World Wide Web (2007) 10:111–112
DOI 10.1007/s11280-007-0023-z

Editorial

Anne H. H. Ngu · Masaru Kitsuregawa · Erich J. Neuhold

Published online: 27 March 2007

The relentless growth in Internet functionality and bandwidth have enabled a new wave of innovations that are transforming the way all types of organizations (businesses, governments, ordinary citizens) collaborate and interact with partners, both within and across organizational boundaries. The Sixth International Conference on Web Information Systems Engineering (WISE 2005) aims at presenting novel topics and approaches to Web engineering in the dynamic, diverse, distributed and ever increasing volume of WWW data and applications. This special issue contains a selection of the best papers of the sixth WISE conference, held in New York City, NY, USA, November 20–22, 2005. The call for papers created a large interest. Thirty full papers were selected from two hundred and fifty nine submissions. Among the full research papers, we invited seven papers to be extended and revised for this special issue. After two rounds of reviews and revisions, four papers were selected for inclusion in this special issue. The first three papers all deal with the issue of data extraction from Web pages. However, they differ in the type of data they deal with. The last paper deals with exciting new Web applications.

The first paper in this issue is titled "Extracting Web Data using Instance-Based Learning" written by Yanhong Zhai and Bing Liu. This paper received the Best Paper award in WISE05 conference. Zhai and Liu proposed an instance-based machine learning technique for extracting structured data from Web pages. The key advantage of their method is that it does not require an initial set of labeled pages to learn extraction rules as in

E. J. Neuhold
University of Vienna, Liebiggasse 4/3-4, A 10080 Vienna, Austria
e-mail: erich.neuhold@univie.ac.at

M. Kitsuregawa
Institute of Industrial Science, University of Tokyo,
4-6-1 Komaba Meguro-ku, Tokyo, Japan
e-mail: kitsure@tkl.iis.u-tokyo.ac.jp

A. H. H. Ngu (✉)
Department of Computer Science, Texas State University-San Marcos,
601 University Drive, San Marcos, TX 78666-4616, USA
e-mail: angu@txstate.edu

existing machine learning-based wrapper induction techniques. Only when a new instance cannot be extracted, then there is a need for labeling. The instance based approach is also very natural because structured data on the Web tend to follow some fixed templates. Pages based on the same template can typically be extracted based on a single instance of the page. The authors proposed a novel technique to match a new instance from the manually labeled instance in order to extract the required data from the new instance. Experimental results show that the technique is efficient and effective based on 1200 Web pages and 24 diverse Web sites.

The second paper is entitled "Towards Deeper Understanding of the Search Interfaces of the Deep Web" by Hai He, Weiyi Meng, Yiyao Lu, Clement Yu and Zonghuan Wu. This paper deals with identification of Web search interfaces which is an important step for the automatic construction of wrappers in integration of Deep Web sources. A Web interface can be considered as containing an interface schema with multiple attributes and rich semantic/meta information; however, the schema is not formally defined in HTML. The authors propose a schema model for representing complex search interface, and then present a layout-expression based approach to automatically extract the logic attributes from search interfaces. Their system, called WISE-iExtractor, has been implemented to automatically construct the interface schema from any Web search interfaces.

The third paper is "Information Extraction from Web Pages using Presentation Regularities and Domain Knowledge" by Srinivas Vadrevu, Fatih Gelgi and Hasan Davulcu. The authors provide a practical solution to the challenging problem of extracting information from unstructured Web data. They proposed an extraction algorithm that is domain independent and that can exploit the presentation regularities of a given Web page and automatically transform it into a weakly annotated semi-structured hierarchical document. Then a statistical domain model is built from the weakly annotated document. This statistical domain model is then used to improve the quality of the annotation and enable more accurate extraction. They demonstrate that such a system can boost the overall accuracy of extraction and recovery from many presentation ambiguities.

The last paper in this issue by Daniel A. Menasce and Vasudeva Akula deals with an interesting and exciting topic of computerized online auctions. The title of the paper is "Improving the Performance of Online Auctions Through Server-side Activity-Based caching". The authors propose server-side caching strategies that leverage the way an online auction operates. More specifically, the auction user behavior is exploited so that the cache management can work more beneficially for the application server. For example, bidding activity on auctions increases considerably after 90% of an auction's life time has elapsed and a very large percentage of auctions have a relatively low number of bids and bidders. The trace-based simulations were used to evaluate their caching strategies and it shows an increase in cache-hit ratios, reduce the auction response time, and reduce I/O activity at the database server.