Printable version (PDF)

Automatic Web Data Extraction for Information Monitoring

Overview

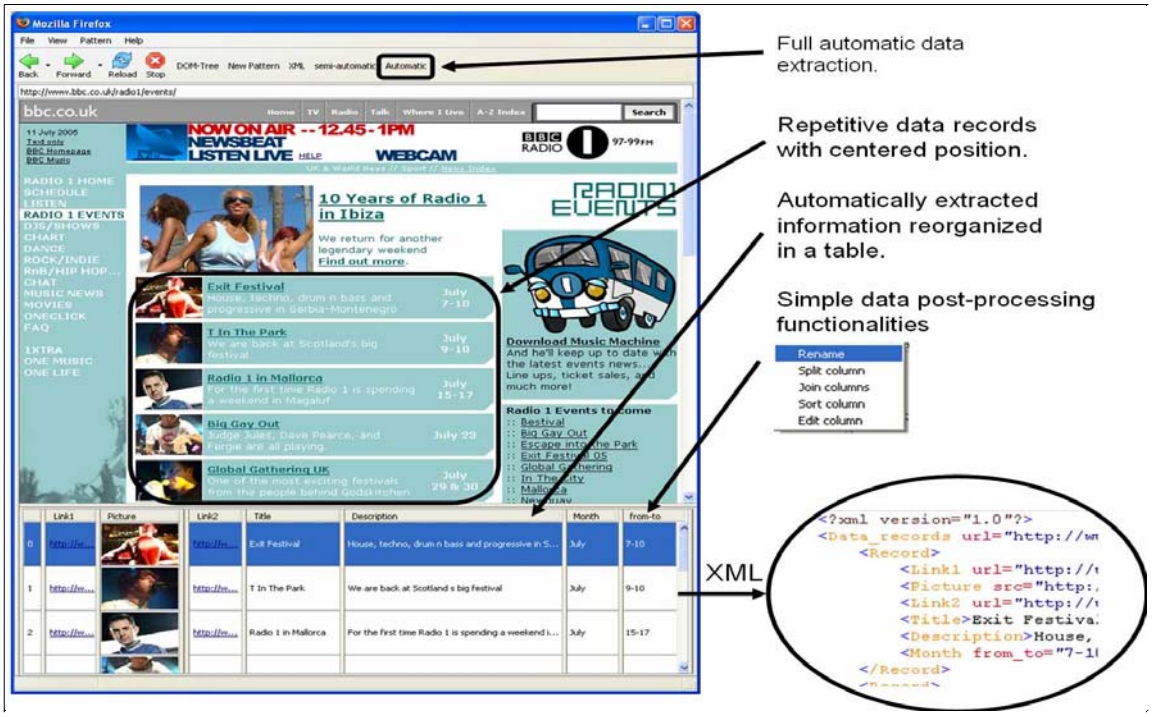
Rife information content available on the World Wide Web is published within representation-oriented semi-structured HTML pages making it difficult for machines to access the content automatically. Therefore tools which are able to unfold the information contained in suchlike resources and transform them into machine-readable and understandable formats are required.

As the vision of the Sematic Web seems to be far away from beeing accomplished, due to the lack of simple mechanisms, we propose a fully automatic wrapper prototype called ViPER (Visual perception-based extraction of records) which is able to extract repetitive structured data records contained in HTML pages with high precision and recall. ViPER tries to identifiy repetitive structures contained in the HTML source code and finally weights and separates the patterns according to the 2D-layout information according to the rendering information of the browser. After the extraction process the pattern with the highest weight becomes aligned in a table, handy to post-process the data. By mapping the data into a structured format, machines are finally able to process the relevant content automatically.

The ViPER system itself has been developed on top of **JREX**, which enables user to access **Mozilla's XPCOM** interface within **Java**. As ViPER is integrated into a meta search engine environment called ASTRO, we are currently working on a plugin-prototype of the system which enables a user to take advantage of the extraction power of ViPER during Web browsing within **Mozilla** or **Firefox**.

The ViPER-plugin for example enables a user to easily generate an agent which monitors a dynamic Web page and sends a notification as soon as the price of an item drops below a certain limit. Where the number of online stores monitored by the agent is not limited.

A sample snapshot of the plugin, demonstrating the automatic extracting functionality applied to a Web site containing repetitive data records, is shown below. In the future we plan to make the plugin freely available for others.



If you like to see further snapshots made during the use of the ViPER plugin take a look at these pictures **1 2 3 4**.

ViPER uses several novel techniques to extract relevant information from dynamic Web Sites. Hereby most of the algorithms for the pattern detection and alignment procedure come from the bioinformatics domain. In order to demonstrate that our approach is able to extract repetitive data records with high precision and recall we compared our system with current state-of-the art systems on several third party data sets with very good results.

	Data Set 2 [1]		Data Set 3 [1]			TBDW Ver. 1.02 [2]	
extraction systems (Asterisk marks results reported in [1])	ViNTs* [1]	ViPER	ViNTs*	MDR* [3]	ViPER	ViNTs	ViPER
#search engines	100		42			51	
#pages per search engine	5	1	1			1	
#**S**earch **R**esult **R**ecords	6905	1390	795			693	
#SRRs extracted	6872	1419	795	479	790	661	686
#SRRs correct	6740	1378	785	420	779	618	676
Recall	97.6%	99.1%	98.7%	52.8%	98.0%	89.2%	97.6%
Precision	98.1%	97.1%	98.7%	87.7%	98.6%	93.5%	98.5%

[1] W. Meng, V. Raghavan, and C. Yu. Fully automatic wrapper generation for search engines. In WWW14, May 2005.

[2] Y. Yamada, N. Craswell, T. Nakatoh, and S. Hirokawa. Testbed for information extraction from deep web. In WWW Alt.'04, pages 346–347, New York, NY, USA, 2004. ACM Press

[3] B. Liu, R. Grossman, and Y. Zhai. Mining data records in web pages. In SIGKDD'03, 2003.

Publications

A scientific paper about ViPER has been submitted for publication to the upcoming CIKM 2005. It describes the ViPER system and its benchmark results in more details. Please feel free to make inquiries in order to obtain most recent, still unpublished work.

Project Members

- **Dipl.-Inf. Kai Simon,**
  PhD student and research fellow, ksimon (at) informatik.uni-freiburg.de.
- **Dr.-Ing. Cai-Nicolas Ziegler,**
  Post-doctoral research fellow, cziegler (at) informatik.uni-freiburg.de.
- **Prof. Dr. Georg Lausen,**
  Head of databases group, lausen (at) informatik.uni-freiburg.de.
- Steffen Kemmerer,
  Studienarbeit, kemmerer (at) informatik.uni-freiburg.de.