The Method of Improving the Specific Language Focused Crawler

Shan-Bin Chan
Graduate School of Fundamental Science and Engineering
Waseda University
chrisjan@yama.info.waseda.ac.jp

Hayato Yamana
Graduate School of Fundamental Science and Engineering
Waseda University
yamana@yama.info.waseda.ac.jp

Abstract

In recent years, more and more CJK (Chinese, Japanese, and Korean) web pages appear in the Internet. The information in the CJK web page also becomes more and more important. Web crawler is a kind of tool to retrieve web pages. Previous researches focused on English web crawlers and the web crawler is always optimized for English web pages. We found that the performance of the web crawler is worse in retrieving CJK web pages. We tried to enhance the performance of the CJK crawler by analyzing the web link structure, anchor text, and host name on the hyperlink and changing the crawling algorithm. We distinguish the top-level domain name and the language of the anchor text on hyperlinks. The method that distinguishes the language of the anchor text on hyperlinks is not used on CJK language specific crawler by other researches. Control experiment is used in this research. According to the experimental results, when the target crawling language is Japanese, the 87% of the crawled web pages are Japanese web pages and improves the efficiency about 0.24% compares to the baseline results. When the target crawling language is Chinese, the 88% of the crawled web pages are Chinese web pages and improves the efficiency about 0.07% compares to the baseline results. When the target crawling language is Korean, the 71% of the crawled web pages are Korean web pages and improves the efficiency about 10% compares to the baseline results.

1 Introduction

The growth of web pages in the internet becomes rapidly in recent years. How to efficiently collect web pages and how to gather more language or topic relative web pages become important. Focused crawling is a kind of method that collects topic specific web pages. (Chakrabarti et al., 1999) Intelligent crawling that can self-learning and predicating makes the focus crawling more efficient. (Chara et al., 2001) Mining for patterns and relations over text, structures, and links is an interesting research. (Neel et al., 2001) In the past few years, the researches are focused on the topic specific focused crawling and optimized the performance of focus crawler for crawling English web pages. Web pages are not only described in English but also in other languages. Our research will be emphasized on the study of language specific focused crawler and how to optimize the crawler for specific language (For example: Chinese, Japanese, and Korean).

Huge amounts of hyperlinks on the CJK web pages link to English web pages. But the hyperlinks on the English web pages almost don't link to the CJK web pages. If we deeply crawl the web pages from CJK seed sets, finally we will gather many English web pages. In this kind of situation, the efficiency of the CJK focus crawler is very worse.

Our research method is that the first we extract the domain name from the hyperlink URL and then determine the top-level domain. For example, if we try to focus crawl the Japanese web pages and the top-level domain in the hyperlink URLs is .jp, this focus crawler will

queue these URLs for the next crawling. If the top-level domain in the hyperlink URLs is not .jp, we will distinguish the language of the anchor text of the hyperlink. If the language of the anchor text is Japanese, we also queue these URLs for the next crawling. Otherwise, we drop the URLs.

This research uses the Nutch as the crawler and uses the Hadoop as the storage. Because of the web pages is enormous, Hadoop is a very efficient tool that can store and process vast amounts of data. We choose the URLs on the DMOZ as the seeds set and extract the URLs by the top-level domain name .cn, .tw, .jp, .kr, .com, .net. After extracting the URLs, we sort these URLs by language (Chinese, Japanese, and Korean). We use these sorted URLs as the seeds set for crawling by Nutch.

The experiment method in our research is control experiment. We divided our experiment to two groups. One is control group and the other one is experimental group. The crawling methods of the control group use the default crawling algorithm in Nutch. The crawling methods of the experimental group use the modified crawling rules provided by us. After crawling by both control group and experimental group, we compare the crawled web pages between baseline results and experimental results. Finally, the results show that we can improve the crawling efficiency by using our modified method.

2 Related Researches

2.1 Web data collection of specific topic

The research of focus crawler is applied in the medical information science. (Thanh et al., 2005) Their research mentions that there are relationships between the hyperlink URLs and the 50 words before and after hyperlinks with the contents in the crawling target pages. And they also mention that if the URL filters are applied in the breadth-first crawling, The crawled results will be better then without URL filters.

2.2 Collection methods of specific language web pages

Tamura (2006) introduces the research of specific language web pages focus crawler. Their research is focused on the collection of Thai language web pages. When crawling, the link selection methods of their research are descript as below:

1. When the authors collect web page j on web server i, they distinguish the language of web page j.
2. If the language of the web page j is Thai, Nr(i) increases 1. (Nr(i) is Thai language web page counts of web server i)
3. Na(i) increases 1. (Na(i) is collected page counts of web server i)
4. Extract the hyperlinks in the web page j.
5. Drop the hyperlink that top-level domain is not Thai.
6. Drop the hyperlink that has been crawled.
7. If the language of web page j is Thai, queue the remained hyperlinks to high priority.
8. If the language of web page j is not Thai, queue the remained hyperlinks to low priority.

The research data set is collected from famous portal web pages in Thailand from July to August in 2004. They download 18,344,217 HTML pages from 574,111 servers, and use this data set for simulating their selective collection method. The estimation methods are as below:

- Server-based filtering, aggressive: No limits when the crawler chooses the new web server.
- Server-based filtering, conservative: Only Thai servers when the crawler chooses the new web server.
- Directory-based filtering, conservative: Only Thai directory when the crawler chooses the directory on the server.
- Hard focused: Drop all the hyperlinks when the web page is not Thai language.
- Soft focused: If the web page is Thai language, queue the hyperlinks with high priority. If the web page is no Thai language, queue the hyperlinks with low priority.
- BFS: Breath-first search.
- Perfect: First, the crawler uses the breath-first search. When the crawler

found a hyperlink, which is Thai, web page, the crawler start to follow this Thai hyperlink and crawl the web pages from this URL.

The result of the estimation method "Perfect" is obvious. While crawling, when the total amount of crawled web pages increase, the cumulative Thai web page ratio increase smoothly from 92% to 99%.

2.3 Seeds set generation method for web crawler

HITS algorithm is used to generate crawler seeds set. (Shervin et al., 2003) Their research considered that it's better to crawl the most important web pages on the resource limited internet. They use the collected web pages to draw a web graph and perform the HITS algorithm to generate seeds set for crawler.

2.4 Focus crawling for dark web forums

Dark web forum is a kind of forum that the contents are associated with cybercrime, hate, and extremism. Fu (2010) developed a focus crawler for crawling dark web forums by using language-independent features, including URL tokens, anchor text, and level features. They also use forum software-specific traversal strategies and wrappers to support incremental crawling. Their system maintains up-to-date collections of 109 forums in multiple languages. Their focus crawler gathers contents from three regions, which are U.S. domestic supremacist, Middle Eastern extremist, and Latin groups. Their human-assisted accessibility mechanism can access identified forums with a success rate of over 90%.

2.5 The differences between previous studies

The previous researches almost focus on the English web page crawlers or the topic specific crawlers. There are few researches about the CJK language specific focus crawlers. The method that judges the top-level domain name of the URLs on hyperlinks has been studied on crawling Thai web pages. And the crawling method of judgment the language of an anchor text on the hyperlink has not studied by other researchers. We consider that if we combine the top-level domain name judgment method and anchor text language distinguish method, we will enhance the efficiency more on crawling CJK web pages. So, we use this combined method to enhance the performance of CJK language specific focus crawler.

3 Methods

The probability is very high that hyperlinks on the CJK web pages link to English web pages. And the probability is very low that hyperlinks on the English web pages link to CJK web pages. If we crawl the web pages deeply from CJK seed sets, finally we will gather huge amounts of English web pages.

We implement the control experiment to resolve the problems that mentioned above. We split our research to 5 steps and show the process in the Figure 1.

Before crawling, a seeds set is very important for the crawler. DMOZ (http://www.dmoz.org) is the biggest web directory service in the world. We use the URLs in the DMOZ as the seeds set. Our URLs extraction method shows as follow.

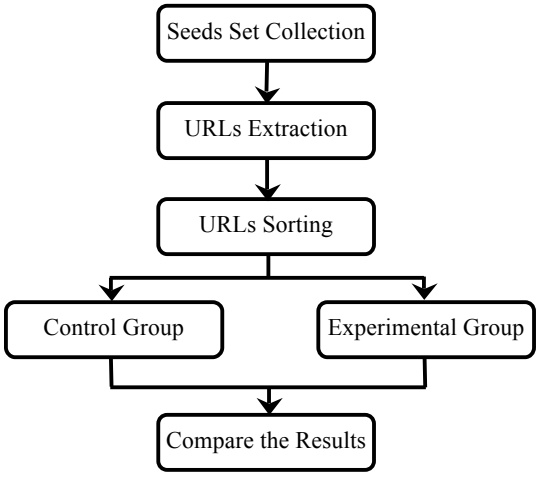

Figure 1. Research Process

- Download the XML formatted web directory data from DMOZ homepage on October 26th 2009.
- Extract URLs form XML formatted web directory data.
- Sort the URLs by top-level domain. (.cn, .tw, .jp, .kr, .com, .net, etc…)

- Choose the .cn, .tw, .jp, .kr, .com, .net top-level domains for language distinguish.
- Use the Perl Lingua::LanguageGuesser which produced by Nakagawa Laboratory of Tokyo University to distinguish the language (Chinese, Japanese, and Korean) of each web pages in the sorted URLs. Perl Lingua::LanguageGuesser is a language distinguisher that is based on N-Gram text categorization. (William et al., 1994)
- Use these sorted seeds sets and perform crawling by using Nutch.

We implement our research by separating into two groups, control group and experimental group. The control group follows the default crawling rules supported by Nutch. And we change the crawling rules in experimental group by importing a URLs queuing replacement plug-in into Nutch. We will explain the modified URLs queuing rules by using the Chinese web pages collection procedures.

- If the top-level domain in the URL is .cn, store this URL to the queue. The reason is that it is high probability that the web page with .cn domain name is a Chinese web page.
- If the top-level domain in the URL is not .cn, distinguish the language of the anchor text on the hyperlink. Then, if the anchor text is Chinese, store the URL to the queue. The reason is that it is high probability that the web page which hyperlink with Chinese anchor text links to is a Chinese a web page.
- Drop the URLs from queue when other situations.

At the final stage, calculates the percentage of each language and compares the results between baseline results and experimental results.

4 Results

URLs extracted from DMOZ that with .com domain is 1,964,053, .net domain is 182,595, .jp domain is 130,125, .cn domain is 14,769, .tw domain is 10,259, .kr domain is 4,910. Figure 2 shows the percentage of each domain.

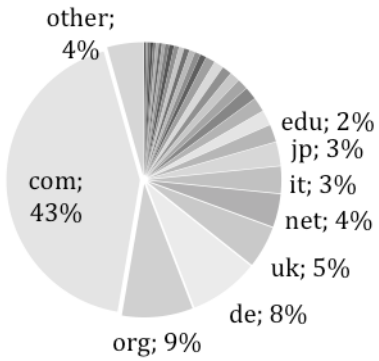

Figure 2. Distribution of DMOZ top-level domains

CJK web page counts in each top-level domain (.com and .net, .cn, .tw, .jp, .kr) are shown in Table 1. We extract 43,216 Chinese URLs, 175,666 Japanese URLs, and 5,252 Korean URLs. We randomly pick 1,000 URLs for each language from these CJK URLs and start to crawl by using these URLs as seeds sets.

We write two functions into URLs queuing replacement plug-in on Nutch. One is html text language distinguisher, and the other one is web page counter. We use the same language distinguisher that supported by Nakagawa Laboratory of Tokyo University with seeds set extraction stage.

Domain	CN	JP	KR
com&net	24,851	56,256	1,975
cn	11,937	40	1
tw	6,220	573	16
jp	147	118,729	14
kr	61	68	3,246
Total	43,216	175,666	5,252

Table 1. CJK web page counts from each top-level domain

Figure 3 shows the crawling process of this research. The original Nutch crawl process is that crawl the web pages from seeds set, parse the html text, extract the URLs from hyperlinks, store the URLs to the queue, and implement the breadth-first crawling. In order to recording the language of URLs, after parsing the html text, we add a Nutch plug-in to judges the language of the html text. Then write the URL of this web page to a language specific file (Chinese, Japanese, Korean, Other) for counting. And then extract all the hyperlinks and store the URLs to the

queue. Control group implements the process described above.

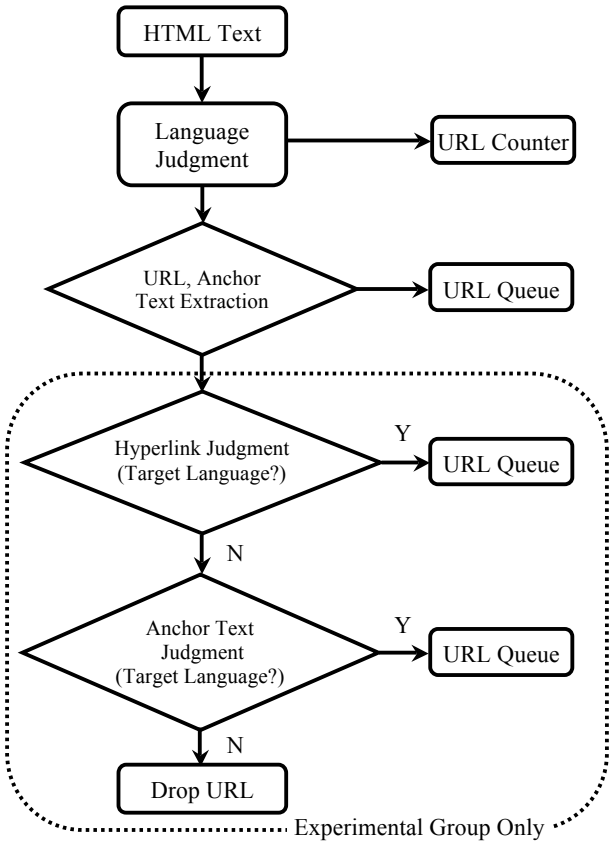

Figure 3. Crawling process

Experimental group also implements the process of control group. After the control group process, we add top-level domain judgment on hyperlinks. When we are focusing on crawling Chinese web pages, and if the top-level domain on hyperlinks is .cn, we store the URL of this hyperlink to the queue. If the top-level domain on hyperlinks is not .cn, we check the anchor text of the hyperlink. If the language of the anchor is Chinese, we also store the URL of this hyperlink to the queue. Otherwise, drop the URLs from queue. We don't prioritize the URLs in queue in this research. In order to not crawling the web pages that are not Chinese, we decide to drop these URLs that may not be Chinese web pages.

We pick a Chinese web page as the crawling example for experimental group. The crawler chooses a URL "http://www.bsc.org.cn/" in the sorted DOMZ seed sets. First, we parse the HTML text of this URL and extract the hyperlinks and anchor text from HTML text. We get a part of the hyperlink and anchor text shown in Table 2.

Anchor text	Hyperlinks
中国科学技术协会	http://www.cast.org.cn
中国科学院生物物理研究所	http://www.ibp.ac.cn
IUPAB	http://www.iupab.org
Asian Biophysics Association	http://www.aba-bp.com
美国生物物理学会	http://www.biophysics.org
Protein and Cell	http://www.protein-cell.org
…	…

Table 2. A part of anchor text and hyperlinks extracted from "http://www.bsc.org.cn/"

Second, according to the URL queuing rules mentioned above, we show a part of matched hyperlinks and anchor text shown in Table 3.

Anchor text	Hyperlinks	Matched	Page lang
中国科学技术协会	http://www.cast.org.cn	Yes	CHS
中国科学院生物物理研究所	http://www.ibp.ac.cn	Yes	CHS
IUPAB	http://www.iupab.org	No	ENG
Asian Biophys-ics Asso-ciation	http://www.aba-bp.com	No	ENG
美国生物物理学会	http://www.biophysics.org	No	ENG
Protein and Cell	http://www.protein-cell.org	No	ENG

Table 3. A part of matched hyperlinks by using experimental group queuing rules from "http://www.bsc.org.cn/"

Finally, we use these matched hyperlinks as the URLs for next crawling loop.

The web page crawled results of control group and experimental group from Feb. 5th to Feb. 12th 2010 show in Table 4. Because of the longer

processing time of language distinguish in experimental group, the total crawl results of experimental results are fewer then baseline results.

	Chinese	Japanese	Korean	Other	Total
KR-C*	12,523	1,926	80,049	36,273	130,771
KR-E	1,757	380	11,328	2,386	15,851
JP-C	6,555	66,235	108	2,838	75,736
JP-E	1,179	11,890	24	465	13,558
CN-C	112,924	2,468	1.052	11,321	127,765
CN-E	10,078	202	99	1,015	11,394

Table 4. The crawled web pages for each language

When the crawling target language is Korean, the language percentage of gathered web pages in baseline results shows as Figure 4. Total 130,771 web pages are crawled and 61.21% are Korean web pages.

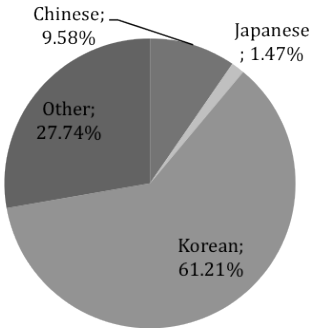

Figure 4. Language distribution of Korean baseline results

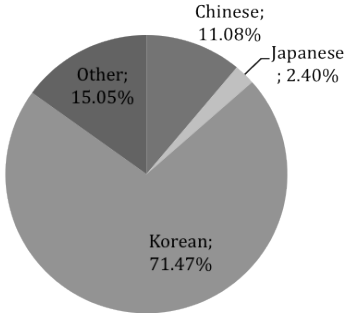

Figure 5. Language distribution of Korean experimental results

When the crawling target language is Korean, the language percentage of gathered web pages in experimental results shows as Figure 5. Total 15,851 web pages are crawled and 71.47% are Korean web pages.

According to the crawled results in Figure 4 and Figure 5, our methods can improve about 10% efficiency in crawling Korean web pages.

When the crawling target language is Japanese, the language percentage of gathered web pages in baseline results shows as Figure 6. Total 75,736 web pages are crawled and 87.46% are Japanese web pages.

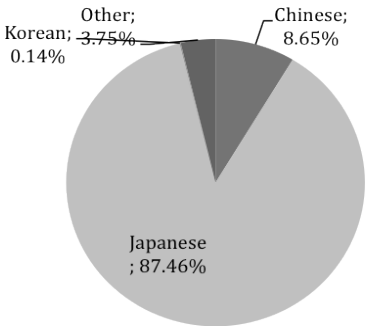

Figure 6. Language distribution of Japanese baseline results

When the crawling target language is Japanese, the language percentage of gathered web pages in experimental results shows as Figure 7. Total 13,558 web pages are crawled and 87.70% are Japanese web pages.

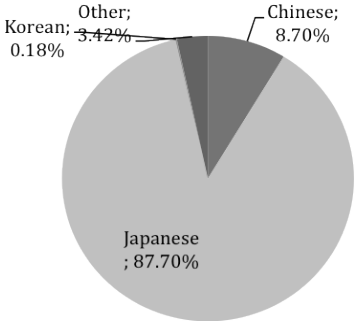

Figure 7. Language distribution of Japanese experimental results

According to the crawled results in Figure 6 and Figure 7, our methods can improve about 0.24% efficiency in crawling Japanese web pages.

* KR(Korean)、JP(Japanese)、CN(Chinese)、C(Baseline Results)、E(Experimental Results)

When the crawling target language is Chinese, the language percentage of gathered web pages in baseline results shows as Figure 8. Total 127,765 web pages are crawled and 88.38% are Japanese web pages.

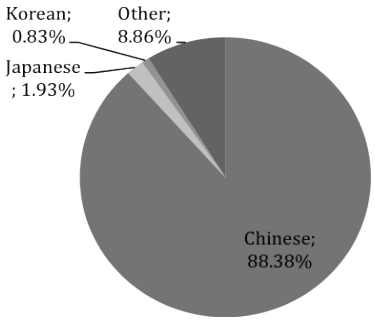

Figure 8. Language distribution of Chinese baseline results

When the crawling target language is Chinese, the language percentage of gathered web pages in experimental results shows as Figure 9. Total 11,394 web pages are crawled and 88.45% are Japanese web pages.

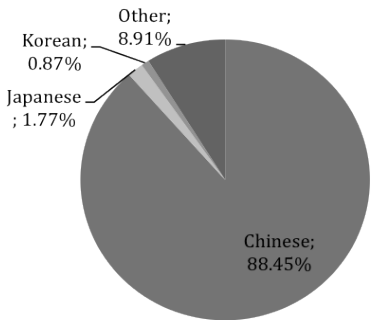

Figure 9. Language distribution of Chinese experimental results

According to the crawled results in Figure 8 and Figure 9, our methods can improve about 0.07% efficiency in crawling Chinese web pages.

The experimental results show that the efficiency improvement of Korean crawling is very obvious. The efficiency improvement of Chinese and Japanese are lower then 1%, actually the improvement is unobvious. We picked some URLs from crawling log to analyze why the improvement of Korean language specific crawler is obvious. We found that many Korean web pages content multi-language links on the main page, such as English, Chinese, Japanese, etc…. It is high probability that the hyperlinks on the Korean web pages link to web pages which are other languages. The languages of anchor text on Chinese or Japanese web pages are always Chinese or Japanese. So following our modified crawling rules, the crawled Korean web pages increase conspicuously and the crawled Chinese or Japanese web pages increase unobvious.

5 Conclusions

The purpose of this research is to enlarge the crawling amounts of language specific crawler. We propose language judgment on anchor text method to enhance the efficiency of the focus crawler. And combine the method which is distinguish the top-level domain name of URLs on hyperlinks to implement a control experiment in this research.

This research also proposes language specific focus crawler by importing a plug-in into Nutch and crawling the web pages with a modified algorithm. This method can easily program, install, and implement modified crawling rules by using plug-in on Nutch.

According to the comparison of control group and experimental group, the efficiency improvement is obvious on Korean focus crawler. And the efficiency improvement is unobvious on Chinese and Japanese focus crawler. It is because many hyperlinks on the Korean web pages link to web pages that are other languages. So that the improvement of Korean web pages crawling is obvious.

Perl language guesser is implemented by using JAVA external call. If we can use the JAVA language guesser, the crawling speed will be improved.

The append function of Hadoop is defective. Only new file can be appended, data can't append to existing file in Hadoop. In order to counting the crawled URLs, immediately append the URLs list to a log file in Hadoop is impossible. So Perl script called by JAVA external call is used to replace the flawed append function in Hadoop. But this kind of method occurs a lot of "Out of memory" and "IOExpection" errors in JAVA. These errors slow down the gather speed of web pages. If the append function in Hadoop

is flawless, the collection speed of web pages will be boosted.

According to the crawled results in experimental group, the percentage of crawled Chinese web pages by Chinese focus crawler is 88%. The percentage of crawled Japanese web pages by Japanese focus crawler is 87%. The percentage of crawled Korean web pages by Korean focus crawler is 71%. We will increase the efficiency of language specific crawler more in the future work.

This research uses DMOZ as the seed sets. We extract CJK URLs from DMOZ data. Actually, the percentage of CJK URLs in DMOZ data is very low. We will try to use the famous CJK portal web sites as the seed sets in the future work.

Acknowledgements

This research is supported by Waseda University Global COE Program “International Research and Education Center for Ambient SoC” and JST Program “Multimedia Web Analysis Framework towards Development of Social Analysis Software."

References

Charu C. Aggarwal, Fatima Al-Garawi, and Philip S. Yu. 2001. *Intelligent Crawling on the World Wide Web with Arbitrary Predicates*, WWW conference.

DMOZ: The Open Directory Project is the largest, most comprehensive human-edited directory of the Web. http://www.dmoz.org

Hadoop: A reliable, scalable, and distributed computing system. http://hadoop.apache.org

HITS: Hyperlink-Induced Topic Search. http://en.wikipedia.org/wiki/HITS_algorithm

Lingua::LanguageGuesser. http://gensen.dl.itc.u-tokyo.ac.jp/LanguageGuesser/LanguageGuesser_ja.html

Neel Sundaresan, and Jeonghee Yi. 2001. *Mining the Web for Relations*, Computer Networks, Volume 33, Issue 1-6: 699-711.

Nutch: A JAVA based open source web crawler developed by Apache Software Foundation. http://nutch.apache.org

Shervin Daneshpajouh, Mojtaba Mohammadi Nasiri, and Mohammad Ghodsi. 2003. *A Fast Community Based Algorithm For Generating Web Crawler Seeds Set*.

Soumen Chakrabarti, Martin van den Berg, and Byron Dom. 1999. *Focused Crawling: A New Approach to Topic Specific Resource Discovery*, WWW Conference.

Tamura Takayuki, Somboonviwat Kulwadee, and Kitsuregawa Masaru. 2006. *A Method for Language Specific Web Crawling and Its Evaluation*, The IEICE transactions on information and systems, J89-D(2):199-209.

Thanh Tin Tang, David Hawking, Nick Craswell, and Kathy Griffiths. 2005. *Focused Crawling for both Topical Relevance and Quality of Medical Information*, CIKM.

Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. 2010. *A Focused Crawler for Dark Web Forums*, Journal of The American Society for Information Science and Technology, 61(6):1213-1231

William B. Cavnar and John M. Trenkle. 1994. *N-gram-based text categorization*, In Symposium On Document Analysis and Information Retrieval:161–176.