Web Content Mining for Information on Information Scientists

Sarah Risse
University of Hildesheim
D-31135, Hildesheim, Germany
sarah_risse@web.de

Abstract

This paper presents a search system for information on scientists which was implemented prototypically for the area of information science, employing Web Content Mining techniques. The sources that are used in the implemented approach are online publication services and personal homepages of scientists. The system contains wrappers for querying the publication services and information extraction from their result pages, as well as methods for information extraction from homepages, which are based on heuristics concerning structure and composition of the pages. Moreover a specialised search technique for searching for personal homepages of information scientists was developed.

1 Introduction

Along with the constant growth of the internet, its importance has continuously increased, while at the same time the problem of information overload has emerged. Finding relevant information in the huge amounts of data to satisfy a precise information need, such as finding information on a currently active scientist, can be quite tedious and time consuming. The area of Web Content Mining comprehends techniques to improve information search and usage on the Web, such as methods for information extraction and integration from web documents and databases, optimisation of search engine results and the development of specialised search engines [Liu and Chang, 2004]. Employing Web Content Mining techniques, a system for searching for information on scientists - homepage URL, email, photo, list of publications and projects - was implemented prototypically for the area of information sciences in German-speaking countries. As resources the publications services DBLP and CiteSeer and the corresponding personal homepage of the target person are used. For the task of finding the personal homepages of information scientists, a specialised search technique was developed.

2 System Overview

As the search term the first and last name of a scientist is entered by the user. This is used to query the publication services DBLP and CiteSeer. From their result pages the publication details are extracted and by comparing the titles of the single publications integrated into one list, thereby eliminating double entries. In the next step a Google-search with the name as search term is performed and the potential homepage of the target person is filtered out of the result list, using heuristics describing typical characteristics of personal homepages in the scientific area. From the homepage the information items academic title, email, photo, list of publications, list of projects and CV are extracted, if present. In the two following sections the methods for information extraction used in the system and the specialised search technique for personal

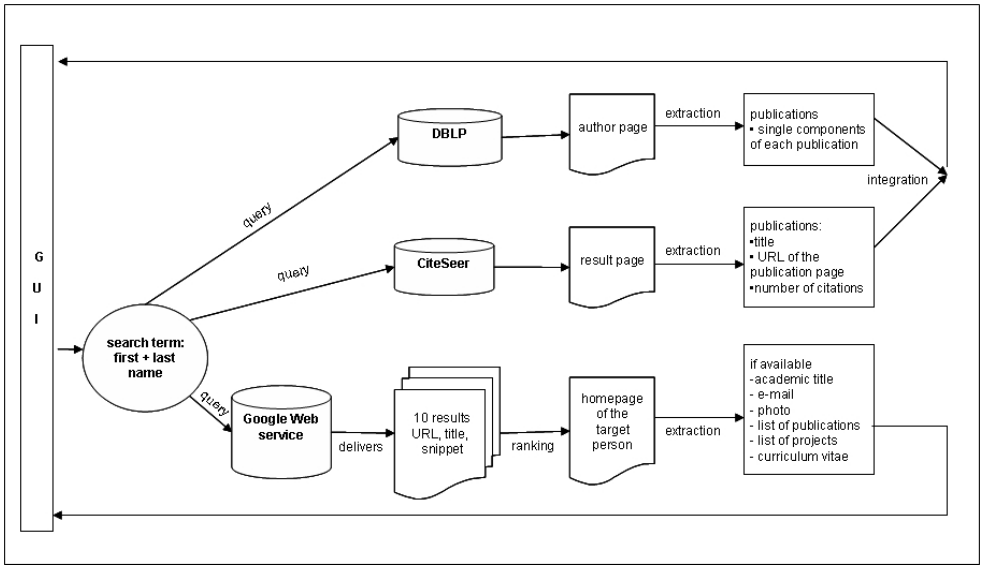

Figure 1: System overview [Risse 2006]

homepages of information scientists are described in more detail.

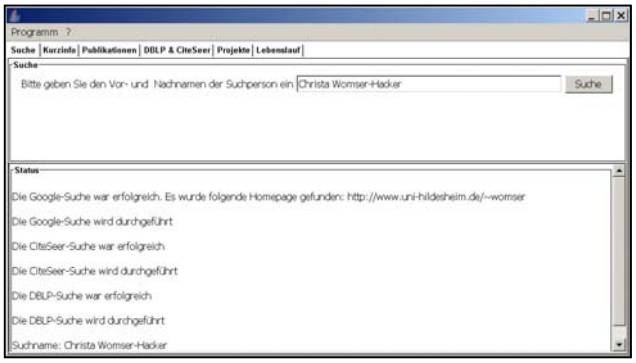

Figure 2: Search page with status information

3 Information Extraction from Source Pages

The manually implemented wrappers for extracting the publication entries from DBLP and CiteSeer make use of the layout and structure of the result pages. Therefore, the relevant HTML-tags are located and the enclosed data extracted.

The methods for locating and extracting the information items from the homepages rely on the observation that the personal homepages of information scientists form a relatively homogeneous set of pages and thus follow certain conventions. These encompass both the structure of the pages and the keywords used for labeling certain content areas. Hence the system assumes that the main page contains the complete name and academic title of the target person, the contact details including the email and a photo of the scientist. Further information items such as a list of publications and research projects as well as a CV are either listed sequentially on a page - separated with subheadings - or located on subpages, that are linked from the main page.

Thus, the academic title, e-mail and photo can be extracted using simple string comparisons and regular expressions - in case of the photo heuristics concerning the metadata provided with the picture are used. The other information items are extracted by using lists of keywords typically used to label these areas, e.g. publications, projetcs, research, CV, curriculum vitae, and the assumption that all subheadings or links to subpages are formatted identically. In the case of a sequentially constructed homepage the extraction rule identifies in this manner all subheadings and extracts the data between two subheadings. In case of a subpage, the rule extracts all data following the respective keyword on that page.

The extracted information items are finally displayed in an integrated way on the result pages of the search system, as figure 3 shows.

4 Specialised Search for Personal Homepages

The developed search technique uses typical characteristics of information scientists' homepages to filter the Google-search results for the homepage of the target person. Following the approach used in HPSearch[1], a specialised search engine for homepages of computer scientists described in [Hoff 2002], URL, title and snippet of information scientists' homepages were analysed. Thus, typical features such as the occurrence of the name of the target person, an academic title or the use of terms as information science identified. These features were transferred into a weighting function, that filters the Google-results, after eliminating e.g. pages from publication services or older pages by comparing the academic titles mentioned.

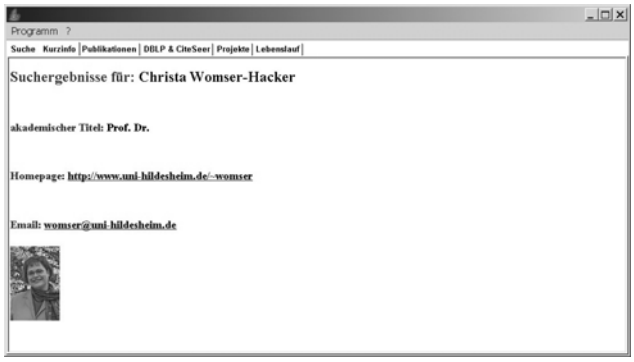

Figure 3: Result page for "Christa Womser-Hacker"

5 Outlook

A first evaluation proved with a recall of 0.65 and a precision of 0.71 for the homepage search module, that the implemented approach is promising. The heuristics used in the weighting function need to be analysed regarding their relevance and their coverage. A ranking approach and a flexible query expansion with frequent terms form the publication list are other potential ways of improvement.

The wrappers for the DBLP and CiteSeer work almost without errors, but a special care needs to be taken for the problem of name disambiguation.

The heuristics applied in the rules for extracting the information items from the personal homepages need to be extended as the evaluation showed that they are not sufficient. Moreover an algorithm to identify content parts in homepages could avoid the extraction of irrelevant data along with relevant information.

References

[Hoff, 2002] Gerd Hoff. Ein Verfahren zur thematisch spezialisierten Suche im Web und seine Realisierung im Prototypen HomePageSearch. *Fachbereich IV der Universität Trier. 2002.*

[Liu and Chang, 2004] Bing Liu and Kevon Cheng-Chuan Chang. Editorial: Special Issue on Web Content Mining. In: *SIGKDD Exploratins Vol. 6(2).* pp 1-4.

[Risse, 2006] Web Content Mining nach Informationen zur wissenschaftlich tätigen Personen im Umfeld der Informationswissenschaft. *Magisterarbeit Universität Hildesheim.*
http://web1.bib.uni-hildesheim.de/edocs/2006/514703415/meta

System URLs:
CiteSeer: http://citeseer.ist.psu.edu
DBLP: http://www.informatik.uni-trier.de/~ley/db

1 The system is available under http://hpsearch.uni-trier.de, but it is not updated anymore.