Language Specific Crawler for Myanmar Web Pages

Pann Yu Mon[1], Chew Yew Choong[2], Yoshiki Mikami[3]

[1]Nagaoka University of Technology
Nagaoka, 940-3106, Niigata, Japan.

[2]Nagaoka University of Technology
Nagaoka, 940-3106, Niigata, Japan.

[3]Nagaoka University of Technology
Nagaoka, 940-3106, Niigata, Japan.

Abstract

With the enormous growth of the World Wide Web, search engines play a critical role in retrieving information from the borderless Web. Although many search engines can search for content in numerous major languages, they are not capable of searching pages of less-computerized languages such as Myanmar due to the use of multiple non-standard encodings in the Myanmar Web pages. Since the Web is a distributed, dynamic and rapidly growing information resource, a normal Web crawler cannot download all pages. For a Language specific search engine, Language Specific Crawler (LSC) is needed to collect targeted pages. This paper presents a LSC implemented as multi-threaded objects that run concurrently with language identifier. The LSC is capable of collecting as many Myanmar Web pages as possible. In experiments, the implemented algorithm collected Myanmar pages at a satisfactory level of coverage. The results of an evaluation of the LSC by two criteria, recall and precision and a method to measure the total number of Myanmar Web pages on the entire Web are also discussed. Finally, another analysis was conducted to determine the location of the servers of Myanmar Web content, and those results are presented.

Keywords: *Language Specific Crawling, Myanmar, Web Search, Language Identification*

1. Introduction

Myanmar language being a member of the Tibeto-Burman language, which is a subfamily of the Sino-Tibetan family of language, is the official language of Myanmar. It is spoken by 32 million people as a first language while a second language by ethnic minorities in Myanmar.

1.1. Lack of Standard Encodings

Since 1988, Myanmar people have been working on the implementation of Myanmar fonts for use in computer systems. The first Myanmar font for GUI environment (Mac) was developed in 1988 and the one for Windows system called Win-Inwa was developed in 1992.

Actually Win-Inwa was modeled using English cod page (technically Latin 1 or Window code page 1252) fonts with English Latin glyphs substituted for the Myanmar glyphs. This meant that the code point allocated for the Latin alphabet represents the Myanmar characters.

Simultaneously, several alternative font families such as MyaZedi, Zawgyi-One and MyMyanmar, have been implemented by different groups of people. Each font family has different mapping and none of these completely follows the Universal Coded-character Set (UCS) or Unicode standard. Since these font families only partially follow the UCS/Unicode standard, they are not yet supported by Microsoft and other major software vendors.

Until approximately 2000, there was no implementation of UCS/Unicode encoding for Myanmar language. Consequently, Myanmar natural language processing task was assigned to Myanmar Unicode & NLP Research Center. The efforts of that center have led to the implementation of a series of UCS/Unicode based fonts: Myanmar1, followed by Myamar2 and then Myanmar3. Those fonts contain not only UCS/Unicode points and glyphs but also the Open Type Layout (OTL) logic and rules. Many Myanmar font makers have also created several alternative UCS/Unicode encoding such as Padauk and Parabaik.

Since given the task, the NLP research center has worked to enable Myanmar language processing to handle all applications. Some examples of Myanmar fonts and their encodings are listed in Table 1.

Table 1: Various Myanmar Fonts and their Encodings

Font Name	***Encoding Scheme***
Myanmar1, Myanmar2, Myanmar3, Padauk, Parabaik	UCS/Unicode: Follows Unicode 5.1.
MyaZedi, MyMyanmar, Zawgyi-One	Partially follows UCS/Unicode: Not fully follows UCS/Unicode, provided under proprietary licenses.
Popular, Win-Inwa	Graphic Encoding: Characters are decomposed into a set of graphic elements, which are mapped to ASCII code points, Provided under proprietary licenses.

1.2. Characteristic of Myanmar Pages on the Web

Myanmar Web pages are minority languages on the Web having multi-encodings. Among these, Zawgyi-One (non-standard Unicode) is widely used in Myanmar Web documents. Web developers of such content hesitate to use any available standards such as Unicode due to their much delayed support of operating systems and rendering the Myanmar scripts. The other encodings are not given any name and can only be identified by their font names. Myanmar Web documents are very few in comparison with English Web pages, and most of the available Web documents are not searchable and, hence, not reachable due to the use of the non-standardized multiple encodings. Moreover, Myanmar Web pages are sparsely scattered over the entire Web. Therefore, finding the Myanmar Web pages from the World Wide Web is a difficult task.

In this study, a language-specific crawler is proposed. This crawler, which is capable of collecting the Myanmar Web pages with the multiple encodings, can be used as a language-specific search engine for Myanmar language.

1.3. Objective of the Study

The goal of the Language Specific Crawler (LSC) is to collect as many relevant pages as possible in the crawling process. LSC is very useful in several Web applications, such as collecting Web pages for language-specific search engines, archiving specific page collection for digital libraries, and so on.

2. Related works

In this section, we survey a number of crawling techniques related to our work. Most of the various techniques proposed are related to topic crawling since only a few are known for language focused crawling [1,2]

In the *Best-First Crawlers*; the best URL in the frontier list is selected first, as shown in the studies of [3] . In their technique, a set of keywords representing a topic guides the crawler by computing the lexical similarity between the current page *p* and the topic.

Fish-Search Crawler by [4] simulates crawling as a group of *Fish* migrating through the Web. The algorithm maintains a sorted list of URLs (*Fishes*) where the relevance of each page is estimated using a simple technique such as keyword match. More URLs are selected (*Fish* breeding) from pages with high relevance (areas with plenty of food) and less URLs are selected from pages of low relevance (fish migrating to areas with no food will die of starvation).

In the *Shark-Search Crawler* by [5], whichis an extension to Fish-Search, the score of a link is influenced by the anchor text, text surrounding the link, and the score of its parent page. *Crawling a country: Better strategies than Breadth First for Web page ordering*[6] describes a method for country-level Web crawling, but the focus of this work is on scheduling issues under network resource constraints, and the crawling targets are based on country level domain (e.g., .gr and .cl).

Pingali et al [7] conducted search engine research that gives users an opportunity to search in the 10 top Indian languages. Their search engine used the language focused crawler. The language identification module they used returns the name of the language for a given Web page. This module is aware of all the proprietary encodings and also uses a bag of words to recognize unknown encodings from META-tag that might be found in Hyper Text Markup Language (HTML) page.

Finally, some strategies were proposed by [8]. To avoid the social and technical difficulties for evaluation of Web crawling strategies, they implemented language-specific crawler by using a Web crawling simulator. For language identification, they used same as in Pingali's specific crawler. They checked character encoding scheme specified in HTML META declaration for a given Web page. In their strategy, they used two simple crawling modes: hard focused mode and soft focused mode. Their simulation results show that 100% crawl coverage in soft focused strategy, but it requires a large amount of memory to maintain the URL queue.

3. System Architecture

Language Specific Crawler (LSC) is very similar to the topic focused crawlers. But the LSC uses the language identification module instead of a classifier and so it is called a language-specific crawler.

LSC has two main components: a *crawler* and *language identifier*. The crawler is responsible for the basic functionality of a Web crawling system e.g.; downloading, link extraction and URL queue management, whereas the language identifier determines relevance of crawl pages and returns the name of the language for the given Web pages. The language identifier is aware of all proprietary encodings for Myanmar Web pages. In this research, it is not applies the technique of checking the META data of HTML pages for identification of Web pages. The language identifier used in this research is called Global Information Infrastructure Laboratory's 2nd version Language Identifier (G2LI). G2LI is explained in more detail in the following section.

3.1. Crawler Algorithm

The LSC fetches Web pages starting from a list of URLs called seed URLs. To create seed URLs for crawling process, well-known Myanmar Web pages are collected manually. These Web pages are encoded with various Myanmar encodings, and are expected to be good entry points to extract other Myanmar pages. LSC uses the breadth-first search (BFS) algorithm for First-in First-out (FIFO) processing. Here, LSC uses two queues, called *seedURLqueue* and

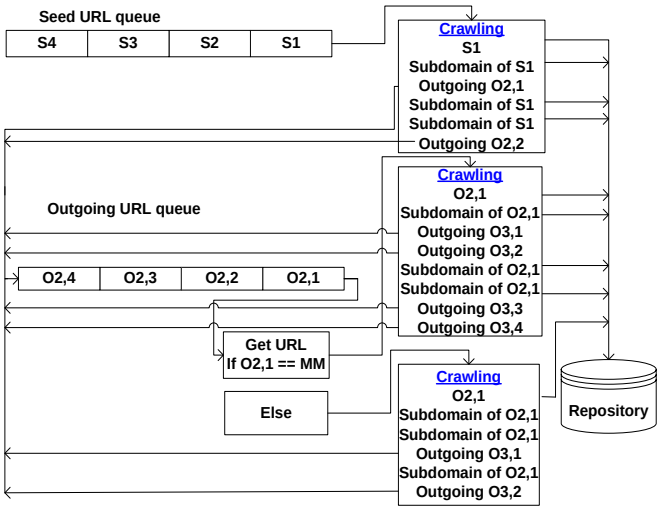
outLinkqueue. All seed URLs are saved in the *seedURLqueu*.as an initial entry.

A general Web crawler downloads all links from Word Wide Web. LSC is used to avoid the collecting of irrelevant pages. Overall architecture of LSC is depicted in Figure 1 (a). As shown in Figure 1 (b), all the Web pages in level 1 are Myanmar Web pages, which are actually the URLs saved in *seedURLqueue*.

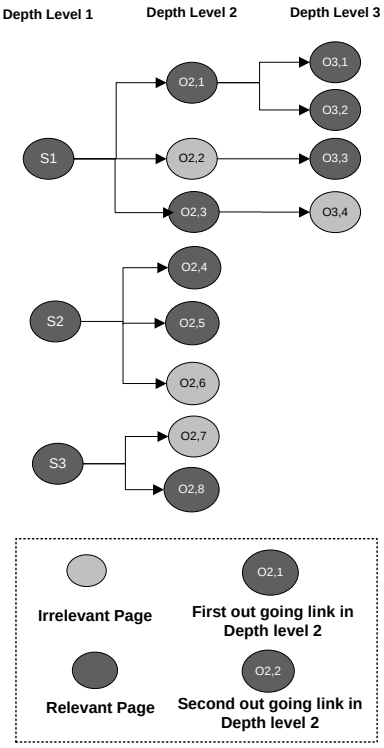
The LSC starts the first link from the *seedURLqueue* and extracts all subdomain links of first URL. If an outgoing link is found, it is saved in *outLinkqueue*. In this way, all of the subdomains in *seedURLsqueue* are put in a database (DB) and all outgoing links are put in *outLinkedqueue*.

Then the crawler starts again from the first link of *outLinkqueue*. At this point, language identifier needs to identify the page pointed outgoing link, because that page can be an irrelevant page. So, after retrieving the first link from *outLinkqueue*, the crawler passes that linked page to language identification module to determine whether it is relevant. If it is relevant, the link itself is saved and the subdomain of that link is also saved in the DB and all outgoing links are saved into the *outLinkqueue*. If it is irrelevant, the subdomain of that link is skipped but all outgoing links are saved in *outLinkqueue*and the link itself is saved in the DB, because these outgoing links can be relevant pages again. If these pages are skipped, the relevant pages in depth level 3, as shown in Figure 1 (b), would inevitably be missed. Myanmar Web page is not a disconnected component of the World Wide Web. According to [9], Web pages in Myanmar mostly link to documents in English.

Here, the accuracy of the language identifier is crucial, and will beexplained in more detail in the next section.



(a)



(b)

Fig. 1 Overall Architecture of LSC

Algorithm 1. Language Specific Crawler

Start with S1 (seed URL queue)
While *(end of seed URL queue)*
{
retrieve subdomain of S1;
if *meet outgoing links(O2,1);*
then *save outgoing links of S1 to outgoing URL queue;*
end
}
end
Start with O2,1 (outgoing URL queue)
{
if *(O2,1== relevant page)*
then
{
save O2,1 to Database;
retrieve subdomain of O2,1;
save outgoing links of O2,1 to outgoing URL queue;
}
end
else
save O2,1 to Database;
save outgoing links of O2,1 to outgoing URL queue
end
}

3.2. Language Identifier

For indexing Myanmar Web pages, a language identification module needs to determine whether a given document is written in Myanmar language. It is the heart of language-specific crawlers that distinguishes them from generic crawlers.

Myanmar language Web pages use different encodings and cannot be specified by checking HTML META declaration or other language detection tools, such as Mozzilla charset detector.

In this research, the n-gram based language identifier is chosen for language identification due to its high accuracy in identification, its resilience to typographical errors, and its minimal data requirement for Training Corpus (TC). The n-gram language identifier employed for this study is called Global Information Infrastructure Laboratory's 2nd version Language Identifier (G2LI) [10]. The G2LI method first creates n-bytes sequences of each training text. Then it checks the n-bytes sequences of the collected page with those of the training text. The language having the highest matching rate is considered to be the language of the Web page.

3.3. Preparation of Training Corpus

To identify correctly all Myanmar fonts and encodings, the G2LI must be trained by using sufficient collections of the training corpus. Therefore, nine text files written in different Myanmar encodings corresponding to almost all of the Myanmar fonts, shown in Table 1, were added to the TC. Not for being biased, all the text files used the contents of the Universal Declaration of Human Rights[1] (UDHR) text. As a result, all text files had the same content but different encodings. During the training stage, the identifier created the byte sequences of the n-gram for every training text.

4. Accuracy of LSC

LSC relies wholly on the technique of language identifier to selectively collect the relevant documents on the Web. To evaluate the performance of G2LI, the language identification result of the first 2,000 pages fetched by the crawler was manually verified page by page. The calculated accuracy rate and the error rate of the LSC are shown in Table 2.

By manual checking, 800 pages were Myanmar, of which 763 pages were verified as correctly identified by G2LI. Based on that account, the *precision* was calculated. The precision is the percentage of relevant pages correctly fetched by the algorithm. The precision of the language identifier was 763/ (763+37) = 95%. The recall for a set of retrieved documents is the ratio of the number of retrieved relevant documents to the total number of pages. In this example, the recall rate was 763/763+106= 87.8%.

Table 2: Identification result of language identifier

G2LI's Guessing	Verified Language		
	Myanmar	Non-Myanmar	Total
Identified as Myanmar	763	37	800
Identified as Non-Myanmar	106	1094	1200
Total	869	1131	2000

This language identification method showed overall accuracy of 93% (= (763+1094)/2000) with 7% of error rate on 2,000 test documents. While investigating the misidentification of the language identifier, two facts were revealed.

Out of 7% error rate, 1.8% of the errors correspond to "false-positive" (identified as Myanmar but actually not). And 5.2% of the errors correspond to "false-negative", most of which were identified as English but actually are found to be bilingual pages written both in Myanmar and English. As we want to collect as many Myanmar pages as possible, those bilingual pages need to be collected in the future. Also, some of the misidentified Web pages used numeric character reference e.g., (ာ&#4156, ့&#4153). These pages were misidentified as English in most cases but are Myanmar pages actually. The numeric character references referred in these cases, "&#4156" and "&#4153", are MYANMAR SIGN ANUSVARA and MYANMAR SIGN VIRAMA.

.

5. Crawling Experience

After continuous crawling of two months (September 2010 to November 2010) by setting 922 well-liked Myanmar Web sites as seed URLs, 1.6 million of Web pages (1.07 million Myanmar pages and 0.53 million non-Myanmar pages) were collected.

The following were set for the crawling parameters: 10 levels of depth and 8 crawling threads running on a server with a Linux operation system with 3GB memory, 512KB CPU.

After continuous crawling of two months (September 2010 to November 2010) by setting well-liked Myanmar Web sites as seed URLs, 1.6 million of Web pages (1.07 Myanmar pages and 0.53 non-Myanmar pages) were finally collected. The following were set for the crawling parameters: 10 levels of depth and 8 crawling threads running on a server with a Linux operation system with 3GB memory, 512KB CPU.

5.1. Crawling experience through its performance

In this section, the data collection performance of LSC is discussed. Here, we introduce a new evaluation parameter "crawling performance". In a crawling process, LSC has to collect non-Myanmar pages as well as Myanmar pages because LSC may

[1] Universal Declaration of Human Rights text translated in over 300 languages.
http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx

happen to encounter Myanmar pages after tracing links stretching from non-Myanmar pages. If LSC neglects all non-Myanmar pages, it may have a risk of missing those Myanmar pages which can be reached only through such links. At the same time, we want to minimize the cost of collecting non-Myanmar pages. For this reason, an evaluation parameter, "Crawling Performance" is defined as the number of relevant pages divided by the number of collected pages. If LSC collected 1,000 pages and the relevant pages were 800 of those, then the Crawling Performance is 0.8.

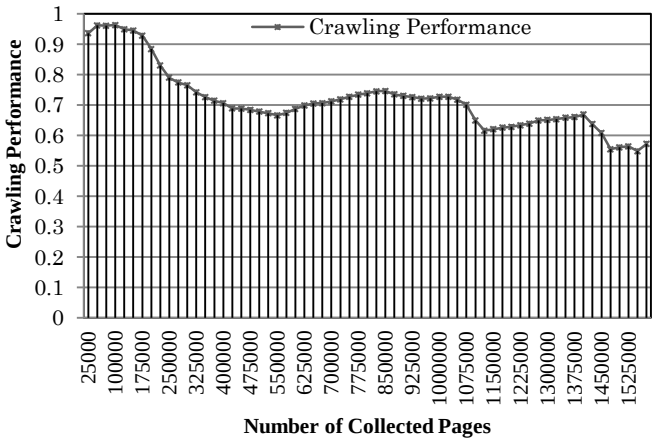


Fig. 2 Crawling performance

In Figure 2, the Crawling Performance of the LSC through the entire crawling process is shown. After crawling 0.2 million pages, the crawling performance of LSC decreased once, but it keeps some level. The nature of the Web crawling cannot be constant. The number of fetched pages in each point depends on the connectivity of Myanmar Web pages to another Myanmar pages. When the overall crawling process is observed, the crawler keeps collecting the relevant pages with Crawling Performance at around 0.6 to 0.7. It seems that the LSC has sufficient ability to collect Myanmar Web pages.

5.2. Evaluation on Crawling Coverage by means of Recall

As stated above, the evaluation parameter "recall" is the ratio of the number of retrieved relevant documents to the number of actual relevant pages on the entire Web. Recall can be viewed as the probability that the crawler retrieves all relevant documents. In other words, recall is a measure of how well the crawler performs in finding relevant documents.

Generally, the recall is hard to compute becauseit is impossible to obtain the exact figures of the relevant documents on the entire Web. To solve this problem, extra experiments were conducted to get an estimate of the total number of Myanmar pages on the Web.

5.3. Estimated total number of Myanmar Pages

This section discusses how to estimate the total number of Myanmar pages on the Web using the methodology proposed by [11]. Based on the result of this experiment, recallrate is calculated.

The proposed method assumes that two words A and B occur independently on the Web. Let N be the total number of Web pages and N_A, N_B, N_{AB} be the number of pages that contain A, B, and $A \cap B$.Then we can assume $N_{A \cap B}/N_A = N_B/N$ if these words occur completely independent manner. Hence N can be estimated by the formula $N = N_A.N_B / N_{A \cap B}$.

Actually, because words do not occur independently, the above estimation always has a bias. To obtain a better estimate, a pair of words that occur as independently as possible should be chosen. The authors tried to find such a pair of words through the following procedure.

In this experiment, 26 keywords were selected from randomly opened pages of "Myanmar-English" dictionary produced by Department of the Myanmar Commission, Ministry of Education, Union of Myanmar. Then occurrence of those words was manually checked using a collection of 92 sample Web documents. Those documents were collected from the URLs as shown in Table 3. Each URL represents different content types and is expected to contain various contexts, which are critical to be to check the occurrence of selected keywords.

Table 3: List of URLs used as seed URL for test document

URLs	***Content Type***
http://www.nineninesanay.blogspot.com	personal blog site
http://www.ainmat.com	personal blog site
http://www.shwepyithu.com	personal blog site
http://www.mysuboo.com	personal blog site
http://www.ald-exile.org	political site
http://www.mogokthar.com	entertainment site
http://www.ayinepan.com	entertainment site
http://www.nldlanorway.com	political site
http://hikmah.info	entertainment site

Then manual check results are summarized in a frequency count matrix f_{ij}, where f_{ij} means the number of occurrence of i^{th} keywords (W_i) in the j^{th} test document D_j.

	W_1 … W_{26}	total
D_1	$f_{1,1}$ … $f_{26,1}$	f_1
:		
D_{92}	$f_{1,92}$ … $f_{26,92}$	f_{92}
total		f

Then, the correlation of those 26 keywords was calculated to find the co-occurrence between them. Finally, the keywords having the lowest correlation were chosen as those considered to be independent of each another.

Table 4: The estimated total number of Myanmar pages on the Web (KW: keyword)

KW1	*KW2*	*the number of Google returned pages for KW1 (A)*	*the number of Google returned pages for KW2 (B)*	*the number of Google returned pages for KW1∩KW2 (A∩B)*	*N=A*B/A∩B*
သူ	ပေါင်းဖော်	218,000	268	120	486,867
(He)	အင်္ကျီ	218,000	11200	7030	347,312
	ဘုရား	218,000	229,000	105,000	475,448
	နတ်သမီး	218,000	37,600	8,890	922,025
	နာရီ	218,000	278,000	195,000	310,790
	လူ	218,000	270,000	241,000	244,232
	လာဘ်စား	218,000	4,110	3,220	278,255
	မှတ်ပုံတင်	218,000	61,000	25,400	523,543
	ကျောင်း	218,000	131,000	49,900	572,305
ပေါင်းဖော်	အင်္ကျီ	268	11200	2	1,500,800
(with together)	ဘုရား	268	229,000	123	498,959
	နတ်သမီး	268	37,600	5	2,015,360
	နာရီ	268	278,000	46	1,619,652
	လူ	268	270,000	65	1,113,231
	လာဘ်စား	268	4,110	1	1,101,480
	ကျောင်း	268	131,000	44	797,909
အင်္ကျီ	ဘုရား	11,200	229,000	7000	366,400
(clothes)	နတ်သမီး	11,200	37,600	14200	29,656
	နာရီ	11,200	278,000	1478	2,106,631
	လူ	11,200	270,000	47896	63,137
	လာဘ်စား	11,200	4,110	14578	3,158
	မှတ်ပုံတင်	11,200	61,000	18821	36,300
	ကျောင်း	11,200	131,000	70000	20,960
ဘုရား	နတ်သမီး	229,000	37,600	2,180	3,949,725
(God)	နာရီ	229,000	278,000	43,300	1,470,254
	လူ	229,000	270,000	60,600	1,020,297
	လာဘ်စား	229,000	4,110	1,300	723,992
	မှတ်ပုံတင်	229,000	61,000	3,190	4,378,997
	ကျောင်း	229,000	131,000	12,600	2,380,873
နတ်သမီး	နာရီ	37,600	278,000	3,110	3,361,029
(angel)	လူ	37,600	270,000	5,610	1,809,626
	လာဘ်စား	37,600	4,110	30	5,151,200
	ကျောင်း	37,600	131,000	1,040	4,736,154
နာရီ	လူ	278,000	270,000	156,000	481,154
(clock)	လာဘ်စား	278,000	4,110	1,470	777,265
	မှတ်ပုံတင်	278,000	61,000	31,400	540,064
	ကျောင်း	278,000	131,000	44,800	812,902
လူ	လာဘ်စား	270,000	4,110	2,630	421,939
(human being)	မှတ်ပုံတင်	270,000	61,000	29,600	556,419
	ကျောင်း	270,000	131,000	31,700	1,115,773
လာဘ်စား	မှတ်ပုံတင်	4,110	61,000	700	358,157
(bribe)	ကျောင်း	4,110	131,000	7200	74,779
မှတ်ပုံတင် (identity card)	ကျောင်း (school)	61,000	131,000	3,320	2,406,928
Average of 43 pairs of Keyword Combination					1,208,417

N= estimated total number of pages

Based on the 10 selected keywords, we obtained 45 pairs of combinations ($_{10}C_2$), as shown in the table 4. If two pairs of keywords (ပေါင်းဖော်, မှတ်ပုံတင်) and (နှတ်သမီး, မှတ်ပုံတင်), which showed the lowest and highest were excluded, the number of pairs of keyword combination was 43. The counting was done on Google search engine on 1 Dec. 2010.

Based on this result, the total number of Myanmar pages on the Web was estimated to be around 1.2 million pages Thus the recall rate, calculated by the ratio of total number of collected Myanmar pages (1.07 million) divided by estimated total number of Myanmar pages, becomes 0.89.

6. DISCUSSION

6.1. Evaluation of Crawling Coverage

As stated in the previous section, the recall ratio of the LSC is calculated to be 0.89. But actually there are many uncertainties on this figure. The estimation of the total Myanmar pages has big variations. The standard deviation of the estimate of the total Myanmar pages (1.32 million) is slightly more than the estimated average (1.2 million). Secondly, taking into account the precision ratio (95%) of G2LI, the real figure of the total number of Myanmar pages would be slightly larger than the original estimate (1.2 million / 0.95 = 1.26 million). But at the same time, collected Myanmar pages may contain 7 percentages of errors in average. Taking into account all these, revised estimate of the LSC's recall rate will become as follows. It means that our LSC collected Myanmar pages with 0.79 recall rate.

$$\frac{1.07\ million\ \times 0.93}{1.2\ million\ /0.95} = 0.79$$

6.2. Where Are the Myanmar Web pages?

6.2.1. By Country Code Top Level Domain (ccTLD)

In order to improve efficiency and coverage of crawling process, we need to know where located Myanmar Web pages. Also we need to know what kind of encodings is most frequently used by Myanmar Web pages. Then we can choose more appropriate seed URLs for next crawling. In the following sections, these information will be introduced.

Table 5: The number of Pages classified by Top-Level Domain

Top-Level Domains	*The Number of Pages Collected*
.com	878,427
.mm	43,581
.net	67,863
.org	81,864
.info	7,828
.com	878,427
Total	1,079,563

The above table shows the distribution collected Myanmar Web pages classified by country code toplevel domains (ccTLD). It is observed that most of the Myanmar Web pages (98%) are outside .mm domain. This is evidence that crawling of Myanmar Web pages in the .mm domain cannot give us good coverage of the Myanmar Web pages.

6.2.2. By Physical Location

With the Internet Address Locator tool of the WHOIS[2] service, the location of Myanmar Web pages can be identified. After downloading the pages, all URLs were filtered to obtain only the unique domains. We obtained 4,158 unique domain names and found the locations of the servers. The servers were geographically located in 26 countries. Figure 3 represents the first five countries having the highest number of servers. Most are located in United States and second most are located in Korea Republic, followed by Germany, Ukraine and Myanmar.

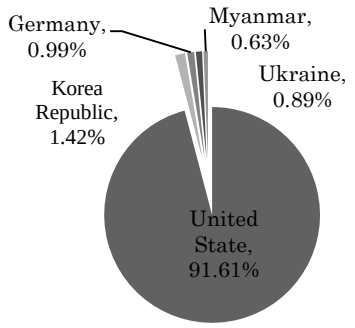


Fig. 3 Location of Web servers of downloaded pages

6.3. Validation of Encodings

6.3.1. Encoding vs. Number of Web Sites

The question what kind of encodings are mostly used in Myanmar Web pages are the big concern for us when preparing the training corpus for the next crawling. According to Figure 4, four different kinds of encodings are used in Myanmar Web pages. Among these, Zawgyi-One encoding is used most often.Very few Web sites publish content using UCS/Unicode encoding.

[2] WHOIS services provide public access to data on registered domain names, which currently includes contact information for Registered Name Holders. http://gnso.icann.org/issues/whois/

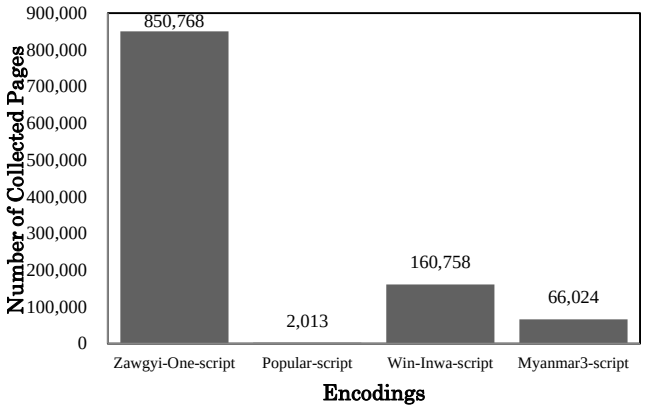


Fig. 4 The relation of encodings and the number of Web pages

6.3.2. ccTLD vs. Encodings

Table 6 shows the use of different encodings of the Myanmar pages on different kinds of gTLDs. Zawgyi-One encoding is used the most in all gTLDs, as shown in the table.

Table 6: The usage of various encodings on different ccTLDs

	.com	***.org***	***.net***	***.info***	***.mm***
Zawgyi-One-script	80.1%	67.8%	68.8%	63.2%	94.2%
Popular-script	0.1%	0.0%	0.0%	0.3%	0.1%
Win-Inwa-script	15.2%	14.9%	19.5%	20.8%	1.7%
Myanmar3-script	4.7%	17.3%	11.7%	15.7%	4.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%

7. Conclusions

In this paper, a language focused crawler designed for Myanmar language is proposed and its architecture and performance are presented. The main feature of LSC is that it can identify the various encodings of Myanmar Web pages. Through experiment, the LSC has successfully downloaded a satisfactory number of Myanmar Web pages. Several measures, such as accuracy rate of language identifier, the performance and the recall rate of the LSC, are presented. To calculate the recall rate, the extra experiment (estimation of total number of Myanmar Web pages) was conducted. Although this study focuses on the Myanmar Web pages, it is assured that the LSC can be easily adapted for constructing language-specific crawling in other languages.

Acknowledgments

We wish to acknowledge and express to team member of Myanmar Unicode and NLP research center for developing the various encoding of Myanmar fonts for training text.

References

[1]. Medelyan, O., Schulzl, S., Paetzold1, J., Poprat12, M., Mark, K., . Language Specific and Topic Focused Crawling.
[2]. Somboonviwat, K., Tamura, T., Kitsuregawa, M., 2006. Finding Thai Web Pages in Foreign Web Spaces.In proceeding of 22nd International conference on data engineering workshops. P.135
[3]. Cho, J., Garcia-Molina, H., Page, L., 1998. Efficient crawling through URL ording. Computer Networks, 30 P(161-172).
[4]. Bra, PME., De., Post, RDJ., 1994. Information Retrieval in the World Wide Web: Making client-based searching feasible. In proceeding of 1st International World Wide Web Conference.
[5]. Hersovici, M., Jacovi, M., Maarek, Y., S., Pelleg, D., Shtalhaim, M., Ur, S., 1998. The shark-search algorithm An application: Tailored Web site mapping. In proceeding of 7th International World Wide Web Conference.
[6]. Baeza-Yates, R., Castillo, C., Mar´ın, M., Rodr´ıguez, A., 2005. Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering. In proceeding of WWW2006 conference.
[7]. Pingali, P., Jagarlamudi, J., Varma, V., 2006. WebKhoj: Indian language IR from Multiple Character Encodings. In proceeding of WWW2006 conference, Edinburgh, Scotland.23-26.
[8]. Somboonviwat, K., Tamura, T., Kitsuregawa, M., 2005. Simulation Study of Language Specific Web Crawling.In proceeding of 21st International conference on data engineering workshops.
[9]. Rizza, C., C., and Mikami, Y., 2008. The Link Structure of Language Communities and its Implication for Language-Specific Crawling.In proceeding of 6th Asian Language Workshop, Hyderabad, India.
[10]. Choong, C., Y., Mikami, Y., "Optimization of N-gram Based Language Identification for Web Documents", Master Thesis. Nagaoka University of Technology, March 2007.
[11]. *Lawrence*,*S.*,*Giles*, C., L., 1998. Searching the World Wide Web, Science Vol.280, No. 5360, P(98-100).

Pann Yu Mon got her bachelor's degree, B.C.Sc, in Computer Science, from the University of Computer Studies Yangon, Myanmar, and her master's degree, ME, in Management Information System from the Nagaoka University of Technology, Japan. She is currently a Ph.D part 3 candidate in the Management and Information System Engineering Department of Nagaoka University of Technology. Her research interests include Web crawling, Web archiving and Web searching.

Chew Yew Choongis a Ph.D student at Information Science and Control Engineering Department of Nagaoka University of Technology, Japan. He holds a master degree of Management and Information Systems Engineering from the same University. His research interests include text algorithm, natural language processing and digital language divide on the Web. He has been working on automatic language identifier to efficiently identify human written language on the Web. He is a member of Language Observatory Project, Country Domain Governance Project and IEICE.

YoshikiMikamiis a professor of management and information science

at Nagaoka University of Technology, Japan. He has initiated several international collaborative projects, such as Language Observatory Project, the Asian Language Resource Network project, and the Country Domain Governance Project. He also serves as a chairman of Joint Advisory Committee for ISO registry of character codes. Mikami received a BE in mathematical engineering from the University of Tokyo and a Ph.D from the Graduate School of Media and Governance at Keio University.