13th International Workshop on the Web and Databases – WebDB 2010 –

Xin Luna Dong
AT&T Labs-Research, USA
lunadong@research.att.com

Felix Naumann
Hasso-Plattner-Institute (HPI), Germany
naumann@hpi.uni-potsdam.de

1. WEBDB HISTORY AND GOALS

The WebDB workshop has been held thirteen times so far: the first WebDB workshop was co-located with EDBT'1998, whereas the other twelve were co-located with the annual SIGMOD/PODS conference. The WebDB workshop provides a forum where researchers, theoreticians, and practitioners can share their knowledge and opinions about problems and solutions at the intersection of data management and the Web. WebDB has had a high impact and has been a forum in which a number of seminal papers have been presented.

Since 2002, the workshop always had a theme. In 2010 WebDB focused on *Quality of Web Data* and on *Linked Data*, but papers on all aspects of the web and databases were solicited, such as unstructured and semi-structured data management, data -extraction, -integration, -cleansing, and -mining, web applications and privacy, search and information retrieval, and distributed data management.

2. REVIEW PROCESS

By the final submission deadline of April 5, 2010, we had received 44 submissions by 105 authors. The program committee consisted of 28 renowned researchers from many different organizations. All papers received three reviews. The discussion phase was quite active and led us to finally accept 14 papers with an acceptance rate of 32%. Our selection emphasized papers on Web-based topics over papers that developed new techniques for XML data management, even if the average overall rating and reviews were slightly lower. The proceedings [3] are available at `http://webdb2010.org/`.

3. WORKSHOP IN INDIANAPOLIS

The workshop itself took place on June 6, 2010 just before the SIGMOD conference. The workshop was attended by 51 participants, who had registered specifically for WebDB 2010. It was thus the largest workshop at SIGMOD that year.

3.1 Invited talk

A special highlight of this year's WebDB was the talk by the invited speaker *Christian Bizer* from Free University of Berlin [1]. Chris Bizer is a co-founder of the DBpedia project [2], which publishes structured and cleansed data extracted from many language versions of Wikipedia. DBpedia is widely used and cited. It has been the kernel for the much larger ambition of creating an enormous web of Linked Data. In the talk, Chris compared the Linked Data movement, which stems from the Semantic Web research area, with research in the field of Dataspaces [4], which stems from the database area. He highlighted their interesting connections, such as the pay-as-you-go approach, and also their differences, such as their scope: while Linked Data research strives for a single global information space, Dataspace research has in mind integrated data sets dedicated to specific domains. The former was paraphrased by Chris in the talk as "somebody-pays-as-you-go".

Throughout the following SIGMOD conference Chris was recognized by several speakers during their talk, as many had used DBpedia for experiments, had attended WebDB and appreciated the chance to speak to one of DBpedia's creators. Such cross-fertilization between the semantic web and the database communities certainly has the potential to advance both fields as proven in this case.

3.2 Flash session and posters

For the workshop program we introduced an innovation, namely a half-hour flash session for all presenters at the beginning of the workshop. Each presenter had the opportunity to give a 1-2 minute pitch for his/her paper. To ease transition, the speakers had to submit one or two slides beforehand. All presenters took this opportunity and handed in one or two slides a few days before the

workshop. In many cases, the presenters chose a humorous take on their subject to pique the curiosity of the audience.

The flash session served multiple purposes. First, it helped introduce the speakers, even if one's actual talk can be scheduled for late afternoon; thus, they were no longer anonymous listeners until their talk, but were already engaged by interested colleagues. Second, it gave participants a preview of the talks to come – the traditional motivation for a flash session. Third, the fast pace of the session was very lively and woke everybody up. Especially regular non-speaker participants were pleasantly surprised by this agenda item. Finally, it ensured that all speakers indeed show up and are present for their talk. We have received very positive feedback for this flash session and can highly recommend it for other workshops.

In addition to the flash session we gave authors another chance to present their work: during coffee and lunch breaks posters could be presented. Seven authors took this chance and presented posters of their work around the coffee area, sparking several interesting discussions.

3.3 Research sessions

The flash session and invited talk were followed by four research sessions spread across the day:

1. Linked data and Wikipedia. The first session of the workshop featured four papers targeted at the workshop theme.

An Agglomerative Query Model for Discovery in Linked Data: Semantics and Approach. Sidan Gao, Haizhou Fu, Kemafor Anyanwu (North Carolina State University). This paper introduces the notion of *agglomerative querying* that knits together information across multiple queries in order to provide users with useful additional information beyond their specific queries.

XML-Based RDF Data Management for Efficient Query Processing. Mo Zhou, Yuqing Wu (Indiana University). This paper proposes decomposing RDF graph into a forest of semantically correlated XML trees, storing them in an XML repository, and rewriting SPARQL queries into XPath/XQuery queries for evaluation.

Querying Wikipedia Documents and Relationships. Huong Nguyen, Thanh Nguyen, Hoa Nguyen, Juliana Freire (University of Utah). This paper presents a new approach for querying Wikipedia content: the proposed approach not only considers documents that match the queries, but also exploits relationships between documents to return richer, multi-document answers.

WikiAnalytics: Disambiguation of Keyword Search Results on Highly Heterogeneous Structured Data. Andrey Balmin (IBM Almaden Research Center), Emiran Curtmola (UC San Diego). This paper proposes WikiAnalytics, which improves search on Wikipedia by clustering the search results based on the record types, fields, and values in the infoboxes.

2. Extraction. The next section of the workshop concentrated on techniques for data extraction from the Web.

Find Your Advisor: Robust Knowledge Gathering from the Web. Ndapandula Nakashole, Martin Theobald, Gerhard Weikum (Max-Planck-Institute fr Informatik). This paper presents a data extraction method that combines generalized patterns automatically learned from seed facts for obtaining *high recall* results with rule-based, declarative reasoning to also ensure *high precision*.

Redundancy-Driven Web Data Extraction and Integration. Paolo Papotti, Valter Crescenzi, Paolo Merialdo, Mirko Bronzi, Lorenzo Blanco (Universit Roma Tre). This paper hypothesizes that underlying sources of the same domain there is a *hidden conceptual relation*, based on which it exploits the redundancy of information to automatically extract and integrate data from the Web.

Using Latent-Structure to Detect Objects on the Web. Luciano Barbosa (AT&T Labs - Research), Juliana Freire (University of Utah). This paper addresses the problem of identifying pages that contain objects with a *latent structure* (*i.e.*, the structure implicitly represented in the page) by automatically extracting statistically significant patterns present *inside* the given set of boot-strapping object instances.

Popularity-Guided Top-k Extraction of Entity Attributes. Matthew Solomon (Columbia University), Cong Yu (Yahoo!), Luis Gravano (Columbia University). This paper proposes for each entity and attribute extracted from Web-accessible text documents, returning the top-k values according to a scoring function that weighs the extraction confidence as well as the "importance" of the documents from where the information originates.

3. Management and mining of large-scaled data. The third session emphasized the fact that much web data is of large scale and discussed solutions in data management and mining for effectively handling such large-scaled data.

Manimal: Relational Optimization for Data-Intensive Programs. Michael Cafarella (University of Michigan), Christopher R (Univ. of Wisconsin-Madison). Manimal combines optimization techniques well known from relational DBMS and ap-

plies them to MapReduce programs through code analysis. While this paper focuses on selection conditions that are present in the code, it also shows the great potential of other typical optimizations for MapReduce programs.

Learning Topical Transition Probabilities in Click Through Data with Regression Models. Xiao Zhang, Prasenjit Mitra (Pennsylvania State University). Users who search through a document collection, such as the Web or a digital library, occasionally switch their topical intent. Search engines that recognize the switch can improve result quality. This paper presents an algorithm to effectively detect these switches and evaluates the algorithm using click-through data from CiteSeerX.

Improved Recommendations via (More) Collaboration. Rubi Boim, Haim Kaplan, Tova Milo, Ronitt Rubinfeld (Tel-Aviv University). Collaborative filtering is a well-known and recognized technique to predict customer ratings and improve product suggestions. The authors of this paper present a distributed method that extends filtering across collaborating organizations, even if they support different domains.

4. Protocols and models for Web data. The final session dug deeper into the foundations of databases on the Web.

Concurrent One-Way Protocols in Around-the-Clock Social Networks. Royi Ronen, Oded Shmueli (Technion). Analysis of social networks, such as Facebook or Twitter, is becoming a hot topic. The authors of this paper present a consistency model for such networks, aiming at resolving inconsistency in scenarios such as the following: users decide about attending a party depending on their friends' decisions, but their friends change their minds later and those decisions become invalid.

Reconciling Two Models of Multihierarchical Markup. Neil Moore (University of Kentucky). While XML documents specify a strict hierarchical model, marking up text documents can result in interleaving markup. The author presents a new and generalized model for such situations, dubbed "range GODDAG", and shows its superiority over previous models, including the ability to unambiguously support update operations.

Tree Patterns with Full Text Search. Maria-Hendrike Peetz, Maarten Marx (University of Amsterdam). Many query languages over XML data have been extended to support text search predicates. The authors show how equivalent but differently specified queries produce different result ranking in most systems. The author's solution is a method to rewrite queries to a normal form before passing them to the query processor.

4. ACKNOWLEDGMENTS

We would like to thank the Hasso Plattner Institute (HPI) for sponsoring the keynote speaker and for providing the proceedings on USB sticks to the participants. We also thank SIGMOD's workshop chair Christian Jensen, SIGMOD's general chair Ahmed Elmagarmid and his team for their support throughout the preparation phase and the workshop in Indianapolis. Finally, we thank Microsoft's CMT team for providing the submission and reviewing platform.

5. REFERENCES

[1] C. Bizer. Web of linked data - a global public data space on the web. In *Proceedings of the 13th International Workshop on the Web and Databases 2010, WebDB 2010, Indianapolis, Indiana, USA*, 2010.

[2] C. Bizer, J. Lehmann, S. A. G. Kobilarov, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia – a crystallization point for the web of data. *Journal of Web Semantics (JWS)*, 7(3), 2009.

[3] X. L. Dong and F. Naumann, editors. *Proceedings of the 13th International Workshop on the Web and Databases 2010, WebDB 2010, Indianapolis, Indiana, USA, June 6, 2010*, 2010.

[4] M. J. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. 34(4):27–33, 2005.