Report on the Third International Workshop on Web Information and Data Management (WIDM'2001)

Ee-Peng Lim
Center for Advanced Information Systems, School of Computer Engineering
Nanyang Technological University, Nanyang Avenue, Singapore 639798
aseplim@ntu.edu.sg

Roger Chiang Hsiang-Li
College of Business Administration, University of Cincinnati
Cincinnati, OHIO 45221, USA
Roger.Chiang@uc.edu

1 Introduction

The third International Workshop on Web Information and Data Management (WIDM'2001) was held at the Doubletree Hotel, Buckhead in Atlanta on November 9, 2001. The workshop was sponsored by ACM SIGIR, ACM SIGMIS and E-Book Systems. Like it predecessors, WIDM'2001 was held in conjunction with the International Conference on Information and Knowledge Management (CIKM'2001).

The objective of the workshop was to bring together researchers, industrial practitioners, and developers to study how the web information can be extracted, stored, analysed and processed to provide useful knowledge to the end users for various advanced database applications. The workshop received 34 submission of research papers, out of which 11 were accepted.

2 Workshop Overview

In WIDM'2001, a keynote address on **Web Intellgence– Mining the Web for Relevant Information: Concepts and Applications** was given by **Jaideep Srivastava**. In his one hour presentation, the speaker shared with the workshop participants his two-year industrial experience with Amazon, Yodlee, and Chingari. Using a web intelligence system framework, the keynote speaker discussed interesting issues in data acquisition, data extraction, privacy and security.

The rest of the workshop was divided into 4 sessions: **Web Data Mining, Web Information Management, Web Performance Optimization,** and **Web Information Integration.** In the following, we will give an overview of the papers in the workshop.

Web Data Mining

The first paper, **Mining Source Coverage Statistics for Data Integration** by **Nie et al.**, proposed the use of association rule mining technique to derive source coverage statistics that are used to select databases to be queried. With the large number of databases available on the web, the proposed solution can guide a query engine to select a small number of relevant databases to be queried. To obtain the data records for mining, query probes were generated and evaluated on the databases. The authors also described some experimental results demonstrating the feasibility of the proposed solution.

In the paper **Effective Personalization Based on Association Rule Discovery from Web Usage Data** by **Mobasher et al.**, association rule mining was applied to clickstream data to support web personalization. The intention is to allow relevant web pages to be recommended to the users by simply examining their clickstream data. This association-based recommendation approach is different from the collaborative filtering approach which involves comparing user records with the historical records of the other users. Experiments had been conducted to demonstrate the performance of the proposed technique.

Web Information Management

The paper, **Keeping Coherence among Web Sources** by **Arcieri et al.**, discussed the mechanism required to keep common web elements shared by different web pages coherent or synchronized over time.

Chidlovskii in his paper **Automatic Repairing of Web Wrapper**, described the use of classifiers and backward wrappers to help a web wrapper cope with changes to web content, context and structure. This wrapper maintenance technique assumed only small changes occurred to web pages. This work was part of a large Iwrap project at the Xerox Research Centre Europe.

A Performance Evaluation of Storing XML Data in Relational Database Management Systems, written by **Khan and Rao**, described the storage of XML data in relational database. The method assumed that DTDs were available and can be used to determine how the XML data can be stored in relations. Queries on the XML data, represented as XPath queries, were translated into SQL queries and their relational results were used to construct XML documents representing the XPath query results.

Web Performance Optimization

Berfield et al. in the paper **Better Client OFF Time Prediction to Improve Performance in Web Information Systems** presented an interesting work on predicting the idle intervals between user requests. Such prediction will help to determine the appropriate time durations for prefetching web pages. Using methods based on neural networks and genetic algorithms, the paper showed that better client off time can be predicted. It was also pointed out under-predicting the off time is better than over-predicting.

Cluster-based Online Monitoring System of Web Traffic was a paper by **Mao et al.** that described a web measurement system for monitoring the web access traffic. The system captured data packets on the network, analysed them, and re-constructed the corresponding HTTP requests. This system was to overcome the limitation of the current web measurement methods of not being able to effectively analyse the access patterns of individual users. The system had been successfully implemented on an Internet backbone of gigabit bandwith.

Cacheda presented a paper on **Superimposing Codes Representing Hierarchical Information in Web Directories** that applied signature file method to identify files in a hierarchical web directory. The main advantage was to allow queries to quickly locate the set of files in a specific directory to be queried.

Web Information Integration

Rapela's paper **Automatically Combining Ranking Heuristics for HTML Documents** presented an automatic method to determine the weights to be used for merging the ranks of web documents returned by different search systems. The features used in determining the weights include the title fields, meta fields, top n terms in the page, highlighted text, all text, and other link information. By combining the ranks, the proposed method aims to provide more accurate overall ranking of the returned web documents.

Automating the Transformation of XML Documents by **Su et al.** focused on the transformation operations used to convert a XML document with the source DTD to another XML document with the destination DTD. Transformation of XML documents is known to be essential during information exchange between two organizations adopting different DTDs. A cost model was proposed for the transformation operations and used in the generation of efficient sequences of transformation operations.

Eguchi and Sugita, in their paper **Automatically Editing Book Reviews on the Web**, proposed an automatic approach to extract book reviews from the web allowing users to obtain more comprehensive information about a book of interest. The web extraction and crawling techniques were presented.

3 Conclusions

On the whole, WIDM'2001 was successfully held. The slides presented during the workshop have been made available at http://www.cais.ntu.edu.sg:8000/~widm01. The next WIDM workshop will be held together with CIKM'2002 in McLean, Virginia.