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: Australia, Brazil, Canada, Czech Republic, China, Chile, France, Germany, Greece, India, Ireland, Israel, Italy, Japan, Korea, Singapore, Spain, Switzerland, Taiwan, Turkey, United Kingdom, and the United States.

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 material (documents, articles, blog entries, etc.) and the search queries performed by users over said material. At an abstract level, the input consists of a stream of documents and a stream of queries. Each query is mapped to the set of documents that are published near the time of the query and that are relevant. This

set is termed the query profile. To detect events, the authors propose to cluster the query profiles based on their similarity. Thus, a cluster represents queries that have selected similar and related documents, and the keywords in the queries can be used to describe the event. The paper evaluates the proposed framework using the query stream on the web site Technorati and documents published at Technorati and Google News.

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 a method to infer a web service wrapper for an HTML form that provides access to a hidden Web source. The method works in a completely unsupervised and automated fashion, requiring only some domain knowledge about the hidden data source. The first step is to annotate the fields for the form with concepts from the underlying domain. This is done by initially matching fields to concept names using lexical information from the form, and then probing the form with instances of the candidate concepts to verify the correspondence. Once the form has been "parsed", the next step is to understand the structure of the HTML pages that result by invoking the form. For this step, the authors propose to use unsupervised machine learning that is bootstrapped with the provided domain knowledge. The paper presents an experimental evaluation of the proposed framework on the HTML forms of ten well known publication servers.

In the paper entitled *PIXSAR: Incremental Reclustering of Augmented XML Trees*, *L. Shnaiderman*, *O. Shmueli* and *R. Bordawekar* propose a clustering-based approach for storing an XML tree across a set of disk pages. The main idea is to maintain access statistics per parent-child and sibling-sibling edge that represent the frequency of accessing the corresponding nodes "close enough" in the evaluation of path expressions. The clustering algorithm partitions the XML tree in disjoint sub-trees so that the intra-cluster weights are maximized, i.e., each cluster represents elements that are frequently accessed together in the evaluation of the workload. The proposed framework updates the statistics continuously and can perform incremental reclustering if the access patterns in the workload change. Thus, the goal is to make the clustering adapt to the characteristics of the workload. The paper presents an experimental study of the proposed framework using the XMark benchmark.

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. The latter can be modeled as follows: users first find relevant documents through appropriate keyword search queries; then, the co-occurrence between expert names and query terms is analyzed to identify the most likely experts. The paper analyzes empirically whether the parameters that affect the first step have any effect on the output of the second step. More concretely, the paper considers the following three parameters: background smoothing, anchor texts, and the in-degree of documents. The empirical study is based on the search topics of the TREC2007 Enterprise Track benchmark.

Finally, *S. Huang, X. Wu* and *A. Bolivar* in their paper entitled *The Effect of Title Term Suggestion on E-commerce Sites* challenge the assumption that sellers in e-commerce sites provide a descriptive title to their products. Based on data collected from eBay, they argue that a significant number of items have a very short title and thus they are missed by customer queries. To address this issue, the authors propose a solution based on 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.

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 IDF (Inverse Document Frequency) of terms at the scale of the web. The IDF value is necessary for ranking a term in a document according to its TF-IDF metric, and computing the precise IDF value over all web documents is clearly infeasible. The authors examine three possible approximation schemes, termed NG, LC, and SC. The NG method uses the Google N-Grams data set and estimates the IDF value as the frequency of the corresponding unigram. The LC method estimates the IDF value 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. The methods are compared empirically by examining the ranking of terms within a set of web pages using each method. The results indicate that the three methods yield similar rankings overall, and in particular they agree in terms of the top ranked terms.

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: (i) a parallel queue that uses binary tournaments upon a complete binary tree in order to identify the top URL, and (ii) 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. The two structures have different complexity guarantees and also differ in terms of their implementation details. The paper presents an experimental study of the proposed parallel queues using a conventional binary heap queue as the baseline data structure. The results indicate 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 relational database management system. The proposed implementation aims to be both portable and efficient, and it supports several common term weighting schemes. Specifically, it relies on recursive queries in order to implement a portable document parser in SQL, and several optimizations are applied in order to generate efficient SQL queries for preprocessing documents and for processing search queries. The paper presents an experimental study of a prototype implementation that evaluates the potential of the proposed techniques.

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. Thus, a query that constrains exactly the same set of levels is answered directly from a single DHT node, and so the cost involves a logarithmic number of messages. HiPPIS resorts to flooding for queries that do not match the selected levels, which in turn implies a linear number of messages. 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. The paper presents an experimental study of HiPPIS using a modified version of the FreePastry simulator.

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 structured peer-to-peer system. Intuitively, query completeness measures the fraction of answers that have been received compared to the total number of answers. A precise computation of this metric is clearly impractical, and hence the authors propose an approximation as the fraction of peers which have provided results. They then describe methods for tracking this metric efficiently as the query is routed in the overlay network, and also describe the derivation of certain and probabilistic bounds for the returned approximation. The paper presents an evaluation of the proposed schemes in the UniStore system running on PlanetLab.

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 an interesting 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 popular tagging portal del.icio.us. The experimental study reveals that by using a simple and easy to deploy tag extraction method, the Web 1.0 tags generated have a significant overlap with the Web 2.0 tags. In addition, a user study to compare the precision of these two approaches of tagging shows that the quality of the tags produced by these two approaches are comparable which suggests that there is a significant degree of equivalence between social tagging and anchortext annotation. Thus, the simple tag extraction method from anchortext can be applied to either bootstrap tagging portals or enrich the set of tags already present 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. This work can contribute to the future development of mashup standards.

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 (e.g., del.icio.us, StumbleUpon). For a given user, a user model 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. The user model is represented by a three-dimensional matrix of co-occurrence values. Using the user model, a new search query is expanded into multiple queries, and the query results are organized into categories for presentation. The paper presents an overview of a search engine, termed *Nereau*, that is developed based on the proposed ideas. The performance of *Nereau* is evaluated using information from the Open Directory Project.

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 (in the form of comments and tags) from social feedback/tagging services. Based on a user study conducted to ana-

lyze of the suitability of using user feedback from various social feedback/tagging services (del.icio.us, Digg, YouTube, Amazon) for the purpose of Web content summarization, the authors found that their proposed method is most suitable when applied to social bookmarking services. The basic idea of the proposed social summarization method operates as follows: 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. The effectiveness of the proposed approach is demonstrated by an experimental study comparing summaries generated manually versus those generated by the proposed approach using tags and user comments from del.icio.us website.

In the final paper of the session, entitled *Granular Modeling 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 two problems, namely, improving the accuracy of web page classification and improving web page ranking. A web page is modeled as a collection of "visual" blocks that are separated by vertical or horizontal separators, and a hyperlink is modeled as a connection from a block within one document to another document. The approach proposed for the first problem is based on the assumption that visual blocks that contain images, referred to as image blocks, contain more significant information about the web page contents. The authors propose an unsupervised algorithm to identify the most informative image blocks and their most relevant terms using an inverse term importance (ITI) metric. The importance of each image block is measured in terms of its coherence within the containing document, and the weight assigned to a term in a document is computed using an extended TF-IDF approach that takes into account of the importance of image blocks in the document. 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. The paper presents an evaluation of the proposed schemes using 10000 web pages from various popular websites.

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 consists of three main steps. First, the style and mood descriptions of music artists are obtained from the All Music Guide website. Using the collected artist-style and artist-mood information, style and mood similarity taxonomies, respectively, are then computed using a hierarchical co-clustering algorithm. The similarity measure for each of style and mood is then 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. The effectiveness of the proposed artist similarity function is evaluated by comparing against artist similarity based on the acoustic features of their music recordings for a collection of music artists.

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. Since a user typically views only the top few results of a search query, acquiring adequate training data (in ther form of relevance judgements for query-result pairs) to train a ranking function would require 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. The effectiveness of the approach is evaluated using the LETOR benchmark dataset.

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 the all documents based on lexical similarity, and then re-ranks the documents based on the preference vector of the user. The effectiveness of the proposed approach is evaluated against other non-personalized ranking methods using data from the CiteSeer website.

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. In the Rocchio algorithm, documents and queries are modeled as TF-IDF vectors; and for adaptive document filtering, a query profile is formulated involving two parameters representing the relative weights when adding positively and negatively tagged articles to the query profile. A document is classified by the Rocchio algorithm as relevant if its cosine similarity with the query profile is above some threshold value. Typically, the weighing schemes for the Rocchio formulation are tuned statically for a specific query and corpus with the assumption that the tuned parameters are optimized for all users. 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. The recommendation performance of eRocchio is experimentally compared with other well-known classifiers for two recommendation tasks based on interestingness and relevance. In addition, the retrieval performance of eRocchio is also compared against other adaptive filters from TREC11.

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 (e.g., searching for web pages on basketball players). The proposed approach consists of the following key steps: given an input set of sample web pages from several distinct websites about some conceptual entity, the system first crawls these websites to collect more web pages about other instances of the conceptual entity. This is performed using the authors' previous work on the *INDESIT* method which relies on the observation that pages that describe different instances of a conceptual entity within a website share a common template. Based on the collected web pages, the system then automatically extracts a description of the entity, which is essentially represented by a set of keywords.

The system then 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 instances of the conceptual entity are stored and used to recursively trigger the search process. The paper presents an experimental evaluation of the proposed approach using four different conceptual entities.