Report of the International Workshop on Business Intelligence and the Web – BEWEB 2011

Jose-Norberto Mazón, Irene Garrigós
University of Alicante, Spain
{jnmazon,igarrigos}@dlsi.ua.es

Florian Daniel
University of Trento, Italy
daniel@disi.unitn.it

Malu Castellanos
HP Labs, USA
malu.castellanos@hp.com

ABSTRACT

The 2nd International Workshop on Business intelligencE and the WEB (BEWEB) was co-located with the EDBT/ICDT 2011 Joint Conference in Uppsala (Sweden) on March 25, 2011. BEWEB intends to be an international forum for researchers and practitioners to exchange ideas on how to leverage the huge amount of data that is available on the Web in BI applications and on how to apply Web engineering methods and techniques to the design of BI applications. This report summarizes the 2011 edition of BEWEB.

1. INTRODUCTION

Over the last decade, we have been witnessing an increasing use of Business Intelligence (BI) solutions that allow enterprise to query, understand, and analyze business data in order to make better decisions.

Traditionally, BI applications allowed business people to acquire useful knowledge from the data of their organization by means of a variety of technologies, such as data warehousing, data mining, business performance management, OLAP, periodical business reports, and the like. Yet, in the very recent years, a new trend emerged: BI applications no longer limit their analysis to the data inside a company. Increasingly, they also source their data from the outside, i.e., from the Web, and complement company-internal data with value-adding information from the Web (e.g., retail prices of products sold by competitors), in order to provide richer insights into the dynamics of today's business.

In parallel to the move of data from the Web into BI applications, BI applications are experiencing a trend from company-internal information systems to the cloud: BI as a service (e.g., hosted BI platforms for small- and medium-size companies) is the target of huge investments and the focus of large research efforts by industry and academia. The idea behind BI as a service is outsourcing the processing and analysis of large bodies of data by consuming BI from the cloud, which will help in enabling the so-called Cloud Intelligence [1].

The International Workshop on Business intelligencE and the WEB (BEWEB) targets the above two moves and creates an international forum for exchanging ideas on how to leverage the huge amount of data that is available on the Web in BI applications, and how to apply Web-related engineering methods and techniques to the design of BI applications, such as BI as a service.

BEWEB 2011 attracted an average attendance of about 20 people, who actively engaged in fruitful and animated discussions. High-quality submissions were received from 7 different countries: Austria, Cuba, Germany, Republic of Korea, Mexico, Spain and USA. Each paper was carefully reviewed by at least three members of the program committee. As result of this process, 5 papers were selected as long papers and 1 as short paper for presentation at the workshop. The program included two invited industrial talks and the opening keynote, which was the highlight of the workshop. The proceedings of this year's edition of BEWEB can be found at http://doi.acm.org/10.1145/1966883.

2. KEYNOTE

The keynote entitled "Dr. Crowdsource: or How I Learned to Stop Worrying and Love Web Data" given by Prof. Felix Naumann (Hasso Plattner Institute, Potsdam, Germany) was magnificent. The title was inspired by a famous Stanley Kubrick movie[1], satirizing the threat of nuclear war. Using this metaphor, Prof. Naumann revealed that dealing with web data can be a threat if its inherent heterogeneity is not tackled with appropriate techniques. He described his experiences with the problems caused by web data heterogeneity and the required daunting tasks and state of the art techniques to overcome them, thus facilitating web data integration: source selection to identify appropriate and high-quality sources, data extraction to obtain relevant structured data, scrubbing to standardize and clean data, entity matching to associate different occurrences of the same entity, and, finally, data transformation and data fusion to combine all data about an entity in a single, consistent representation.

3. INVITED TALKS FROM INDUSTRY

With the aim of bringing together researchers from academy and industry, we were proud of having two ani-

---

[1] http://en.wikipedia.org/wiki/Dr._Strangelove

mated speakers from industrial research labs: Sihem Amer-Yahia (Yahoo! Research) and Xin Luna Dong (AT&T Labs-Research).

Sihem's talk was entitled "I am complex: Cluster Me, Don't Just Rank me". In her lively talk, Sihem argued that Web search over high-dimensional and structured data should go beyond the "10-blue links experience", i.e., a ranked list of results to a keyword-based query. She postulated that an alternative to ranking is to cluster results by means of two approaches: persona-driven search and rank-aware clustering.

In her interesting talk entitled "SOLOMON: Seeking the Truth via Copying Detection", Luna presented the SOLOMON system, whose core detects copying between data sources. She delved into the techniques to effectively detect copying relationships between data sources, leverage the results in various aspects of data integration, and provide a user-friendly interface to facilitate identifying sources that best suit their information needs.

4. RESEARCH SESSIONS

Accepted papers were organized into two research sessions: (i) BI with Web Data and (ii) Engineering Web-Enabled BI.

4.1 BI with Web Data

In the last decade, the amount and complexity of data available on the Web has been growing rapidly. As a consequence, designers of BI applications making use of data from the Web have to deal with several issues. Among the most interesting challenges we find, for instance, the extraction and integration of heterogeneous data sources. But there are many other interesting research challenges that arise when the Web is seen as data repository: developing Web warehousing solutions, tackling data quality issues, leveraging semantic Web technologies, employing Web mining, extending BI to unstructured data (e.g., text) or semi-structured data (e.g., XML), and so on. In addition, Web Intelligence, which explores the use of Artificial Intelligence in conjunction with or in relation to Web technologies, has emerged as a new area that imposes new research challenges. The following three papers were presented in this session:

Self-supervised Web search for any-k complete tuples, by Alexander Löser, Christoph Nagel, Stephan Pieper, Christoph Boden (University of Technology Berlin, Germany). This paper highlights the importance of querying structured information from Web pages. The authors define a query processor that (i) transforms a structured query into a set of keyword queries that are submitted to a search engine; (ii) forwards search results to relation extractors; and then (iii) combines relations into result tuples. This novel query processor completes tuples returned by the relation extractors by systematically discovering any-k relations from Web search results.

Toward total business intelligence incorporating structured and unstructured data, by Byung-Kwon Park (Dong-A University, South Korea), Il-Yeol Song (Drexel University, USA). This paper surveys existing approaches that consolidate both unstructured and structured data for realizing the so-called total business intelligence. After reviewing existing work, the authors present an architecture for total business intelligence in which information retrieval, text mining, and information extraction technologies are integrated with relational OLAP technologies.

Integrating Web feed opinions into a corporate data warehouse, by Lisette García-Moya, Shahad Kudama, Maria Jose Aramburu Cabo, Rafael Berlanga (Universitat Jaume I, Spain). This paper presents an approach to integrate sentiment data extracted from Web opinion feeds into the corporate data warehouse where company analytical data and models are stored. This approach allows BI applications to perform new analysis tasks by using the traditional OLAP-based data warehouse operators.

4.2 Engineering Web-Enabled BI

The move of BI applications from company-internal information systems to applications that are accessible over the Web implies the need for web-specific design competencies. In this context, Web engineering methodologies and technologies represent a large body of knowledge and expertise that could be very useful in the design of applications that allow decision makers to access BI data and functionalities over the Web. Good Web engineering is one of the key foundations in the design of real-time BI and business performance management applications, as Web applications provide access to data from anywhere, at anytime, and via any media. However, BI on the Web implies a plethora of new research challenges that are specific to the BI context, e.g., using Web mashups and RIA for BI development, BI as a service, usability and accessibility for BI applications, etc. This session featured the following three papers:

Capturing data quality requirements for Web applications by means of DQ_WebRE, by César Guerra (UPSLP, Mexico), Ismael Caballero, Mario Piattini (University of Castilla-La Mancha, Spain). This paper presents a model-driven Web engineering approach for considering data quality requirements in the development of Web applications for BI. This approach is based on two artifacts: a metamodel and a UML profile for the management of data quality software requirements for Web applications called DQ_WebRE.

Model-driven restricted-domain adaptation of question answering systems for business intelligence, by Katia Vila (University of Matanzas, Cuba), Antonio Ferrández (University of Alicante, Spain). This paper presents an approach for adapting question answering (QA) systems to restricted domains, such as those related to specific business areas (e.g., healthcare, agriculture, transportation, etc.), with the

aim of providing BI applications with actionable information from unstructured sources (e.g., data from the Web, etc.). QA systems have been applied in an interesting fashion for obtaining concise answers to questions formulated in natural language from a collection of text documents, thus supporting the decision maker in the analysis of textual data sources.

Towards TomTom like systems for the Web: a novel architecture for browser-based mashups, by Emilian Pascalau (Hasso Plattner Institute, University of Potsdam, Germany). This paper introduces a new architecture for browser-based mashups based on the TomTom navigation systems. The described architecture is capable of addressing issues such as BI on demand and instant use, and offers the same degree of generality as the browser.

5. OUTLOOK

In previous editions of BEWEB the focus of the papers has been on synergies between BI and the Web that leverage heterogeneous and semantically rich Web data in BI applications, and use Web-related engineering methods for designing BI as a service. In future editions, we plan to continue this focus but additionally we would like to foster research on methods, models and technologies for realizing the BI-aided Web engineering, i.e., how to acquire, analyze, and manage actionable BI information from Web usage data (e.g., logs, data streams, click streams, etc.) to support the development of Web applications (e.g., to achieve ad-

vanced levels of personalization in websites). Likewise, we would like to consider other BEWEB related topics such as big data, data visualization, social networks, streaming data, data quality, privacy and security, and adaptive, contextualized and personalized Web applications.

Finally, to encourage alignment between academic research and industry, we plan to include a session devoted to present innovative industrial products, services, experiences and case studies.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Pedersen, T. B. Research challenges for cloud intelligence. In *Proc. of the 1st International Workshop on Business Intelligence and the Web, BEWEB 2010, DOI* = http://doi.acm.org/10.1145/1754239.1754247