.

The book should be of interest to students, researchers, and practitioners of this emerging topic.

Marie-Francine Moens, Juanzi Li, and Tat-Seng Chua

This page intentionally left blank

Editors

Marie-Francine Moens: Marie-Francine (Sien) Moens is a professor at the Department of Computer Science of the Katholieke Universiteit Leuven, Belgium. She holds an M.Sc. and a Ph.D. degree in computer science from this university. She is head of the Language Intelligence and Information Retrieval (LIIR) research group (http://www.cs.kuleuven.be/groups/∼liir/), and is a member of the Human Computer Interaction unit. Her main interests are in the domain of automated content retrieval and extraction from text using a combination of statistical, machine learning, and symbolic techniques, and exploiting insights from linguistic and cognitive theories. Dr. Moens is author of more than 240 international publications among which are two monographs published by Springer. She is coeditor of 15 books or proceedings, coauthor of 40 international journal articles, and 29 book chapters. She is involved in the organization or program committee (as PC chair, area chair, or reviewer) of major conferences on computational linguistics, information retrieval, and machine learning (ACL, COLING, EACL, SIGIR, ECIR, CORIA, CIKM, ECML-PKDD). She teaches courses on text-based information retrieval and natural language processing at KU Leuven. She has given several invited tutorials in summer schools and international conferences (e.g., tutorial "Linking Content in Unstructured Sources" at the 19th International World Wide Web Conference, WWW 2010), and keynotes at international conferences on the topic of information extraction from text. She participates or has participated as a partner or coordinator of numerous European and international projects, which focus on text mining or the development of language technology. In 2011 and 2012, Dr. Moens was appointed as chair of the European Chapter of the Association for Computational Linguistics (EACL) and was a member of the executive board of the Association for Computational Linguistics (EACL). She is a member of the Research Council of the Katholieke Universiteit Leuven. E-mail: marie-francine.moens@cs.kuleuven.be

Juanzi Li: Juanzi Li is a full professor at Tsinghua University. She obtained her Ph.D. degree from Tsinghua University in 2000. She is the principal of the Knowledge Engineering Group at Tsinghua University. Her main research interest is to study semantic technologies by combining natural language processing, semantic Web, and data mining. She is the Vice Director of the Chinese Information Processing Society of the

Chinese Computer Federation in China. She is principal investigator of the key project cloud computing based on large-scale data mining supported by the Natural Science Foundation of China (NSFC), she is also the PI of many national basic science research programs and international cooperation projects. Dr. Li took the important role in defining Chinese News Markup Language (CNML), and developed the CNML specification management system which won the “Wang Xuan” News Science and Technology Award in 2009 and 2011. She has published about 90 papers in many international journals and conferences such as WWW, TKDE, IJCAI, SIGIR, SIGMOD, and SIGKDD. E-mail: lijuanzi2008@gmail.com

Tat-Seng Chua: Tat-Seng Chua is the KITHCT chair professor at the School of Computing, National University of Singapore. He was the acting and founding dean of the school from 1998–2000. Dr. Chua’s main research interest is in multimedia information retrieval, question-answering (QA), and live social media analysis. He is the director of a multimillion-dollar joint center between NUS and Tsinghua University in China to develop technologies for live social media searches. The project will gather, mine, search, and organize user generated content within the cities of Beijing and Singapore. Dr. Chua is active in the international research community. He has organized and served as the program committee member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He was the conference cochair of ACM Multimedia 2005, ACM CIVR 2005, and ACM SIGIR 2008. He serves on the editorial boards of: ACM Transactions of Information Systems (ACM), Foundation and Trends in Information Retrieval (NOW), the Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He is a member of the steering committee of ICMR (International Conference on Multimedia Retrieval) and Multimedia Modeling conference series; and is member of the International Review Panel of two large-scale research projects in Europe. E-mail: chuats@comp.nus.edu.sg

Contributors

Kenji Araki: Kenji Araki received BE, ME, and Ph.D. degrees in electronics engineering from Hokkaido University, Sapporo, Japan, in 1982, 1985, and 1988, respectively. In April 1988, he joined Hokkai-Gakuen University, Sapporo, Japan, where he was a professor. He joined Hokkaido University in 1998 as an associate professor in the Division of Electronics and Information Engineering and became a professor in 2002. Presently, he is a professor in the Division of Media and Network Technologies at Hokkaido University. Dr. Araki's research interests include natural language processing, spoken dialogue processing, machine translation, and language acquisition. He is a member of the AAAI, IEEE, JSAI, IPSJ, IEICE, and JCSS. E-mail: araki@media.eng.hokudai.ac.jp

Manuel Eduardo Ares Brea: M. Eduardo Ares Brea is a Ph.D. candidate and a research and teaching assistant at the IRLab of the University of A Coruña. His main areas of research are semisupervised learning, Web mining, and applications of natural language processing. E-mail: maresb@udc.es

Shenghua Bao: Shenghua Bao is a research staff member at IBM Research–China. He obtained a Ph.D. degree in computer science at Shanghai Jiao Tong University in 2008. His research interests lie primarily in Web search, data mining, machine learning, and related applications. He received an IBM Ph.D. Fellowship in 2007 and was named IBM Master Inventor in 2012. Currently, Dr. Bao serves as an editor of CCF Technews, is a PC member of conferences like WWW, EMNLP, and WSDM, and a reviewer of several journals, including, IEEE TKDE, ACM TALIP, and IPM. E-mail: baoshhua@us.ibm.com

Roi Blanco: Roi Blanco is a senior research scientist at Yahoo! Labs Barcelona, where he has been working since 2009. Dr. Blanco is interested in applications of natural language processing for information retrieval, Web search and mining, and large-scale information access in general, and publishes at international conferences in those areas. He also contributes to Yahoo! products such as Yahoo! Search. Previously, he taught computer science at A Coruña University, where he received his Ph.D. degree (cum laude) in 2008. E-mail: roi@yahoo-inc.com

Kalina Bontcheva: Kalina Bontcheva is a senior research scientist and the holder of an EPSRC career acceleration fellowship, working on text summarization of social media. Dr. Bontcheva received her Ph.D. on the topic of adaptive hypertext generation from the University of Sheffield in 2001. Her main interests are information extraction, opinion mining, natural language generation, text summarization, and software infrastructure for NLP. She has been a leading developer of GATE since 1999. Dr. Bontcheva is also leading the Sheffield NLP research teams in the TrendMiner (http://www.trendminer-project.eu/) and uComp (http://www.ucomp.eu/) research projects, working respectively on mining and summarization of social media streams and crowdsourcing of NLP resources. E-mail: k.bontcheva@dcs.shef.ac.uk

Jia Chen: Jia Chen received double bachelor degrees in mathematics and computer science at Shanghai JiaoTong University in 2008. He is now a Ph.D. candidate in the Department of Computer Science and Engineering in Shanghai JiaoTong University. His research interests are in image annotation, content-based image retrieval, and machine learning. E-mail: chenjia@apex.sjtu.edu.cn

Constantin Comendant: Constantin Comendant holds an M.Sc. degree in media informatics from RWTH Aachen and the Bonn-Aachen IT Center (B-IT) in Germany. In his master thesis, he treated the topic of link prediction in networks. He is currently a Ph.D. student in the group of Jan Ramon, where he works on models for random graphs. E-mail: constantin.comendant@cs.kuleuven.be

Pawel Dybala: Pawel Dybala was born in Ostrow Wielkopolski, Poland in 1981. He received his MA in Japanese studies from the Jagiellonian University in Krakow, Poland in 2006, and a Ph.D. in information science and technology from Hokkaido University, Japan in 2011. Dr. Dybala is a director and general project manager at Kotoken Language Laboratory in Krakow. Currently, he is a JSPS postdoctoral research fellow at the Otaru University of Commerce, Otaru, Japan. His research interests include natural language processing, humor processing, metaphor undestanding, HCI, and information retrieval. E-mail: paweldybala@res.otary-uc.ac.jp

Mostafa Haghir Chehreghani: Mostafa Haghir Chehreghani received his B.Sc. in computer engineering from Iran University of Science and Technology (IUST) in 2004, and his M.Sc. from the University of Tehran in 2007. In January 2010, he joined the Department of Computer Science as a Ph.D. student, Katholieke Universiteit Leuven. His research interests include data mining and network analysis. E-mail: mostafa.haghirchehreghani@cs.kuleuven.be

Xianpei Han: Xianpei Han is an associate professor in the IR Laboratory at the Institute of Software Chinese Academy of Sciences. Prior to joining the IR Laboratory, he received his Ph.D. degree in the NLPR, Institute of Automation, Chinese Academy of Sciences in 2010. His research interests include natural language processing and information extraction. E-mail: xianpei@nfs.iscas.ac.cn

Noam Koenigstein: Noam Koenigstein received a B.Sc. degree in computer science (cum laude) from the Technion–Israel Institute of Technology, Haifa, Israel, in 2007 and an M.Sc. degree in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 2009. Currently, he is working toward a Ph.D. degree in the School of Electrical Engineering, Tel-Aviv University. In 2011, he joined the Xbox Machine Learning research team of Microsoft, where he developed the algorithm for Xbox recommendations serving more than 50 million users worldwide. His research interests include machine learning, information retrieval, and large-scale data mining, with a specific focus on recommender systems. E-mail: noamk@eng.tau.ac.il

Christina Lioma: Christina Lioma is an assistant professor and Freja research fellow in the Department of Computer Science, University of Copenhagen, Denmark. Her research focuses on the computational processing of language, mainly in the areas of information retrieval and computational linguistics. She publishes internationally in these areas, often coauthoring with collaborators from a widespread network of researchers. She is broadly engaged in program committees and reviewing in main journals and conferences within the areas covered by her research interests. E-mail: liomca@gmail.com

Chin-Yew Lin: Chin-Yew Lin is a research manager of the Knowledge Mining group at Microsoft Research Asia. His research interests are knowledge mining, social computing, question-answering, and automatic summarization. Dr. Lin is developing technologies to automatically learn social interaction knowledge from large-scale real-world data and transform unstructured and semistructured Web data into structured data to enable semantic computing. He has developed automatic evaluation technologies for summarization and machine translation. In particular, he created the ROUGE automatic summarization evaluation package. ROUGE was the official automatic evaluation package for Document Understanding Conferences and has become the de facto standard in summarization evaluation. He is a member of the Editorial Board of Computational Linguistics (2013–2015) and an action editor of the Transactions of the Association for Computational Linguistics. He was the program cochair of ACL 2012 and program cochair of AAAI 2011 AI & the Web Special Track. E-mail: cyl@microsoft.com

Kang Liu: Kang Liu received his Ph.D. degree from NLPR, Institute of Automation, Chinese Academy of Sciences in 2010. Before that, he received his M.Sc. and B.Sc. degrees from Xidian University in 2002 and

2005, respectively. Currently, he is working as an assistant professor in NLPR, Institute of Automation, Chinese Academy of Sciences. His current research interests include natural language processing, information extraction, question-answering, opinion mining, and so on. He has authored/coauthored more than 20 papers in leading conferences, including ACL, IJCAI, EMNLP, and CIKM. E-mail: kliu@nlpr.ia.ac.cn

Yang Liu: Yang Liu is a fourth-year Ph.D. candidate in NLPR, Institute of Automation, Chinese Academy of Sciences. He received his bachelor's degree from Harbin Institute of Technology in 2009. His research interest is information extraction. E-mail: liuyang09@nlpr.ia.ac.cn

Claudio Lucchese: Claudio Lucchese (http://hpc.isti.cnr.it/∼claudio) received his master's degree in computer science (summa cum laude) from Ca' Foscari University of Venice in October 2003, and Ph.D. in computer science from the same university in 2007. Currently, he is a researcher at the Italian National Research Council (CNR). Dr. Lucchese research activity focuses on large-scale data mining techniques for information retrieval. He has published more than 40 papers on these topics in peer-reviewed international conferences and journals. He has participated in several EU-funded projects, and served as program committee member in numerous data mining and information retrieval conferences. E-mail: claudio.lucchese@isti.cnr.it

Jacek Maciejewski: Jacek Maciejewski studied for an M.Sc. degree in computer science at the University of Adam Mickiewicz, Poznan, Poland. He was awarded a scholarship to Hokkaido University, Japan, for the period 2008–2010. During his scholarship, he participated in research activities at the Graduate School of Information Science and Technology, Hokkaido University. His research interests include software engineering, natural language processing, Web mining, and information retrieval. E-mail: jacek.maciejewski@gmail.com

Yoshio Momouchi: Yoshio Momouchi was born in 1942 in Hokkaido, Japan. He obtained a master's degree and a doctorate in engineering from Hokkaido University. He was a member of the Division of Information Engineering in the Graduate School at Hokkaido University from 1973 to 1988. Since 1988, Dr. Momouchi has been a professor in the Faculty of Engineering at Hokkai-Gakuen University. He fulfilled duties as dean of the Graduate School of Engineering at Hokkai-Gakuen University during the years 2005–2008. He specializes in intelligent information processing, computational linguistics, and machine translation. He is a member of IPSJ, ANLP, ACL, MLSJ, JCSS, and JSAI. E-mail: mouchi@mo/eli.hokkai-su.ac.jp

Cristina Ioana Muntean: Cristina Ioana Muntean graduated in business information systems at Babes-Bolyai University, Cluj-Napoca, where she also received a Ph.D. in cybernetics and statistics in 2012. She is currently a research fellow at HPC Lab, ISTI-CNR, Pisa. Her interests are tourist recommender systems, machine learning, and information retrieval applied to Web and social network data. E-mail: cristina.muntean@econ.ubbcluj.ro

Raffaele Perego: Raffaele Perego (http://hpc.isti.cnr.it/∼raffaele) is a senior researcher at ISTI-CNR, where he leads the High Performance Computing Lab (http://hpc.isti.cnr.it/). His main research interests include data mining, Web information retrieval, query log mining, and parallel and distributed computing. He has coauthored more than 100 papers on these topics published in journals and in the proceedings of peer-reviewed international conferences. E-mail: raffaele.perego@isti.cnr.it

Michal Ptaszynski: Michal Ptaszynski was born in Wroclaw, Poland in 1981. He received an MA degree from the University of Adam Mickiewicz, Poznan, Poland, in 2006, and a Ph.D. in information science and technology from Hokkaido University, Japan in 2011. Currently, Dr. Ptaszynski is a JSPS postdoctoral research fellow at the High-Tech Research Center, Hokkai-Gakuen University, Japan. His research interests include natural language processing, dialogue processing, affect analysis, sentiment analysis, HCI, and information retrieval. He is a member of ACL, AAAI, IEEE, HUMAINE, AAR, SOFT, JSAI, and NLP. E-mail: ptaszynski@media.eng.hokudai.ac.jp

Zhenyou Qi: Zhenyou Qi is a last year Ph.D. candidate in NLPR, Institute of Automation, Chinese Academy of Sciences. He received his bachelor's degree from the University of Science and Technology Beijing in 2007. His research interest is information extraction. E-mail: zyqi2013@163.com

Jan Ramon: Jan Ramon obtained his Ph.D. in 2002 from the KU Leuven, Belgium, on clustering and instance-based learning in first-order logic. Currently, he is a senior researcher at KU Leuven. Dr. Ramon's current research interests include statistical and algorithmic aspects of graph mining and machine learning with structured data. He also has a strong interest in applications, among other things in medical domains and computational biology. E-mail: jan.ramon@cs.kuleuven.be

Dominic Rout: Dominic Rout is working toward a Ph.D. in summarization of social media at the University of Sheffield as part of an EPSRC-funded project on this topic. He has been a member of the Natural Language Processing Research Group since 2011, working as part of the GATE team in research and training. He is currently also involved part time in the EC-funded TrendMiner project (http://www.trendminer-project.eu/), where his summarization research is tested with users in the political and financial domains. His other research interests include user interest modeling, user and content geolocation, and content recommendation and ranking.

In addition to his research work, he is passionate about teaching and has been an assistant in a number of courses, as well as developing and teaching outreach classes. E-mail: d.rout@sheffield.ac.uk

Rafal Rzepka: Rafal Rzepka received an MA degree from the University of Adam Mickiewicz, Poznan, Poland, in 1999, and a Ph.D. from Hokkaido University, Japan, in 2004. Currently, he is an assistant professor in the Graduate School of Information Science and Technology at Hokkaido University. His research interests include natural language processing, Web mining, commonsense retrieval, dialogue processing, language acquisition, affect analysis, and sentiment analysis. He is a member of AAAI, ACL, JSAI, IPSJ, IEICE, JCSS, and NLP. E-mail: kabura@media.eng.hokudai.ac.jp

Markus Schedl: Markus Schedl graduated in computer science from the Vienna University of Technology. He earned his Ph.D. in computational perception from the Johannes Kepler University Linz, where he is employed as assistant professor in the Department of Computational Perception. He further studied international business administration at the Vienna University of Economics and Business Administration as well as at the Handelshgskolan of the University of Gothenburg, which led to a master's degree. He has coauthored 80 refereed conference papers and journal articles (among others, published in ACM Multimedia, SIGIR, ECIR, IEEE Visualization, Journal of Machine Learning Research, ACM Transactions on Information Systems, Springer Information Retrieval, IEEE Multimedia). Furthermore, he serves on various program committees and has reviewed submissions to several conferences and journals (among others, ACM Multimedia, ECIR, IJCAI, ICASSP, IEEE Visualization, IEEE Transactions of Multimedia, Elsevier Data & Knowledge Engineering, ACM Transactions on Intelligent Systems and Technology, Springer Multimedia Systems). His main research interests include Web and social media mining, information retrieval, multimedia, music information research, and user interfaces. Since 2007, he has been giving several lectures, for instance, "Music Information Retrieval," "Exploratory Data Analysis," "Multimedia Search and Retrieval," and "Learning from User Generated Data." Dr. Schedl further spent several guest lecturing stays at the Universitat Pompeu Fabra, Barcelona, Spain; the Utrecht University, the Netherlands; the Queen Mary, University of London, UK; and Kungliga Tekniska Hgskolan, Stockholm, Sweden. E-mail: markus.schedl@jku.at

Fabrizio Silvestri: Fabrizio Silvestri (http://pomino.isti.cnr.it/~silvestr) is currently a researcher at ISTI-CNR in Pisa. He received his Ph.D. from the Computer Science Department of the University of Pisa in 2004. His research interests are mainly focused on Web information retrieval with a particular focus on efficiency-related problems such as caching, collection partitioning, and distributed IR, in general. In his professional activities he is a member of the program committee of many of the most important conferences in IR. Dr. Silvestri is author of more than 60 publications in highly relevant venues spanning from distributed and parallel computing to IR and data mining related conferences. E-mail: fabrizio.silvestri@isti.cnr.it

Mohamed Sordo: Mohamed Sordo is a postdoctoral researcher at the Music Technology Group of the Universitat Pompeu Fabra in Barcelona, Spain. He obtained his Ph.D. at the Music Technology Group in 2012, with a thesis entitled "Semantic Annotation of Music Collections: A Computational Approach," mainly devoted to the topic of music automatic tagging. Dr. Sordo's research areas involve music text/Web mining, music information retrieval, and machine learning. He has participated in a number of European-funded projects, including Variazioni, Pharos, and CompMusic. He is currently involved in the latter, developing systems to extract semantically and musically meaningful information from Web data. E-mail: mohamed.sordo@upf.edu

Zhong Su: Zhong Su is a senior technical staff member at IBM Research China (CRL) and senior manager of the Information Analytics Department. He joined CRL after receiving his Ph.D. degree in computer science at Tsinghua University in 2002. Dr. Su has been involved in many projects at CRL including text analytics, NLP, rich media analysis, and information integration. He has led a number of research projects which were awarded the Technical Accomplishment of IBM research many times and Outstanding Technical Accomplishment of IBM research in 2008 and 2010. Dr. Su was awarded the IBM Master Inventor in 2007 and currently chairs the Invention and Disclosure Board in CRL. He has published more than 50 papers in top international conferences/journals, with more than 40 patents or patents pending. Dr. Su is guest professor of APEX Lab, Shanghai JiaoTong University. He is also Vice Chairman of Technical Expert Council IBM Greater China Group. E-mail: suzhongatcn.ibm.com

Hossein Vahabi: Hossein Vahabi received his bachelor's and master's degrees in computer engineering from the University of Modena and Reggio Emilia, respectively, in 2006 (summa cum laude) and 2008 (summa cum laude). In 2012, he received a Ph.D. with European honors in computer science and engineering at the IMT Institute for Advanced Studies Lucca, Italy. Currently, Dr. Vahabi is leading his own private company H.V. His research interests cover many topics in the field of recommender systems: query recommendation, tourism point of interests recommendation, tag recommendation, tweet recommendation, and large-scale recommendation. E-mail: hossein.vahabi@imtlucca.it

Rossano Venturini: Rossano Venturini is currently a researcher at the Computer Science Department, University of Pisa. He received a Ph.D. from the Computer Science Department of the University of Pisa in 2010, with his thesis titled "On Searching and Extracting Strings from Compressed Textual Data." Dr. Venturini's research interests are mainly focused on the design and analysis of algorithms and data structures with special attention to problems of indexing and searching large textual collections. E-mail: rossano.venturini@isti.cnr.it

Haofen Wang: Haofen Wang has rich research experience in data management, Web search, and semantic Web fields. He has published more than 40 high-quality papers in the related international top conferences and journals including ISWC, WWW, SIGIR, SIGMOD, ICDE, CIKM, and Journal of Web Semantics. He has also won the best paper award of the 3rd China Semantic Web Symposium (CSWS 2009). He has been serving as reviewer for VLDB Journal, Journal of Web Semantics, and IEEE Transaction of Knowledge and Data Engineering. Dr. Wang has also served on program committees for international conferences such as ESWC 2008, ISWC 2009, EDBT 2009, ESWC 2009, ESWC 2010, ISWC 2010, EDBT 2010, ESWC 2011, ISWC 2011, ESWC 2012, ISWC 2012, WWW 2011, and WWW 2012. As a cochair, he has organized SemSearch 2009, SemSearch 2010, and SemSearch 2011, collaborated with WWW 2009, WWW 2010, and WWW 2011, respectively. As one of the deputy local chairs, he hosted the 9th International Semantic Web Conference (ISWC 2010). As a project leader, he took charge of several innovative research projects like scalable semantic search, semantic enterprise portal, and large-scale semantic data query answering using cloud computing. His education background includes a two-year IBM Ph.D. fellowship, fellowships at Hong Kong Science and Technology University and the University of Karlsruhe, Germany, and double bachelor degrees in computer science and math at Shanghai Jiao Tong University. E-mail: whfcarter@apex.sjtu.edu.cn

Yuyi Wang: Yuyi Wang received an M.Sc. in computer science in the School of Computing of the National University of Singapore, Singapore. His master's thesis concerns the nonlinearity of Boolean functions. He is currently a Ph.D. student in the MiGraNT project researching graph mining from a linear algebra point of view. E-mail: yuyi.wang@cs.kuleuven.be

Udi Weinsberg: Udi Weinsberg is a researcher and associate fellow at Technicolor Research in Palo Alto, California. He studies privacy and security, focusing on enabling practical privacy-preserving machine learning algorithms in recommender systems. He received his Ph.D. from Tel-Aviv University, Israel, School of Electrical Engineering in 2011, from where he also received his M.Sc. in 2007. During his Ph.D., Dr. Weinsberg was a member of the NetDIMES group, studying the structure and evolution of the Internet. E-mail: udi.weinsberg@technicolor.com

Ning Yu: Ning Yu is an assistant professor of the School of Library and Information Science at University of Kentucky. She received her Ph.D. in information science and a Ph.D. minor in cognitive science with an emphasis on computational linguistics at Indiana University. Dr. Yu's research interests lie broadly in information retrieval and text mining, with a focus on opinion mining (also known as sentiment analysis). She has collaborated on several information retrieval projects that produced top runs in various retrieval tasks at the Text Information Retrieval Conference (TREC), including high-accuracy document retrieval and opinion retrieval. She also has rich experience in adopting a hybrid approach to leverage human expertise and machine learning techniques for sentiment analysis on various data domains: news articles, blogs, reviews, and suicide notes. Her recent studies on semisupervised sentiment analysis have proven to be promising in solving two fundamental problems for sentiment analysis: insufficient training data and domain adaption. E-mail: nyu.yuning@gmail.com

Yong Yu: Yong Yu received his master's degree at the Computer Science Department of East China Normal University. He began to work in Shanghai Jiao Tong University in 1986. He is now the Ph.D. candidate tutor and the chairman of the E-Generation Technology Research Center (SJTU-IBM-HKU). He was the teacher of the course computer graphics and human machine interface and the course next generation Web infrastructure. As the head coach of SJTU ACM-ICPC team, he and his team have won the 2002, 2005, and 2010 ACM ICPC Championships. His research interests include semantic Web, Web mining, information retrieval, and computer vision. Email: yyu@apex.sjtu.edu.cn

Jun Zhao: Jun Zhao is a professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in computer science from Tsinghua University in 1998. Prior to joining NLPR, he was a postdoctoral in Hong Kong University of Sciences and Technology. Dr. Zhao's research interests include natural language processing, information extraction, and question-answering. He has served on over 10 program committees of the major international conferences in the field of natural language processing and knowledge engineering, and also has served as an associate editor for ACM Transactions on Asian Language Information Processing (TALIP). He has won the second prize of KDD-CUP 2011. In recent years, he has published more than 20 papers in top conferences including ACL, SIGIR, IJCAI, CIKM, EMNLP, and COLING. E-mail: junzhao1001@gmail.com

Guangyou Zhou: Guangyou Zhou is an assistant professor in the NLPR, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in computer science from NLPR in 2013. Before that, Dr. Zhou received his bachelor's degree from Northeast Normal University in 2008. His research interests include natural language processing and information retrieval. E-mail: gyzho@nlpr.ia.ac.cn

This page intentionally left blank

List of Reviewers

We are very grateful for the assistance of the following reviewers. Their valuable remarks were very helpful to improving the quality of the chapters.

Hadi Amari: Institute for Infocomm Research, Singapore

Steven Bethard: University of Alabama, United States

Tat-Seng Chua: National University of Singapore, Singapore

Xue Geng: National University of Singapore, Singapore

Karl Gyllstrom: ANATAS, Australia

Lei Hou: Tsinghua University, China

Juanzi Li: Tsinghua University, China

Zhixing Li: Tsinghua University, China

Banyong Liang: Microsoft China

Marie-Francine Moens: KU Leuven, Belgium

Jialie Shen: Singapore Management University, Singapore

Xuemeng Song: National University of Singapore, Singapore

Ivan Vulić: KU Leuven, Belgium

Zhigang Wang: Tsinghua University, China

Jianxing Yu: Institute for Infocomm Research, Singapore

This page intentionally left blank

List of Figures

This page intentionally left blank

List of Tables

This page intentionally left blank

Part I

Introduction

This page intentionally left blank

Chapter 1

Mining User Generated Content and Its Applications

Marie-Francine Moens

KU Leuven, Belgium

Juanzi Li

Tsinghua University, China

Tat-Seng Chua

National University of Singapore, Singapore

1.1 The Web and Web Trends

1.1.1 The Emergence of the World Wide Web (WWW): From Connected Computers to Linked Documents

Joseph Carl Robnett Licklider formulated the earliest ideas of a global computer network in August 1962,[1] known as the *Galactic Network*. He explained it as a set of *computers* that would be globally inter *connected* so

[1] http://en.wikipedia.org/wiki/J._C._R._Licklider

people could access data or programs when they wanted, which contained almost everything that the Internet is today.

Twenty-seven years later, the World Wide Web was proposed by Tim Berners-Lee. As defined in Wikipedia now,[2] it is a system of inter*linked*, hypertext *documents* that runs over the Internet. With a Web browser, a user views Web pages that may contain text, images, and other multimedia and navigates between them using hyperlinks. The proposal was meant for a more effective *European Organization for Nuclear Research* (CERN) communication system but Berners-Lee eventually realized that the concept could be implemented throughout the world. Figure 1.1 shows the architecture of the WWW that was drawn by him in the proposal.[3]

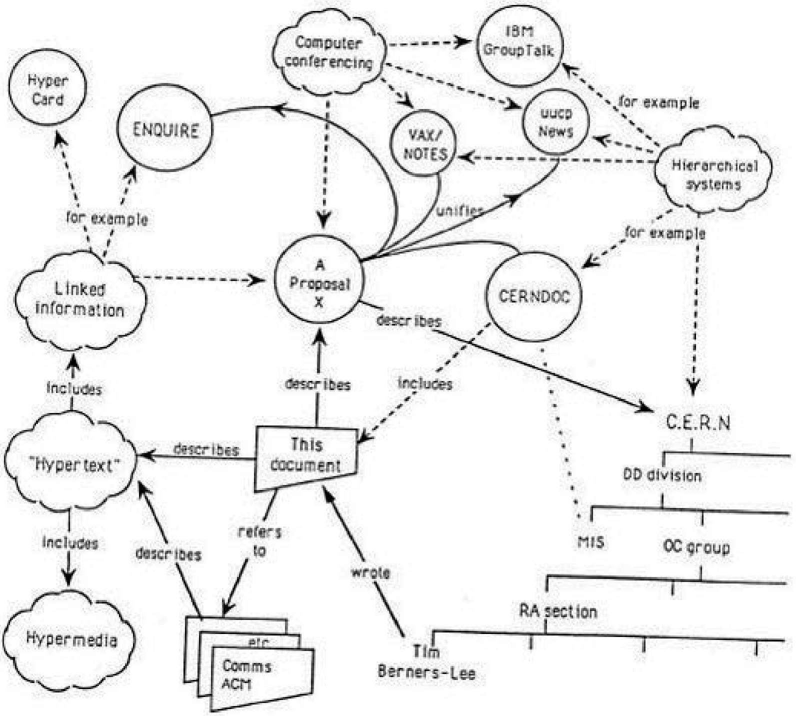

FIGURE 1.1: The architecture of the WWW in Berners-Lee's proposal.

Berners-Lee wrote the code for the WWW after that, and some essential technologies were listed as follows:

- **Hypertext and Hyperlink:** It is the key difference between documents

[2] http://en.wikipedia.org/wiki/World_Wide_Web

[3] http://www.w3.org/History/1989/proposal.html