Wrangling Web Data

BY BOB SKIPPER

THE INTERNET OFFERS A VAST AMOUNT OF INFORMATION, MOST OF WHICH IS PRESENTED IN XML (EXTENSIBLE MARKUP LANGUAGE) FORMAT, THE STANDARD FOR WEB PUBLISHING AND DATA EXCHANGE. THE WIDE USE OF XML HAS PROMOTED THE NEED FOR ALGORITHMS AND TOOLS TO EFFICIENTLY MANIPULATE WEB DOCUMENTS.

Dr. Guangming Xing

Dr. Guangming Xing, an assistant professor of computer science at Western Kentucky University, has developed a way to automatically transform this information so that

it can be stored in a database. He is also implementing a system that will more quickly and efficiently classify information contained in Web documents.

"There is a huge amount of information on the Internet. How do we manage this vast amount of web data? That is our main goal," Dr. Xing said. "Before the Web age, most was stored in a relational database. It is very difficult for us to store web data in a relational database because web data are essentially tree-structured."

Tree structures can be of infinite depth — think of what your family tree looks like on paper — but a relational database has fixed schema, he explained. "It's just a table. Putting different trees into a fixed table — now that's a difficult task."

Dr. Xing and his students have implemented an XML database management system and used a relational database as the back end to store data. The system will automatically push a set of XML documents into a relational database. "The storage process is very, very efficient," he said. "We used much less space than other systems, and our queries were much faster in experimental studies. And together with my students, we have published several papers on this."

The system performed well at the XML Mining Challenge at the University of Paris, Xing said.

The transformation of XML documents from one format to another has many applications, such as information filtering (delivering classified information to a relevant party), and web document cleaning (preparing the vast amount of Web data for efficient retrieval). "Given a collection of source and target documents, and the correspondence between the source and target, the question is how to find the rules of the mapping between the source and the target, such that additional source documents can be transformed in target format automatically," Dr. Xing said.

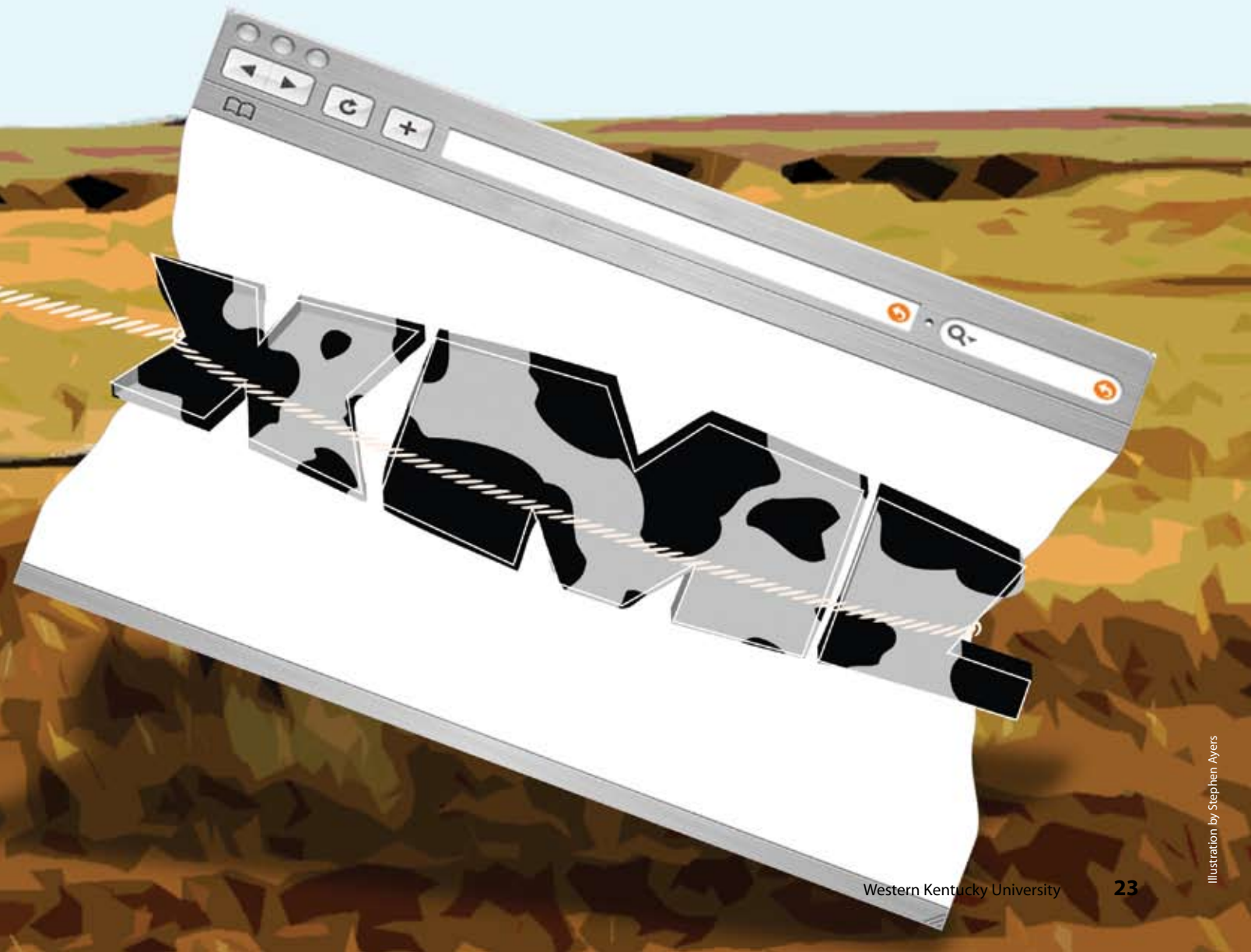

Illustration by Stephen Ayers

"The successful completion of this project will greatly help automate the process of transforming XML documents. It is expected that our methods can also be used in document classification, Web data integration, management of digital libraries, information management for the health-care industry, and many other ways," said Dr. Xing.

"There is a huge amount of information on the Internet. How do we manage this vast amount of web data? That is our main goal."

Dr. Xing and his students have also implemented a system to classify Web documents. "It can be used in a lot of applications," he said. "Whenever you have a large collection of documents, in order to retrieve the useful information or data from these documents, it is important to put them into different books to make the search space much smaller. That is the goal of classification. We only concentrate on the relative information."

He has found an application for his research in the WKU Center for Water Resources in the area of environmental informatics. Dr. Xing and his students have developed two software programs: a legacy document digitalization system and a semantic e-mail system.

"The Water Center has a huge number of legacy documents just on paper," he said. Those documents are scanned and Optical Character Recognition (OCR) software is used to extract the textual information. That information is converted to XML format, and their system is used to store the XML information in a relational database. The system cuts down on the time needed to manually input the information and improves its accuracy, he said.

The semantic e-mail system can automatically process reports from small water utility companies by placing database capability behind the e-mail system. "E-mail today is made for a person to read," he said. "We can use e-mail for other purposes. Right now, the water quality reports are hard copies. If there is a problem, the agency will find it and notify the utility and tell them they need to take some action. That process takes a long time. It really doesn't make a lot of sense."

With the semantic e-mail system, the report is sent via e-mail and is automatically processed, Dr. Xing said. "The information in the e-mail is not just text. It has semantics because it's not just for a human to read; the computer can understand it. Proper actions can be taken," he said.

Dr. Xing has found a use for his system that is even closer to home: "I've been using that as a system to collect homework assignments from my students."

Plans are under way to continue the research. "There's still more work to be done," he said. "Right now we are concentrating on the structure of Web documents because the data can be classified by the structure and also by the topic. Search engines use topic-based searches more often and the structure is used for storage." The next step is researching how to combine the structure information and the topic information together. "That would probably take at least two years."

Dr. Xing's training, including a Ph.D. from the University of Georgia and a B.S. from Nankai University, China, is in theoretical computer science, and he began using his interest in automata theory and implementation to research this area about 2002. He said he began producing results in about eighteen months.

The research has "a strong theoretical connection because it involves trees, which are a mathematical concept, and grammar from theoretical computer science," he said. "We have developed an algorithm that can find the closest distance between a tree and a grammar. We use this new distance to classify how one document is related to another."

Dr. Xing said the algorithm runs in $O(p \times \log p \times n)$ where p is the size of the schema (grammar) and n is the size of the XML document (tree). Experimental studies have also shown that the running time of the algorithm is linear with respect to the size of the XML document when normalized regular hedge grammar is used to specify a schema.

"How to transform an XML document so that it conforms to a schema is not only theoretically interesting, but critical to a lot of applications like document classification, document integration, and information extraction," Dr. Xing wrote in the article "Fast Approximate Matching Between XML Documents and Schemata" (APWeb 2006).

Dr. Xing's research has been partially supported by a grant from the WKU Research Incentive Fund, and by the Kentucky Science and Engineering Foundation Research and Development Excellence fund. ■