

The Doctoral College "Computational Perception" in collaboration with the WIE (Women in Engineering) Group of IEEE Austria Section and femOVE cordially invite you to the following talk by Prof. Birgit Pröll.

Web Information Extraction

Acquiring Structured Information from Websites

Prof. Birgit Pröll

When? **Tuesday, 23.11.2010**
11:00 - 12:00

Where? **Seminar Room 188/2**
1040 Wien, Favoritenstraße 9-11
4. Floor, Staircase 3

Abstract

Information extraction (IE) is commonly defined as extracting structured data out of unstructured data, as it is provided, e.g., in textual documents. During the last decade IE heavily gained in importance not least to the massive and permanently growing amount of unstructured data, which is available online. There is a wide range of techniques to cope with this challenging task, which is partly based on information retrieval methods and techniques and, due to its addiction to the natural language, subject to linguistic research.

Web information extraction (Web IE) takes as input Web pages instead of local textual documents and addresses the given peculiarities of this domain, e.g., semi-structured data, distributed text sources, and design issues. Techniques range from screen scrapping tools, relying on structural and layout tags of Web pages, to NLP based and machine learning approaches. Even if some general approaches exist, e.g., text engineering frameworks, the majority of application systems are domain dependent, relying on a domain specific vocabulary and grammar.

The current lecture provides an introduction into Web IE on hand of an example and discusses some aspects which we deal with in current projects.

CV

Birgit Pröll studied computer science at the Johannes Kepler University (JKU) Linz, Austria. Since 1991, she has been employed with FAW (Institute for Applied Knowledge Processing) at JKU. She has been engaged in industrial and research projects in the areas of expert systems, Web based-information systems and Web information extraction.

From 1995 to 2000, she had managed the development of the Web-based tourism information systems TIS@WEB and TIScover at FAW.

In 2003, she received her habilitation (venia docendi) for applied computer science from the Faculty of Natural Sciences and Engineering of JKU. Her current research interests and fields of teaching comprise Web information retrieval, Web information extraction, Web engineering, Web sciences, and e-commerce.