Report on the 9th International Workshop on Web Information and Data Management (WIDM 2007)

Irini Fundulaki *
ICS-FORTH, Greece

Neoklis Polyzotis
University of California-Santa Cruz, USA

1 Introduction

The 9th *ACM International Workshop on Web Information and Data Management (WIDM 2007)* was held in Lisbon, Portugal on November 9, 2007 and was co-located with the 16^{th} International Conference on Information and Knowledge Management (CIKM). The main objective of the workshop was to bring together people from various communities to study how Web information can be extracted, represented, stored, analyzed, and processed.

The program committee accepted 20 papers from a total of 80 papers (an acceptance rate of 25%). The proceedings were published by ACM Press and distributed during the workshop. The papers accepted at the workshop addressed a number of subjects from diverse areas of research for the Web.

The papers were grouped in the following subject areas: *XML and Semi-Structured Data*, *Peer-to-Peer and System Design Issues*, *Personalization*, *Knowledge Mining* and finally *Web Metadata and Search*.

2 Research Papers

XML and Semi-Structured Data The paper by *J. Coelho* and *M. Florido* entitled *XCentric: Logic Programming for XML Processing* introduces a logic programming language similar to Datalog for querying and transforming XML data. To handle the extraction of XML elements, XCentric provides a pattern matching mechanism that is built on the typed unification of terms with functors of variable arity.

R. Ronen and *O. Shmueli* in their paper entitled *Evaluation of Datalog Extended with an XPath Predicate* propose the integration of XPath primitives in Datalog. They allow variables to range over XML elements, and introduce built-in predicates that test for the different XPath axes. The paper examines two techniques for evaluating the new primitives that are suitable for batch and ad-hoc query evaluation.

In the paper entitled *An approach to XML path matching*, *A. Vinson, C. A. Heuser, A. Silva* and *E. Silva de Moura* examine the problem of evaluating the similarity of XML paths. The authors propose a similarity function that treats each path as a sequence of labels, and applies an edit distance to compare the two sequences. The computation of the edit distance is coupled with a string similarity function that assigns a cost to the operation of substituting labels in the two sequences, thus taking into account the possibility that different tags refer to the same underlying concept.

F. Mesquita, D. Barbosa, E. Cortez and *A. Silva* in their paper *FleDEx: Flexible Data Exchange* describe a lightweight framework for data exchange that is suitable for non-expert users sharing data on the Web or through P2P systems. The proposed framework is based on FDM, a semi-structured data model, that enables a unified representation model for potentially diverse data sources. Given the FDM schemata of the source and target databases, a schema mapping is derived by first matching the leaf nodes in the two schemata and then inductively generalizing the matches to internal schema nodes. Source and target schema sample instances are considered to discover potentially interesting mappings.

Peer-to-Peer and System Design Issues *H. Kurasawa, H. Wakaki, A. Takasu* and *J. Adachi* in their paper entitled *Data Allocation Scheme Based on Term Weight for P2P Information Retrieval* discuss the implementation of a P2P system for the indexing of large text corpora. The system employs a document index to store the association between terms and the documents in which they appear. The index itself is distributed over all the nodes of the P2P system through a Distributed Hash Table (DHT). To reduce the number of indexed terms, the index only stores

*The author at the time of the organization of the workshop was a Research Fellow at the Database Group of the University of Edinburgh.

the most important terms for each document. Each document is broken in overlapping chunks using erasure codes, and the chunks are stored in the index in the entries of the corresponding document terms. The use of erasure codes guarantees that the document can be reconstructed by retrieving only a fraction of the total set of chunks stored in the index.

In the paper entitled *Distributed Monitoring of Peer to Peer Systems*, *S. Abiteboul* and *B. Marinoiu* introduce a framework for specifying monitoring tasks over a P2P system. Monitoring tasks are described in a new declarative language, called P2PML, that essentially specifies continuous queries over event streams. A P2PML program is compiled in a distributed algebraic plan that is evaluated on the nodes of the P2P system. The plan is amenable to optimization prior (or during) its execution through the application of rewrite rules.

In their paper entitled *Self-optimizing Block Transfer in Web Service Grids*, *A. Gounaris, C. Yfoulis, R. Sakellariou* and *M. Dikaiakos* examine the scenario where an application needs to transfer a large amount of data over the network, and consider the problem of tuning the block size of the data transfer in order to maximize efficiency. They propose to develop a controller that adjusts the block size automatically and continuously, based on observations of the system's recent performance. The paper examines two possible approaches for the realization of the controller, namely, numerical optimization (essentially, Newton's method) and extremum control.

The paper entitled *Load Balancing Distributed Inverted Files* by *M. Marin* and *C. Gomez* considers the problem of query scheduling in the context of a large-scale, parallel search engine. The goal of the authors is to examine, through extensive simulations, the performance of well known scheduling algorithms. An interesting point of the simulation methodology is the adoption of a simplified system architecture in order to model the parallel operation of the search engine. The results of the simulations indicate that simple algorithms, like round-robing and least-loaded-processor-first, perform better than more complicated algorithms in a wide variety of scenarios, thus arguing in favor of simplicity in the design of large-scale parallel query processors.

Personalization In the paper entitled *Supporting personalized top-k skyline queries using partial compressed skycube*, *J. Lee, G.-W. You, I.-C. Sohn, S.-W. Hwang, K. Ko* and *Z. Lee* examine the computation of top-k skyline objects based on a ranking function specified by the query. The paper considers the class of ranking functions that define a preference sequence over the attributes of the objects, and describes a query evaluation algorithm that utilizes the compressed sky-cube.

The paper entitled *Toward Editable Web Browser: Edit-and-Propagate Operation for Web Browsing* by *S. Nakamura, T. Yamamoto* and *K. Tanaka* discusses the idea of attaching user-generated annotations to Web pages to facilitate the discovery of interesting information. In a nutshell, the Web browser enables the user to mark part of the content in a web page as interesting or uninteresting, which affects the display of the remaining content. Several possible uses of the mechanism are discussed: filtering of messages and reviews, removal of advertisements, re-ranking of search results, and refinement of snippets in the results of search engines.

In their paper: *Mining User Navigation Patterns for Personalizing Topic Directories*, *T. Dalamagas, P. Bouros, T. Galanis, M Eirinaki* and *T. Sellis* examine the use of personalized recommendations to assist users in the browsing of topic directories. The idea is to cluster the visits of recent users based on the similarity of navigational patterns, and to identify within each cluster the dominant patterns of the following types: back-and-forth navigation among topics, frequent navigation patterns, and frequent navigation patterns that involve only topics whose sub-trees are visited frequently. This knowledge is used to predict topics of interest to users that fall within a specific user group, and to introduce short-cuts in the hierarchy when the same users browse the topic directory.

In the paper entitled *An Online PPM Prediction Model for Web prefetching*, *Z. Ban*, *Z. Gu* and *Y. Jin* discuss the use of an online *Prediction by Partial Mapping (PPM)* model to capture the evolving navigational patterns of users that visit a specific web site. The model is stored as a non-compact suffix tree that records the user's requests over a sliding window of the user's history. Predictions on future requests are generated by taking into account the patterns in the model and information on the accuracy of recent predictions.

Knowledge Mining *A. Schuth, M. Marx* and *M. de Rijke* in their paper: *Extracting the Discussion Structure in Comments on News-Articles*, discuss how to *collect, store,*

enrich and *discover* the *implicit structure* of discussions related to online newspaper articles. Four different but complementary methods that extract the *reacts on* relation between the comments are proposed. Standard information retrieval measures are used to evaluate the proposed methods that have low recall but high precision.

The paper entitled *Adaptive Web-page Content Identification* by *J. Gibson, S. Lubar* and *B. Wellner* discusses how to *detect identical* and *near duplicate* news articles from a set of Web pages. The approach undertaken in the paper is based on the idea of *breaking the document into sequences of blocks* and labeling each of the blocks as *Content* or *NotContent* by employing three different statistical machine learning methods, the best being Conditional Random Fields with very encouraging effectiveness results.

I. Horie, K. Yamaguchi and *K. Kashiwabara* in the paper: *Pattern Detection from Web using AFA Set Theory* propose an approach based on the Anti-Foundation-Axiom (AFA) Set Theory which discovers common substructures in webpages that belong to a single website. Webpages in a Web site are represented as a graph which is viewed as a membership graph of the AFA set theory. The proposed techniques overcome the limitations of the naive application of the AFA set theory: the first considers as potentially common substructures only the ones that appear more than once in a website; the second detects and removes unimportant links around the index pages and the third uses Galois lattices formed by the identified patterns.

In the paper: *Web Based Linkage*, *E. Elmacioglu, M.-Y. Kan, D. Lee* and *Y. Zhang* study the *entity resolution (record linkage)* problem. The basic assumption of the proposed technique is that if an entity is a duplicate of another and the first appears together with some information on the Web, then the latter may appear frequently with the same information on the Web (called *representative data* for the entity). The authors propose a new approach based on Information Retrieval metrics but takes into account the frequency information from the Web to identify the representative data of an entity.

Web Metadata and Search In their paper entitled *Using Neighbors to Date Web Documents*, *S. Nunes, C. Ribeiro* and *G. David* discuss how to use the *neighbors of a Web document* to determine its *last modification date*. The last modification date of a document is determined by looking at its Web-related features, and most specifically its neighboring documents. More precisely, the authors look at the pages that have an incoming link (in-links) and those that are pointed to by the page in question (out-links), in addition to page assets such as its images, objects, CSS and JavaScript files.

In the same context, *A. Jatowt, Y. Kawai* and *K. Tanaka* in the paper *Detecting Age of Page Content* discuss a novel approach for extracting *approximate creation dates* of content elements in webpages. The approach is based on searching inside page histories to discover the creation date of a page element, estimated as the most probable point in time at which the content was inserted in the page (as it can be approximated from past data). Page histories are reconstructed by automatically selecting and downloading past snapshots of pages from existing Web archives.

The paper entitled *On Improving Wikipedia Search using Article Quality* by *M. Hu, E.-P. Lim, A. Sun, H. W. Lauw* and *B.-Q. Vuong* discusses the development of quality-aware search methods that determine the quality of the Wikipedia articles automatically without interpreting the actual content of the article. The approach is based on *associating the quality of each article with the authority of their contributors*: an article has high quality if it is contributed by high authority authors, and an author has high authority if she contributes to high quality articles. Two new models are proposed that take into account the relation between the article quality and the authority of the authors. The empirical results showed that quality-aware search methods have encouraging performance over Wikipedia's full text search engine.

A. G. Lages, F. C. Delicato, P. F. Pires and *L. Pirmez* in the paper: *SATYA: A Reputation-based Approach for Service Discovery and Selection in Service Oriented Architectures*, use the reputation values of Web services to discover and select such services in the context of SOA. Reputation values are used to represent reliability of SOA-based systems in SATYA. These are assigned to each service provider regarding each QoS parameter. The authors carried out a set of experiments that proved that SATYA is effective in guaranteeing a high level of consumer satisfaction, and at the same time keeping the overhead of the system lower than traditional Service Level Agreements-based systems.