Searching People on the Web According to Their Interests

Bing Liu and Chee Wee Chin

School of Computing
National University of Singapore
3 Science Drive 2, Singapore 117543
{liub, chinchee}@comp.nus.edu.sg

ABSTRACT

Due to lack of structural data and information explosion on the World Wide Web (WWW), searching for useful information is becoming increasingly a difficult task. Traditional search engines on the Web render some form of assistance, but perform sub-optimally when dealing with context sensitive queries. To overcome this, niche search engines serving specific Web communities evolved. These engines index only pages of high quality and relevance to a specific domain and make use of context information for searching. This poster presents BullsI Search (publicly available at http://dm2.comp.nus.edu.sg), a fielded domain specific search engine that helps Web users locate computer science faculty members' homepages and email addresses by specifying a research interest, teaching interest or name. The system is able to automatically extract and index research interests, teaching interests, owners' email addresses and names from a set of discovered homepages. Experimental evaluation shows that the proposed algorithms are very accurate and are independent of structures in different homepages.

Keywords: Domain specific search engine, hypertext classification, information extraction, Web mining.

1. INTRODUCTION

As we march into the digital information age, data overload on the Web becomes an eminent problem. Currently, there are billions of Web pages distributed over millions of sites and the actual numbers become archaic everyday. Although today's generic search engines are able to tackle this crisis reasonably well, problem surfaces when handling queries with answers buried within the semantics of Web pages. Generic search engines use vacuum cleaning crawlers, which suck in all documents that come in their way. Their indexing operation works only on raw text, without trying to assign any meanings to them. This approach works pretty well for generic queries. However, its relevancy suffers when searching using terms that are context sensitive (see [2]).

In this poster, we present a fielded domain specific search engine, called BullsI Search. The aim of this system is to assist computer science professors, researchers and others in searching for homepages and email addresses of people by specifying research or teaching interests as the query terms. In building of this system, information such as research and teaching interests, names and email addresses are extracted from a set of discovered computer science faculty members' homepages. At present, we have homepage information of approximately 25,000 faculty members across 664 Computer Science Departments situated in USA, United Kingdom, Canada, Australia, New Zealand, Israel, Hong Kong and Singapore

Specifically, our system's targeted audience includes, but is not limited to, people in the following categories:

- People active in computer science research or industrial applications. People in this category often need to look for information in homepages of those with related research interests.
- Faculty members in universities or other institutions teaching computer science subjects. University professors and lecturers often need a great deal of information when preparing course materials (*e.g.* when drafting lecture slides and assignments).
- Companies in commercial sector. Commercial companies can use the system to assist them in targeted marketing. For example, a book publishing company that is about to launch a new book on Artificial Intelligence (AI) can use the system to search for people working or teaching in the field of AI.

The ultimate goal of this project is to discover and crawl all personal homepages on the Web and extract personal interests, names and email addresses of their owners to provide a people-oriented search engine, i.e., an homepage search engine that allows searching by professional and personal interests, names, or email addresses. Clearly, such a service is very useful for targeted marketing based on the interests of the people. It is also important for customer profiling and segmentation. Our current system provides a strong evidence indicating that such a system is feasible.

2. SYSTEM ARCHITECTURE

The overall system architecture of our system is shown in figure 1. Each hexagon represents a module while the block arrows depict Web pages flowing between modules.

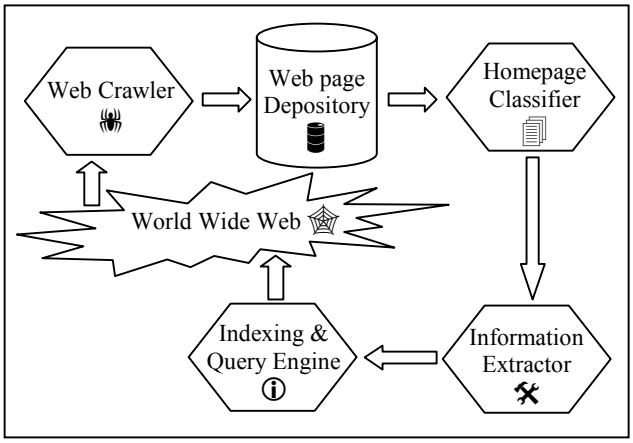

Figure 1: Overall system architecture

The Web crawler essentially crawls the pages from Computer Science departments' faculty listings, which were obtained from Google Web Directory [6] and Yahoo! Web Directory [7]. Next, we need to decide whether a page is a faulty personal homepage. For this, two methods are employed. The first method checks if the URL contains a '?' character. A '?' character in the URL indicates that it is a dynamic Web page. In the second method, HTML tags are first removed from the Web page. A number of other heuristics were also incorporated to filter out links that are not homepages. Details can be found in [1]. Below, we briefly discuss information extraction from homepages, indexing and querying.

2.1 Information Extraction

Research and teaching interests: We employed several techniques to extract research and teaching interests. First, a segmentation technique splits a homepage into research or teaching segments using presentation styles and keyword features. Second, we make use of hyperlink anchor information to detect hyperlinks leading to research or teaching page. Third, we apply some linguistic heuristics for the extraction task (*e.g.* My *research* interest includes… I am *teaching*…). Lastly, pattern matching is used to detect course codes (*e.g.* CS4240 COM5225) about teaching.

Email addresses and names: A different methodology is applied to extract email addresses and names of professors. It is based on the following observations:

1. A faculty member's email address is usually very similar to his/her homepage URL.
2. The email address userid is usually very similar to his/her name.

To determine string similarity, we used a normalized version of Levenshtein Distance [4]. Levenshtein Distance computes the differences between two strings, where we would count a difference not only when strings have different characters but also when one has a character whereas the other does not. The '@' character is used to detect all email addresses occurring within a homepage. The email address that is most similar to the homepage URL is extracted, while the rest are discharged.

To extract the name of a professor, we rely on the userid of the email address extracted. We first generate a list of candidate names by extracting sequences of initially capitalized words (since initial capitalization is a good indication of names). Again, the normalized Levenshtein Distance metric is used to detect the most probable candidate names (most similar to the userid of the email address).

2.2 Indexing and Querying

After the extraction process, the final task is to index the extracted information and to set up the search engine on the Web. A public domain text retrieval system, MG [5], is used to index the extracted information. Figures 2 and 3 show the query interface and a partial search result of people teaching machine learning courses at Carnegie Mellon University (CMU) respectively. A user begins by entering a search term and selecting a search category. He/she then chooses the country followed by the university or institution of interest.

Figure 2: Query interface of BullsI Search

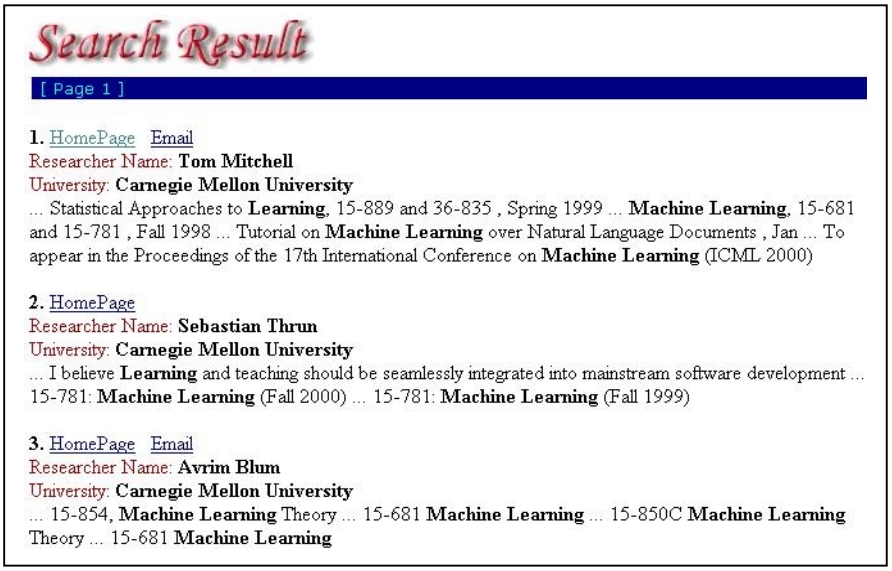

Figure 3: Partial search results of people teaching Machine Learning at CMU

3. PRELIMINARY RESULTS

Each of the information extraction tasks is evaluated using precision and recall [3]. We tagged a homepage as relevant to each of the extraction task if it contains the respective information (name, email address, research, and teaching) in textual forms. The results are shown in Table 1. Column 1 shows ten randomly selected universities that are sampled. Column 2 displays the size (the number of faculty members) of each department. The rest of the columns display the precision and recall of each extraction task respectively. The final row shows the average results. We observe that our extraction techniques are very effective and accurate.

University	Size	Email Address Extraction		Name Extraction		Research Interest		Teaching Interest	
		Prec.	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*	*Prec.*	*Recall*
Yale University (USA)	24	100.00	91.67	100.00	100.00	100.00	91.30	81.82	100.00
Massachusetts Inst. of Tech. (USA)	80	95.52	91.43	98.70	97.44	98.55	94.44	97.62	80.39
Univeristy. Of Waterloo (Canada)	78	100.00	92.96	94.87	94.87	100.00	97.10	98.15	98.15
Uni. of British Columbia (Canada)	40	100.00	100.00	100.00	100.00	100.00	100.00	96.30	96.30
Brunel University (UK)	44	100.00	100.00	93.02	93.02	100.00	96.15	87.50	77.78
Kings College, London (UK)	25	100.00	95.83	100.00	100.00	95.24	95.24	84.62	84.62
Uni. of New South Wales (AU)	77	98.48	95.59	95.89	95.89	94.12	100.00	100.00	75.51
Deakin University (AU)	34	100.00	100.00	96.97	96.97	62.50	90.91	90.91	83.33
Hong Kong Uni. of Sci. & Tech.	43	97.44	97.44	97.67	97.67	100.00	97.14	78.26	69.23
National. Uni. of Singapore	99	98.86	98.86	97.98	97.98	100.00	94.05	92.96	89.19
Average	54	99.03	96.38	97.51	97.38	95.04	95.63	90.81	85.45

Table 1: Results of various information extraction tasks
(All precision and recall are shown in percentage)

4. CONCLUSION AND FUTURE WORK

In this poster, we presented a homepage-oriented search system, called BullsI Search, which enables users to search for homepages satisfying a specified query. In particular, it allows users to search for computer science faculty members' homepages and email addresses by specifying a research interest, a teaching interest or a name. Currently, the system encompasses homepages of faculty members of major computer science departments in 8 countries. The system is not only useful for people active in academic research, but also to companies in the commercial sector. In our future work, we plan to discover all personal homepages on the Web and extract useful information from them to expand the services of BullsI Search. We will also investigate the possibility of building targeted marketing tools based on the framework and techniques presented.

REFERENCES

[1] Bing L. & Chin C.W. “BullsI Search: Searching People on the Web by Sniffing Their Interests” In Technical Report 2001.

[2] Lawrence S. "Context in Web Search", *IEEE Data Engineering Bulletin Vol 23 (3) pp. 25-32*, 2000.

[3] Lehnert W. & Sundheim B. "A Performance Evaluation of Text Analysis Technologies" *AI Magazine pp. 81-94*, 1991.

[4] Levenshtein V. I. "Binary Codes Capable of Correcting Spurious Insertions and Deletions of Ones (Orginal in Russian)", *Russian Problemy Peredachi Informatsii, 1:12-25*, 1965.

[5] Witten I. H., Alistair M. & Bell T. "Managing Gigabytes: Compression and Indexing Documents and Images", *Morgan Kaufmann Publisher*, 1999.

[6] Google Web Directory: http://directory.google.com/Top/Computers/Computer_Science/Academic_Departments/

[7] Yahoo Web Directory: http://dir.yahoo.com/Science/Computer_Science/College_and_University_Departments/