Report on the 7th ACM International Workshop on Web Information and Data Management (WIDM 2005)

Angela Bonifati
Icar CNR, Italy
bonifati@icar.cnr.it

Dongwon Lee
Penn State University, USA
dongwon@psu.edu

Abstract

In this report, to our best recollection, we provide a summary of the 7th ACM International Workshop on Web Information and Data Management (WIDM 2005), which took place at the Hilton Bremen Hotel, in Bremen, on November 5, 2005, in conjunction with the 14th ACM Int'l Conf. on Information and Knowledge Management (CIKM).

1 Workshop Overview

The WIDM [2] (pronounced like "Widom") was originally started by Fereidoon Sadri (University of North Carolina at Greensboro) and Cyrus Shahabi (USC) in late 90s, and has reached its 7th in its series (all held in conjunction with ACM CIKM 2005). Initially, WIDM was one of few workshops that bridge DB and IR community to the (then) emerging Web community. About the same time as WIDM in 1998, for instance, WebDB has also started with the similar research themes.

This year, WIDM implemented a one-day program, and attracted 44 submissions from 15 countries, making the selection very competitive. All the papers were reviewed by 3 members of the Program Committee. After a discussion phase on a few papers with conflicting reviews, 12 papers were accepted, among which 8 full papers and 4 short papers. This represents a highly competitive acceptance rate of 27%. The covered topics include among the others Web Mining, Web and XML Data Management, Semantic Web, Web Commerce, Advanced Web Applications, Web Exploration, and Performance of Web Applications. Submissions covered most of the above topics with a predominance of papers on Web ranking and retrieval, XML data management and Web discovery, Web clustering and filtering. These happen to be some of the "hottest" topics that the community has been working on these days.

2 Technical Program

The 12 accepted papers were divided into three technical sessions. The papers can be downloaded from the workshop website [1], which has also additional information about the program.

2.1 Session 1: Web Ranking and Retrieval

The first paper "*Web Path Recommendations Based on Page Ranking and Markov Models*" [4] tried to overcome the shortcomings of current techniques to predict users' navigational behavior. The authors proposed the use of a PageRank-style algorithm for assigning prior probabilities to the web pages based on their importance in the web site graph, and experimentally showed that their approach yielded more objective and representative predictions than the ones produced from the pure usage-based approaches.

The second paper "*Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web*" [5] studied semantic similarity methods using WordNet. Despite the arrival of the Semantic Web era and a growing support of metadata, it is still unclear how to measure the "similarity" of two ontologies that bear little lexicographic similarity. In the paper, the authors investigated approaches that map terms (or concepts) to the ontology graph of WordNet and examined their relationships.

The last paper "*DirectoryRank: Bringing Order to Web Directories*" [6] presented DirectoryRank, a ranking framework that orders the web pages within a given topic according to how informative they are about the topic.

2.2 Session 2: XML Data Management and Web Discovery

In the first paper "*Exploiting Native XML Indexing Techniques for XML Retrieval in Relational Database Systems*" [7], the authors aim at reducing the gap between native XML database and XML-to-RDBMS systems. To that purpose, the authors examined how native XML indexing techniques can boost the retrieval of XML stored in an RDBMS, and proposed the Relational CADG (RCADG), an adaptation of several native indexing approaches to the relational model.

The second paper "*Query Translation Scheme for Heterogeneous XML Data Sources*" [8] investigated the query and data schema mismatching problem, and proposed an inclusion mapping algorithm that decides how compatible the schema of the query and that of the target XML documents are.

In the third paper "*Impact of XML Schema Evolution on Valid Documents*" [9], the authors studied the problem of the evolution of an XML Schema, and proposed to minimize document revalidation by detecting the document parts potentially invalidated by the schema changes.

The problem of discovering and composing the right web services from a pool of many is an important problem. In this context, the last paper "*A Framework for Semantic Web Services Discovery*" [10] proposed a framework for ontology-based flexible discovery of semantic web services. That is, their approach relied on user-supplied, context-specific mappings from an user ontology to relevant domain ontologies used to specify web services.

2.3 Session 3: Web Clustering, Filtering and Applications

The first paper "*Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora*" [11] investigated the significance of narrative text classification in the task of automatic key phrase extraction in Web document corpora. That is, according to their findings, key phrases extracted from the narrative text only are significantly better than those obtained from all plain text of Web pages.

The second paper "*On improving Local Website Search Using Web Server Traffic Logs: A preliminary Report*" [12] presented their study on the importance of web server traffic logs to improve "local" search, in addition to pure link counts as in PageRank.

The "shilling attack" can be defined as malicious user or set of users attempting to change the behavior of the system to suit their own needs. In the third paper "*Preventing Shilling Attacks in Online Recommender Systems*" [13], the authors proposed several metrics for analyzing rating patterns of malicious users, and suggested an algorithm for protecting recommender systems against shilling attacks.

The fourth paper "*Looking at Both the Present and the Past to Efficiently Update Replicas of Web Content*" [14] concerns the problem of keeping the copies of web pages in sync with their original copies. To solve this problem, they proposed a new approach that learns to predict the change behavior of web pages based both on the static features and change history of pages, and refreshes the copies accordingly.

Clustering the results of a search helps the user to overview the information returned. In the last paper of the workshop "*A Search Result Clustering Method using Informatively Named Entities*" [15], the authors regarded the clustering task as indexing the search results. Their solution first extracted named entities as labels, and used the importance in the search result and the relation between terms and search queries in selecting the right labels. In their prototype, the proposed solution showed higher performance than existing methods.

2.4 Keynote Address

This year's keynote address was given by Prof. Donald Kossmann from ETH Zurich, Switzerland with the title "*A Web of Data; New Architectures for New Technology?*" [3]. The development and evolution that the Web has undergone this last decade has not been coupled with an evolution of the way we build Web applications. Imperative programming languages (e.g., Java or C#) and middleware architectures are still dominant in industry. His talk discussed why these old architectures are problematic and spurred interesting ideas for novel software architectures to build Web applications. The keynote talk was immediately followed by several questions from the workshop audience on the major research issues behind this technology and its future directions. In particular, the talk highlighted the importance of new infrastructures for Web applications, and the emerging needs of several IT institutions. These are indeed experimenting the use of XML data in their applications, and in general of new cutting-edge technologies. An interesting issue emerged during the talk was the high flexibility of XQuery, the standard XML query language to be used as a full-fledged web programming language as well. Being Turing-complete, XQuery is unsurprisingly powerful and could in the long term be coupled with the Java technology in the web development process. To such a purpose, the keynote speaker has advocated his teaching experience with XQuery courses, and the observed fast learning curve, not to mention his consulting experience within the major financial institutions in Switzerland.

3 Final Thoughts

WIDM 2005 was very successful. It proved the effectiveness of a one-long-day workshop with high quality of talks and papers resulted in a lively and interesting discussion carried through the entire workshop. The proceedings of the workshop have been published by ACM [2]. The next WIDM, the 8th one in the series, will take place in conjunction with ACM CIKM 2006, in Arlington (VA) on November 10, 2006.

Acknowledgements

WIDM 2005 would have not been successful without the contributions of all the members of the organizing committee and the technical program committee. WIDM 2005 was sponsored by ACM SIGIR in co-operation with ACM SIGMOD, and Geselschaft für Informatik e.V. (GI).

References

[1] WIDM 2005 workshop web site, WIDM 2005 Program, WIDM 2005 Organizing and Program Committees, November 2005. "URL: http://pike.psu.edu/widm05/".

[2] A. Bonifati, and D. Lee. (Eds.). "7th ACM Int'l Workshop on Web Information and Data Management (WIDM 2005)". Bremen, Germany, November 4, 2005. ISBN 1-59593-194-5

[3] D. Kossmann. "A web of data: new architectures for new technology". In Proc. of WIDM 2005, page 1.

[4] M. Eirinaki, and M. Vazirgiannis and D. Kapogiannis. "Web path recommendations based on page ranking and Markov models". In Proc. of WIDM 2005, pages 2-9.

[5] G. Varelas, and E. Voutsakis, and P. Raftopoulou, and E.G.M. Petrakis, and E.E. Milios: "Semantic similarity methods in wordNet and their application to information retrieval on the web". In Proc. of WIDM 2005, pages 10-16.

[6] V. Krikos, and S. Stamou, and P. Kokosis, and A. Ntoulas, and D. Christodoulakis. "DirectoryRank: ordering pages in web directories". In Proc. of WIDM 2005, pages 17-22.

[7] F. Weigel, and K.U. Schulz, and H. Meuss. "Exploiting native XML indexing techniques for XML retrieval in relational database systems". In Proc. of WIDM 2005, pages 23-30.

[8] C.X. Chen, and G.A. Mihaila, and S. Padmanabhan, and I. Rouvellou. "Query translation scheme for heterogeneous XML data sources". In Proc. of WIDM 2005, pages 31-38.

[9] G. Guerrini, and M. Mesiti, and D. Rossi. "Impact of XML schema evolution on valid documents". In Proc. of WIDM 2005, pages 39-44.

[10] J. Pathak, and N. Koul, and D. Caragea, and V. Honavar. "A framework for semantic web services discovery". In Proc. of WIDM 2005, pages 45-50.

[11] Y. Zhang, and A.N. Zincir-Heywood, and E.E. Milios. "Narrative text classification for automatic key phrase extraction in web document corpora". In Proc. of WIDM 2005, pages 51-58.

[12] Q. Cui, and A. Dekhtyar. "On improving local website search using web server traffic logs: a preliminary report". In Proc. of WIDM 2005, pages 59-66.

[13] P-A. Chirita, and W. Nejdl, and C. Zamfir. "Preventing shilling attacks in online recommender systems". In Proc. of WIDM 2005, pages 67-74.

[14] L. Barbosa, and A.C. Salgado, and F. de Carvalho, and J. Robin, and J. Freire. "Looking at both the present and the past to efficiently update replicas of web content". In Proc. of WIDM 2005, pages 75-80.

[15] H. Toda, and R. Kataoka. "A search result clustering method using informatively named entities". In Proc. of WIDM 2005, pages 81-86.