# EVALUATING GLOBAL LINK STRUCTURE OF THE WEB FOR FOCUSED CRAWLING IN THE GENOMICS AND GENETICS DOMAINS

Ari Pirkola and Tuomas Talvensaari
*Department of Information Studies, University of Tampere, Finland*

Keywords: Focused web crawling, Genomics, Genetics.

Abstract: A focused crawler is a program that fetches Web pages that are relevant to a pre-defined domain. In this paper we consider focused crawling in the domains of genomics and genetics. Crawling is often started with seed URLs that point to central North-American and European universities, research institutions, and other organizations in North-America and Europe. We investigate how strongly this central region of the Web is connected to other large geographical regions of the Web: Australia (top level domain .au), China (.cn), and five South-American countries (.ar, .br, .cl, .mx, and .uy). We consider what implications the observed global link structure has for the selection of seed URLs for focused crawling. The results showed that the proportion of out-links from the North-American and European region to the other regions is low whereas pages in the other regions often point to the central region. We also found that two focused crawling processes, one started from the central region and the other from another large region, overlap only to a small extent. Overall, the results suggest that the effectiveness of focused crawling can be improved considerably if crawling is started with a geographically heterogeneous seed URL set.

## 1 INTRODUCTION

Web crawling refers to the process of gathering data from the World Wide Web. *Focused crawlers* are programs that selectively download Web pages, restricting the scope of crawling to a pre-defined domain or topic (Bergmark et al., 2002; Castillo, 2004; Chakrabarti et al., 1999; Talvensaari et al., 2008; Tang et al., 2005; Zhuang et al., 2005). Depending on the purpose of focused crawling (FC), different methods are applied to process the downloaded pages, e.g. they can be indexed for a domain specific search engine or a digital library. Focused crawlers can even be used as personal search agents. The benefits of focused crawling are that it is able to find a large proportion of relevant pages on that particular domain and it is well able to keep up with the change of the Web.

Crawling starts with a set of *seed URLs*. The crawler connects to servers and downloads pages 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 is met. URLs are extracted from the pages and are added to the URL queue which determines the order in which new pages are downloaded. A focused crawler differs from a general crawler in two main points. First, it judges whether the visited pages and the pages pointed to by the URLs are relevant for the pre-defined domain. Domain identification is based, for example, on domain vocabularies or topical hierarchies. Second, focused crawlers reorder the URL queue based on the probabilities that the downloaded pages deal with the defined domain or topic. The pages assessed to be very relevant are downloaded first.

In a current project we are examining the connections between different geographical regions of the Web and the effects of geographically categorized seed URL sets on the effectiveness of FC. So far, we have completed one set of experiments, the results of which are reported in this position paper.

In FC, a common practice is to retrieve the seed URLs by a Web search engine. The returned URLs typically point to pages of central information providers in the field, such as universities, journals, and research institutions. These are mainly North-American and European Web pages. We investigate in this paper how strongly this central region is

Pirkola A. and Talvensaari T. (2009).
EVALUATING GLOBAL LINK STRUCTURE OF THE WEB FOR FOCUSED CRAWLING IN THE GENOMICS AND GENETICS DOMAINS.
In *Proceedings of the International Conference on Health Informatics*, pages 499-502
DOI: 10.5220/0001777004990502

connected to other large geographical regions of the Web, and what implications the observed global link structure has for the selection of seed URLs for FC.

# 2 METHODS AND DATA

## 2.1 Web Regions

We call Web pages with generic and sponsored top level domains (gTLDs and sTLDs, see http://en.wikipedia.org/wiki/Generic_top-level_ domain), e.g. *.com, .edu, .gov,* and *.org*, as well as North-American and European country code top level domains (ccTLDs), e.g. *.ca, .de, .es, .fr, .it, .pt,* and *.uk* collectively the *Major region of the Web.* The Major region is contrasted to Australian (ccTLD: *.au*), Chinese (ccTLD: *.cn*), and South-American (five ccTLDs: *.ar, .br, .cl, .mx,* and *.uy*) regions. These are here called *Minor regions of the Web*. It should be noted that a small portion of TLDs defined here as Major region TLDs are registered outside North-America and Europe. Therefore, Major region does not exactly correspond to North-America and Europe. It should also be noted that the terms *major* and *minor* refer to the relative size dimensions of the Web. For simplicity, Mexico (*.mx*) is here called a South-American country.

## 2.2 Test Topics, Seed URLs, and Crawling

We experimented with two kinds of *test topics* in the domains of *genomics* and *genetics*: *specific* topics (an example: *regulatory targets of nkx genes*), and *general topics* (an example: *hereditary diseases*). As specific topics we used five TREC (http://trec.nist.gov) Genomics Track 2004 topics (the topic numbers 1, 10, 20, 30, and 40). For the Genomics Track, see Hersh et al. (2005). There were also five general topics. They were created by one of the authors who has expertise in health informatics. For seed URL retrieval, *queries* containing synonyms and morphological variants of the topic words were constructed based on the topics. Statistical information, such as term and document frequencies and the total number of hits for a query in the Medline database (http://www.ncbi.nlm. nih.gov/pubmed/) are some measures to determine topic specificity. We used the latter measure.

Four crawls were performed for each topic with the seed URLs from the Major region, Australia, China, and five South-American countries (Argentina, Brazil, Chile, Mexico, and Uruguay). Each seed URL set contained 50 URLs. The seeds of the Major region were retrieved by means of the basic Google (http://www.google.com) whereas the seeds of the other regions were retrieved by Google's local versions (e.g. http://www.google.cl). The South-American URL sets were formed by taking top ten URLs from each five local Google. Most of the Chinese seed URL pages were bilingual Chinese-English pages.

The majority of the Major region seed URLs were of the type *.com, .edu, .gov, .org, .de,* and *.uk*. The original Major region seed URL sets contained (only) a few Australian, Chinese, and South-American URLs which were removed from the final sets to allow us to investigate the defined research questions (presented in Section 3).

The *Nalanda iVia Focused Crawler* (http://ivia.ucr.edu) was used in the experiments. It is based on the work of Chakrabarti et al. (1999). At the start of a crawl, the Nalanda crawler initializes a priority queue of URLs with a set of seed URLs. One by one, each page *u* pointed to by the URLs in the queue are fetched. The probability Pr(t|u), i.e., the probability of *u* being about the wanted topic *t* is calculated with a text classification algorithm. The probabilities Pr(t|u) were estimated with a *logistic regression classifier* (Zhang et al., 2003) that, for every topic, was trained with positive and negative instances of the topic in question.

There were 40 seed URL sets in total: 10 topics and for each topic 4 URL sets representing different regions. Accordingly, we performed 40 crawls. In each case crawling was stopped after 20 000 pages had been downloaded. Thus, each result list contained 20 000 pages.

For evaluation, the fetched pages were indexed with the *Terrier search engine* (http://ir.dcs.gla.ac.uk/terrier/) that ranked the pages based on their probability of relevance to the entered query. The same queries that were used in searching for seed URLs were used to represent the topics, however they were modified to fit Terrier's query language. Of course, the probabilities calculated by the classifier could have been used to rank the pages, but Terrier was used to provide stronger evidence.

# 3 RESEARCH QUESTIONS AND EVALUATION

We denote by *T(S)* the situation where pages in the region of *T(arget)* are downloaded (or are considered in calculations) in a crawling process that starts with seed URLs in the region of *S(eed)*. The regions considered in this study are denoted as follows: *M* (Major), *A* (Australia), *C* (China), *SA*

(South-America), and *O* (other). The category of *other* includes regions not used as *seed URLs* in this study as well as indeterminate TLDs. A combination of the symbols is marked as, for example, *A,C,O,SA(A)*. This refers to the case where crawling is started with seeds in the region of A, and pages in the Minor regions A, C, O, and SA are downloaded.

We investigate how strongly the Major region of the Web is connected to the other geographical regions of the Web. We are interested in both directions: from the region of M to the regions of A, C, O, SA, and from A, C, SA to M. If, as expected, the former direction is weak and the latter one strong, FC starting with seed URLs only from the Major region may lose a substantial amount of relevant information. First, it loses a considerable number of pages inside the Minor regions. Second, if FC is started from a Minor region, it is likely that this would find a significant number of Major region pages that are not within the crawling scope of FC starting from the Major region. This is because, rather than being a true web the Web is a community of communities that are isolated or only loosely connected to each other (Toyoda and Kitsuregawa, 2001). It is therefore likely that FC starting from two remote areas finds pages from different communities.

To explore how strongly the Major region is connected to the Minor regions we calculated, first, the proportion of pages downloaded from a target region $T_j(S_i)$ among all downloaded pages $T_{all}(S_i)$: $T_j(S_i)$ / $T_{all}(S_i)$. Naturally, in most cases the seed URLs point to the target regions indirectly through URLs extracted from the pages downloaded during crawling. This measure is called *seed-to-target (ST) rate*. It was calculated for the following test cases: Major → Major; includes the case *M(M)*

- Major → Minor; includes the case *A,C,O,SA(M)*
- Minor → Major; includes the cases *M(A), M(C), M(SA)*
- Minor → Minor; includes the cases *A,C,O,SA(A), A,C,O,SA(C), A,C,O,SA(SA)*

Second, we calculated for all three Minor seed regions *overlap rate*, i.e., the percentage of identical URLs downloaded for the seed regions Major and Minor. Generally, high overlap indicates that two focused crawling processes with different starting points (Major and Minor) operate mainly in the same Web communities while low overlap shows that they operate mainly in different communities.

Both ST and overlap rates were measured at two relevance probability values assigned by Terrier to the downloaded pages. The *thresholds* were TR1 > 0.0 and TR2 > 5.0. The higher TR is, the more relevant pages there are in a page set.

## 4 FINDINGS

The seed-to-target (ST) rates are presented in Table 1. As expected, the highest ST rates were obtained for Major→ Major region. The figures are very high: 94.7% or more. This means that the majority of North-American and European pages dealing with genomics and genetics are connected to other North-American and European pages. It was also expected that the direction of Major→ Minor is very weak: ST ranges from 2.9% to 5.3%. The opposite direction, Minor→ Major is much stronger, with ST ranging from 40.1% to 85.1%. For Minor→ Minor, ST is in the range (14.9%, 59.9%). We did not consider the Minor→ Minor cases where S is the same as T. However, it is obvious that in most cases Minor region target pages are of the same type as Minor region seeds. For example, Australian seeds find Australian pages rather than Chinese and South-American pages. Table 1 also shows that in the case of Minor region seeds, the seeds for general topics point more often to the Minor region than the seeds for specific topics.

The results of the overlap calculations are reported in Table 2. In each case the first column shows the absolute numbers of downloaded pages for a region pair (identical URLs in single result lists were first removed), and the second column shows the overlap percentages. For example, FC / Major and FC / Australia with TR=0.0 gave 21575 pages in total. Of these pages 1.2% (N=267) shared the same URL. In all cases the overlap rates are very low, 1.4% or less. Overall, the overlap results show that the crawling results of two FC processes, one started from the Major region and the other from a Minor region, overlap only to a small extent.

## 5 CONCLUSIONS

The results revealed a biased link structure in that North-American and European seeds point primarily to other North-American and European pages whereas Minor region seeds point both to the Major and Minor regions. The results also showed that pages downloaded using Major and Minor region seeds overlap only to a small extent. Overall, the results suggest that the effectiveness of FC can be improved considerably if crawling is started from different geographical regions. The domains of genomics and genetics are typical scientific

Table 1: Seed-to-Target rates (%).

| *Topic type and relevance threshold* | *Major → Major* | *Major → Minor* | *Minor → Major* | *Minor → Minor* |
|---|---|---|---|---|
| **General, TR=0.0** | 94.7 | 5.3 | 76.8 | 23.2 |
| **General, TR=5.0** | 96.0 | 4.0 | 40.1 | 59.9 |
| **Specific, TR=0.0** | 97.1 | 2.9 | 85.1 | 14.9 |
| **Specific, TR=5.0** | 96.6 | 3.4 | 65.7 | 34.3 |

Table 2: Overlap rates (%) for the Major and Minor regions.

| *Topic type and relevance thr.* | *Major+ Australia N* | *Major+ Australia Overlap%* | *Major+ China N* | *Major+ China Overlap%* | *Major+ South-America N* | *Major+ South-America Overlap%* |
|---|---|---|---|---|---|---|
| **General TR=0.0** | 21575 | 1.2 | 23342 | 1.4 | 19326 | 0.3 |
| **General TR=5.0** | 3430 | 0.7 | 3349 | 0.9 | 2857 | 0.2 |
| **Specific TR=0.0** | 18845 | 0.2 | 21227 | 0.6 | 18926 | 0.0 |
| **Specific TR=5.0** | 2349 | 0.0 | 2837 | 0.4 | 1914 | 0.0 |

domains. We therefore assume that the obtained results are generalizable to other scientific domains.

## REFERENCES

Bergmark, D., Lagoze, C. and Sbityakov, 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.

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

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.

Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M. and Kraemer, D. F., 2005. TREC 2004 genomics track overview. *Proceedings of the Thirteenth TExt REtrieval conference (TREC-13)* (Gaithersburg, MD). http://trec.nist.gov/pubs/trec13/t13_proceedings.html

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

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.*

Toyoda, M. and Kitsuregawa, M., 2001. Creating a Web community chart for navigating related communities. *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, Århus, Denmark, August 14 - 18.

Zhang, J., Jin, R., Yang, Y. and Hauptmann, A., 2003. Modified logistic regression: An approximation to svm and its applications in large-scale text categorization. *Proceedings of the 20th International Conference on Machine Learning (ICML)*, Washington, DC.

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.