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

**Irini Fundulaki**[*]
ICS-FORTH, Greece
*fundul@ics.forth.gr*

**Neoklis Polyzotis**
University of California-Santa Cruz, USA
*alkis@cs.ucsc.edu*

## 1 Introduction

The 9th *ACM International Workshop on Web Information and Data Management (WIDM 2007)* was held in Lisbon, Portugal, in conjunction with the $16^{th}$ International Conference on Information and Knowledge Management (CIKM), on November 9, 2007.

Continuing the tradition of the previous WIDM workshops, the main objective of the workshop was to bring together researchers, industrial practitioners, and developers to study how Web information can be extracted, stored, analyzed, and processed to provide useful knowledge to the end users for various advanced database applications.

The call for papers resulted in the submission of 80 papers from 32 countries: Australia, Austria, Belgium, Brazil, Canada, Chile, China, Cuba, Cyprus, Czech Republic, France, Germany, Greece, India, Iran, Israel, Italy, Japan, Mauritius, Mexico, Netherlands, Norway, Pakistan, Poland, Portugal, Singapore, South Korea, Sweden, Switzerland, Taiwan, United Kingdom and the United States.

The program committee accepted 20 papers that 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 Paper Presentations

### 2.1 XML and Semi-Structured Data

The paper by *J. Coelho* and *M. Florido* entitled *XCentric: Logic Programming for XML Processing* introduces the XCentric logic programming language that is suitable for querying and transforming XML

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

data. Similar to Prolog, an XCentric program is built as a set of rules that combine logical predicates. To enable the extraction of elements from the XML data, the language provides a pattern matching mechanism that is realized through the typed unification of terms with functors of variable arity. The expressiveness of power matching is essentially regulated by the typing system, which includes regular expressions. This property lends great flexibility to XCentric, and as illustrated in the paper through a series of examples, it enables the concise specification of complex data transformations.

*R. Ronen* and *O. Shmueli* in their paper entitled *Evaluation of Datalog Extended with an XPath Predicate* propose the integration of XPath primitives in the well known datalog query language. More specifically, the authors allow variables that range over XML elements, and introduce built-in predicates that test for the different XPath axes. The goal is to enable the combined processing of relational and XML data in a seamless fashion. The paper examines two techniques for query evaluation, namely static and dynamic. The static technique pre-processes the XPath predicates in the datalog program and materializes relations containing their results. (The pre-processing is done in isolation with respect to the remaining predicates.) The datalog program is then rewritten to employ the materialized relations instead of evaluating the XPath predicates on-the-fly. The dynamic method incorporates the evaluation of XPath predicates directly in the processing logic of the datalog program (specifically, the semi-naive evaluation algorithm). As shown in the paper, the overhead of the static method can be amortized across the evaluation of several queries, provided that there is some locality in the query predicates. On the other hand, the dynamic approach exhibits better scalability when the workload contains queries with low locality.

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 paper introduces a similarity function that takes into account the hierarchical nature of XML data and also the possibility that different tags refer to the same information. (The latter arises frequently due to the use of synonyms, or simply due to errors in the data.) The main idea behind the similarity function is to treat each path as a sequence of labels, and apply 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. Intuitively, the string similarity function assigns a lower cost when a label is substituted with a similar label, and a high score otherwise. The paper presents experimental results showing that the proposed function works well compared to a previously proposed baseline technique.

Finally, *F. Mesquita, D. Barbosa, E. Cortez* and *A. Silva* in their paper entitled *FleDEx: Flexible Data Exchange* describe a lightweight framework for data exchange that is suitable for non-expert and casual users sharing data on the Web or through peer-to-peer systems. To accommodate the potential diversity of data sources, the authors base their framework on a semi-structured data model termed FDM. An FDM instance is essentially a node-labeled tree, where each node describes a real-world entity and the edges denote the nesting between entities. Nodes can also carry attributes whose values describe properties of the corresponding real-world entities. FDM has a direct correspondence to the XML data model, which makes it straightforward to translate a DTD or an XML document to FDM. Given the FDM schemata of the source and target databases, and perhaps sample instances, the derivation of a schema mapping proceeds in a bottom-up fashion, as follows. First, matchings are identified between source and target attributes that have similar names and/or content, and subsequently source and target nodes are matched based on similarities in the corresponding sub-trees and their attributes. The authors also present an algorithm that performs the translation given a specific mapping. The paper concludes with an empirical study that evaluates the performance of the framework on test data.

## 2.2 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 index is essentially an inverted list of terms, distributed over the P2P system using the services of a Distributed Hash Table. To reduce the number of index terms, the system only stores the most important terms for each document, under the assumption that these terms are most likely to select the document in user queries. To make the retrieval of documents more efficient and more resilient to failures, the authors employ a document chunking technique. In a nutshell, a document is broken in overlapping chunks using erasure codes, and the chunks are distributed among the index entries that reference the document. The allocation of chunks is performed in accordance to the importance of terms in the document, i.e., the index entry for an important term is assigned a larger subset of the chunks. The use of erasure codes guarantees that the document can be reconstructed by retrieving only a fraction of the total set of chunks. Thus, the fragments for a specific document can be downloaded in parallel from its different index entries, and not all fragments need to be available for reconstruction.

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. An interesting property of this approach is that the algebraic plan is amenable to optimization prior (or during) its execution through the application of rewrite rules. The paper pays particular attention to the Filter algebraic operator, which is responsible for routing a specific event stream to several monitoring tasks. The idea is that a monitoring task is notified of an event in a stream only if the event matches specific conditions. The authors propose an implementation, based on the YFilter algorithm, that can sustain a high rate of events and a high number of subscribed monitoring plans, each placing its own conditions over the stream. The paper also discusses the sharing of work among several monitoring plans through the re-use of event streams. The proposed approach is to maintain a distributed database of all the available event streams, and then to rewrite the algebraic plans of new monitoring tasks in order to re-use existing event streams.

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. The proposed solution is 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, and evaluates them empirically for a wide set of parameters. The results indicate that the self-regulated controller can yield substantial improvements in system performance. Moreover, the results suggest that the extremum control controllers are more robust in the sense that they are less sensitive to fluctuations in the run-time environment.

In the final paper of the session, entitled *Load Balancing Distributed Inverted Files*, *M. Marin* and *Carlos Gomez* consider 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 operation of the parallel computer. The simplified model is termed the bulk-synchronous parallel computer, and, as shown in the paper, it yields an accurate enough approximation for the purpose of the specific evaluation. Using this model, the authors put to the test a variety of previously proposed scheduling algorithms. Perhaps somewhat surprisingly, 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.

## 2.3 Personalization

In the paper entitled *Supporting Personalized Top-k Skyline Queries in High-dimensional Web Data*, *J. Lee, G.-W. You, S.-W. Hwang* and *I.-C. Sohn* 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. Existing solutions for this problem need to maintain information that grows at an exponential rate with the number of dimensions. To address this issue, the authors introduce a query evaluation algorithm that utilizes the compressed sky-cube (proposed by Xia and Zhang in SIGMOD 2006). The algorithm materializes only the parts of the sky-cube that are relevant to the query, thus leading to a solution that is efficient both computationally and in terms of space requirements.

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 in order to facilitate the discovery of interesting information. In a nutshell, the Web browser enables the user to mark some content in a Web-page as interesting or uninteresting, which in turn affects the display of the remaining content. For instance, if the user is looking at reviews of a specific product and marks one particular review as uninteresting, the browser automatically hides all other reviews that are similar (and vice versa). The authors discuss several possible uses of the annotations depending on the marked content, namely, filtering of messages and reviews, removal of advertisements, re-ranking of search results, and refinement of snippets in the results of search engines. The paper also reports the results of a user-based study that evaluates the effectiveness of this methodology in isolating the interesting bits of information on a webpage. The presentation of the paper at the workshop included a demo of a prototype browser that implements the proposed functionality.

In their paper entitled *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. A topic directory is defined as a tree of category nodes, and a user is assumed to explore the tree one node at a time. The main idea is to form a database of user navigation patterns extracted from the visits of recent users. The database is organized in terms of user groups that represent users with similar navigation patterns. For each group, the authors propose to extract a set of navigation patterns from the following classes: back-and-forth navigation among categories, frequent navigation patterns, and frequent navigation patterns that involve only categories whose sub-trees are visited frequently. This knowledge is used to predict topics of interest to users that fall within a specific user group. The identified topics are then presented as shortcuts in the topic hierarchy in order to facilitate their easy discovery. The authors consider the application of this idea in two contexts: off-line, where the shortcuts are the same for all the users in the same user group, and on-line, where the shortcuts are tailored to the most recent actions of each user. The paper presents a user study of the proposed technique on the Open Directory Project hierarchy.

In the final paper of the session 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 changing user request patterns. PPM is a commonly used technique in Web prefetching where decisions are made based on historical user requests in a Markov prediction tree. The problem with the available PPM models is that they might become too large to fit in memory since *i)* the Web changes constantly and *ii)* user requests are continuously available. Therefore it is highly desirable to perform the update of the PPM model *online* and *incrementally*. The authors propose to use an *online PPM model* based on the non-compact suffix tree where new user requests are inserted and old ones are deleted. To control the number of requests in the PPM model the authors propose the use of a sliding window. In this way, the model fits in main memory and at the same time captures changing patterns. The concept of entropy is used to select the nodes that can be used for predicting the Web pages that should be fetched and stored for later use. The prediction function combines smaller entropy, greater prediction accuracy rate and the longest match to predict the user's next request. Experimental results show that the proposed model gives the best result when compared with existing PPM models.

## 2.4 Knowledge Mining

In their paper entitled *Extracting the Discussion Structure in Comments on News-Articles*, *A. Schuth, M. Marx* and *M. de Rijke* discuss how to *collect, store, enrich* and *discover* the *implicit structure* of discussions related to online newspaper articles. The paper discusses data collection and storage: specialized wrappers are used to collect the data which is then stored in a relational database. The authors look at properties of the collected newspaper articles such as the amount of comment threads in an article, the article's publication time etc.. Properties of comments such as the length of the comment, the number of sentences per comment and the length of the sentences are also discussed. The paper presents the identifying properties of a commenter (such as the name, e-mail address) used in extracting the discussion structure in comments. The authors propose and experiment with four different but complementary methods to extract the *reacts on* relation between the comments using a definition that takes into account the comment's *i)* publication date and *ii)* authors. Standard information retrieval measures (such as recall, precision and F1) are used to evaluate the proposed methods that have low recall but high precision. Nevertheless, the authors show that when those methods are combined recall improves without affecting precision.

*J. Gibson, S. Lubar* and *B. Wellner* in their paper entitled *Adaptive Web-page Content Identification* discuss how to *detect identical* and *near duplicate* news articles from a set of Web pages. The authors provide supporting evidence on the limitations of custom extractors that fail to identify the content of Web pages when, for example, the layout of a page changes. 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. This method can identify all of the blocks in the document and label them accordingly. More specifically, the content from unseen Web pages and from unseen Web sites can be identified over 80% of the time. Under less strict conditions, the document level accuracy improves to 98%. The experimental results prove that the use of machine learning-based content extraction is a good alternative to custom made content extractors.

The paper *Pattern Detection from Web using AFA Set Theory* by *I. Horie, K. Yamaguchi* and *K. Kashi-*

*wabara* proposes an approach based on the Anti-Foundation-Axiom (AFA) Set Theory which discovers common substructures in webpages that belong to a single website. The authors motivate the discovery of common substructures with the need of maintaining the consistency of a Web site when webpages change. The approach presented in this paper is based on the use of *higher-order rank* for set theory (based on the Anti-Foundation Axiom (AFA)). The idea is to represent the webpages in a Web site as a graph, and view it as a membership graph of the AFA set theory. The authors propose three new techniques to overcome the limitations of a naive application of the AFA set theory (which results in missing certain common structures, and identifying non-common ones as common). The first technique consists in considering as potentially common substructures only the ones that appear more than once in a website (these substructures are called *patterns*). The second one detects and removes unimportant links around the index pages and the third one uses Galois lattices that are formed by the identified patterns.

*E. Elmacioglu, M.-Y. Kan, D. Lee* and *Y. Zhang* in their paper: *Web Based Linkage*, study the *entity resolution (record linkage)* problem. The basic assumption of the proposed technique is that if an entity is a duplicate of another entity and the first appears together with some information on the Web, then the latter may appear frequently with the same information on the Web. This information is defined as the *best representative data piece* for the entity, that is, the data that distinguishes the entity from others. Once the representative data for an entity $e$ is collected, in order to decide whether an entity $e'$ is a duplicate of entity $e$, the authors need to decide whether $e'$ appears frequently with the representative data for $e$. To identify the representative data of an entity, the authors use conventional approaches from Information Retrieval (like *tf*, *tf*idf* etc.) and propose a new one that takes into account the frequency information from the Web. Once this data is identified, Web pages for $e$ and $e'$ are collected by asking the following queries: one query that contains $e$ with its representative data, and another that contains $e'$ along with the representative data of $e$. The authors then consider the numbers of Web pages returned, the similarities between the URLs and the content of these pages. The purpose of the experimental study is to verify whether and which Web based linkage schemes work as well as how the results are affected by the choice of the search engine.

## 2.5 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* in order to determine its *last modification date*. The authors recognize that incorporating temporal information in Web resources is a first step towards enabling *temporal sensitive ranking*. Information such as HTTP headers is already used towards determining the last modification date, but often this information is either unavailable or unreliable. The current work explores ways of determining this value by looking at Web-related features of the document, 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 files and JavaScript files. An average value for the last modification date is computed for each of these sets, and associated with the last modification date of the Web page. An interesting correlation was observed between a document's last modification date and the average value for the last modification date obtained from its neighbors: the correlation was stronger with the document's out-links, followed by the page assets and then with the in-links. The experiments conducted showed that when the document's neighborhood is considered, the proposed method is able to determine the last modification date of 86% of the sample documents. On the other hand, when neighboring information is not considered, then the last modification date of only 52% of the sample documents can be established.

*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. Timestamps or other temporal information that indicate the creation time and date of a page are not always available or cannot always be trusted. The approach proposed in this paper provides a method for the automatic generation of temporal metadata by determining the age of content elements in Web pages. The approach is based on searching inside page histories to discover the creation date of a page element, that is 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). The interesting point of the work is that in order to determine the age of page content, the page histories are reconstructed by automatically selecting and downloading past snapshots of pages from exising Web archives (named *past Web*). Then, these snapshots are searched (by applying a sequential or multi-binary search method) to discover the point in time at which the current content was inserted in the Web page.

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 novel approaches for designing and evaluating quality aware search methods for Wikipedia articles using both *relevance* and *quality* incorporated in the ranking of search results. Wikipedia is a very successful online encyclopedia with a continuously increasing number of articles that are contributed by both experienced and inexperienced users and are often of low quality. These articles are searched using either external search engines or the Wikipedia built-in search engine but none of the available engines ranks the articles based on their quality: it simply uses their relevance to the user query. This work aims at developing 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*: that is, 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. The authors propose the *Basic* and *PeerReview* models for measuring the quality of the articles, which are then compared with the naive way of measuring the quality of an article using only its length. The former takes into account the mutual dependency between article quality and the authority of the contributor. The latter extends the former by taking into account the fact that an author can also review articles. Hence, content that is reviewed and approved by high authority authors, will be of high quality. The objective of the experimental study was to examine the effectiveness of the proposed models in obtaining articles that are relevant to the query and are of high quality. The results showed that incorporating quality-aware search methods have encouraging performance over Wikipedia's full text search results.

*A. G. Lages, F. C. Delicato, P. F. Pires* and *L. Pirmez* in their paper entitled *SATYA: A Reputation-based Approach for Service Discovery and Selection in Service Oriented Architectures*, discuss a novel approach based on reputation values of Web services that can be used to discover and select such services in the context of SOA. SATYA augments the reliability of SOA-based systems which is represented through reputation values assigned to each service provider regarding a QoS parameter. These reputation values are computed from the subjective evaluation values issued by service consumers, and the objective evaluation values provided by objective monitoring entities. These values are compared to validate subjective evaluations, minimize the degree of subjectivity of computed reputation values and finally discover consumer preferences in terms of QoS metrics. In case of discrepancies, and if the objective values are out-of-date, then the system dynamically adopts the probing sequence to reflect more realistically the current state of the provider. This is a novel point of SATYA, that has the advantage of increasing the scalability of the whole system. 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 that traditional Service Level Agreements-based systems.