EFFECTS OF CRAWLING STRATEGIES ON THE PERFORMANCE OF FOCUSED WEB CRAWLING

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

Keywords: Focused crawling, Web crawling.

Abstract: Focused crawlers are programs that selectively download Web documents (pages), restricting the scope of crawling to a specific domain or topic. We investigate different focused crawling strategies including the use of data fusion in focused crawling. Documents in the domains of genomics and genetics were fetched by Nalanda iVia Focused Crawler using three crawling strategies. In the first one, a text classifier was trained to identify relevant documents. In the latter two strategies, the identification of relevant documents was based on query-document matching. In experiments, the crawling results of the single strategies were combined to yield fused crawling results. The experiments showed, first, that different single strategies overlap only to a small extent, identifying mainly different relevant documents. Second, a query-based strategy where the words of the link context were weighted gave the best coverage (i.e., number of relevant documents) after 10 000 and 40 000 documents had been downloaded. The combination of the two query-based strategies was the best fused strategy but it did not perform better than the best single strategy.

1 INTRODUCTION

Focused crawling (FC) refers to the process of gathering Web documents (pages) dealing with a specific domain or topic (Castillo, 2004; Chakrabarti et al., 1999; Novak, 2004; Tang et al., 2005; Zhuang et al., 2005). Depending on the purpose of focused crawling, 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. There are different *strategies* to identify relevant documents during crawling. Two prevailing ones are classifier-based and query-based approaches. In the former approach relevant documents are identified by a text classifier that is trained for the domain or topic in question. In the latter approach identification is based on query-document matching.

In many information processing areas *data fusion* has yielded noticeable performance improvements. The term refers to the use of techniques that combine data from different complementary sources (http://en.wikipedia.org/ wiki/Data_fusion). Fusion is performed in the hope that the fused system is more effective than the single systems separately. Even though the use of data fusion does not guarantee effectiveness improvements (Beitzel et al., 2004), many studies, for example in the field of information retrieval, have shown that data fusion helps to improve the effectiveness of systems (e.g. Braschler, 2004; Manmatha et al., 2001; Montague and Aslam, 2002). Braschler (2002) applied data fusion successfully in multilingual information retrieval. Manmatha et al. (2001) developed an unsupervised probabilistic method for combining retrieval results from different retrieval systems. They showed that the method performed as well as the best current combination techniques. Montague and Aslam (2002) were able to improve retrieval results by combining document ranking functions. In FC the use of data fusion is an unexplored area. It is therefore interesting and important to study whether the effectiveness of FC can be improved by combining the crawling results obtained by different focused crawling systems / strategies.

In this paper we investigate whether *fused focused crawling strategies* produce better performance than the component single crawling strategies separately, and which of the single strategies selected for testing in the study yields the best performance. We report the first results of a current research project investigating different approaches to focused crawling. In crawling experiments we used the *Nalanda iVia Focused Crawler* (http://ivia.ucr.edu/), created by Chakrabarti et al. (1999). In the system, page relevance probabilities are computed based on a logistic

Pirkola A. and Talvensaari T.
EFFECTS OF CRAWLING STRATEGIES ON THE PERFORMANCE OF FOCUSED WEB CRAWLING.
DOI: 10.5220/0002037603760381
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

regression classifier. Three single strategies were investigated in the present study: a classifier-based approach and two query-based approaches. In the experiments the crawling results of the single strategies were combined to yield fused crawling results. The performance of the fused strategies was then compared to that of single strategies, and the performance of the single strategies was compared to each other. The results reported in this paper are based on 30 different crawls in the domains of *genomics* and *genetics.*

The rest of this paper is organized as follows. Section 2 discusses the motivation of the study and presents the research questions. The methodology and data is presented in Section 3. Section 4 describes the evaluation measures. Section 5 contains the findings and Section 6 conclusions.

2 MOTIVATION AND RESEARCH QUESTIONS

There are two main factors that decrease the effectiveness of FC. *First*, FC is only capable of finding documents that are located in the same Web community as one or more of the seed URLs entered into the FC program. In other words, FC can only find documents that through some routes are linked to at least one of the seed URLs (start URLs). Documents outside these routes are missed. To address this problem it is possible to use a large set of seed URLs. However, effort costs of selecting large seed URL sets by hand are prohibitive. If seed URLs are selected automatically, then increasing the number of relevant seed pages also increases the number of irrelevant seed pages. This in turn decreases FC effectiveness. One approach that would be interesting to study would be the use of very heterogeneous seed URL sets, for example URLs from many different countries and URLs representing different types of top level domains. Such a seed set could guide crawling to many different relevant routes.

Second, identification of links that point to relevant pages or good routes is always incomplete, and it is likely that even the most effective FC program or strategy misses a large number of relevant links and documents even though the documents are located inside the same Web community as the seed URLs. However, FC research has still developed several methods that improve the effectiveness of FC, such as tunneling (Bergmark et al. 2002), reinforcement learning (Rennie and McCallum, 1999) context graphs (Diligenti et al., 2000), and the utilization of the hierarchical structure of Web documents (Chakrabarti et al. 2002). Also link popularity statistics, e.g. PageRank (Brin and Page, 1998) can be used in FC (in this study it was not used). This paper contributes to the issue by investigating the effectiveness of different FC strategies. The specific research questions are as follows:

(1) What is the degree of *overlap* between single focused crawling strategies? This issue provides background information for the two research questions below.

(2) Which of the three single strategies examined in the study yields the highest *coverage* (number) of relevant documents?

(3) Does a fused strategy yield a better coverage in comparison to a coverage yielded by the best single strategy? Which combination of the three strategies yields the highest coverage?

3 METHODS AND DATA

3.1 Crawling Program and the Three Strategies

Next we describe the program and the three crawling strategies. The first strategy (*strategy 1* in Findings section) was based on *Nalanda's* basic features, i.e., unlike other strategies we did not modify the program. For each downloaded page *u,* Nalanda calculates Pr(t|u), i.e., the probability of relevance of *u* to the topic *t*. Nalanda extracts the outlinks <*u,v*> on the page, and assigns the yet-unseen page *v* the same probability of relevance Pr(t|v) = Pr(t|u). The URL of *v* is inserted into the URL queue with priority Pr(t|v). The probabilities are assigned with a logistic regression classifier that, for every topic, was trained with positive and negative instances of the topic in question. The same Google queries that were used in retrieving the seed URLs (Section 3.2) were used in acquiring the positive examples. For each topic, about 300 on-topic pages were used in training the classifier. The negative examples were taken partly from a sample of "random" pages, and partly from the positive examples of other topics. The random pages were retrieved by querying Google with a query generated by a random phrase generator (http://watchout4snakes.com/creativitytools/RandomWord/RandomPhrase.aspx). The number of negative examples was also 300.

The pages fetched by Nalanda were indexed with the *Lemur search engine* (http://www.lemurproject.org/) 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 Lemur's query language. Of course, the probabilities calculated by the classifier could have been used to rank the pages, but Lemur was used to provide stronger evidence. Also, Lemur was also used in other strategies, and it is important to use compatible probability scores when the crawling results are combined.

The second strategy (*strategy 2* in Findings section) was based on a modified version of the Nalanda crawler. Instead of a text classifier, the probabilities were assigned directly using the Lemur search engine. Specifically, the structured query mode of Lemur was used. The topic of each crawl was "translated" into the query language of the search engine. The probabilities Pr(t|u) and Pr(t|v) were determined by matching a query to the page *u.* The similarity score given by the search engine was used as the probability. Further, in the second strategy, the context of each link <*u, v*> was also used in determining Pr(t|v). The context words of each link, meaning words that appeared no more than 5 DOM tree nodes apart from the link, were matched against the topic query. That is, Pr(t|v) = Pr(t|c_{uv}), where c_{uv} is the context of the link <*u,v*>. However, the context probability was used only if it exceeded Pr(t|u). If it applied that Pr(t|c_{uv}) < Pr(t|u), then Pr(t|v) = Pr(t|u). In this way, it was ensured that each unseen page was assigned a minimum probability of relevance of the linking page.

The third strategy (*strategy 3* in Findings section) differed from the second one only slightly. In the third strategy, the words of the link context were weighted based on their distance from the link node. The words that appeared in the link node itself were given the highest weights while words further apart were given linearly decreasing weights. We purposefully selected two nearly similar strategies since it is interesting to observe how similar crawling results they produce.

3.2 Test Topics and Seed URLs

In experiments we used two kinds of *test topics* in the domains of genomics and genetics: specific topics and general topics. As specific topics we used five TREC (http://trec.nist.gov) Genomics Track (Hersh et al., 2005) test topics from the year 2004 (the topic numbers 1, 10, 20, 30, and 40). TREC is an annual research forum to develop and evaluate information retrieval methods and systems. Genomics Track focuses on genomics information retrieval. There were also five general topics. They were created by one of the authors who has expertise in biology and genomics information retrieval. 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.

For seed URL retrieval, *queries* containing synonyms and morphological variants of the topic words were constructed based on the topics. The seed URLs were retrieved by means of the Google search engine (http://www.google.com). The first 50 seed URLs returned by Google in response to the constructed queries were selected as a seed URL set for each topic. The majority of the seed URLs were of the type *.com, .edu, .gov, .org, .de,* and *.uk*.

Below are two example topics. The first one represents a general topic and the second one a specific topic.

Topic: Find information on hereditary diseases
Seed URL Query: hereditary (disease OR diseases)
#Medline_hits: 28 074

Topic: NEIL1. Find articles about the role of NEIL1 in repair of DNA.
Seed URL Query: (neil1 OR "nei endonuclease" OR flj22402 OR fpg1 OR nei1 OR hfpg1) dna (repair OR lesion OR lesions)
#Medline_hits: 54

There were 10 topics, and accordingly 10 seed URL sets, each containing 50 URLs. In all, 30 crawls were performed (three strategies x 10 topics). Crawling was stopped after 40 000 documents had been downloaded. So each result list contained 40 000 documents. Due to the rapid change of the Web all crawls were performed within a time period of two weeks, i.e., there was no great space of time between the crawls.

4 EVALUATION

In Section 2 we defined the three research questions. In this section we describe the evaluation measures for the research questions.

Overlap was measured using the measure of *overlap rate*. It refers to the percentage of identical URLs downloaded for each pair of single strategies and for all three strategies. Overlap shows the degree to which different single strategies find the same pages. We expect the overlap rates not to be very high because small differences may accumulate to

large differences during crawling. This may hold also for the two query-document matching strategies examined in this study that do not differ much from each other.

Coverage was measured using the measure of *coverage rate after N downloaded documents*. It refers to the number of downloaded documents above the selected relevance threshold (see below) after N documents (=10 000 and 40 000) have been downloaded. The sooner relevant documents are downloaded during the crawling process, the better and more efficient the crawling system / method is (Baeza-Yates et al., 2005). We compare the coverage of each single strategy to each other and against that of each fused strategy, i.e., coverage rates of strategies 1, 2, and 3 are compared to each other and to those of strategies 1+2, 1+3, 2+3, and 1+2+3. We are interested in the following question: When an equal number of documents have been downloaded for two strategies which one of these has been fetched more probably relevant documents (i.e., documents with probabilities above the selected threshold). In all comparison situations coverage was calculated after 10 000 and 40 000 documents were downloaded. In the case of fused strategies the result lists of the two / three single strategies were combined by taking from each list the first 5000 / 3333 (3334) documents (=10 000 in total), and the first 20 000 / 13 333 (13 334) documents (=40 000 in total). In this manner we can directly compare the coverage rates of different strategies.

We took samples of documents at different positions of the ranked crawling result lists, assessed the relevance of the sample documents, and selected a reasonable probability *threshold* on the basis of the assessment. In this manner we were able to anchor the probabilities to the actual relevance of the documents. Documents above the selected threshold were taken for the evaluation of the crawling results. The selected probability threshold was prob. > 0.45. The probability scores of the Lemur search engine are in the range of 0.4 – 1.0 where 0.4 is zero probability; scores > 0.6 do no appear very often. On average 84% of documents above the threshold of 0.45 were relevant. This figure is based on the assessment of 450 documents and it represents all 10 topics (45 documents per topic) and both high and low rank document positions in the result lists of prob. > 0.45. The relevance of documents was evaluated by one of the authors who has expertise in biology and genomics information retrieval (this is the reason why genomics and genetics were selected as test domains). Documents that discussed the topic, contained facts about the topic (e.g. database entries and Web forms), or contained relevant literature references were considered relevant.

5 FINDINGS

Table 1 reports the overlap results. The results are categorized based on the topic type. As described before, strategy 1 refers to the classifier-based crawling, strategy 2 to the first query-based and strategy 3 to the second query-based crawling. In each case the first column shows the absolute number of downloaded pages for a strategy pair / the three strategies, and the second column shows the overlap percentages. For example, crawling based on the strategies 1 and 2 gave 797 documents for the general topics. Of these pages 7.1% shared the same URL.

As can be seen from Table 1, in all cases overlap rates are low. In all but one case the figures are less than 10%. The two query-based strategies have the highest overlap rates: 9.4% and 13.6%. However, even these figures are low. It should be noted that the query-based strategies were very similar: The only difference was whether or not the context words of links were weighted. Overall, the results show that different FC strategies overlap only to a small extent, identifying mainly different relevant documents.

Table 2 and Figure 1 report the coverage results. Table 2 shows the number of probably relevant documents for 10 000 documents and Figure 1 for 40 000 documents. In Table 2 the highest figures are highlighted in a bold font. As can be seen, the second query-based strategy, i.e., strategy 3, and the combination of the two query-based strategies (2+3) are the best strategies, and they perform almost equally well. The fused strategy performs slightly better in the case of general topics whereas in the case of specific topics strategy 3 provides a slightly higher coverage. After 10 000 documents had been downloaded crawling based on the combination of strategies 2 and 3 yielded 7163 documents, and strategy 3 yielded 6944 documents. In the case of 40 000 documents strategy 3 outperforms the fused strategy of 2+3 (Figure 1).

Strategy 1 (classifier-based approach) does not perform well. This is probably due to the fact that it is often difficult to adequately model the off-topic class, which can impair the performance of the classifier. It is often easier to present the topic as a query than to train a classifier, especially in narrow, easily defined topics. Effectiveness and implementability were the primary reasons for investigating the query-based strategies in this study.

Table 1: Overlap rates (%) between the single strategies. Probability threshold 0.45.

Topic type	*1+2 N*	*1+2 %*	*1+3 N*	*1+3 %*	*2+3 N*	*2+3 %*	*1+2+3 N*	*1+2+3 %*
General	797	7.1	787	6.1	2028	9.4	677	3.0
Specific	421	4.5	455	4.0	2649	13.6	376	1.9

Table 2: Coverage (=number of documents with prob > 0.45) after 10 000 downloaded documents.

Topic type	*1*	*2*	*3*	*1+2*	*1+3*	*2+3*	*1+2+3*
General	1057	3637	3695	2784	2786	**3924**	3117
Specific	561	2815	**3249**	2063	2031	3239	2686
All	1618	6452	6944	4847	4817	**7163**	5803

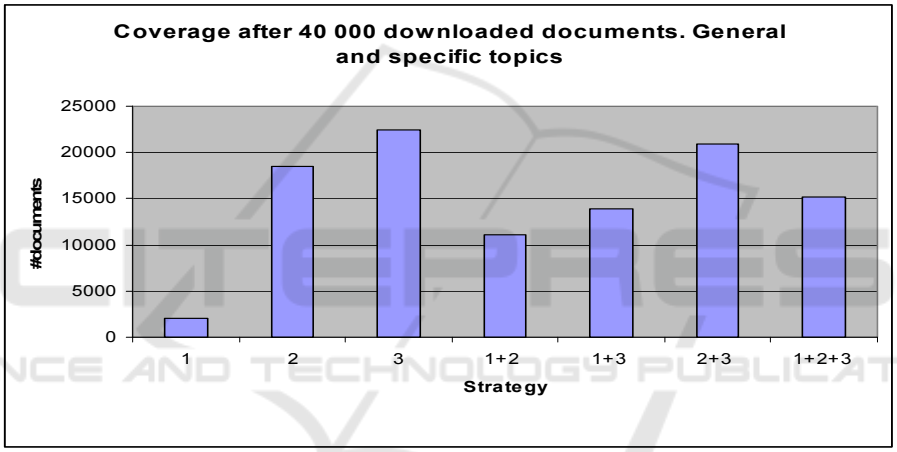

Figure 1: Coverage for 40 000 documents.

6 CONCLUSIONS

In this paper we reported the first results of a current research project on different approaches to focused Web crawling. The results showed that the combination of the two query-based strategies was the best fused strategy but it did not perform better than the best single strategy, i.e., a query-based approach where the words of the link context were weighted. Comparing the results of this study to those in the literature is difficult due to many different evaluation measures and definitions of page relevance in the field of focused crawling. Srinivasan et al. (2005) also discuss this problem and point out that this is typical of a field that is still in its most creative phase.

There are not many studies that have compared different crawlers / strategies. Perhaps the most comprehensive work is that of Srinivasan et al. (2005). The researchers developed a general framework to evaluate focused (topical) crawlers, and used it to evaluate four off-the-shelf crawlers: BreadthFirst (a simple strategy for crawling); BSF1 and BSF256 that are variations of best-first search; InfoSpiders where the identification of relevant links is based on keyword vectors and neural nets. The study showed that BSF1 and BSF256 yielded the best performance for 4000-page crawls. For a long crawl (50 000 pages) InfoSpiders performed best.

There are many other FC strategies / approaches than those investigated in this paper. For example, in the *Context Graph* approach (Diligenti et al., 2000) a graph of several layers deep is constructed for each page and the distance of the page to the target pages is computed. The context graphs are used to train a classifier with features of the paths that lead to

relevant pages. We are presently investigating the issues presented in this paper more extensively. In addition to new strategies, we consider the effects of a search engine used to rank the pages and seed URL sets on crawling performance. One aim of this study is to find an effective strategy for multilingual focused crawling (Pirkola and Talvensaari, 2009).

ACKNOWLEDGEMENTS

This study was funded by the Academy of Finland (research projects 125679 and 129835).

REFERENCES

Baeza-Yates, R., Castillo, C., Marin, M. and Rodriguez, A., 2005. Crawling a country: better strategies than breadth-first for web page ordering. *Proc. of the 14th International conference on World Wide Web / Industrial and Practical Experience Track*, Chiba, Japan, pp.864-872.

Beitzel, S., Jensen, E., Chowdhury, A, Grossman, D., Frieder, O. and Goharian. N., 2004. Fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology*, 55(10): 859-868.

Bergmark, D., Lagoze, C. and Sbityakov, A., 2002. Focused crawls, tunneling, and digital libraries. *Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, Rome, Italy, September 16-18, pp. 91 – 106.

Braschler, M., 2004. Combination approaches for multilingual text retrieval. *Information Retrieval*, 7 (1-2): 183-204.

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.

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. *Proc. of the Eighth International World Wide Web Conference*, Toronto, May 11 - 14.

Chakrabarti, S., Punera, K. and Subramanyam, M., 2002. Accelerated focused crawling through online relevance feedback. *Proc. of the 11th International Conference on World Wide Web*, Honolulu, Hawaii, May 7 - 11, pp. 148-159.

Diligenti, M., Coetzee, F. .M., Lawrence, S., Giles, C.L. and Gori, M., 2000. Focused crawling using context graphs. *Proc. of the 26th International Conference on Very Large Databases (VLDB)*, pp. 527-534.

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

Manmatha, R., Feng, F. and Rath, T., 2001. Using models of score distributions in information retrieval. *Proc. of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana.

Montague, M. and Aslam, J. 2002: Condorcet fusion for improved retrieval. *Proc. of the Eleventh International Conference on Information and Knowledge Management*, McLean, VA, November 4-9, pp. 538-548.

Novak, B., 2004. A Survey of focused Web crawling algorithms. *Proc. of SIKDD 2004 at Muticonference IS*, Ljubljana, Slovenia, October 12-15, pp. 55–58.

Pirkola, A. and Talvensaari, T. 2009. Developing a system for multilingual focused crawling. Submitted to *WWW'2009 - 18th International World Wide Web Conference*, Madrid, Spain, April 29-24, 2009. Poster manuscript.

Rennie, J. and McCallum, A., 1999.Using reinforcement learning to spider the web efficiently. *Proc. of the Sixteenth International Conference on Machine Learning* (ICML).

Srinivasan, P., Menczer, F., Pant, G. 2005. A general evaluation framework for topical crawlers. *Information Retrieval*, 8(3): 417-447.

Tang, T., Hawking, D., Craswell, N. and Griffiths, K., 2005. 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*.

Zhuang, Z., Wagle, R. and Giles, C.L., 2005. What's there and what's not?: focused crawling for missing documents in digital libraries. *Proc. of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, CO, pp. 301 – 310.