Report on the 10th International Workshop on Web Information and Data Management (WIDM)*

Chee-Yong Chan
National University of Singapore, Singapore
chancy@comp.nus.edu.sg

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

1 Introduction

The 10th *ACM International Workshop on Web Information and Data Management (WIDM 2008)* was held in Napa Valley, California, USA, in conjunction with the 17^{th} International Conference on Information and Knowledge Management (CIKM), on October 30, 2008.

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 53 papers from 22 countries. The program committee accepted 20 papers that were grouped in the following subject areas: *Data Mining and Clustering*, *Systems Issues*, *Web 2.0 and Social Networks*, and finally *Ranking and Similarity Search*.

2 Paper Presentations

2.1 Data Mining and Clustering

The paper by *M. Hu*, *A. Sun*, and *E.P. Lim* entitled *Event Detection with Common User Interests* deals with the problem of identifying events that can be detected through the publication of online documents (articles, blog entries, etc.) and the search queries posed over said documents. The proposed solution correlates the stream of queries to the stream of documents and maps each query to a set of documents such that (a) the documents are published near the time of the query, and (b) the documents are relevant for the query. Events are detected by computing clusters of queries based on the similarity between the corresponding document-sets, where each cluster represents queries on the same event.

P. Senellart, *A. Mittal*, *D. Muschick*, *R. Gilleron*, and *M. Tomassi* in their paper entitled *Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge* propose an automated method to infer a web service wrapper for an HTML form. The method works in two stages, using solely knowledge about the domain of the data source accessed through the form. First, the form's fields are matched to concepts based on lexical information and the provided domain knowledge, and the matching is verified by probing the form with instances of the concepts. Subsequently, a unsupervised machine learning algorithm is used (bootstrapped with the domain knowledge) in order to understand the structure of the HTML pages returned by invoking the form.

In the paper entitled *PIXSAR: Incremental Reclustering of Augmented XML Trees*, *L. Shnaiderman*, *O. Shmueli* and *R. Bordawekar* propose a clustering-based approach for the physical storage of an XML document. The proposed method maintains access statistics during query processing per parent-child and sibling-sibling edge, and then partitions the XML tree in disjoint sub-trees (each stored on the same page) representing elements that are frequently accessed together in the evaluation of the workload. The statistics are updated continuously and incremental reclustering may be performed if the access patterns in the workload change.

The paper entitled *A Study of the Relationship between Ad Hoc Retrieval and Expert Finding in Enterprise Environment* by *J. Zhu* evaluates how the results of search queries affect the task of expert finding. Specifically, given a search query, the experts are identified by examining the co-occurrence between query terms and expert names in the documents most relevant to the query. The paper analyzes empirically whether the parameters that affect the relevance of documents also have an effect on the identification of experts. The following three parameters are considered: background smoothing, anchor texts, and the in-degree of documents.

Finally, *S. Huang, X. Wu* and *A. Bolivar* in their paper entitled *The Effect of Title Term Suggestion on E-commerce Sites* question the assumption that sellers in e-commerce sites provide a descriptive title to their products. Data collected from eBay shows that a significant number of items have a very short title and are thus missed by customer queries. To address this issue, the authors employ the idea of query expansion: the title of an item is pre-processed at the time of registration, and a set of additional terms are suggested based on terms found in the query logs and titles of other related items. The seller can then select which of the suggested terms should be included in the title in order to increase the chance of successful customer searches.

*http://widm2008.comp.nus.edu.sg/

2.2 System Issues

M. Klein and *M. Nelson* in their paper entitled *A Comparison of Techniques for Estimating IDF Values to generate Lexical Signatures for the Web* evaluate methods for estimating the inverse document frequency (IDF) of terms at the scale of the Web. The authors examine three possible approximation schemes: the NG method uses the Google N-Grams data set and estimates the IDF as the frequency of the corresponding unigram; the LC method estimates the IDF based on a sample of web pages downloaded from the Internet Archive and the Open Directory Project; finally, the SC method googles the term and scrapes the screen of results for the reported document frequency. An empirical comparison shows that the three methods yield similar approximations results.

In their paper entitled *High-Performance Priority Queues for Parallel Crawlers*, *M. Marin, R. Paredes* and *C. Bonacic* examine efficient data structures for prioritizing the URLs downloaded by a highly parallel crawler. The authors propose two new data structures that can be implemented efficiently in a parallel system: (a) a parallel queue that uses binary tournaments upon a complete binary tree in order to identify the top URL, and (b) the Quick Heap structure (QH for short) that uses Hoare's QuickSelect algorithm to perform a partial sorting and identify efficiently the top k URLS, where k is a parameter in the system. An experimental study demonstrates that the new queues can enable significant performance improvements in a parallel crawler.

C. Garcia-Alvarado and C. Ordonez in their paper entitled *Information Retrieval from Digital Libraries in SQL* describe the implementation of an IR framework mostly in standard SQL, with the motivation of supporting ad-hoc information retrieval on top of a conventional RDBMS. The proposed implementation aims to be both portable and efficient, and it supports several common term weighting schemes.

The paper *HiPPIS: An Online P2P System for Efficient Lookups on d-Dimensional Hierarchies* by *K. Doka, D. Tsoumakos* and *N. Koziris* describes a DHT-based index for relational data sets conforming to a star schema. The indexing scheme, termed HiPPIS, indexes the tuples in the DHT by fixing a specific level on each dimension. A query that constrains exactly the same set of levels is answered directly from a single DHT node, whereas any other type of query is answered by flooding. To amortize the cost of flooding, HiPPIS creates soft-state indices that cache the location of tuples for recently flooded queries. Moreover, HiPPIS is able to adjust dynamically the set of indexed dimension levels in order to track changes in the workload.

Finally, *M. Karnstedt, K.U. Sattler, M. Ha, M. Hauswirth, B. Sapkota* and *Roman Schmidt* in their paper entitled *Approximating Query Completeness by Predicting the Number of Answers in DHT-based Web Applications* propose the metric of query completeness as an indicator of query progress in a DHT-based peer-to-peer system. Intuitively, query completeness measures the fraction of answers that have been received compared to the total number of answers. The authors propose and analyze several approximation schemes for this metric that can be computed as the query is routed in the overlay network.

2.3 Web 2.0 and Social Networks

The paper entitled *From Web 1.0 to Web 2.0 and back - How did your Grandma use to tag?* by *S. Kinsella, A. Budura, G. Skobeltsyn, S. Michel, J. Breslin* and *K. Aberer* presents a study to compare the relationship between "Web 1.0 tags" that are extracted from Web 1.0 anchortext and metadata and "Web 2.0 tags" that are obtained from the tagging portal del.icio.us. The study reveals that the Web 1.0 tags generated from a simple tag extraction method have a significant overlap with the Web 2.0 tags. Thus, the simple extraction method can be applied to bootstrap tagging portals or enrich the set of tags in tagging portals.

In the paper entitled *Modeling the Mashup Space*, *S. Abiteboul, O. Greenshpan* and *T. Milo* introduce a formal framework for specifying mashups. A mashup is modeled as a dynamic network of interacting *mashlets*, which are the basic components of the proposed model. Mashlets can query data sources, import other mashlets, use external Web services, and specify complex interaction patterns between its components. The state of a mashlet consists of a set of relations and its logic is expressed in terms of Datalog-style active rules. The concepts of the model is illustrated with a personal health information system demonstrating its expressiveness and usefulness.

In their paper entitled *Nereau: Query Expansion Using Social Bookmark*, *C. Biancalana, A. Micarelli* and *C. Squarcella* present a new approach to enhance query expansion with personalization by exploiting tag information from social bookmarking services. A user model (in the form of a three-dimensional matrix of co-occurrence values) is first built by analyzing the user's previous search queries and visited urls and deriving relevant terms from the visited web pages using the tag information from social bookmarking services. Using the user model, a new search query is expanded into multiple queries with the results organized into categories.

J. Park, T. Fukuhara, I. Ohmukai, H. Takeda and *S.-G. Lee* in their paper entitled *Web Content Summarization Using Social Bookmarks: A New Approach for Social Summarization* propose a novel Web content summarization technique to create text summaries by exploiting user feedback from social bookmarking services. First, representative words are extracted from user comments, which are then used to extract sentences that contain the representative words; the sentences are then scored and a summary is then formed using the top-k sentences.

In the final paper of the session, entitled *Granular Mod-*

eling of Web Documents: Impact on Information Retrieval Systems, *E. Fersini, E. Messina* and *F. Archetti* examine the use of a granular representation of web pages for improving the accuracy of web page classification and improving web page ranking. The approach proposed for the first problem is based on the assumption that a web page's blocks that contain images, referred to as image blocks, contain more significant information about the web page contents. An unsupervised algorithm is proposed to identify the most informative image blocks and their most relevant terms using an inverse term importance metric. In the second problem, the authors exploits the semantic relationships among document blocks for page ranking computation, where the probability of clicking a hyperlink is estimated by the degree of textual coherence between the source and destination web pages through the block containing the hyperlink.

2.4 Ranking and Similarity Search

In their paper *Quantify Music Artist Similarity based on Style and Mood*, *B. Shao*, *T. Li*, and *M. Ogihara* discuss the use of style and mood aspects to quantify music artist similarity. Their proposed approach operates by first obtaining the style and mood descriptions of music artists from the All Music Guide website, which are then used to compute style and mood similarity taxonomies based on a hierarchical co-clustering algorithm. The similarity measure for each of style and mood is derived by taking the average of four normalized similarity values computed using known approaches. The final combined artist similarity function is computed as a weighted sum of the mood similarity and style similarity.

In the paper entitled *Boosting the Ranking Function Learning Process using Clustering*, *G. Giannopoulos, T. Dalamagas, M. Eirinaki* and *T. Sellis* examine the problem of how to increase the training input for ranking function learning systems without requiring more explicit or implicit user feedback. Acquiring adequate training data (in the form of relevance judgements for query-result pairs) to train a ranking function typically requires a long training period or involve a large number of users. The basic idea of the proposed approach is to expand the initial set of relevance judgements obtained from implicit user feedback by first clustering the search result documents based on their content similarity. After removing clusters that have low coherence in terms of the distribution of the relevance judgement of the cluster documents, each of the remaining clusters is labeled a relevance judgement based on the majority of the relevance judgements in the cluster. In this way, relevance judgements are estimated for documents that have not been viewed thereby increasing the set of training data.

Y. Sun, H. Li, I. Councill, J. Huang, W.-C. Lee and *C. Lee Giles* in their paper entitled *Personalized Ranking for Digital Libraries Based on Log Analysis* propose a personalized ranking method that is based on user preference models to improve the accuracy of predicting user actions. A user preference is modeled as a vector, termed the user preference vector, in the document feature space. The user preference vectors are obtained by training on implicit user feedback extracted from web log mining results, where each user feedback is represented by a document pair indicating that the user prefers the first document over the second one. The level of relevance of a document for a user is defined as the inner product of the document vector and the user preference vector. When a user submits a query to the system, the system first retrieves all the documents based on lexical similarity, and then re-ranks the documents based on the user's preference vector.

R. Pon, A. Cardenas and *D. Buttler* in their paper entitled *Online Selection of Parameters in the Rocchio Algorithm for Identifying Interesting News Articles* study how to dynamically adapt parameter values for the Rocchio algorithm to improve recommendation performance for a news articles filtering application. To enable more effective document filtering for different users, the authors propose an enhanced approach, termed *eRocchio*, where each incoming document is evaluated by multiple instantiations of the Rocchio formulation in parallel. Each Rocchio instantiation has its own unique weight parameter value and adaptive thesholder to optimize its corresponding instantiation. The best instantiation is then selected using a F-measure metric, and the recommendation outcome is used to adaptively update each instantiation.

In the final paper of the session, entitled *Supporting the Automatic Construction of Entity Aware Search Engines*, *L. Blanco, V. Crescenzi, P. Merialdo* and *P. Papotti* present a domain-independent approach to automatically search the web for pages that are publishing data about instances of a conceptual entity of interest. Given an input set of sample web pages from several distinct websites about some conceptual entity, the proposed approach first crawls these websites to collect more web pages about other instances of the conceptual entity. This is performed assuming that pages that describe different instances of a conceptual entity within a website share a common template. The system then automatically extracts a description of the entity using a set of keywords and launches web searches to look for new pages about the conceptual entity. By analyzing the returned web pages using the extracted entity description, new pages that represent the conceptual entity are identified and used to recursively trigger the search process.