Finding Thai Web Pages in Foreign Web Spaces

Kulwadee Somboonviwat[1] Takayuki Tamura[1, 2] Masaru Kitsuregawa[1]
[1] *Institute of Industrial Science, The University of Tokyo*
4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
[2] *Information Technology R&D Center, Mitsubishi Electric Coporation*
5-1-1 Ofuna, Kamakura-shi, 247-0056 Japan
Email: {kulwadee, tamura, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

This paper proposes language specific web crawling (LSWC) as a method of creating large-scale language specific 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 the results of static analyses of the Thai Web graph. We evaluated the performance of the LSWC strategy using a web crawling simulator.

1. Introduction

While the Web has been recognized as a culturally valuable social artifact, many nations endeavor to create national Web archives for long term preservation e.g. the WARP project [1] 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 Thai national domain name, i.e. the '.th' domain name.

However, due to the internationalization effort 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 analysis on the Thai dataset (see Section 3), we found that more than half of Thai web pages (in this paper, Thai web pages means web pages that are written all or partially in the Thai language) are outside the .th domain name. It can be seen clearly that the domain name based restriction approach will become less useful the greater there are Thai web pages outside the .th domain name. Therefore a more flexible web crawling strategy is necessary for the construction of the large-scale national Web archives.

The web crawling strategy that is suitable for the construction of the national Web archives is the one in which it is possible to obtain web pages which are relating to the country from virtually anywhere on the Web, and the method must be scalable to the tremendous size of the continuously growing WWW.

This paper proposes a *"language specific web crawling" (LSWC)* as a method of creating large-scale Web archives for countries with linguistic identities such as Thailand, and Japan. In the LSWC approach, domain name independence and scalability are being 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 web crawler consisted of three major mechanisms: (1) automatic language classification of web pages, (2) discarding of irrelevant URLs from the URL queue, and (3) prioritization of URLs to be downloaded. Language classification of web pages was implemented using TextCat (an n-gram based language guessing tool) [3]. Discarding of irrelevant URLs and URL prioritization were designed based on the knowledge obtained from results of Thai Web graph analysis. According to the evaluation results obtained from the web crawling simulator [4], the LSWC strategy achieves the highest harvest rate and comparatively good crawl coverage.

2. Dataset and Thai Web Graph Analysis

As a dataset for the Thai Web graph analysis and for the evaluation of the LSWC strategy, we have collected about 14 million web pages by starting from following three websites which are considerably

popular websites in Thailand: http://www.sanook.com/ (a Thai web directory), http://www.siamguru.com/ (a Thai search engine), http://www.matichon.co.th/ (a Thai newspaper website).

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%.

2.1 Languages of Web Pages in the Dataset

The language identification method used in this paper is relying on the metadata of html documents and the language classification results from TextCat [3]. The TextCat is an implementation of an n-gram based text classification proposed by Cavnar and Trenkle [2].

2.1.1 Language Identification Method.

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 output from TextCat = 'unknown' THEN *textcat_lang* ='unknown' ELSE IF output 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 the output

3) From the values of *metacs_lang* and *textcat_lang*, determine the language of the web page using Table 1.

According to the evaluation on the 2,000 test documents, our language identification method achieves 94% accuracy, with 91% precision and 94% recall.

2.1.2 Language Guessing against the Thai Dataset.

After applying the language identification method to the dataset, we found that the top 3 major languages in the .th domain name are Thai, English, and Japanese, respectively. Table 2 shows the result of language identification classified by domain name.

It can be seen from Table 2 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 a more efficient method for crawling Thai web pages outside .th domain, especially .com and .net, is necessary.

Table 1 Determining language of a web page

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 2 Languages of web pages classified by top-level domain name

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%

2.2 Thai Web Graph

After identifying the languages of web pages in the dataset, we extracted linkage information and derived various statistics about the graphical structure of the Web graph. 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.

2.2.1 Distance between the nearest Thai pages.

Figure 1 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. According to our observation on these kinds of paths (i.e. 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.

2.2.2 Linguistic locality of outlink. Table 3 shows the ratio of Thai destination pages that can be reached from Thai and non-Thai source pages. From Table 3, 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.

3. Language Specific Web Crawling

The LSWC strategy was designed based on the analysis results obtained in the previous section, and can be described 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.

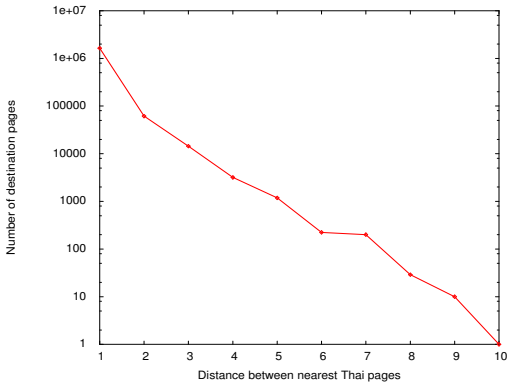

Figure 1 Distribution of distances between the nearest Thai pages

Table 3 Linguistic locality of outlink

	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%

4. Evaluation

4.1 Evaluated Web Crawling Strategies

The following web crawling strategies were evaluated on the web crawling simulator [4] with the input dataset as described in Section 2.

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

4.2 Simulation Result

First, let us consider the coverage trace in Figure 2. 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, all relevant documents will be obtained.

From Figure 3, 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. The high harvest rate of the LSWC results from the URL prioritization effort which employs the various results of the statistical analyses on the Thai Web graph. Although the soft focused strategy also has the URL prioritization mechanism, its URL prioritization is quite simple. In that, it is based only on the languages of the parent pages.

Nevertheless, the crawl coverage of the LSWC strategy is not so impressive and there is still the room left for more improvement (as can be seen from the gap between the crawl performance of the perfect case and the LSWC strategy). To improve the crawl coverage, we will conduct more static analyses on the Web graph structure and content so that we can obtain more insights on the characteristics of the Thai Web.

Other open problems are the effect of the two threshold parameters of the LSWC strategy on the crawling performance: 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 [4]. To understand the effect of the S parameter, we need more statistics such as the distribution of the number of Thai web pages in the web servers and the behavior of the depth first crawling strategy (DFS).

5. Conclusion

We proposed a language specific web crawling (LSWC) strategy for the construction of the large scale language specific Web archives. The LSWC strategy was evaluated on the web crawling simulator, and was compared with the BFS, the hard focused, and the soft focused strategies. According to the simulation results, the LSWC strategy achieves the highest harvest rate with comparatively good crawl coverage. The performance gained by the LSWC strategy is mainly caused by its URL prioritization mechanism.

Although in this paper we are focusing our study on the collection of Thai web pages, we think that the LSWC strategy can be easily adapted and will also be useful for the construction of the large scale language specific Web archives for other languages.

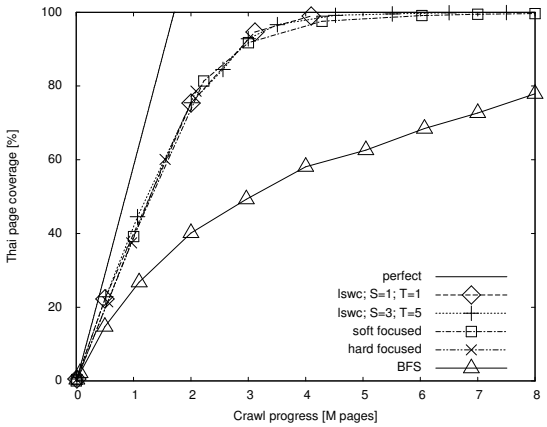

Figure 2 Coverage

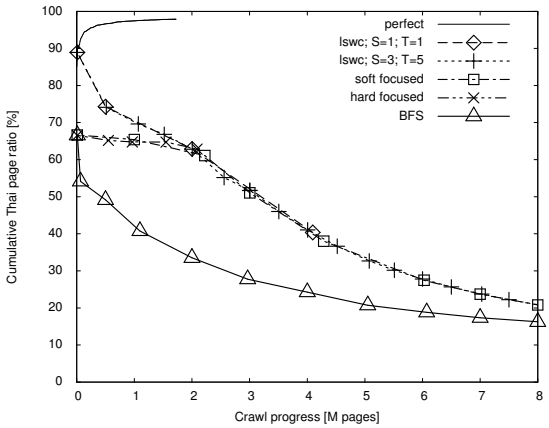

Figure 3 Harvest Rate

6. References

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

[2] 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)

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

[4] K. Somboonviwat, T. Tamura, M. Kitsuregawa: "Simulation Study of Language Specific Web Crawling”, Proceedings of the International Special Workshop on Databases for Next Generation Researchers in Memoriam of Prof. Yahiko Kambayashi (SWOD2005), pp.142-145.