

Effects of Start URLs in Focused Web Crawling

Ari Pirkola and Tuomas Talvensaari

University of Tampere, Finland

ari.pirkola@uta.fi, tuomas.talvensaari@uta.fi

INFORUM 2009: 15th Conference on Professional Information Resources
Prague, May 27-29, 2009

Abstract.

Web crawling refers to the process of gathering data from the Web. Focused crawlers are programs that selectively download Web documents (pages), restricting the scope of crawling to a pre-defined domain or topic. The downloaded documents can be indexed for a domain specific search engine or a digital library. In this paper, we describe the focused crawling technique, review relevant literature, and report novel experimental results. Crawling is often started with URLs that point to the pages of central North-American and European universities, research institutions, and other organizations in North-America and Europe. In the experiments we investigated, first, how strongly this central region of the Web is connected to three 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). Test topics were selected from the domains of genomics and genetics which are typical scientific fields. We found that two focused crawling processes, one started from the central region and the other from the region of Australia / China / South-America, overlap only to a small extent, identifying mainly different relevant documents. Document relevance was assessed (1) by a human judge and (2) by assigning probability scores to documents using a search engine. Second, we investigated the coverage (number) of relevant documents obtained for different focused crawling processes started with URLs from the four different geographical regions. The results showed that all regions considered in this study are good starting points for focused crawling in the domains of genetics and genomics since each of them yielded a high coverage. As genomics and genetics are typical scientific domains we assume the obtained results to be generalizable to other scientific domains. We discuss what implications the observed results have for the selection of crawling approach in scientific focused crawling tasks.

1. INTRODUCTION

Web crawling refers to the process of gathering data from the World Wide Web. *Focused crawlers* are programs that selectively download Web documents (pages), restricting the scope of crawling to a pre-defined domain or topic (Castillo, 2004; Chakrabarti et al., 2002; Pirkola and Talvensaari, 2009; 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.

Web crawling usually starts with a set of start (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. Focused crawlers differ from general crawlers in that they judge whether the documents pointed to by the URLs

are relevant for the domain or topic in question. The URL queue is ordered based on the relevance probabilities, and the pages assessed to be very relevant for the specific domain or topic are downloaded first.

The selection of a start URL set is a crucial step in FC. Despite this it has remained an unexplored area in FC research. A common practice is to retrieve the start URLs by a Web search engine or from a subject directory. 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, first, how strongly this central region is connected to other large geographical regions of the Web and, second, the coverage (number) of relevant documents obtained for different FC processes started with URLs from the different geographical regions. We discuss what implications the observed results have for the selection of crawling approach in FC. The results are based on 80 different crawls in the domains of *genomics* and *genetics*.

The rest of this paper is organized as follows. Section 2 reviews literature on focused crawling. Section 3 presents the methodology and data. Research questions and evaluation measures are presented in Section 4. Section 5 contains the findings and Section 6 discussion and conclusions.

2. RESEARCH ON FOCUSED CRAWLING

A central issue in FC is how to identify links and pages that are relevant to the specific domain or topic in question. Here domain specific words are a necessary resource. Besides these, approaches to domain identification and the ordering of the URLs in the URL queue include the determination of the authoritativeness or popularity of pages, the utilization of the hierarchical structure of Web documents, and the application of formal models, such as Context Graphs.

The *HITS algorithm* (Kleinberg, 1998) searches for authoritative pages based on the number of links pointing to pages. The popularity of documents is usually determined using the well-known PageRank algorithm (Brin and Page, 1998). PageRank rewards documents that are pointed to by documents that themselves are popular documents.

Chakrabarti et al. (2002) utilized the *Document Object Model* (DOM) of Web pages in their FC. A DOM tree represents the hierarchical structure of a page: the root of the tree is usually the HTML element which typically has two children, the HEAD and BODY elements, which further are divided into sub-elements. The leaf nodes of the DOM tree are typically text paragraphs, link anchor texts, or images. In addition to the usual bag-of-words representation of Web pages, the FC proposed by Chakrabarti et al. (2002) represented a hyper-link as a set of features $\langle t, d \rangle$ where t is a word appearing near the link, and d its distance from the link. The distance is measured as the number of DOM tree nodes that separates t from the link. These features were used to train a classifier to recognize links leading to relevant pages. The links with low distance to relevant text are considered to be more important than links that are far from the relevant text.

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. In the beginning of crawling a set of start URLs is entered in the FC program. Pages that point to the start URLs, i.e., parent pages, and their parent pages (etc.) form a context graph. The context graphs are used to train a classifier with features of the paths that lead to relevant pages.

In Talvensaaari et al. (2008) we used focused crawling as a means to acquire German-English and Spanish-English comparable corpora in biology. The acquired corpora were aligned at a paragraph level, and the alignments were employed in statistical translation for cross-language information retrieval. The technique was capable of providing correct translations for most of the words used in the translation experiments.

3. METHODS AND DATA

3.1 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.

3.2 Test Topics and Start URLs

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 TREC Genomics Track, see (Hersh et al., 2005). There were also five general topics. They were created by one of the authors who has worked in the field of medical informatics. For start 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. The test topics, queries, and the number of Medline hits are presented in Appendix.

To minimize the effects of the particular crawling program and to increase the reliability of results two experiments with different crawling strategies were carried out. The crawling program and the two experiments are described in a separate subsection below (Section 3.3).

In both experiments four crawls were performed for each topic with the start URLs from (1) the Major region, (2) Australia, (3) China, and (4) five South-American countries (Argentina, Brazil, Chile, Mexico, and Uruguay). Each start URL set contained 50 URLs. The start URLs of the Major region were retrieved by means of the basic Google (<http://www.google.com>) whereas the start URLs 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 start URL pages were bilingual Chinese-English pages.

The majority of the Major region start URLs were of the type *.com*, *.edu*, *.gov*, *.org*, *.de*, and *.uk*. The original Major region start 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 4).

Figure 1 illustrates our experimental approach. Crawling was started from four regions: Major, Australia, China, and South-America. The target category of *other* includes regions not used as *start URLs* in this study, e.g. African countries, as well as indeterminate TLDs. Africa was not considered because we were not able to obtain 50 African start URLs for all topics. For each four start URL region, the FC program fetched pages from all five target regions. There were 40 start URL sets in total: 10 topics and for each topic four start URL sets representing the four different start URL regions. Accordingly, we performed 40 crawls in both experiments, and 80 crawls in total. In the first experiment, crawling was stopped for each crawl after 20 000 pages had been downloaded. In the second experiment, crawling was stopped after 40 000 pages had been downloaded. Thus, each result list contained either 20 000 or 40 000 pages.

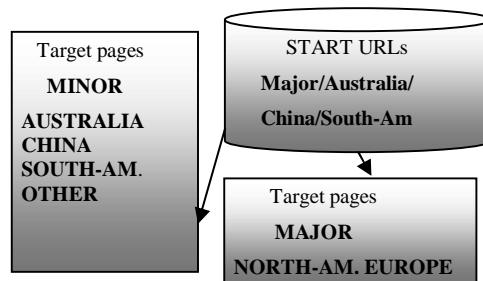


Figure 1. Experimental approach

3.3 Crawling Program and Crawling

The two experiments differed in the points shown in Table 1. A detailed description of the crawling program and methods used in crawling are described below.

Table 1. Differences between the two experiments

Feature	First Experiment	Second Experiment
Method to identify relevant documents	A test classifier was trained	Query-document matching
Probabilities were assigned by	Terrier search engine	Lemur search engine
# downloaded documents	20 000	40 000

In the first experiment, the *Nalanda iVia Focused Crawler* (<http://ivia.ucr.edu>) was used. It is based on the work of Chakrabarti et al. (1999). 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 $\langle u, v \rangle$ 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* (Zhang et al., 2003) 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 start URLs 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 negative examples were manually examined to prune out positive pages of other topics that were also relevant to the topic in question. The number of negative examples was also 300.

In the first experiment, 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 start 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.

In the second experiment, a modified version of the Nalanda crawler was used. Instead of a text classifier, the probabilities were assigned directly using a search engine. We used the *Lemur search engine* (<http://www.lemurproject.org/>) and specifically its structured query mode. 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 experiment, the context of each link $\langle u, v \rangle$ 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 $\langle u, v \rangle$. However, the context probability was only used 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 query-based approach has some advantages over the classifier-based approach. Most importantly, 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.

4. RESEARCH QUESTIONS AND EVALUATION

Relying only on probabilities without looking at documents is not a very strong approach to assess document relevance, because there is no *anchor point* that would anchor probabilities to the actual relevance of documents. On the other hand, human assessment of the relevance of documents of long result lists would be a tremendous task. Our solution to this widely known dilemma was to take samples of documents at different positions of the ranked result lists, to assess the relevance of the sample documents, and to select reasonable probability thresholds on the basis of the assessment. Documents above the selected thresholds were taken for the evaluation of the crawling results. The selected probability thresholds were prob. > 5.0 for Terrier and prob. > 0.45 for Lemur. In the latter case, probability scores are in the range of 0.4 – 1.0, where 0.4 is zero probability (scores > 0.7 do not appear very often). On average, 83% of documents above these thresholds were relevant. This percentage is based on the assessment of 500 documents. The relevance of documents was evaluated by one of the authors who has worked in the field of medical informatics and genomics information retrieval. Documents that discussed the topic, contained facts about the topic (e.g. database entries and Web forms), or contained relevant links or literature references were considered relevant. Our evaluation

approach can be considered to be satisfactory because in this manner we were able to anchor the probabilities.

We evaluated the crawling results using the following two measures: (1) *coverage rate for a region* and (2) *overlap rate between the Major region and each Minor region*. Both measures were measured above the two relevance probability thresholds mentioned above, assigned by Terrier and Lemur to the downloaded pages. Moreover, to provide stronger results both (1) and (2) were also measured for all documents with probabilities higher than zero.

(1) *Coverage rate for a region* refers to the number of downloaded documents. It was measured for all regions. We expect the Major region to receive the highest coverage. We are interested in the magnitude of coverage of each Minor region. For example, do Chinese start URLs give approximately the same number of probably relevant documents (i.e., documents with probabilities above the selected thresholds) as the Major region start URLs or, for example, ten times less?

(2) *Overlap rate* refers to the percentage of identical URLs downloaded for the start URL regions Major and Minor. By measuring overlap rate between the Major region and each Minor region we will find out how strongly the Major region is connected to the other geographical regions of the Web. 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. We expect the overlap rates to be low because, rather than being a true web, the Web is a community of communities that are isolated or only loosely connected to each other (Kleinberg, 1998; Toyoda et al., 2001). It is therefore likely that two FC processes starting from two remote regions find pages from different communities.

5. FINDINGS

The results of the two coverage experiments are shown in Tables 2 and 3. In the first experiment, the Major region received the highest coverage. Then come the Australian and Chinese regions. The South-American region received the lowest coverage. For example, in the case of the first experiment, general topics and prob. > 5.0 coverage rates are, respectively, 2501, 934, 853, and 361 (Table 2). In other situations the Chinese region received a higher coverage than the Australian region. The coverage figures obtained in the second experiment are somewhat different. Now the Major region does not clearly outperform the other regions. Overall, both tables show that all four regions are good starting points for FC in the fields of genomics and genetics. These are typical scientific domains, therefore the obtained results are assumed to be generalizable to other scientific domains. It should be noted that we are now talking about crawling English documents. In the case of, for example Spanish FC, the results would probably be quite different.

The results of the overlap rates in the first experiment are reported in Table 4. In each case the first column shows the absolute number 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, Major and Australia with prob. > 0.0, gave 21 575 pages in total. Of these pages 1.2% (N=267) shared the same URL. As shown, in all cases the overlap rates are very low, 1.4% or less. The figures shown in Table 5 are quite similar. Again, overlap rates are low. In the case of specific topics / prob. > 0.45 the figures are from 7.1% to 9.7%, but in other cases no more than 5.0%. 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.

High coverage rates of Minor regions and low overlap rates show that FC starting with URLs from the Major region only, loses a considerable number of relevant pages. On the other hand, the results show that crawling starting from different geographical regions will receive a high coverage. Therefore, when the aim is to build a large topic-specific document collection, one should use several crawling processes with geographically, or otherwise different, start URLs.

Table 2. Coverage rates for the Major and Minor regions. The first experiment. For each cell the number of downloaded documents is 100 000 (5 topics x 20 000 documents).

Topic type and probability threshold	Major	Australia	China	S-Amer.
General, prob. > 0.0	12937	8638	10405	6389
General, prob. > 5.0	2501	934	853	361
Specific, prob. > 0.0	12607	6238	8620	6319
Specific, prob. > 5.0	1480	869	1357	434

Table 3. Coverage rates for the Major and Minor regions. The second experiment. For each cell the number of downloaded documents is 200 000 (5 topics x 40 000 documents).

Topic type and probability threshold	Major	Australia	China	S-Amer.
General, prob. > 0.0	110475	107104	108006	110952
General, prob. > 0.45	9956	11956	12797	11523
Specific, prob. > 0.0	100371	96327	95983	99062
Specific, prob. > 0.45	8599	6263	8512	8497

Table 4. Overlap rates (%) in the first experiment.

Topic type and probability thr.	Major+ Australia	Major+ Australia	Major+ China	Major+ China	Major+ South-America	Major+ South-America
	N	Overlap%	N	Overlap%	N	Overlap%
General, prob. > 0.0	21575	1.2	22802	1.4	19326	0.3
General, prob. > 5.0	3435	0.7	3354	0.9	2862	0.2
Specific, prob. > 0.0	18845	0.2	21227	0.6	18926	0.0
Specific, prob. > 5.0	2349	0.0	2837	0.4	1914	0.0

Table 5. Overlap rates (%) in the second experiment.

Topic type and probability thr.	Major+	Major+	Major+	Major+	Major+	Major+
	Australia N	Australia Overlap%	China N	China Overlap%	South- America N	South- America Overlap%
General, prob. > 0.0	217579	2.6	218481	2.2	221427	2.6
General, prob. > 0.45	21912	4.5	22753	3.6	21479	5.0
Specific, prob. > 0.0	196698	3.9	196354	3.4	199433	3.4
Specific, prob. > 0.45	14862	9.7	17111	9.0	17096	7.1

6. DISCUSSION AND CONCLUSIONS

The selection of start URLs is a crucial step in focused crawling. Despite this it has remained an unexplored area in FC research. Daneshpajouh et al. (2008) developed a start URL extraction algorithm for general Web crawling where the aim is to download a large number of Web pages with high PageRank scores. These results are not directly applicable to focused crawling where PageRank is, at most, of secondary importance.

FC is a useful approach to Web information retrieval and a necessary method to build domain and topic specific document collections. However, FC has its limitations. It is likely that even the most effective FC program misses a considerable number of relevant documents. This is because FC is only capable of finding documents that, through some routes, are linked to the starting point, i.e., the set of start URLs entered into the FC program, and documents outside these routes are missed. The start URL set used in crawling greatly affects the crawling results.

In this paper we investigated the selection of start URLs for FC from the geographical point of view of the Web. We investigated the connections between different geographical regions of the Web and the coverage rates of documents obtained for different focused crawling processes started from the different geographical regions. Queries were formulated on the basis of the test topics dealing with genomics and genetics, and start URLs were retrieved from four regions of the Web: North-America and Europe (Major region), and Australia, China, and South-America (Minor regions). Using URLs from the four regions as starting points documents were downloaded from the Web using the *Nalanda iVia Focused Crawler*. To minimize the effects of the crawling program and to increase the reliability of results two experiments with different crawling methods were performed.

The results showed, first, that all Minor regions are good starting points for FC since each of them yielded a high coverage. The results also showed that two FC processes, one started from the Major region and the other from a Minor region (Australia / China / South-America), overlap only to a small extent. This is reasonable and expected since the Web is a community of communities that are isolated or only loosely connected to each other (Kleinberg, 1998; Toyoda et al., 2001), and two FC processes started from two remote regions can be expected to mainly find pages from different communities.

High coverage rates of Minor regions and low overlap rates between the Major region and Minor regions show that FC starting with URLs from the Major region only, loses a considerable number of pages inside the Major and Minor regions. On the other hand, the results indicate that several crawling processes starting from different geographical regions will receive a high coverage if the crawling results are combined. The domains of genomics and genetics are typical scientific domains. We therefore assume the obtained results to be generalizable to other scientific domains.

An alternative method to an approach where several crawling processes with geographically different start URL sets are used to obtain a high coverage, would be the use of a geographically heterogeneous start URL set. Overall, the results of this study suggest that the effectiveness of FC can be improved remarkably if crawling is based on different types of start URL pages (categorized, for example, on the basis of geographical regions, organizations, and top level domain names). One important question is what types of combinations of start URLs will yield the best crawling performance. These are some issues that future research could shed light on.

7. ACKNOWLEDGMENTS

This study was funded by the Academy of Finland (research projects 119600, 124630, 125679 and 129835).

8. REFERENCES

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

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.

Daneshpajouh, S., Nasiri, M. and Ghodsi, M. 2008. A fast community based algorithm for generating Web crawler seeds set. *Proceedings of 4th International Conference on Web Information Systems and Technologies, WEBIST*, Funchal, Portugal, May 4-7, pp. 98-105.

Diligenti, M., Coetzee, F.M., Lawrence, S., Giles, C.L. and Gori, M. 2000. Focused crawling using context graphs. *Proceedings of 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

Kleinberg, J. 1998. Authoritative sources in a hyperlinked environment. *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 668–677.

Pirkola, A. and Talvensaaari, T. 2009. Effects of crawling strategies on the performance of focused Web crawling. *WEBIST – 5th International Conference on Web Information Systems and Technologies*. Lisbon, Portugal, March 23-26, 2009.

Talvensaaari, 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.

APPENDIX

GENERAL TOPIC 1

Description: Find information on mouse gene mutations

Query: (mouse OR mice OR "mus musculus") AND (gene OR genes) AND (mutation OR mutations)

#Medline_hits: 55 382

GENERAL TOPIC 2

Description: Find information on human genome

Query: human (genome OR genomics)

#Medline_hits: 352 024

GENERAL TOPIC 3

Description: Find information on hereditary diseases

Query: hereditary (disease OR diseases)

#Medline_hits: 28 074

GENERAL TOPIC 4

Description: Find information on chromosomal abnormalities

Query: (chromosomal OR chromosome OR chromosomes) (abnormality OR abnormalities OR deletion OR deletions OR duplication OR duplications OR translocation OR translocations OR inversion OR inversions)

#Medline_hits: 121 374

GENERAL TOPIC 5

Description: Find information on genetic engineering

Query: "genetic engineering"

#Medline_hits: 18 961

SPECIFIC TOPIC 1 (TREC TOPIC 1)

Description: Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans.

Query: ("ferroportin 1" OR ferroportin1 OR slc40a1 OR fpn1 OR hfe4 OR ired1 OR mtp1 OR slc11a3) (human OR humans)

#Medline_hits: 175

SPECIFIC TOPIC 2 (TREC TOPIC 10)

Description: NEIL1. Find articles about the role of NEIL1 in repair of DNA.

Query: (neil1 OR "nei endonuclease" OR flj22402 OR fpg1 OR nei1 OR hfpg1) dna (repair OR lesion OR lesions)

#Medline_hits: 54

SPECIFIC TOPIC 3 (TREC TOPIC 20)

Description: Substrate modification by ubiquitin. Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins?

Query: ubiquitin ("covalent attachment" OR "covalent binding" OR "isopeptide bond" OR ubiquitination OR ubiquitylation) substrate

#Medline_hits: 855

SPECIFIC TOPIC 4 (TREC TOPIC 30)

Description: Regulatory targets of the Nkx gene family members. Documents identifying genes regulated by Nkx gene family members.

Query: nkx gene (regulate OR regulates OR regulation OR regulatory)

#Medline_hits: 217

SPECIFIC TOPIC 5 (TREC TOPIC 40)

Description: Antigens expressed by lung epithelial cells. To identify the antigens expressed by lung epithelial cells and the antibodies available.

Query: (antigen OR antigens) (lung OR lungs OR pulmonary OR bronchus OR bronchi OR bronchial) (epithelium OR epithelial)

#Medline_hits: 7 067