COMMIX: Towards Effective Web Information Extraction, Integration and Query Answering*

Tengjiao Wang Shiwei Tang Dongqing Yang Jun Gao Yuqing Wu [+] Jian Pei [?]

Dept. of Computer Sciences & Technology, National Laboratory on Machine Perception

Peking University, Beijing, China

[+] University of Michigan, Ann Arbor, USA [?]Simon Fraser University, Canada

Email: tjwang@db.pku.edu.cn, {tsw, dqyang}@pku.edu.cn,

gaojun@db.pku.edu.cn, yuwu@eecs.umich.edu, peijian@cs.sfu.ca

ABSTRACT

As WWW becomes more and more popular and powerful, how to search information on the web in database way becomes an important research topic. COMMIX, which is developed in the DB group in Peking University (China), is a system towards building very large database using data from the Web for effective information extraction, integration and query answering. COMMIX has some innovative features, such as ontology-based wrapper generation, XML-based information integration, view-based query answering, and QBE-style XML query interface.

1. SYSTEM DESCRIPTION

COMMIX (**C**ontent-**O**riented **M**assive infor**M**ation **I**ntegration based on **X**ML) builds large databases for effective Web information extraction, integration and query answering. The innovative features of COMMIX are highlighted as follows.

Feature 1. COMMIX has an ontology-based, highly adaptive wrapper.

Instead of developing one wrapper per data source, COMMIX has a template database and an ontology. The template database contains a set of templates learnt from an extensive webpage training set, and each template is derived from a concept in the ontology. When extracting a webpage, COMMIX selects the most matchable template from the template database according to the ontology, by analyzing the keywords and structure of the page as well as related hyperlinks.

Feature 2. COMMIX organizes and stores webpages in a database in XML format.

The wrappers in COMMIX extract webpages and store them in an XML database. The extraction phase is based on the ontology, which could be converted into XML DTD in data integration. COMMIX provides a powerful language, an extension to XML-QL, which provides primitives, such as union, join and some aggregate functions. It is capable of integrating information from multiple XML documents in the database.

Feature 3. COMMIX provides a QBE-style XML query interface and builds semantic views for query results.

Formal XML query languages could be too complicated for casual users. Instead, COMMIX introduces a graphic user interface for QBE-style XML queries, and builds semantic views for query results to exhibit the complex data to users friendly.

So far, we have used COMMIX to extract and integrate more than 1.5 million HTML webpages into two databases: one for e-business and the other for research literature. Currently, we are in the progress of extracting the e-business information from all webpages in China (approximate 30 million pages).

2. DEMONSTRATION

The demonstration consists of three parts.

Part 1. System Design and Performance Study

We will illustrate the challenges and strategies of designing a very large database with Web data for information extraction, integration and query answering. Particularly, Some performance data will be shown in tables and figures. The results will help audience to understand the challenges of processing web data.

Part 2. Data Extraction and Integration

We will demonstrate the process of data extraction and integration from webpages using COMMIX. The performance and accuracy of the system will be illustrated. Furthermore, we will analyze the pages extracted incorrectly and explore some issues for future research.

Part 3. Query over Very Large Webpage Databases

We plan to bring the e-business database to the conference. Audiences are encouraged to have a hand-on experience of COMMIX by using the query interface of COMMIX and compare the results obtained from COMMIX with those from publicly available search engines.

*This work is supported by the NKBRSF of China (973) under grant No.G1999032705, the National '863' High-Tech Program of China under grant No. 2001AA114040 and the National Natural Science Foundation of China.