A Report on The Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011)

Irwin King
The Chinese University of Hong Kong
Dept. of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, NT, Hong Kong
king@cse.cuhk.edu.hk

Hang Li
Microsoft Research Asia
Beijing Sigma Building
Beijing 100080, China
hangli@microsoft.com

1. INTRODUCTION

The Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011) was held in Hong Kong, during February 9-12, 2011.

WSDM (pronounced "wisdom") was held first at Stanford University in 2008 and subsequently in Barcelona, Spain and New York, USA. This is the first time the conference series is being located in the Asia Pacific, reflecting the growth of web search and data mining research interest in this region. Although it has only been a few years since its inception, WSDM has rapidly grown to become the premier international conference on Web search and Web data mining. Different from WWW, WSDM focuses on search and data mining. It also differs from SIGIR and SIGKDD with its emphasis on the Web. WSDM is sponsored by ACM SIGIR, SIGKDD, SIGMOD, and SIGWEB.

WSDM emphasizes on participation and communication for both academia and industry. It invites fundamental and theoretical research on Web search and Web data mining, as well as practical applications of this area. It is well known that Web search and Web data mining has developed to become one of the most popular research areas in computer science and engineering. It is changing people's work and life constantly and rapidly, and has a profound and tremendous influence on the whole society. For this reason, WSDM will exert greater influence and receive more and more attention around the world.

The founders and previous WSDM conference organizers are all well-known scholars related to search and data mining on the Web. WSDM publishes original, high quality papers which represent the most cutting-edge research results. The quality of WSDM papers is directly comparable to those from WWW, SIGIR, and SIGKDD conferences. This is one of the reasons why WSDM has grown rapidly to become an international conference with important academic, industrial and social values.

2. CONFERENCE OVERVIEW

The General Chair of WSDM 2011 is Irwin King (The Chinese University of Hong Kong, Hong Kong), and the Program Co-Chairs are Wolfgang Nejdl (L3S and University of Hannover, Germany) and Hang Li (Microsoft Research Asia, China). There are 19 Senior Program Committee (SPC) members and 131 Program Committee (PC) members.

WSDM 2011 has received a record of 372 submissions and accepted a total of 83 papers (22.3% acceptance rate) full-length papers. Due to its concentrated research areas and a high number of participants, WSDM 2011 is a single track conference. Thirty-two papers (8.6% acceptance rate) were selected for oral presentation, and all accepted papers were presented in poster sessions over a two-day period. This also marks the first time the WSDM conference added the poster session to increase interactions among delegates. The accepted papers are from 13 countries and areas, including USA, China, UK, Japan, Italy, Israel, Hong Kong, etc. Noticeably, the mainland China contributed up to 11 papers. All accepted papers are from 42 institutions, including Microsoft Research, Yahoo! Labs Research, Stanford University, The Chinese University of Hong Kong, etc. These papers are equally divided into three areas: Web search, Web data mining, and Social Web search and mining. They cover a variety of aspects related to search and data mining on the Web and the Social Web, such as theories, models, algorithms, evaluations, experiments and applications.

WSDM 2011 lasted a total of four days. On the first day, four tutorials and two workshops were held, then it was followed by the 3-day main conference. There are approximately 240 participants, half from the academia community and half from the industry, e.g., from top Internet companies such as Microsoft, Yahoo!, eBay, Google, Yandex, Baidu, Facebook, etc. About 40% of delegates were from the United States and Asia, respectively and the remaining 20% were from Europe and other regions. Moreover, out of the 120 academics there are over 40 scholars and students from mainland China, representing a major increase from previous years. In particular, several professors from Tsinghua, Peking, Renmin University, Chinese Academy of Science were invited to take part in this conference.

Christos Faloutsos (Professor, Carnegie Mellon University) and Harry Shum (Global Vice President, Microsoft) gave the keynotes at the conference. In the keynote entitled "Mining Billion-node Graphs: Patterns, Generators and Tools", Professor Faloutsos introduced their recent research results on rule and law mining in billion-node graphs (e.g., the Web), including a comprehensive list of static and temporal laws, some generators which naturally match all of the known properties of real graphs, and tools oddBall and PEGASUS for discovering anomalies and patterns in large-scale graphs. Dr. Harry Shum gave a talk named "Bing Dialog Model: Intent, Knowledge and User Interaction". He

outlined Microsoft's future decision engine called the Bing Dialog Model. This new search paradigm can better understand user intent, organize and manage information and knowledge. By engaging users, the novel approach could further clarify intent and facilitate task completion.

The best paper award was presented to the paper entitled "Unbiased Offline Evaluation of Contextual Bandit-based News Article Recommendation Algorithms", by Lihong Li, Wei Chu, John Langford, and Xuanhui Wang from Yahoo! Labs Research. In addition, the best student paper award was presented to the paper entitled, "Correcting for Missing Data in Information Cascades", by Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina from Stanford University. The best paper proposed a new evaluation metric for online recommender system learning algorithms, while the best student paper proposed a new estimation method for missing data in information cascades in the Social Web.

3. HOTSPOT: SOCIAL COMPUTING AND CROWDSOURCING

Similar to previous WSDM conferences, there are many papers related to search and data mining this year. Personalized search, search result diversification, evaluation for search results, relevance ranking, and topic models are still interesting and popular research topics. However, we observe a tremendous boost of social search and social data mining papers. For example, there are four papers on microblogging, covering many aspects such as automatic microblog search, identification of topical authorities in microblogs, etc.

Teevan et al. reported their analysis of microblog search. They collected a large amount of user's microblog search and Web search data through the browser toolbar, and compared these data. Subsequently, they reached the following conclusion: microblog search users prefer to retrieve and access timely information, including fast-breaking news, real-time reports, and current situations and dynamics. The search queries are in general shorter and more popular, and they are often searched repeatedly. Pal and Counts proposed an approach to identifying topical authorities in microblogs. There are many topical authorities in microblogs. Although some users may have many followers, they are not necessarily authorities for specific topics. In this approach, Pal and Counts collected microblogs sharing the same topic (e.g., all microblogs that have the keyword "oil spills"). Then they extracted features of authorities in these microblogs, and clustered authorities into two classes based on features. Finally they ranked authorities in major classes to find the topical ones. An important contribution of their work is the proposed useful features for topical authority identification.

In addition, Crowdsourcing appeared at the conference as a new research direction, as evidenced by one tutorial and one workshop. Crowdsourcing is to outsource data labeling tasks to the crowd on the Web. Due to its low cost, crowdsourcing can facilitate data labeling and obtain useful knowledge. This is why crowdsourcing will have far-reaching effects on Web search and data mining.

Omar Alonso and Matt Lease gave a tutorial on crowdsourcing. They systematically and detailedly introduced current situations, methods, advantages, challenges and opportunities about crowdsourcing. Currently, there are more than 10 crowdsourcing websites such as Mechanical Turk, Crowdflower, CloudCrowd. Everyone can register as a member and play the role of a requester or a worker on these websites. The requester distributes labeling tasks and lists detailed labeling methods, standards, and payments. The worker selects his interesting job, and performs the labeling task to obtain the payment. The requester often distributes labeling tasks for training and testing sets. There are a variety of tasks, such as evaluation of search results, misspelled error correction, document translation, and image recognition, etc. The requester often employs a large number of workers for the task to improve the quantity and quality for labeling data. The payment is often low, even for several cents. The worker treats the task as a chance to entertain, learn and exchange. The payment is of course one attractive reason. The crowdsourcing websites will benefit from royalty. Currently crowdsourcing is a hotspot, for example, there are 10 thousand completed tasks from January 2009 to April 2010 on Mechanical Turk. Alonso and Lease introduced how to use Mechanical Turk as a tool for researchers. They also illustrated how to design labeling standards and build labeling systems. They emphasized on the importance of quality management repeatedly and introduced management methods. Alonso and Lease showed advantages of crowdsourcing, such as low cost and convenient usage. They also mentioned its shortcomings: crowdsourcing is only suitable for simple labeling tasks. They believed that the combination of crowdsourcing and cloud computing will be a direction for future research.

The website of WSDM 2011 is available at http://www.wsdm2011.org/. The WSDM2011 Conference Committee would like to thank ACM SIGIR, SIGKDD, SIGMOD, & SIGWEB as sponsors and Microsoft, Google, Yahoo, Yandex, K.C. Wong Education Foundation, Arnetminer, and VeriGuide as supporters for making the conference a great success. The next conference in its series, WSDM 2012, will be held in Seattle, USA in February, 2012.

Irwin King
Lead Member of Computational Advertising Group, AT&T Labs Research, San Francisco, USA and Professor at The Chinese University of Hong Kong, Hong Kong. His research interests include machine learning, social computing, data mining, information retrieval and multimedia information processing.

Hang Li
Senior researcher and research manager of Web Search and Data Mining Group, Microsoft Research Asia, China. His research interests include information retrieval, natural language processing, statistical machine learning, and data mining.