Building Domain-Specific Web Collections for Scientific Digital Libraries: A Meta-Search Enhanced Focused Crawling Method

Jialun Qin, Yilu Zhou
Dept. of Management Information Systems
The University of Arizona
Tucson, AZ 85721, USA
+1 520 621 3927
{qin, yilu}@u.arizona.edu

Michael Chau
School of Business
The University of Hong Kong
Pokfulam, Hong Kong
+852 2859 1014
mchau@business.hku.hk

ABSTRACT
Collecting domain-specific documents from the Web using focused crawlers has been considered one of the most important strategies to build digital libraries that serve the scientific community. However, because most focused crawlers use local search algorithms to traverse the Web space, they could be easily trapped within a limited sub-graph of the Web that surrounds the starting URLs and build domain-specific collections that are not comprehensive and diverse enough to scientists and researchers. In this study, we investigated the problems of traditional focused crawlers caused by local search algorithms and proposed a new crawling approach, meta-search enhanced focused crawling, to address the problems. We conducted two user evaluation experiments to examine the performance of our proposed approach and the results showed that our approach could build domain-specific collections with higher quality than traditional focused crawling techniques.

Categories and Subject Descriptors
H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital Libraries – *collection*

General Terms
Design, Experimentation.

Keywords
Digital libraries, domain-specific collection building, focused crawling, meta-search, Web search algorithm.

1. INTRODUCTION
The Web, containing more than 3 billion pages, has made available a large amount of information and resources that can be useful in various scientific research areas, such as papers reporting research results and patents describing industrial innovation. Collecting domain-specific documents from the Web has been considered one of the most important strategies to build digital libraries that serve the scientific community. Since late 1990s, there has been much research on different tools to build domain-specific Web collections and currently the most popular and widely-used tool is *focused crawler* [5].

Focused crawlers are programs designed to selectively retrieve Web pages relevant to a specific domain for the use of domain-specific search engines and digital libraries. Unlike the simple crawlers behind most general search engines which collect any reachable Web pages in breadth-first order, focused crawlers try to "predict" whether or not a target URL is pointing to a relevant and high-quality Web page before actually fetching the page. In addition, focused crawlers visit URLs in an optimal order such that URLs pointing to relevant and high-quality Web pages are visited first, and URLs that point to low-quality or irrelevant pages are never visited. There has been much research on algorithms designed to determine the quality of Web pages. However, most focused crawlers use local search algorithms such as *best- first search* to determine the order in which the target URLs are visited.

Scientific digital libraries with collections built by focused crawlers can provide search results with high precision and greatly alleviate users' information overload problem [3], a problem in which a search employing a general search engine such as Google can result in thousands of irrelevant hits. However, various scientific fields, such as bioinformatics and nanotechnology, have experienced tremendous growth over the past several years. Now, a discipline often encompasses a diversity of research perspectives and application areas. Such high speed and diversity of knowledge creation and information generation in scientific domains further complicate the issue of collection building. The use of local search algorithms and the existence of Web communities [9, 11, 12, 22] often limit the scope of focused crawlers within the topics that the starting URLs are related to. Thus the collections built by existing focused crawling techniques often result in low recall and are not diverse enough to serve the scientific society.

In this research, we studied some major problems in existing focused crawler design, especially the problems caused by using local Web search algorithms. We also proposed to use a meta-search enhanced focused crawling techniques to address the above problems.

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM

The rest of the paper is structured as follows. Section 2 reviews related research on focused crawler techniques and their strengths and weaknesses. Section 3 describes our research questions. Section 4 describes our proposed domain-specific collection building approach. In Section 5, we discuss our evaluation methodology and present some experimental results. In Section 6 we discuss our conclusions and suggest some future directions.

2. LITERATURE REVIEW

In this section, we review previous research on the algorithms used in focused crawlers, the limitations of focused crawlers, and the potential solutions to address these limitations.

2.1 Algorithms Used in Focused Crawlers

Focused crawlers rely on two types of algorithms to keep the crawling scope within the desired domain. *Web analysis algorithms* are used to judge the relevance and quality of the Web pages pointed to by target URLs and *Web search algorithms* determine the optimal order in which the target URLs are visited.

2.1.1 Web Analysis Algorithms

Many different Web analysis algorithms have been proposed in previous studies. In general, they can be categorized into two types: content-based Web analysis algorithms and link-based Web analysis algorithms. Content-based analysis algorithms analyze the actual HTML content of a Web page to obtain relevance information about the page itself. For example, key words or phrases can be extracted from the body text by using document indexing techniques to determine whether the page is relevant to a target domain. Web pages also can be compared to *Standard Documents* that are already known to be relevant to the target domain using the *Vector Space Model* [20]. The Vector Space Model has been used in many existing focused crawlers [1, 14, 19].

Previous studies have shown that the link structure of the Web represents a considerable amount of latent human annotation and offers some important information for analyzing the relevance and quality of Web pages [11]. For example, when there is a direct link from page *A* to page *B*, it often means that the author of page *A* recommends page *B* because of its relevant contents. Moreover, similarly to *Citation Analysis* in which frequently cited articles are considered to be more important, Web pages with more incoming links are often considered to be better than those with fewer incoming links. Co-citation is another concept borrowed from the citation analysis field that has been used in link-based analysis algorithms. Web pages are co-cited when they are linked to by the same set of parent Web pages and heavily co-cited pages are often relevant to the same topic. Co-citation is particularly helpful in finding relevant pages in some domains where pages with similar contents avoid linking to each other (e.g., commercial domains where providers of similar online contents are competitors). The most popular link-based Web analysis algorithms include PageRank [4] and HITS [12].

2.1.2 Web Search Algorithms

Web search algorithms are used in focused crawlers to determine an optimal order in which the URLs are visited. Many different search algorithms have been tested in focused crawling. Among them, *Breadth-first Search* and *Best-first Search* are the two most popular ones. Some other more advanced search algorithms, such as *Spreading Activation* [6] and *Genetic Algorithm* [8], also have been proposed in Web searching.

Breadth-first search is one of the simplest search algorithms used in Web crawling. It does not utilize heuristics in deciding which URL to visit next. All URLs in the current level will be visited in the order they are discovered before URLs in the next level are visited. Although breadth-first search does not differentiate Web pages of different quality or different topics, some researchers argued that breadth-first search also could be used to build domain-specific collections as long as only pages at most a fixed number of links away from the starting URLs or starting domains are collected (e.g., [18, 21]). This method assumes that pages near the starting URLs have a high chance of being relevant. However, after a large number of Web pages are fetched, breadth-first search starts to lose its focus and introduces a lot of noise into the final collection. Other researchers have tried to use breadth-first search and Web analysis algorithms together in focused crawling [10]. In their approach, Web pages are first fetched in a breadth-first order, and then irrelevant pages are filtered from the collection using a Web analysis algorithm. This method can avoid adding irrelevant pages into the final collection. However, since a lot of irrelevant pages are fetched and processed by Web analysis algorithms during the crawling process, this method suffers from low efficiency.

Best-first search is currently the most popular search algorithm used in focused crawlers [1, 10, 13, 14, 19]. In best-first search, URLs are not simply visited in the order they are discovered; instead, some heuristics (usually results from Web analysis algorithms) are used to rank the URLs in the crawling queue and those that are considered more promising to point to relevant pages are visited first. Non-promising URLs are put to the back of the queue where they rarely get a chance to be visited. Clearly, best-first search has advantages over breadth-first search because it "probes" only in directions where relevant pages locate and avoids visiting irrelevant pages. However, best-first search also has some problems. In [2], it has been pointed out that using best-first search the crawlers could miss many relevant pages and result in low recall of the final collection, because best-first search is a *Local Search Algorithm*. By local search algorithm, we mean that best-first search can only traverse the search space by probing neighbors of the nodes previously visited.

In addition to the most popular Web search algorithms, previous studies also introduced some more advanced search algorithms into the focused crawling domain. Chau and Chen [6] used a parallel search algorithm called *Spreading Activation Algorithm* in building domain-specific collections. In their algorithm, the Web is viewed as a Hopfield Net that is a single-layered, weighted neural network. Nodes (Web pages) are visited in parallel and activation relevance judgments from different sources are combined for each individual node until the relevance scores of nodes on the network reach a stable state (convergence). The advantage of this search algorithm is that content-based and link-

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM

based Web analysis algorithms can be effectively combined to avoid many shortcomings of using either one of them alone. Experiment results showed that crawlers using spreading activation algorithm can build domain-specific collections with higher precision and recall than crawlers using breadth-first search or best-first search algorithms. However, as spreading activation algorithm is also a local search algorithm, it shares the limitations of other local search algorithms.

2.2 Limitations of Focused Crawlers

As reviewed above, most existing focused crawlers use local search algorithms in Web searching. While many domain-specific search engines and digital libraries have been built by using focused crawlers, the problems caused by local search have been largely overlooked.

Local search algorithms are algorithms that traverse the search space by visiting the neighbors of previously visited nodes. Using such local search algorithms, a focused crawler will miss a relevant page if there does not exist a chain of hyperlinks that connects one of the starting pages to that relevant page. Furthermore, unless the hyperlinks on the chain all point to relevant pages, the crawler will give up searching in this direction before it reaches the final target. Because of this limitation, crawlers using local search algorithms can only find relevant pages within a limited sub-graph of the Web that surrounds the starting URLs and any relevant pages outside this sub-graph will be ignored, a problem usually referred to as *being trapped with local optimal*.

The shortcomings of using local search algorithms become even more obvious after recent Web structural studies revealed the existence of *Web communities* [9, 11, 12, 22]. Researchers found that Web pages are naturally organized into different groups by special hyperlink structures. Inside such groups, called Web communities, the member pages are all relevant to the same topic of interest. To perform focused crawling is similar to fetching all and only those Web pages that belong to relevant Web communities. However, researchers have found that three structural properties of Web communities make local search algorithms not suitable for building collections for scientific digital libraries.

First, instead of directly linking to each other, many pages in the same Web community relate to each other through co-citation relationships [9, 22]. This is particularly true in the commercial domains where competition is involved. For example, major news agency Websites all provide similar types of information, but they almost never include hyperlinks pointing to each other. In this case, focused crawlers could miss some relevant pages even though they are in the same relevant Web community as the starting URLs.

Figure 1 illustrates the problem described above. As shown in the figure, starting with relevant page, *P1*, a focused crawler is supposed to fetch all the pages in a relevant Web community, *C1*. However, because relevant pages *P5* and *P6* are related to *P1* through co-citation relationships only, these two relevant pages, and all the relevant pages linked to by them, would be missed by the focused crawler.

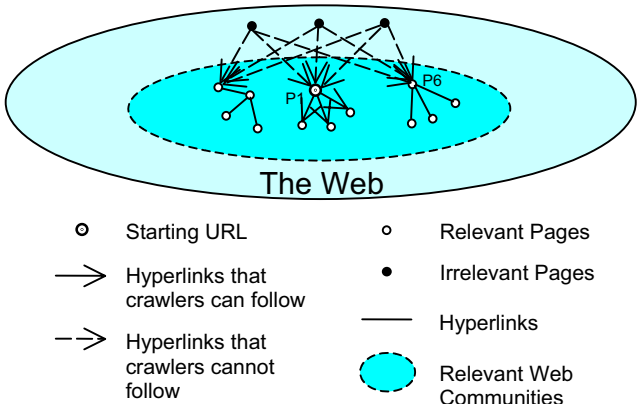

Figure 1: Relevant pages with co-citation relationships missed by focused crawlers

Second, Web pages relevant to the same domain could be separated into different Web communities by irrelevant pages. Through a study of 500,000 Web pages, Bergmark [2] found that most pages that are relevant to the same target domain are separated from at least 1, to a maximum of 12, irrelevant pages. The number of irrelevant pages between two relevant ones commonly is 5. Kumar et al. [15, 16] reported that they identified more than 100,000 distinct Web communities from a large snapshot of the Web, many of them relevant to similar topics. Because focused crawlers using local search algorithms will give up searching when they encounter irrelevant pages, they will not be able to explore relevant Web communities which are separated from the initial communities containing the starting URLs. This problem is illustrated in Figure 2.

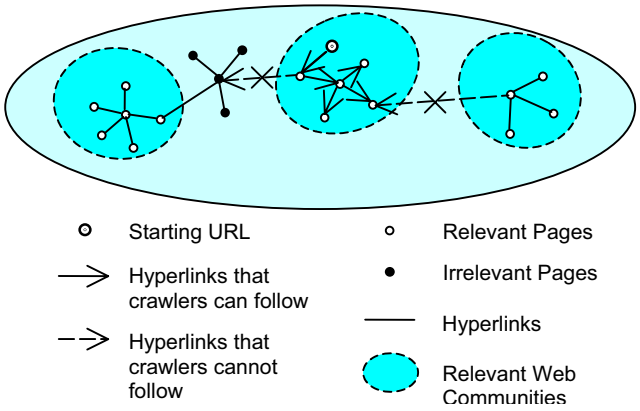

Figure 2: A focused crawler trapped within the initial community

Third, researchers found that sometimes when there are some links between Web pages belonging to two relevant Web communities, these links may all point from the pages of one community to those of the other, with none of them pointing in the reverse direction [22]. For example, consider two Web communities, *F* and *T*, which are relevant to basketball games. Community *F* contains basketball fan club pages and community *T* contains basketball team official Web pages. Intuitively, pages in *F* will contain links pointing to pages in both *F* and *T*. However, pages in *T* may only contain links pointing to other pages in *T*, but no links pointing to pages in *F*. In this case, if the starting URL was in *F*, then relevant pages in *T* could still be

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM

fetched by a focused crawler using local search algorithms. But if the starting URL was in *T*, then relevant pages in *F* would be missed.

2.3 Potential Solutions

As most previous focused crawling studies used local search algorithms, researchers have suggested several strategies to alleviate the problems of local search.

One of the simplest strategies is to use more starting URLs. The assumption is that the more starting URLs one uses in the crawling process, the more comprehensive the final collection will be. However, composing a list of high-quality starting URLs is an expensive and time-consuming task. Also, considering the tremendous size of the Web, the effect of increasing the number of starting URLs could be very limited.

Bergmark [2] proposed to use *Tunneling* technique to address the problems of local search. Tunneling is a heuristic-based method that solves simple global optimization problem. In the focused crawling scenario, a focused crawler using Tunneling will not give up probing a direction immediately after it encounters an irrelevant page. Instead, it continues searching in that direction for a pre-set number of steps. This allows the focused crawler to travel from one relevant Web community to another when the gap (number of irrelevant pages) between them is within a limit. Experiment results showed that focused crawlers using Tunneling can find more relevant pages than those without Tunneling. However, this method cannot completely solve the problem as it does not change the local search nature of focused crawling. Furthermore, Tunneling may introduce noise into the collection and lower efficiency by forcing the crawler to visit irrelevant pages.

Outside the focused crawling domain, some research has provided insights into addressing the problems caused by local search. In their famous study on the size of the Web, Lawrence and Giles [17] found that the overlap between the search indexes of major search engines is actually very small and the combined top results from multiple search engines have high coverage over the Web. They suggested that anyone seeking comprehensive and diverse information about a topic should meta-search multiple search engines and get the combined top results. Although it has not been tested in building domain-specific collections, we believe that meta-searching multiple search engines could be integrated into focused crawling as a potential solution to the problems caused by local search.

3. RESEARCH QUESTION

As mentioned above, most existing focused crawling techniques have difficulty building comprehensive domain-specific collections for scientific digital libraries because they adopted local search algorithms such as best-first search. Several methods have been suggested to alleviate the problem of local search to some extent, but they have not fully addressed the problem. A promising algorithm to address the problem, meta-search, has not been used in focused crawlers before. Thus, in this study, we pose the following research question: How can meta-search be used in focused crawling to build domain-specific collections for scientific digital libraries with higher quality, in terms of precision and recall, when compared with traditional crawling techniques?

The remainder of the paper presents our work in studying this question.

4. PROPOSED APPROACH

To build collections for scientific digital libraries with both high precision and high recall, we propose to use a meta-search enhanced focused crawling approach to address the problems of traditional focused crawling.

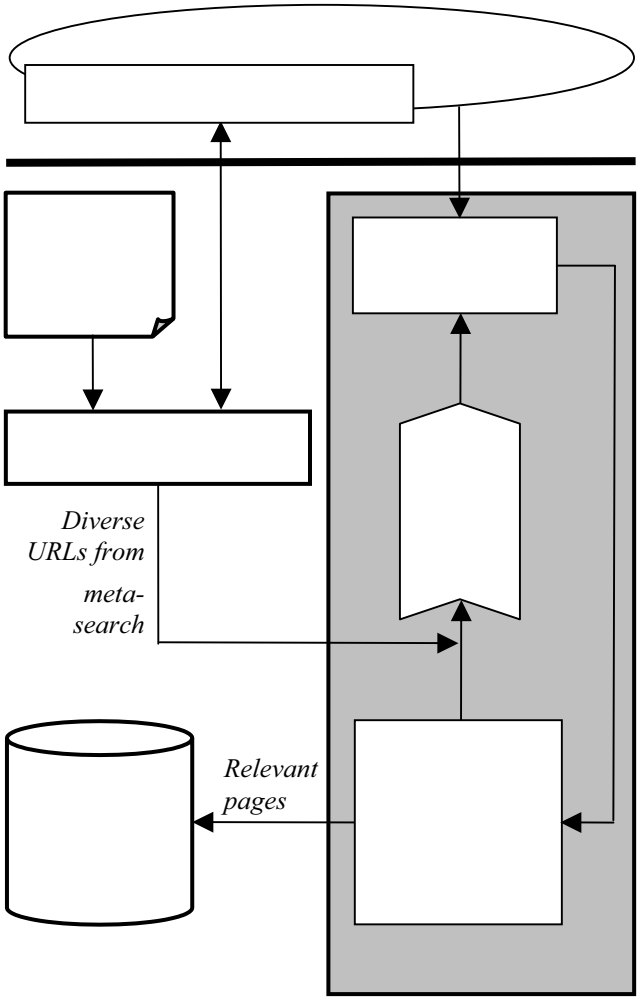

Figure 3: The Meta-search Enhanced Focused Crawling Method

Figure 3 illustrates the idea of the meta-search enhanced focused crawling method. Similarly to traditional focused crawlers, our crawler starts with a set of starting URLs and fetches relevant pages back based on the content- and link-based analysis results. Outgoing links in the relevant pages are extracted and put into the URL queue. At the same time, a meta-searching component keeps drawing queries from a domain-specific lexicon, retrieving diverse and relevant URLs by querying multiple search engines, and combining their top results. Given the fact that the search indexes of different major search engines have little overlap and their combination covers a very large portion of the Web, it is highly likely that the meta-search component retrieves diverse URLs from many different relevant Web communities. These

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM

diverse URLs are then added into the URL queue such that new communities on the Web could be explored by the crawler. Furthermore, as major search engines often include highly co-cited URLs in their search results [4], the meta-searching can make the exploration of individual relevant communities more comprehensive by adding those co-cited URLs into the collection. This combination of meta-searching and focused crawling provides both diversity and relevance for our collection.

Compared to a previously suggested approach which used more starting URLs, the proposed meta-search enhanced approach has advantages. In the proposed approach, only a list of domain-specific queries is required; this is much easier to compose than a list of high-quality starting URLs. Furthermore, the list of domain-specific queries can be updated by adding frequently used queries found in the domain-specific search engine's search log. This will not only make the collection building process easier but also allow the final collection to address the users' information needs more effectively.

The proposed approach also shows advantages over the Tunneling technique. Tunneling technique extends the reach of focused crawlers without changing their local search nature. Tunneling also introduces noise into the collection by forcing the focused crawlers to visit irrelevant pages. By retrieving and combining diverse URLs from multiple search engines, the meta-searching allows the proposed crawler to find new relevant Web communities without any distance limit and does not introduce noise into the collection.

5. EVALUATION

In order to examine the performance of our proposed crawling approach, it was implemented as the backend crawler of a Nanoscale Science and Engineering (NSE) domain-specific Web portal called NanoPort [7]. The speed and scope of NSE development make it urgent for researchers to get timely and high-quality NSE-related information from a domain-specific digital library. The NSE domain also encompasses a diversity of research perspectives and application areas such as nanoscale physics, nanoscale medicine, and nanoscale electronics, which makes comprehensiveness a very important issue in building such a NSE domain-specific digital library. It is thus an ideal domain for testing our proposed meta-search enhanced crawling approach.

Two user evaluation experiments were conducted to evaluate and compare our approach to other existing approaches. In the first experiment, we asked domain experts to judge and compare the precision of the search results from NanoPort to those from two other commercial search engines: Google and NanoSpot (www.nanospot.org). To gain further insights into how the meta-search enhancement could help a focused crawler improve the collection quality, we conducted a second user evaluation experiment in which we built a collection for NanoPort by using traditional focused crawlers and compared the results from this collection to those from the collection built by the meta-search enhanced focused crawler.

Based on our research questions, we aimed to test the following hypotheses:

- H1: When compared to Google and NanoSpot, the retrieval results from NanoPort are of higher precision.
- H2: When compared to the collection built by a traditional focused crawler, the retrieval results from the collection built by a meta-search enhanced focused crawler are of higher precision.

5.1 User Evaluation Experiment 1

5.1.1 Experiment Design

In our first user evaluation experiment, we let domain experts judge and compare the search results from NanoPort to those from two benchmark systems: Google and NanoSpot. We chose these two benchmark systems because Google is currently known as the best general search engine and NanoSpot is currently one of the best NSE domain-specific search engines. The collection of NanoPort was built by a meta-search enhanced focused crawler with 137 expert-selected starting URLs and 387 expert-defined NES-related queries. The final collection contains 996,028 pages and about 1/3 of the pages were obtained through meta-search. Three senior Ph.D. students with NSE-related training were recruited as our domain experts and they provided 22 NSE-related queries of interest to them. The top 20 results for these 22 queries were retrieved from NanoPort and the other two benchmark systems. Then experts were asked to judge whether or not the result pages were relevant to the queries. Then the major measure used to compare the three systems was defined as:

$$\text{Precision} = \frac{\text{Total number of relevant pages in the results}}{\text{Total number of result pages}}$$

5.1.2 Experimental Results

The results on precision are summarized in Table 1. The NanoPort system had a precision of 50.23%, compared with 42.73% and 36.36% obtained by Google and NanoSpot respectively.

Table 1. Results of the First User Evaluation Experiment

	# of relevant results	Total # of results	Precision
NanoPort	221	440	50.23%
Google	188		42.73%
NanoSpot	160		36.36%

The *t*-test results showed that the NanoPort system achieved a significantly higher precision than both Google and NanoSpot (*p*-values are 0.036 and 0.019 respectively) and H1 was supported. While Google achieved a higher precision than NanoSpot in the experiment, the difference was not significant (*p*-value = 0.11).

5.2 User Evaluation Experiment 2

5.2.1 Experiment Design

To gain further insights into how the meta-search could help the crawlers improve the quality of the collection, we conducted a second user evaluation experiment to directly compare the meta-search enhanced focused crawler to a traditional focused crawler. We disabled the meta-search component in our crawler such that it would behave exactly like a traditional focused crawler. Then this focused crawler was used to build a new collection for NanoPort using the same set of 137 starting URLs as those used in experiment 1. This new collection contains 997,632 pages, roughly the same size as the one built by the meta-search enhanced crawler. Two senior Ph.D. candidates with NSE-related training were recruited as our domain experts and each of them

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM

provided 5 NSE-related queries of interest to them. The top 10 results for these 10 queries were retrieved from the two collections using the same retrieval procedure. Then the experts were asked to give each of the result pages a relevance assessment score in the range of 1 to 4, where 4 meant most relevant. Then the average relevance scores of the results from the collections were compared.

5.2.2 Experimental Results

The results from the collection built by the meta-search enhanced focused crawler achieved an average relevance score of 2.77, significantly higher than the score of 2.51 obtained by the results from the collection built by the crawler without meta-search and H2 was supported (p-value < 0.00001). Furthermore, among the total 100 results from the collection built by the meta-search enhanced crawler, 26 were obtained through meta-search and these achieved significantly higher relevance scores (average 3.22) than the rest of the results did (average 2.61) (p-value = 0.000103). In general, the results of the user evaluation suggest that meta-search helped improve the quality of the collection.

6. CONCLUSION AND FUTURE DIRECTIONS

As scientific research domains and the Web are fast evolving, it is difficult to build Web collections with both high precision and high diversity by using traditional focused crawlers. In this research, we proposed a new domain-specific collection building approach, the meta-search enhanced focused crawling, to address the limitations of traditional approaches. We also conducted two experiments to evaluate our proposed approach and got some encouraging results. Our first experiment showed that an NSE domain-specific Web portal built by the proposed approach, NanoPort could provide results with higher precision than benchmarking search engines Google and NanoSpot. Furthermore, our second experiment showed the meta-search component could help crawlers improve the quality of the collections.

Our future work will be carried out in several directions. First, we plan to conduct more experiments in different scientific domains to further validate our approach. We will also investigate other measures that can be used to represent the comprehensiveness of Web collections to make our experiments more meaningful. We will also explore other potential solutions to address the limitations of focused crawling such as integrating global search algorithms into focused crawlers.

ACKNOWLEDGMENTS

This research has been supported in part by the following grants:

- NSF Digital Library Initiative-2 (PI: H. Chen), "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," IIS-9817473, April 1999 – March 2002;
- NSF/NSE/SGER, "NanoPort: Intelligent Web Searching for Nanoscale Science and Engineering," CTS-0204375, February 2002 – November 2002.

We would like to thank Dr. Mihail Roco and Dr. Steve Goldstein of NSF for their support and advices throughout the project. We would also like to thank the AI Lab team members who developed the AI Lab SpidersRUs toolkit and the Meta-Searching Toolkit. Finally, we also want to thank the domain experts who took part in the evaluation study.

REFERENCES

[1] Bergmark, D. (2002a). "Collection Synthesis," in *Proc. of Joint Conference on Digital Libraries 2002,* Portland, Oregon, USA.

[2] Bergmark, D., Lagoze, C. and Sbityakov, A. (2002b). "Focused Crawls, Tunneling, and Digital Libraries", in Proc. of the 6th European Conference on Digital Libraries, Rome, Italy.

[3] Bowman, C., Danzig, P., Manber, U. and Schwartz, M. (1994). "Scalable Internet Resource Discovery: Research Problems and Approaches," *Communications of the ACM*, 37(8), 98-107.

[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).

[5] Chakrabarti, S., van den Berg, M. and Dom, B. (1999). "Focused crawling: a new approach to topic-specific Web resource discovery," in *Proc. of the 8th International World Wide Web Conference*, Toronto, Canada.

[6] Chau, M. and Chen, H. (2003). "Comparison of Three Vertical Search Spiders," *IEEE Computer*, 36(5), 56-62.

[7] Chau, M., Chen, H., Qin, J., Zhou, Y., Sung, W., Chen, Y., Qin, Y., McDonald, D., Lally, A. M. and Landon, M. (2002). "NanoPort: a web portal for nanoscale science and technology" in *Proc. of Joint Conference on Digital Libraries 2002,* Portland, Oregon, USA.

[8] Chen, H., Chung, Y., Ramsey, M. and Yang, C. (1998). "A Smart Itsy-Bitsy Spider for the Web," *JASIS*, 49(7), 604-618.

[9] Dean, J. and Henzinger, M. R. (1999). "Finding Related Pages in the World Wide Web," in *Proc. of the 8th International World Wide Web Conference*, Toronto, Canada.

[10] Flake, G. W., Lawrence, S. and Giles, C. (2000). "Efficient Identification of Web Communities," in *Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, Massachusetts, USA.

[11] Gibson, D., Kleinberg, J. and Raghavan, P. (1998). "Inferring Web Communities from Link Topology," in *Proc. of the 9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, Pennsylvania, USA.

[12] Kleinberg, J. M. (1998). "Authoritative Sources in a Hyperlinked Environment," in *Proc. of the ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, California, USA.

[13] Kleinberg, J. M., Kumer, R., Raghavan P., Rajagopalan, S. and Tomkins, A. (1999). "The Web as a Graph: Measurements, Models, and Methods," in *Proc. of the 5th International Computing and Combinatorics Conference*, Tokyo, Japan.

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM

[14] Kluev, V. (2000). “Compiling Document Collections from the Internet,” *SIGIR Forum*, 34(2).

[15] Kumar, R., Raghavn, P., Rajagopalan, S. and Tomkins, A. (1999). “Extracting Large-Scale Knowledge Bases from the Web,” in *Proc. of the 25th International Conference on Very Large Data Bases Conference*, Edinburgh, Scotland, UK.

[16] Kumar, R., P Raghavan,., Rajagopalan, S. and Tomkins, A. (1999). “Trawlling the Web for Emerging Cyber-Communities,” in *Proc. of 8th International World Wide Web Conference*, Toronto, Canada.

[17] Lawrence, S. and Giles, C. L. (1998). “Searching the World Wide Web,” *Science*, 280(5360):98.

[18] Manber, U., Smith, M., and Gopal, B. (1997). “WebGlimpse: Combining Browsing and Searching,” in *Proceedings of the USENIX 1997 Annual Technical Conference*, Anaheim, California, Jan 1997.

[19] McCallum, A., Nigam, K., Rennie, J. and Seymore, K. (1999). “Building Domain-Specific Search Engines with Machine Learning Techniques, “ in *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*.

[20] Salton, G. (1989). “Automatic Text Processing,” MA: Addison-Wesley, 1989.

[21] Sumner, R. G., Jr., Yang, K., and Dempsey, B. J. (1998). “An Interactive WWW Search Engine for User-defined Collections,” in *Proceedings of the 3rd ACM Conference on Digital Libraries*, Pittsburgh, Pennsylvania, USA, Jun 1998, pp. 307-308.

[22] Toyoda, M. and Kitsuregawa, M. (2001). “Creating a Web Community Chart for Navigating Related Communities,” in *Proc. of ACM Conference on Hypertext and Hypermedia*, Århus, Denmark.

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL’04)
1-58113-832-6/04 $ 20.00 © 2004 ACM