A Web Browsing System for Retrieving Scholarly Pages

Ari Pirkola
School of Information Sciences
University of Tampere
Finland
ari.pirkola@uta.fi

***Abstract*—The major Web search engines are useful tools to find information from the Web but their commercial nature, coupled with some other factors, makes it difficult to find scholarly pages on a specific topic. Many users prefer browsing to searching because it is easier. Even though browsing subject directories is an important method to retrieve information from the Web, such directories are not suitable for specific scientific retrieval tasks. In this paper, we present a Web browsing system focused on scholarly pages related to a specific topic. The first implementation is dedicated to the topic climate change, but the method to construct the system is a general method that can be applied to any reasonable topic. The climate change browsing system provides access through links to the thematic Web pages of scientific organizations engaged in climate change research, as well as to the pages of organizations that are linked to them. Each link in the system is categorized under a given index term based on the occurrences of phrases related to climate change (keyphrases) on the target pages of the links. In browsing, the user clicks the desired index term and the system returns a list of links to the pages associated with the index term. This paper also presents the crawler used to fetch pages for the browsing system and a keyphrase dictionary used in indexing the pages included in the system.**

Keywords - browsing; climate change; focused crawling; information retrieval; scientific organizations

I. INTRODUCTION

The Web contains tens or even hundreds of billions of documents (pages). Access to its huge content is dominated by commercial search engines, and a concern has been raised about the commercially biased search results. Common experience shows that the pages of companies and other commercial organizations often populate the top ranks of the search results while, for example, many highly informative pages of scientific organizations do not appear in the results. Obviously, this bias is caused by the commercial nature of the search engines and specifically by the search engine optimization (SEO), which means that Web page publishers can promote the ranking of their Web pages in the search results. Commercial organizations are willing and have resources to invest in SEO whereas in the scientific organizations such practice seems not to be common. It has also been discussed where to draw a line between ethical and unethical SEO. The above facts suggest that there is need for such retrieval systems that satisfy the information needs of users who search for scholarly information rather than products or online shops.

The major Web search engines are query-based retrieval systems. Querying is an effective method when the information need can be expressed using a relatively simple query and the search term is clear, e.g., when the searcher looks for information on the disease whose name he or she knows. However, it is common that the searchers do not know or remember the appropriate search terms. Complex information needs are also difficult to formulate as a query. In situations such as complex work tasks the information need itself may be vague and ill defined [1]. Browsing subject directories (e.g., Open Access Directory) is an alternative way to retrieve information from the Web, and it is preferred by many users because it is easier, requiring less mental work than formulating a query. Such directories are created manually and usually in (more or less) unsystematic fashion. Their target pages may be both commercial and informative pages. Even though the subject directories are helpful tools, these features make them unsuitable for specific scientific retrieval tasks. Automatically constructed and systematically organized browsing systems focused on scholarly pages are missing from the Web. In this paper, we present a method to construct such a system.

Using the proposed method we implemented a browsing system (a pilot system) for the topic *climate change*. We present the existing climate change browsing system in this paper, but the proposed method is a general method and can be applied to any reasonable topic. The climate change browsing system contains links to the Web pages of universities, research organizations, and their units (e.g., departments, centers, and institutes) investigating climate change, as well as to the pages of organizations that are linked to them. The system provides information that is difficult or impossible to obtain by the major Web search engines and not included in scientific publications, such as information on current and completed research projects, observational field data and new, not (yet) published research findings. The system allows exploring systematically different aspects of climate change research. In goal-oriented browsing the system provides answers to the questions such as: What organizations conduct research in the field of climate change and what is the specific research area of each organization. Where can I find information related to a specific research area of climate change research? There are also numerous more specific questions where the system may be helpful.

The rest of this paper is organized as follows. Section II presents the main features of the climate change browsing system. The method to construct the browsing system is presented in Section III. Section IV describes the tools needed to construct the system: a focused Web crawler and a keyphrase dictionary of climate change. Section V contains the conclusions.

II. CLIMATE CHANGE BROWSING SYSTEM

In this section, we present the main features of the climate change browsing system. The system provides access to the *thematic pages* of universities, research organizations, and their units, as well as to the thematic pages of organizations that are linked to them (e.g., online journals). Non-thematic pages, such as *about us* and *contact* are pruned out by applying a stop-word list for the URLs of the pages (Section III). We differentiate between two types of thematic pages: *project pages* that describe the research areas of the organizations, or present ongoing or completed research projects, or include some information related to the research projects, and *findings pages* that describe recent research findings (from the present back a few years).

The following simple example illustrates the structure of the browsing system:

melting glaciers 11.7
↓ ↓
Link to page A →
Link to page B → project pages A-C
Link to page C →

Link to page D →
Link to page E → findings pages D-F
Link to page F →

Page titles serve as links, and they are categorized under index terms, such as *melting glaciers*, *climate models*, or *sea level rise* based on the occurrences of the phrases related to climate change (keyphrases) on the pages.

The index terms describe the research areas of the organizations in the field of climate change research, and their source is the climate change keyphrase dictionary (Section IV B). Each index term has an importance score, which reflects the significance of the term in the context of climate change research. As shown in the example above, the index term *melting glaciers* has an importance score of 11.7. The first set of the links (A-C in the example) point to the project pages while the second set (D-F) point to the findings pages. In browsing, the user clicks the desired index term and the system returns the second page through which the user can access the pages discussing the issue represented by the index term.

The first implementation of the browsing system (the pilot system) will be published at [2] in the spring 2012. This site also contains a climate change search system and the keyphrase dictionary, both of which were developed in our earlier project. The search system is described in [3] and the keyphrase identification and extraction method in [4].

III. METHOD

Fig. 1 depicts the crawling and page processing stages involved in constructing the pilot browsing system. In the first stage, relevant pages were crawled for the search system from the scientific Web sites using a focused crawler developed in our earlier project. The crawler is described in Section IV A. Focused crawlers are programs aiming to fetch Web pages that are relevant to a pre-defined domain or topic [5, 6, 7]. The crawled pages were indexed using the Apache Lucene programming library (http://lucene.apache.org/). The constructed search system covers 95 819 (public version 73 194) Web pages.

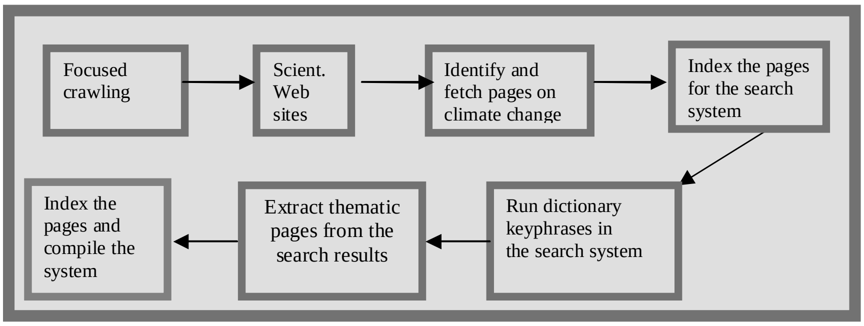
Focused crawling
Scient. Web sites
Identify and fetch pages on climate change
Index the pages for the search system
Run dictionary keyphrases in the search system
Extract thematic pages from the search results
Index the pages and compile the system

Figure 1. The crawling and page processing when constructing the browsing system.

A set of keyphrases contained in the keyphrase dictionary were run as queries in the search system. The purpose of this stage was to retrieve pages dealing the issues represented by the keyphrases. For findings pages, the keyphrases were required to appear in the titles or URLs of pages. In other words, if the title or URL contains the keyphrase, the page is regarded as a findings page. These restrictions were not applied for project pages because their titles and URLs may only be loosely descriptive. If the keyphrase appears in the title or URL we can be highly confident that the page deals with the issue represented by the keyphrase. A drawback of this method is that it excludes part of relevant documents. However, it includes highly relevant documents, which usually are the most important for the user.

The search results were processed to separate the thematic pages from non-thematic pages and to separate the two types of the thematic pages from each other. The identification of the thematic pages is based on the fact that different organizations apply descriptive and similar names for similar items in URLs. Specifically, the URLs of the project pages typically contain the word *research*, *project*, or *science*, e.g., *research-highlights*, *current_projects*, or *research.php*. The URLs of the finding pages in turn often contain the word *news*, sometimes *press* (e.g., *press_release*). We also applied a stop-list of some 50 words indicating non-thematic content (such as *student*, *course*, *contact*, and *grant*) for the URLs to exclude such pages as *www.university.edu/research_grants*.

In the next stage in constructing the browsing system, the URLs of the thematic pages were extracted from the search results, and each page was indexed under the query keyphrase that returned the page (now the keyphrases are called index terms). Finally, the browsing system was compiled by creating appropriate links.

In the keyphrase dictionary, different variant forms of the phrase are grouped together into the same synonym set (Section IV B). Synonymy was taken into account, so that pages containing synonymous keyphrases are under the same index term in the browsing system (i.e., the keyphrase with the highest IS among the synonyms).

IV. THE CRAWLER AND THE KEYPHRASE DICTIONARY

This section presents the crawler used to collect pages for the search system and for source data for the climate change keyphrase dictionary, and describes the dictionary.

A. *Crawler*

We developed a focused crawler that is used to fetch pages dealing with climate change and that can be easily tuned to fetch pages on other topics. The crawler determines the relevance of the pages during crawling by matching a topic-defining query against the retrieved pages using a search engine. It uses the Lemur search engine (http://www.lemurproject.org/) for this purpose. The pages on climate change were crawled using the following search terms in the topic-defining query: *climate change*, *global warming*, *climatic change*, *research*. We used the core journals in the field to find relevant start URLs. When pages on some other topic are fetched only the search terms and the start URL set need to be changed. So, applying the crawler to a new topic is easy.

To ensure that the crawler fetches mainly scholarly documents its crawling scope is limited, so that it is only allowed to visit the pages on the start URL sites and their subdomains (for example, research.university.edu is a subdomain of www.university.edu), as well as sites that are one link apart from the start domain. If needed, this restriction can be relaxed so that the crawling scope is not limited to these sites.

A focused crawler does not follow all links on a page but it will assess which links to follow to find relevant pages. Our crawler assigns the probability of relevance to an unseen page v using the following formula, which gave the best results in an experiment where we compared four different methods [3]:

$$Pr(T|v) = (\alpha * rel(u) * (1/\log(Nu))) + ((1 - \alpha) * rel(<u,v>)), \alpha = 0.3$$

where Pr(T|v) is the probability of relevance of the unseen page v to the topic T, α is a weighting parameter ($0 < \alpha < 1$), rel(u) is the relevance of the seen page u, calculated by Lemur, N_u the number of links on page u, and rel(<u,v>) the relevance of the link between u and the unseen page v. The relevance of the link is calculated by matching the context of the link against the topic query. The context is the anchor text, and the text immediately surrounding the anchor. The context is defined with the help of the Document Object Model (DOM): all text that is within five DOM tree nodes of the link node is considered belonging to the context. The Document Object Model is a convention for representing and interacting with objects in HTML, XHTML and XML documents (http://en.wikipedia.org/wiki/Document_Object_Model).

As can be seen, Pr(T|v) is a sum that consists of two terms: one that depends on the relevance of the page, and one that depends on the relevance of the link. The relative importance of the two terms is determined by the weight α. Based on our crawling experiment we apply for the α parameter the value of $\alpha = 0.3$. Also, the number of links on page u inversely influences the probability. If rel(u) is high, we can think that the page "recommends" page v. However, if the page also recommends lots of other pages (i.e., N_u is high), we can rely less on the recommendation.

B. *Keyphrase Dictionary*

To be able to systematically explore a given scientific topic, a necessary requirement is to have a terminology assistance (e.g., ontology, dictionary, thesaurus) that is used to divide the topic into meaningful subtopics or concepts. We constructed the keyphrase dictionary of climate change, which is used to index the organizations for the browsing system. Originally, the dictionary was developed for use as a search assistance to support query formulation in Web searching, but it can be used in document indexing as well. The dictionary contains some 5 500 phrases related to climate change. Most of the phrases represent different aspects or research areas of the climate change research, some are more technical in nature. The phrases were extracted from the Web pages of scientific organizations discussing climate change issues fetched by the crawler described in Section IV A.

Each phrase is assigned a frequency-based *importance score*, which reflects the significance of the phrase in the context of climate change research. Different variant forms of the same phrase, such as *sea level rise*, *sea level rising*, and *rising sea level*, are grouped together into the same entry (synonym set) using approximate string matching.

When devising the dictionary the first challenge was to determine which sequences of words are phrases in the crawled pages. Here we applied the phrase identification method by Jaene and Seelbach [8]. The main point of the technique is that a sequence of two or more consecutive words constitutes a phrase if it is surrounded by small words (such as *the*, *on*, *if*) but do not include a small word (except for *of*).

The other main challenge besides phrase identification was to develop a method to identify the *keyphrases* among all phrases in the relevant pages and prune out out-of-topic phrases. To address this problem, we calculated the importance scores (ISs) for phrases as described below. The most obvious out-of-topic phrases receive a low score and are not accepted in the dictionary. The remaining phrases are regarded as keyphrases and are included in the dictionary.

The IS is calculated on the basis of the frequencies of the phrases in the corpora of various densities of relevant text, and in a non-relevant corpus. We determined the IS using four different corpora. The relevant corpora are built on the basis of the occurrences of the topic title phrase (i.e., climate change) and a few known keyphrases related to climate change in the original corpus crawled from the Web. Assumedly, a phrase which has a high frequency in the relevant corpora and a low frequency in the non-relevant corpus deserves a high score. Therefore, the importance score is calculated as follows:

$$IS(P_i) = \ln(F_{DC(1)}(P_i) * F_{DC(2)}(P_i) * F_{DC(3)}(P_i) / F_{DC(4)}(P_i)); \quad (F_{DC} > 0)$$

$F_{DC(1)}(P_i)$... $F_{DC(4)}(P_i)$ = the frequencies of the phrase P_i in the four corpora.
DC(1) = Highly dense corpus
DC(2) = Very dense corpus
DC(3) = Dense corpus
DC(4) = non-relevant corpus

The presented method allows us to indicate the importance of the phrase in the Web texts discussing the topic in question, and separate between the keyphrases and out-of-topic phrases based on the fact that the relative frequencies of keyphrases decrease as the density decreases.

Synonyms were identified using the digram approximate matching technique. Phrases were first decomposed into digrams, i.e., substrings of two adjacent characters in the phrase (for n-gram matching see [9]). The digrams of the phrase were matched against the digrams of the other phrases in the list of the relevant phrases generated in the phrase identification phase. Similarity between phrases was computed using the Dice formula, and the phrase pairs that had the similarity value higher than the threshold (SIM=0.75) were regarded as synonyms.

Table 1 presents three example entries in the dictionary. The dictionary is organized alphabetically, and each phrase acts as a head phrase in its turn.

V. CONCLUSIONS

In this study, we developed a method to construct a scholarly Web browsing system. The system allows a systematic exploration of a particular scientific topic through browsing the pages of the scientific organizations, and is also a helpful tool in goal-oriented browsing.

Our plan is to extend the existing (but not yet published) pilot browsing system and construct similar systems for new topics. We also want to evaluate the browsing system in a user study. Naturally, many important pages cannot be included in the browsing system if pages are only collected using a crawler. We therefore also plan to implement an upload option for the system, so that the users can suggest and upload URLs to be added to the system.

TABLE I. THREE EXAMPLE ENTRIES IN THE KEYPHRASE DICTIONARY

Head phrase	IS	Synonyms	IS
permafrost thaw	7.3	thawing permafrost	7.9
satellite observation	6.3	satellite observations	11.0
greenhouse gas	25.3	green house gases	7.8
		greenhouse gases	23.9
		greenhouse gasses	13.7
		greenhouses gases	8.1

ACKNOWLEDGMENT

This study was funded by the Academy of Finland (research projects 130760, 218289).

REFERENCES

[1] P. Ingwersen and K. Järvelin. The Turn: Integration of Information Seeking and Retrieval in Context. Heidelberg: Springer, 2005.

[2] http://searchclimatechange.com/

[3] A. Pirkola. "A Web search system focused on climate change." Digital Proceedings, Earth Observation of Global Changes (EOGC). Munich, Germany, April 13-15, 2011.

[4] A. Pirkola. "Constructing topic-specific search keyphrase suggestion tools for Web information retrieval." Proc. of the 12th International Symposium on Information Science (ISI 2011). Hildesheim, Germany, March 9-11, 2011, pp. 172-183.

[5] T. Talvensaari, A. Pirkola, K. Järvelin, M. Juhola, and J. Laurikkala. "Focused Web crawling in the acquisition of comparable corpora." Information Retrieval, 11(5), 2008, pp. 427-445.

[6] S. Chakrabarti, M. van den Berg, and B. Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." Proc. of the Eighth International World Wide Web Conference. Toronto, Canada, May 11-14, 1999.

[7] T. Tang, D. Hawking, N. Craswell, and K. Griffiths. "Focused crawling for both topical relevance and quality of medical information." Proc. of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05). Bremen, Germany, October 31-November 5, 2005, pp. 147-154.

[8] H. Jaene and D. Seelbach. Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten. Report ZMD-A-29. Beuth Verlag, Berlin, 1975.

[9] A.M. Robertson and P. Willett. "Applications of n-grams in textual information systems." Journal of Documentation, 54(1), 1998, pp. 48-69.