Fakultät für Angewandte Wissenschaften

ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

Lehrstuhl
Datenbanken und Informationssysteme

OntoGather: Intelligent search in dynamic data streams

Introduction

Classic search engines (e.g. Google) crawl the Web in intervals and build up indexes from vast amounts of data to be able to answer user queries within a reasonable time frame. This was (and still is) well-suited for static Web pages that remain stable for a longer period of time, but today many resources on the Web are in constant flux. Prominent examples for this are virtual market places or real-time information systems, e.g. stock-exchange price services. The underlying volatile nature of the aforementioned domains necessitates a dynamic approach to support user queries of the form what is the cheapest price for an IXUS digital camera *(at the moment)*?

To solve this problem we propose an approach that relies on dynamic integration of information sources which are accessed at query time. The core of our system is Florid [1], a deductive object-oriented database system based on F-Logic [2], which operates on top of a domain-specific background ontology. It serves as inference engine used for Web resource selection and evaluates a user query on up-to-the-minute information. In this project we focus on formation extracted from HTML pages with the aid of a wrapper tool. Web resources that can be accessed via Web Service interfaces, such as WSDL APIs or RSS feeds have not been considered yet, but can be easily integrated into our system. Wrapper tools range from semi-automatic approaches,such as LIXTO [3], to fully-automatic.

In our scenario a fully-automatic approach is most suitable, because the wrapper generation and maintenance effort is negligible, which suits the requirements of our dynamic world scenario best. Therefore we use our fully-automatic extraction system ViPER [4], that is able to extract up-to-date information from arbitrary HTML pages consisting of data records, which have a similar structure.

System overview

Figure 1 shows the information flow between the abovementioned components of our system that take place while query processing. A detailed description of the different steps can be found in [5].

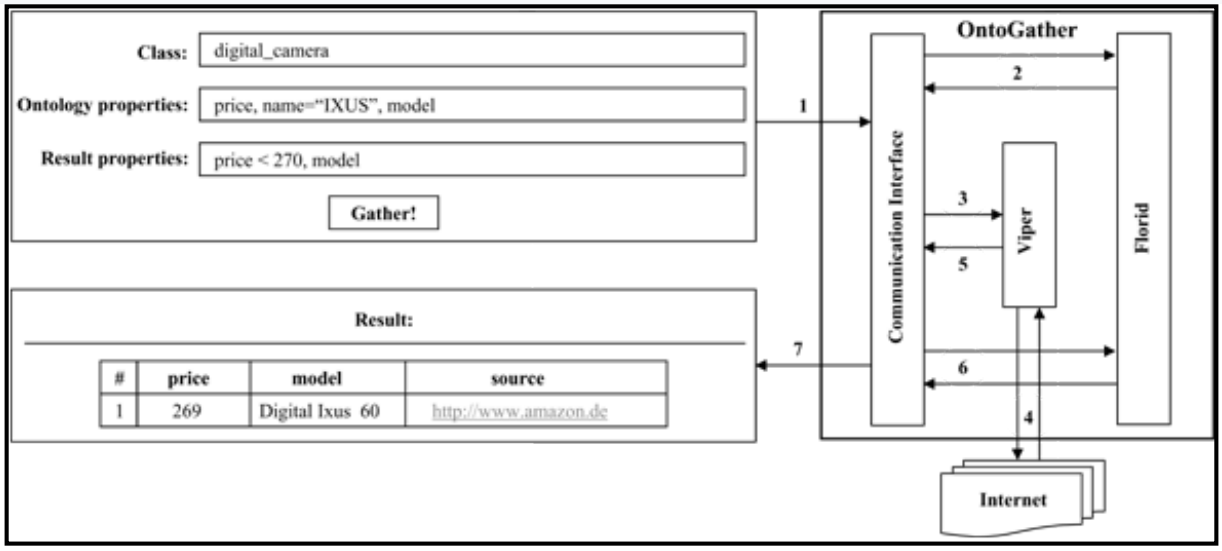

Figure 1: OntoGather system overview

References

The following list identifies our own and related work of interest in this area. Please feel free to make inquiries in order to obtain most recent, still unpublished work.

[1] **Florid: A prototype for F-Logic**; J. Frohn, R. Himmeröder, P.T. Kandzia, G. Lausen, C. Schlepphorst; *Proceedings of the 1997 International Conference on Data Engineering (ICDE '97, Exhibition Program),* April 7-11, 1997, Birmingham, U.K.

Download: **[PS]**

[2] **Logical foundations of object-oriented and framebased languages**; M. Kifer, G. Lausen, J. Wu; *Journal of ACM 1995,* vol. 42, p. 741-843

Download: **[PS]**

[3] Visual web information extraction with Lixto; R. Baumgartner, S. Flesca, G. Gottlob; *Proceedings of the 2001 International Conference on Very Large Data Bases (VLDB '01),* September 11-14, 2001, Rome, Italy

[4] **ViPER: Augmenting Automatic Information Extraction with Visual Perceptions**; K. Simon, G. Lausen; *Proceedings of the 2005 ACM International Conference on Information and Knowledge Management (CIKM '05),* October 31 - November 05, 2005, Bremen, Germany

Download: **[PDF]** **[BibTeX]**

[5] **Information Gathering in a Dynamic World**; T. Hornung, K. Simon, G. Lausen; *Principles and Practice of Semantic Web Reasoning Workshop (PPSWR '06)* June 10-11, 2006, Budva, Montenegro.

Download: **[PDF]** **[BibTeX]**

Project Members

Dipl.-Inf. Thomas Hornung,
PhD student and research fellow, hornungt (at) informatik.uni-freiburg.de.

Dipl.-Inf. Kai Simon,
PhD student and research fellow, ksimon (at) informatik.uni-freiburg.de.

Prof. Dr. Georg Lausen,
Head of databases group, lausen (at) informatik.uni-freiburg.de.