Finding Thai Web Pages in Foreign Web Spaces

Kulwadee SOMBOONVIWAT[†] Takayuki TAMURA[†, ‡] Masaru KITSUREGAWA[†]

† Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505 Japan
‡ Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura-shi, 247-0056 Japan
E-mail: † {kulwadee, tamura, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract While the Web has been increasingly recognized as a culturally valuable social artifact, many nations endeavor to create national Web archives for long term preservation. However, due to its borderless-ness, gathering information for a specific nation from the Web is challenging. This paper proposes *language specific web crawling (LSWC)* as a method of creating Web archives for countries with linguistic identities such as Thailand. The LSWC strategy for selectively gathering Thai web pages from virtually anywhere on the Web is derived based on static analyses of the Thai Web graph. Then, the LSWC strategy is evaluated on a crawling simulator with large dataset.

Keyword 情報検索, 性能評価, Web とインターネット, Web アーカイブ,フォーカストクローリング,言語判定, Web グラフ

1. Introduction

While the Web has been increasingly recognized as a culturally valuable social artifact, many nations endeavor to create national Web archives for long term preservation e.g. Kulturawa3 project [3] of the National Library of Sweden and WARP project [4] of the National Diet Library of Japan. An important method for building such large scale Web archives as those of the national Web archiving projects, is national domain name based restriction web crawling. For example, to construct a Thai Web archive, a web crawler will be used to collect as much as possible all web pages belonging to the Thailand national domain name, i.e. the '.th' domain.

However, due to internationalization force in our modern society and economical reasons, web pages relating to a country are frequently being put on web servers outside the national domain name. In the case of Thailand, according to our analyses on the Thai dataset (see Section 4), we found that more than half of Thai web pages (i.e. web pages written all or partially in the Thai language) are outside the .th top-level domain name. It can be clearly seen that the domain name based restriction approach will become less useful the greater there are Thai web pages outside the '.th' domain. Therefore a more flexible approach is necessary for the construction of the large scale national Web archives.

The crawling strategy that is suitable for building the large scale national Web archives is the one in which it is possible to obtain web pages relating to the country from virtually anywhere on the Web, and the method must be scalable to the tremendous size of the rapidly growing World Wide Web.

This paper proposes a *"language specific web crawling" (LSWC)* as a method for creating large-scale Web archives for countries with linguistic identities such as Thailand, and Japan. In the LSWC approach, domain name independence and scalability, as mentioned earlier, are addressed by crawling as much as possible web pages written all or partially in the language of interest while at the same time crawling as less as possible web pages written in other languages. The LSWC crawler will selectively seek out web pages that are written in the user-specified language by following a URL with the highest probability of leading to other relevant pages. The probability of leading to other relevant pages of a URL is assigned based on attributes of parent pages (e.g. the language of the parent page), the attributes of the URL itself (e.g. its server and domain), and the graphical structure of the Web discovered so far.

In this paper, we are focusing our study on the crawling of Thai web pages. So, we need a tool for automatically classifying the languages of downloaded web pages. The language classifier used in this paper is implemented based on TextCat [6], which is an n-gram based language guesser tool. We made some customization to TextCat so that our language classifier can detect Thai web pages encoded in both non-utf8 and utf-8 character sets.

The knowledge about the graphical structure of the Thai Web is derived from the analyses of the Thai Web graph on a large dataset. The dataset used in our analyses and experiments was created by crawling some portion of the Thai Web and its neighborhood with a limit on the distance from the start seed URLs. The LSWC crawler for Thai web pages, which incorporates the Thai language classifier and the knowledge about the Thai Web graph structure, is evaluated using a web crawling simulator (proposed in [8]). According to the evaluation results, the LSWC strategy achieves the highest harvest rate and comparatively good crawl coverage.

The paper is organized as follows: Section 2 reviews related works. Section 3 describes the language identification method and its evaluation. Section 4 presents the results of the Thai Web graph analyses. Section 5 describes the LSWC strategy. Subsequently, in Section 6, we report the simulation-based evaluation results. Finally, Section 7 concludes the paper.

2. Related Works

This section gives a brief description of focused crawling (originally proposed by Chakrabarti et al. [9]) and an application of focused crawling to the construction of large scale topic-specific web collections. Then, web crawling strategies for a large scale setting and a technique for collecting language specific web pages based on automatic generation of Web-search queries will be discussed.

Focused crawling is a method for selectively seeking out web pages that are relevant to a pre-defined set of topics specified by exemplary documents. The focused crawling addresses the problem of scalability (which is caused by the unprecedented explosive growth of the World Wide Web) faced by general-purpose web crawlers. The focused crawling system consists of three main components:

1) *A Bayesian classifier*: makes relevance judgments on web pages crawled to a node in a hierarchical topic-directory.
2) *A distiller*: identifies pages with a large number of links to relevant pages by using a modified version of Kleinberg's algorithm [11].
3) *A crawler*: with a dynamically reconfigurable priority control which is governed by the classifier and the distiller.

The focused crawler can be configured to work in two modes: hard focused and soft focused modes. In the hard focused mode, the crawler will follow URLs found on a web page *p* if and only if *p* was classified to a leaf node with any relevant ancestors (i.e., *p* is relevant). In the soft focused mode, the crawler will follow all URLs found on web pages in order according to the relevant scores of parent pages of the URLs.

Recent studies on focused crawling include [10, 15, 17, 18]. Focused crawlers have been widely applied to build the collections of web pages for large scale digital libraries, topic-specific search engines and web portals [19, 20]. In [19], focused crawling based on the adaptive cutoff of crawl paths was proposed as a method for efficient topic-specific collection synthesis for large scale digital libraries.

A comprehensive study on performance of various page ordering strategies for large-scale web crawling was presented in [7]. The performance of different page ordering strategies were compared based on how quickly the crawler can download the most important pages early during the crawl. The experiment was done on a crawling simulator with two datasets corresponding to web pages under the .cl (Chile) and .gr (Greek) top-level domain. According to [7]'s evaluation result, the larger-site-first strategy was very good in obtaining most important pages earlier; and the use of historical information , e.g. Pagerank values of the previous crawl, to guide the crawl might be helpful in a large scale web crawling in the order of billions of pages.

Techniques and a system for automatically collecting language specific resources from the WWW based on automatic generation of Web-search queries were proposed and described in [21]. The proposed system (which is called the CorpusBuilder) creates a query by selecting words from two document sets, i.e. relevant and non-relevant, to be used as inclusion and exclusion terms of the query respectively. The new query is sent to a search engine and the highest ranking document is retrieved, classified as relevant or non-relevant, and added to the corresponding document set. The process is then repeated until a sufficient number of relevant documents have been obtained.

3. Language Identification of Web Pages

Our language identification method was tailored to discriminating the Thai web pages. We implemented the language identification method based on a language guesser tool, called TextCat [6]. Before discussing the language identification method, let us take a look at an introduction to TextCat.

3.1 TextCat

TextCat [6] is an implementation of the n-gram based classification technique proposed by Cavnar and Trenkle [5]. The n-gram based classification technique requires the minimal computation resources, can be trained with a small training set (a training set of 20K bytes in length is enough), and is able to achieve good performance even with a short input document (a document with >= 300 characters is enough). These advantages make the n-gram based classification ideally suitable for language identification of web pages.

The n-gram is an n-character slice of a longer string [5]. For example, the string “CRAWL” would be composed of the following n-grams (“_” is used to represented blanks):

uni-grams: _, C, R, A, W, L, _
bi-grams: _C, CR, RA, AW, WL, L_
tri-grams: _CR, CRA, RAW, AWL, WL_, L_ _

In the n-gram based classification, an input document with an unknown category (or language) is represented by a profile of top *N* highest frequency n-grams extracted from the document. Likewise, each known category is represented by a profile of top *N* highest frequency n-grams extracted from the training document. After the language profile of the input document has been created, the input language profile will be compared to the profiles of all languages known by the system by using some distance measures such as the Out-of-Place measure (see Figure 1, borrowed from [5]). Then, the languages with the lowest Out-of-Place value(s) (i.e. languages that the input document is most similar to) will be selected as the classification outputs. In this paper, we made the following customization to TextCat:

- Add a language profile for Thai utf8 character set.
- Comment out some rarely used languages e.g. drents, basque etc.

The problem of automatically identifying the language of a given Web document was studied in detail by [22] and [23]. In [22], language guessing of web pages was done using an n-gram based algorithm complemented with heuristics for the categorization of web documents, such as weighting n-grams according to HTML mark up. The study in [23] proposed an iterative cross-training (ICT) for automatically identifying Thai web pages. The ICT combines two classifiers: a word segmentation classifier and a naïve Bayes classifier, and uses unlabeled examples to iteratively train each classifier. The ICT can achieve 100% precision and 98.89% recall of Thai Web pages without the help of human in labeling the examples.

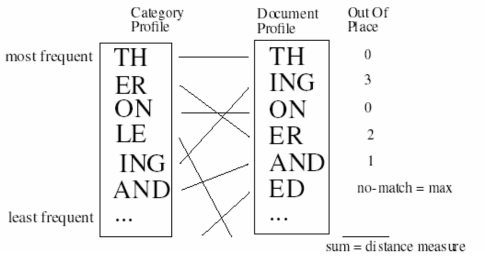

Figure 1 Similarity between an input document (Document Profile) and a known language profile (Category Profile) is measured using the Out-of-Place distance measure.

3.2 Language Identification Method

The procedure for the identification of the language of web pages used in this paper can be described as follows.

1) Find *metacs_lang*:
- Extract charset name from html’s meta-tag,
- Infer language from the extracted charset.
(For example, if charset = ‘windows-874’ or ‘tis-620’ then *metacs_lang*=’thai’)

2) Find *textcat_lang*:
- Remove all html tags from a web page, and submit the remaining text to TextCat
- IF result from TextCat = ‘UNKNOWN’ or ‘SHORT’

```
 THEN textcat_lang =’unknown’
 ELSE IF result from TextCat contains ‘thai’
       THEN textcat_lang=’thai’
 ELSE IF result from TextCat contains ‘english’
       THEN textcat_lang=’english’
 ELSE textcat_lang=first element in TextCat output
```

3) The language of a web page can be determined from *metacs_lang* and *textcat_lang* as shown in Table 1.

metacs_lang	*textcat_lang*	Language of the web page
thai	X	thai
X	thai	thai
Y	unknown	Y
unknown	Z	Z
unknown	unknown	unknown
Y	Z	Y

Table 1 Determining language of a web page

3.3 Language Identification Method Evaluation

Table 2 shows the performance of our method in detecting Thai documents. The method has 94% accuracy, with 91% precision and 94% recall in the evaluation using 2,000 test documents.

We investigated the documents that were misclassified by our method. For the false-negative cases, we found that those documents were (1) html pages using frames, (2) web pages having a lot of numeric data, (3) web pages written in both Thai and English languages, and (4) web pages using numeric character references, e.g. &#3590, to represent Thai characters (The values of numeric character references for Thai characters are ranging from &#3585 to &#3675). For more information on numeric character reference, see section 5.3.1 of the HTML 4.01 Specification [13].

In the false-positive cases, the misclassifications are all in the type of the assignment of the English web pages as the Thai web pages. Further inspection indicates that, while meta-charset information is misleading, TextCat is able to classify all of those documents correctly. We will use these observations to improve our language identification method in the future. Although our method cannot achieve as high performance as the ICT method discussed earlier, it is a lot easier to implement, requires less computational resources, and can give quite good performance. In the following section, we will apply our language identification method to the Thai dataset and present various interesting statistics about the dataset and the Web graph associate with it.

	Thai web pages	non-Thai web pages
Assigned as Thai	764	71
Assigned as non-Thai	50	1115

Table 2 Discriminating Thai web pages

4. Thai Dataset

As a dataset for the evaluation of the language specific web crawling strategy, we have collected about 14 million web pages by starting from the following three websites which are considerably popular websites in Thailand.

- http://www.sanook.com/
- http://www.siamguru.com/
- http://www.matichon.co.th/

The dataset contains web pages from 668,934 servers. And, the proportions of web pages in each top-level domain name are: .com 46.4%, .jp 11.0%, .th 8.2%, .de 6.7%, .net 6.2%, .org 5.7%.

In the following subsections, we will present the results of the language guessing, and discuss about the structure of the Web graph associated with the Thai dataset.

4.1 Languages of Web Pages in the Dataset

Based on the language identification method described in Section 3.2, we found that the top 3 major languages presented in the .th domain are Thai, English, and Japanese, respectively. Table 3 shows the result of language identification classified by domain name.

It can be seen from Table 3 that most Thai web pages (65%) are outside .th domain. This is the evidence that comprehensively crawling of Thai web pages in the .th domain cannot give us good coverage of the Thai Web and an efficient method for crawling Thai web pages outside .th domain, especially .com and .net, is necessary.

However, there are a large number of web pages whose languages cannot be identified (i.e., unknown). After checking a handful of unknown-language pages, we found that they were pages with very little text. As a treatment of the unknown-language pages, we will guess that the language of an unknown-language web page is the same as the language of a web page pointing to it (its parent).

Language	**Domain**				**Total**
	.th	.com	.net	Other	
Thai	588,082	903,792	70,777	143,587	1,706,238 12.2%
English	344,679	1,199,555	158,065	784,461	2,486,760 17.8%
Other	25,841	130,767	27,979	292,502	477,089 3.4%
Unknown	182,957	4,251,728	612,909	4,282,166	9,329,760 66.6%
Total	1,141,559	6,485,842	869,730	5,502,716	13,999,847 100%

Table 3 Language of web pages classified by top-level domain

4.2 Thai Web Graph

After identifying the languages of web pages in the dataset, we extract linkage information and derived various statistics about the graphical structure of the Web graph imposed by the Thai dataset. The derived Thai Web graph consists of 39,078,797 nodes (13,999,847 crawled nodes + 25,078,950 uncrawled nodes), and there are 1,706,238 relevant nodes (Thai web pages) in the graph. The number of directed links is equal to 123,836,342.

4.2.1 Distance from Seed URLs

Figure 2 shows the ratio of the Thai web pages within a distance from the seed URLs. The ratio of the Thai web pages decreases while going farther from seeds. This suggests that crawling with a breadth first strategy with some limited distance from the seed URLs is not suitable for the collection of language specific web documents because it cannot achieve a reasonable level of efficiency.

Note that, Thai pages ratio was unchanging after distance=7 because we were using the limitation of the crawl radius as a stopping condition of the crawl.

4.2.2 Distance between nearest Thai pages

Figure 3 shows the distribution of distances between the nearest Thai pages. Distance 1 represents the case that a Thai page is being linked directly to another Thai page; Distance > 1 means that it is necessary to traverse through at least (distance-1) non-Thai pages in order to reach another Thai page. As expected, most Thai web pages are linked directly to other Thai pages. The number of the Thai destination pages exponentially decreases while the distance increases. From Figure 3, the farthest Thai page is at distance 10. This means that we have to download at least 9 non-Thai pages before obtaining one Thai page. Therefore, to prevent missing Thai pages, it is necessary to download some non-Thai pages.

According to our observation on these kinds of crawl paths (Thai → non-Thai → Thai*; distance = 2 to 4*), the following patterns are found.

- The non-Thai page is an English html document using frames or flash or html image maps, and is the homepage of a Thai website.
- The non-Thai page is a Thai-English html document classified as English.

4.2.3 Linguistic Locality of Outlink

Table 4 shows the ratio of Thai destination pages that can be reached from Thai and non-Thai source pages. From Table 4, the ratio of links (or URLs) found on Thai web pages pointing to Thai web pages is about 70%, and this ratio increases when a URL is of the same server (or the same domain) as the source pages. But, the ratio of Thai destination is very small (only 3%) in the case of links from non-Thai web pages.

4.2.4 Anchor Text and the Destination Page Languages

The identification of the language of an anchor text was done by applying a Thai Word Separator Program (cttex [12]) to the anchor text. The cttex will return the number of Thai words found in the input byte stream. The anchor text will be classified as Thai if and only if the number of Thai word returned from cttex is greater than 0, otherwise; the anchor text will be classified as non-Thai. The result of our analysis is as shown in Table 5.

It can be seen from Table 5 that the difference of the ratios of links with the Thai anchor text and the non-Thai anchor text which are pointing to Thai destination pages is small (about 10%). This means that the anchor text is not a good choice for guessing the language of an unknown URL. Lastly, it should be note that the number in Table 5 corresponding to "source=non-Thai, anchor text language = Thai" (73,328) represents the source pages that are misclassified as non-Thai by the language classifier.

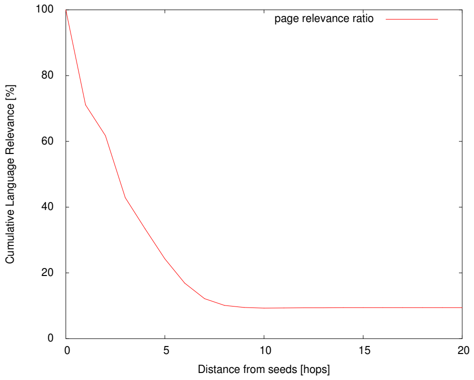

Figure 2 Ratio of Thai pages within a distance from seed URLs

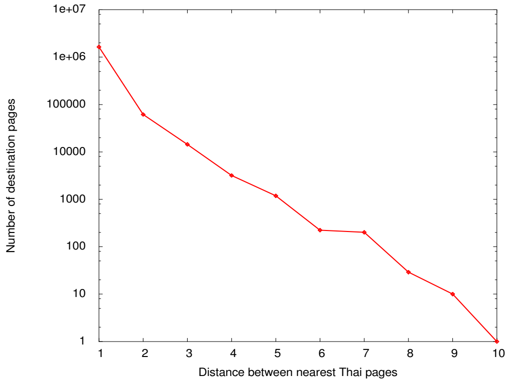

Figure 3 Distribution of distances between the nearest Thai pages

	Ratio of Thai Destination			
Source	same domain	different domain	same server	different server
Thai	71.3% (27,542,015)			
	83.0%	30.7%	84.1%	38.5%
non-Thai	3.1% (1,760,768)			
	2.3%	6.2%	3.0%	3.2%

Table 4 Linguistic locality of outlink

Source	Anchor Text Language	Ratio of Thai Destination
Thai	Thai	81.1% 9,090,883
	non-Thai	66.0% 20,328,096
non-Thai	Thai	2.6% 73,328
	non-Thai	2.7% 1,787,553

Table 5 Relationship between the languages of the anchor text vs. the languages of the destination page

5. The Proposed Strategy: Language Specific Web Crawling

The main findings found in the previous section can be summarized as follows.

- In order to get a Thai page, sometimes, it is necessary to follow *N* non-Thai pages (*N* ranges from 2 to 9). And for *N* = 2 to 4, the non-Thai web pages of the crawl path frequently reside in the same server.
- Following the links extracted from Thai web pages has a better chance of obtaining Thai web pages than following the links extracted from non-Thai web pages.

Based on the above findings, we derive the language specific web crawling strategy (LSWC) as follows.

(1) Discard URLs with the distance from the latest relevant parent page *d*>T, where T is a distance threshold.

(2) Discard URLs of the irrelevant servers.

- A server is *irrelevant* when the crawler cannot find any relevant pages on it after S consecutive downloads, where S is a server traversal depth threshold.
- A server is *relevant* when the crawler finds the first relevant page on it.

(3) Prioritization of URL downloading: the crawler selects the URLs from the URL queue in the following order.

① URLs of the relevant servers

② URLs from the relevant parent pages

- The server of the URLs and the server of the parent pages are same.
- The server of the URLs and the server of the parent pages are different.

③ URLs from the irrelevant parent pages

- Order by the distance from the latest relevant parent page.
- If the previous two consecutive parent pages of the URL are Thai and English respectively (i.e. Thai→English→the URL), then increase the priority of the URL one step higher.

6. Simulation Evaluation

6.1 Web Crawling Strategies to be evaluated

The following crawling strategies are evaluated on the web crawling simulator [8] which uses the input dataset in Section 4 to create the Web graph for the crawlers.

- *LSWC* with S=1 and T=1
- *LSWC* with S=3 and T=5
- *Hard focused*: discards links from non-Thai pages.
- *Soft focused*: follows links extracted from Thai pages first.
- *BFS*: a breadth first crawling strategy.
- *Perfect*: follows only links that lead to Thai pages. (The links that lead to Thai pages were determined in advance by breadth first crawling.)

6.2 Result

The performance measures, used for the evaluation of the crawling strategies, are:

- *Coverage*: the fraction of relevant pages that is found by the crawler.
- *Harvest rate*: the rate at which relevant pages are acquired and how effectively irrelevant pages are filtered out.

Figure 4 and Figure 5 respectively show the traces of the coverage and the harvest rate for each strategy.

First, let us consider the coverage trace in Figure 4. In the case of LSWC with S=1 and T=1, the crawl stops after downloading about 4M documents and we obtain almost all relevant documents. And, if we allow some relaxation on the filtering condition by setting S=3 and T=5, we will obtain all relevant documents.

The LSWC strategy focuses itself to the Web region that is more likely to be relevant by using the prioritization of the URLs in the URL queue based on the linguistic locality of outlinks (Section 4.2.3), the distance to the latest relevant URLs (Section 4.2.2), and the relevance of web servers detected when any Thai web pages can be found on them. The effect of these prioritizations can be clearly seen in the harvest rate graph in Figure 5.

From Figure 5, the harvest rate of the LSWC strategy during the first 1M crawl progress is about 80% in average, which is much higher than the BFS, the hard focused, and the soft focused strategy. Although the hard focused and soft focused strategies also have the prioritization mechanism, the prioritization mechanism, employed by these two strategies, is quite simple. Namely, it is based only on the idea that relevant pages are likely to link to other relevant pages.

Although the URL prioritization of the LSWC strategy results in the increase of the harvest rate of the crawl, the crawl coverage is not so impressive. To improve the crawl coverage, more analyses on the Thai Web graph structure and content are necessary.

The remaining open problems are the effect of the two parameters of the LSWC strategy: T (the distance threshold) and S (the server traversal depth threshold). The effect of the T parameter may be understandable by studying the behavior of a pure distance based strategy (which was previously proposed and studied in [8]). For the S parameter, we need to do more analyses to study its effect (e.g. the distribution of the number of Thai web pages in the web servers, and the behavior of the depth-first (DFS) crawler).

Although in this paper we are focusing our study on the collection of Thai web pages, we think that the LSWC can be easily adapted and will also be useful for constructing large scale language-specific Web archives for other languages. The expected advantages of using the LSWC in constructing large scale language specific Web archives are 1) being able to obtain web pages related to the country from anywhere on the WWW, 2) scalability, 3) saving of computation resources and time, and 4) maintaining the freshness of the archives will be easier.

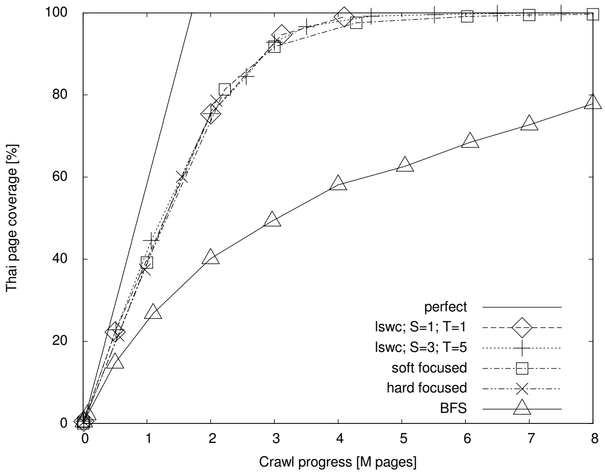

Figure 4 Coverage

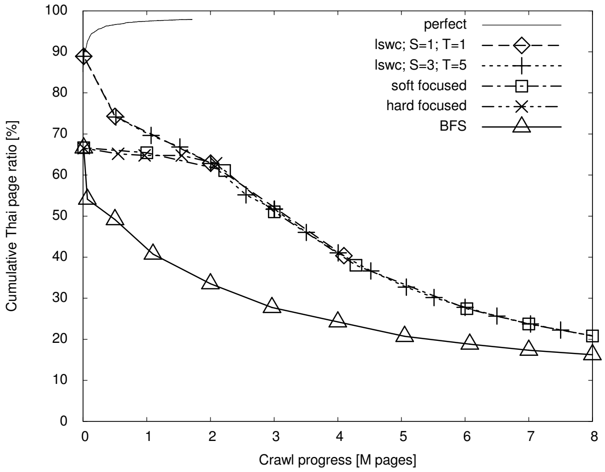

Figure 5 Harvest Rate

7. Conclusion

The great diversity and tremendous amount of Web information has attracted a lot of interest in collecting and preserving the WWW. Many nations are trying to collect web pages from the WWW for their future generations. Collecting information specific to a country and/or culture from the borderless-ness information space like the WWW is not easy. In this paper, we have proposed a language specific web crawling (LSWC) strategy for the collection building of the national web archiving projects.

The LSWC strategy has been evaluated on the crawling simulator, and compared its performance with the BFS, the hard focused, and the soft focused strategies. According to the simulation result, the LSWC strategy achieves the highest harvest rate with comparatively good crawl coverage.

Nevertheless, to apply the LSWC strategy in the real web crawling situation, more studies are needed, for example, the effect of network bandwidth on crawl performance, and the restriction on accessing the remote web servers. In order to estimate these effects, we need more elaboration on the crawling simulator. We will continue our study on these topics in the future.

References

[1] International Internet Preservation Consortium: “netpreserve.org”. http://netpreserve.org/

[2] Internet Archive: “About IA”. http://www.archive.org/about/about.php

[3] National Library of Sweden: “Kulturawa3”. http://www.kb.se/kw3/ENG/

[4] National Diet Library of Japan: “Web Archiving Project (WARP)”. http://warp.ndl.go.jp/

[5] W. B. Cavnar and J. M. Trenkle: “N-gram-based text categorization”, Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175 (1994)

[6] WiseGuys Internet B.V.: “libTextCat – lightweight text categorization” (2003). http://software.wise-guys.nl/libtextcat/index.html

[7] R. Baeza-Yates, C. Castillo, M. Marin and A. Rodriguez: “Crawling a country: better strategies than breadth-first for web page ordering”, WWW ’05: Special interest tracks and posters of the 14th international conference on World Wide Web, New York, NY, USA, ACM Press, pp. 864–872 (2005)

[8] K. Somboonviwat, M. Kitsuregawa, and T. Tamura. "Simulation Study of Language Specific Web Crawling," *icde*, p. 1254, 21st International Conference on Data Engineering (ICDE'05), 2005.

[9] S. Chakrabarti, M. van den Berg and B. Dom: "Focused crawling: A new approach to topic-specific web resource discovery", Proc. of WWW8 (1999)

[10] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles and M. Gori: "Focused crawling using context graphs", 26th Intl. Conference on Very Large Databases, VLDB 2000, Cairo, Egypt, pp. 527–534 (2000)

[11] J. Kleinberg, "Authoritative sources in a hyperlinked environment", in: Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.

[12] Thai Word Separator Program: "cttex" http://webls.ex.nii.ac.jp/~vuthi/files/

[13] D. Raggett, A. Le Hors, I. Jacobs, "HTML 4.1 Specification": http://www.w3.org/TR/html4/

[14] B. D. Davison. (2000) "Topical Locality in the Web", in: Proc. of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000), July 24-28, pages 272-279.

[15] C. C. Aggarwal, F. Al-Garawi, and P.S. Yu. Intelligent crawling on the World Wide Web with arbitrary predicates. In Proc. of the 10th Intl. World Wide Web Conference, WWW 10, Hong Kong, China, May 2001.

[16] J. Cho, H. Garcia-Molina, and L. Page. "Efficient crawling through URL ordering", Computer Networks, 30(1-7):161-172, 1998.

[17] S. Chakrabarti, K. Punera, and M. Subramanyam, "Accelerated focused crawling through online relevance feedback". In Proc. of the 11th Intl. World Wide Web Conference, WWW11, Hawaii, USA, May 2002.

[18] J. Johnson, K. Tsioutsiouliklis, and C. L. Giles. "Evolving strategies for focused web crawling", In Proc. of the 20th Intl. Conference on Machine Learning (ICML2003), Washington DC, 2003.

[19] D. Bergmark, C. Lagoze, and A. Sbityakov, "Focused Crawls, Tunneling, and Digital Libraries", in Proc. 6th European Conf. Research and Advanced Technology for Digital Libraries. pp. 91-106, 2002.

[20] G. Almpanidis, C. Kotropoulos, I. Pitas. "Focused Crawling Using Latent Semantic Indexing - An Application for Vertical Search Engines". In Proc. 9th European Conf. Research and Advanced Technology for Digital Libraries pp. 402-413, 2005.

[21] R. Ghani, R. Jones, D. Mladenic. "Mining the Web to Create Minority Language Corpora". In Proc. of the 10th Intl. Conference on Information and Knowledge Management (CIKM 2001): 279-286

[22] Bruno Martins, Mário J. Silva, "Language identification in web pages". In Proc. of the 2005 ACM Symposium on Applied Computing (SAC 2005): 764-768

[23] B. Kijsirikul, P. Sasiphongpairoege, N. Soonthornphisaj, S. Meknavin, "Supervised and Unsupervised Learning Algorithms for Thai Web Pages Identification". In Proc. of Pacific Rim International Conference on Artificial Intelligence, (PRICAI 2000): 690-700