

Focused Crawling: A Means to Acquire Biological Data from the Web

Ari Pirkola
Department of Information Studies
University of Tampere
Finland
pirkola@cc.jyu.fi

ABSTRACT

Experience paper. World Wide Web contains billions of publicly available documents (pages) and it grows and changes rapidly. Web *search engines*, such as Google and Altavista, provide access to indexable Web documents. An important part of a search engine is a *Web crawler* whose function is to collect Web pages for the search engine. Due to the Web's immense size and dynamic nature no crawler is able to cover the entire Web and to keep up with all the changes. This fact has pushed the development of *focused crawlers*. In contrast to crawlers used by the general search engines focused crawlers selectively download Web documents, restricting the scope of crawling to a predefined domain. The downloaded documents can be stored and used as a source for data mining. In this paper we describe the main features of focused crawling, discuss the research on focused crawling conducted by the research group of the author, and discuss the problem areas associated with focused crawling not discussed in the literature which our work so far has revealed. Our test data consisted of Web documents in the *genomics* domain.

1. INTRODUCTION

Currently World Wide Web contains billions of publicly available documents (pages). Besides its huge size the Web is characterized by its rapid growth and change rates. It grows rapidly in terms of new servers, sites, and documents. The URL addresses of documents and their contents are changed, and documents are removed from Web sites. As more information becomes available on the Web it is more difficult to find relevant information from it.

Web search engines, such as Google and Altavista, provide access to Web documents. A search engine's *crawler* collects

Web documents and periodically revisits the pages to update the index of the search engine. Due to the Web's immense size and dynamic nature no crawler is able to cover the entire Web and to keep up with all the changes. This fact has motivated the development of *focused (topical) crawlers* [1, 2, 3, 5, 6, 12, 13]. The focused crawlers are designed to download Web documents that are relevant to a predefined domain (e.g. *genomics* or *immunology*), and to avoid irrelevant areas of the Web. Depending on the purpose of focused crawling different methods are applied to process the downloaded documents, e.g., they can be indexed for a domain specific search engine or they can be stored and used as a source for data mining. The benefits of the focused crawling approach are that focused crawling is able to find a large proportion of relevant documents on that particular domain, it is able to effectively discard irrelevant documents, and it is well able to keep up with the change of the Web.

In this experience paper we describe the main features of focused crawling, discuss the research on focused crawling conducted by the research group of the author, and discuss the problem areas associated with focused crawling not discussed in the literature which our work so far has revealed.

2. GENERAL VS. FOCUSED CRAWLING

Figure 1 shows the main processes of a general Web crawler used by a Web search engine. Crawling usually starts with a collection of seed URL addresses. The crawler connects to servers and downloads documents from the servers. Crawling starting from a given URL continues until it comes to a dead end or until some restriction defined in the crawling policy of a search engine is met. Words and URLs are extracted from the documents, the words are indexed and the URLs are added to the URL queue which determines the order in which new documents are downloaded. The huge size of the Web implies that a crawler of a general purpose search engine cannot download all the indexable Web documents, but it has to prioritize which portions of the Web to download. Usually also other types of information, such as PageRank scores (see below), is stored to make crawling and information retrieval more effective.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Database Endowment. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permissions from the publisher, ACM.

VLDB '07, September 23-28, 2007, Vienna, Austria.

Copyright 2007 VLDB Endowment, ACM 978-1-59593-649-3/07/09.

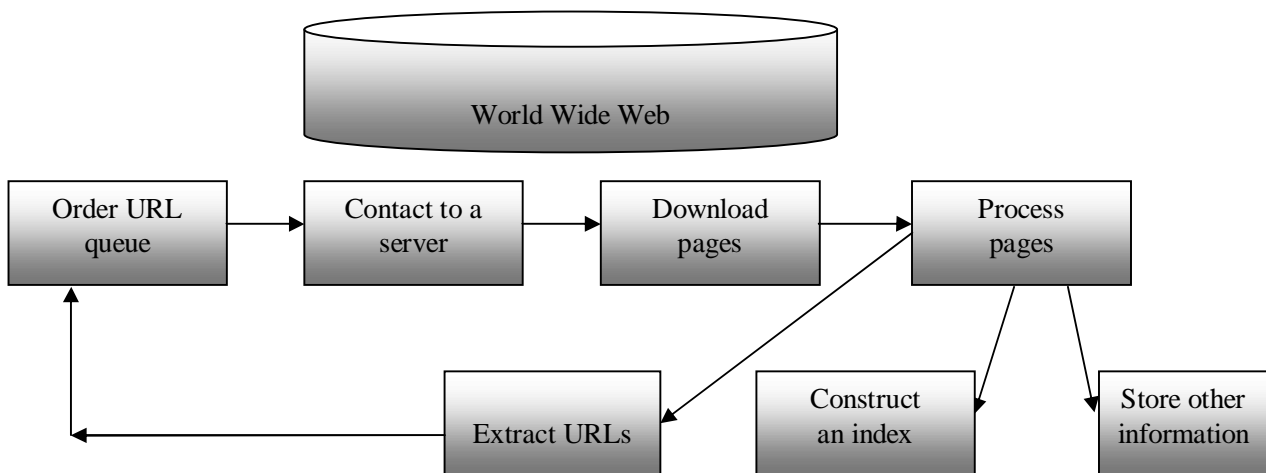


Figure 1. Basic processes of a Web crawler of a search engine

A focused crawler differs from a general crawler in two main points. First, it judges whether the visited documents and the documents pointed to by the URL addresses are relevant for the predefined domain. Domain identification is usually based on the use of domain vocabularies. Second, focused crawlers reorder the URL queue based on the relevance probability (i.e., the likelihood that a document deals with a topic in the domain). The documents assessed to be very relevant for the specific domain are downloaded first. General crawlers reorder the queue based on the authoritativeness or popularity of documents. The HITS algorithm [7] searches for authoritative documents based on the number of links pointing to documents. The popularity of documents can be determined e.g. using a PageRank algorithm [4] which rewards documents that are pointed to by documents that themselves are popular documents. However, also in focused crawling the authoritativeness and popularity of documents can be taken into account to improve the quality of the results of crawling.

The effectiveness of focused crawling is often evaluated using the measures of *harvest rate* and *coverage*. Harvest rate refers to the proportion of documents relevant to the domain to all downloaded documents. Coverage refers to the number of obtained relevant pages at time point T. The domain relevance of the documents is judged by human assessors.

3. RESEARCH ON WEB CRAWLING BY THE RESEARCH GROUP

The research group of the author has implemented two Web Crawlers. The first one (implemented in C language) is a general crawler which is used for the purpose of constructing word frequency lists in different languages [8]. The Crawler first downloads Web documents. Then the language of each sentence in each document is identified and sentences in a desired language are extracted from the documents. Finally word

frequency lists are constructed on the basis of the extracted sentences. We have also used the Crawler to construct a vocabulary in the field of *immunology*. In this case the domain of documents was identified by using a (long) list of URLs known to point to documents discussing immunology.

In a recent research [11] we built a Focused Crawler (coded in Perl) which was used as a tool to acquire comparable corpora in German-English and Spanish-English in the *genomics* domain (for data see [10]). The acquired corpora were used as a statistical translation tool in cross-language information retrieval (CLIR) experiments (CLIR refers to the retrieval of documents in other languages than the language of queries, see e.g. [9]).

Prior to the actual crawling phase seed URLs were retrieved using a Web search engine. A set of genomics terms in English, German, and Spanish were used as search terms to find relevant seed URLs. For each language, a few dozen seed URLs were used in the crawling phase. The genomics term lists played a key role also in the crawling phase - they were used to filter genomics-specific content from the documents. Since Web documents contain lots of noise from domains' point of view (e.g. personal contact information) the Crawler extracted text paragraphs from the downloaded documents. Thus text paragraphs rather than the entire documents were used to construct the comparable corpora. In a few days we were able to collect 149 500 English paragraphs, 84 200 German paragraphs and 30 800 Spanish paragraphs.

4. PROBLEM AREAS

The quality of the downloaded data and the effectiveness of focused crawling vary considerably depending on many factors. One important factor is the method how probably relevant links are identified, e.g. on the basis of the content of documents or on the basis of link anchor texts [2]. The second major factor is how irrelevant documents should be handled [1, 3]. It is quite

common that a relevant document points to an irrelevant document which points to another relevant document.

Based on our experiments we have so far found two significant problem areas not discussed in the literature:

(1) The source of seed URLs. Focused crawlers usually start with a core of URL addresses obtained by means of general search engines, thus being unable to find documents outside the crawling scope of the general search engines. Many of the documents that are distant from the starting points as measured by the number of intermediate links are missed. A geographical location of a server may also affect the performance of a focused crawler. In an ongoing study, we use two sets of seed URLs to investigate the effects of seed URL sources on the performance of focused crawling: URLs obtained from the Web search engines, and URLs obtained from the Web search engines and additional sources, such as literature references and databases (e.g. the Medline database). Electronic additional sources are identified so that URL extraction can be automated.

(2) Plain monolingual crawling vs. monolingual crawling through foreign language documents. An example of the former case is German crawling where only German documents are traversed, and an example of the latter case is German crawling through documents written in English and other foreign (i.e. non-German) languages. A German document may point to an English document which may point to another German document. These kinds of cases are common and they may lead to a performance loss when plain monolingual crawling is used as a crawling strategy.

5. CONCLUSIONS

The Web is a significant source of biological data and focused crawling is a means to download data from the Web for data mining and other purposes. The potential benefits of focused Web crawling in biological sciences are considerable. We have experimented with focused crawling and have identified two important problem areas - the source of seed URLs and the issues related to the multilinguality of the Web - which may have considerable effects on the quality of the downloaded data and the effectiveness of focused crawling. In an ongoing study we are investigating these problems. In particular the multilinguality of the Web poses major challenges for focused crawling.

6. ACKNOWLEDGMENTS

This research was supported by the Academy of Finland.

7. REFERENCES

- [1] Alpanidis, G., Kotropoulos, C. and Pitas, I. 2007. Combining text and link analysis for focused crawling - an application for vertical search engines. *Information Systems*, 32(6), 886-908.
- [2] Barfouroush, A. A., Nezhad, H.R.M., Anderson, M. L. and Perlis, D. 2002. Information retrieval on the World Wide Web and active logic: a survey and problem definition. *Technical Report, CS-TR-4291*, University of Maryland, Computer Science Department. <http://www.cs.umd.edu/Library/TRs/>
- [3] Bergmark, D., Lagoze, C. and Sbitiyakov, A. 2002. Focused crawls, tunneling, and digital libraries. *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, Rome, Italy, September 16-18, pp. 91 – 106.
- [4] Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.
- [5] Castillo, C. 2004. Effective Web crawling. *Ph.D. Thesis*. University of Chile, Department of Computer Science, 180 pages. <http://www.chato.cl/534/article-63160.html>
- [6] Chakrabarti, S., van den Berg, M. and Dom, B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Proceedings of the Eighth International World Wide Web Conference*, Toronto, May 11 - 14.
- [7] Kleinberg, J. 1998. Authoritative sources in a hyperlinked environment. *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 668–677.
- [8] Pirkola, A., Toivonen, J., Keskustalo, H. and Järvelin, K. 2008. Frequency-based identification of correct translation equivalents (FITE) obtained through transformation rules. *ACM Transactions on Information Systems (TOIS)*, 26(1). To appear.
- [9] Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. and Järvelin, K. 2003. Fuzzy translation of cross-lingual spelling variants. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '03)*, Toronto.
- [10] Pirkola, A. 2005. TREC 2004 Genomics track experiments at UTA: The effects of primary keys, bigram phrases and query expansion on retrieval performance. *The Thirteenth Text REtrieval conference*, Gaithersburg, MD. http://trec.nist.gov/pubs/trec13/t13_proceedings.html
- [11] Talvensaaari, T., Pirkola, A., Järvelin, K., Juhola, M. and Laurikkala, J. 2007. Focused Web crawling in the acquisition of comparable corpora. Submitted to *Information Retrieval*.
- [12] Tang, T., Hawking, D., Craswell, N. and Griffiths, K. 2005. Focused crawling for both topical relevance and quality of medical information. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management CIKM '05*.
- [13] Zhuang, Z., Wagle, R. and Giles, C.L. 2005. What's there and what's not?: focused crawling for missing documents in digital libraries. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, CO, pp. 301 – 310.