MYANMAR WEB PAGES CRAWLER

Su Mon Khine, Yadana Thein

University of Computer Studies, Yangon
sumon5.8.1986@gmail.com, yadana@ucsy.edu.mm

ABSTRACT

Nowadays web pages are implemented in various kinds of languages on Web and web crawlers are important for search engine. Language specific crawlers are crawlers that traverse and collect the relative web pages using the successive URls of web page. There is very little research area in crawling for Myanmar Language web sites. Most of the language specific crawlers are based on n-gram character sequences which require training documents, the proposed crawler differ from those crawlers. The proposed system focused on only part of crawler to search and retrieve Myanmar web pages for Myanmar Language search engine. The proposed crawler detects the Myanmar character and rule based syllable threshold is used to judgment the relevant of the pages. According to experimental results, the proposed crawler has better performance, achieves successful accuracy and storage space for search engines are lesser since it only crawls the relevant documents for Myanmar web sites.

KEYWORDS

Language specific crawler, Myanmar Language, rule base syllable segmentation.

1. INTRODUCTION

The Internet provides valuable resource of all types and web area is grown exponentially day by day. Web pages are added by different site holders every times. Gathering the web pages manually for language specific search engine is not possible and realistic. Therefore search engine mainly rely on crawlers to create and maintain indices for the web pages. Web crawlers are short software codes also called wanderers, automatic indexers, Web robots, Web spiders, ants, bots, Web scatters [2].To collect the set Myanmar Web pages for search engine, crawlers, which traverses Web by following the hyperlinks and stored the download pages in a repository and used then by indexer component to index the web pages, are needed.

In comparison to general purpose crawlers which traverse all the pages on Web, language specific crawlers are collected only for specific languages on Web. Most of the language specific crawlers were implemented using n-gram character sequences to detect language, encoding schemes and scripts of training corpus, which is the basic method for text categorization and required trained documents in prior to classify language of web pages.[7] Some researchers detected language of web pages on Urls of top domain. Eda BayKan, Monka Henzinger, Ingmar Weber [5]determined the language of web pages using its URL of the country code of the top level domain by using machine learning classifiers such as Naïve Bayes, Decision Tree, Relative Entropy, Maximum Entropy and experimented English, German, French, Spanish and Italian Languages. Takayuki Tamura, Kulwadee Somboonviwat and Masaru Kitsuregawa [8] identified language of the Thai web pages by content type of HTML META tag firstly. If the content types are missed, checked then the content of web pages based on TextCat, a language guesser based on n-gram statistics. Myanmar web pages can't detect exactly language of web pages by checking the character set of HTML META tags since most of the web sites developers are not definitely identified for Myanmar character set in META tag. Furthermore, web pages can't identified its languages by using Urls of top domain since

Myanmar languages web pages are mostly distributed on other top level domain such as .com, .info, .net rather than .mm which is refer to Myanmar country. Therefore this proposed system relies on content of web pages for crawling in order to download the Myanmar web pages and the judgment of relevancy is easily determined by proposed rule based syllable percentage threshold. The crawling process in this system is based on crawler4j[1] and extends the crawler to collect only Myanmar web pages for further process of web search engine for Myanmar Language.

This paper is organized into seven sections. Literature reviews are discussed in the next section2. Section 3 describes various types of crawlers and some open source general web crawlers. Myanmar scripts, fonts and encoding on web are explained in Section 4. Section 5 describes the proposed crawler. Experimental results will discuss in section 6 and proposed system will be concluded in section 7.

2. LITERATURE REVIEWS

In this section, the topics related to this proposed crawler are discussed. AltaVista search engine introduced a crawling module named as Mercator [4], which was scalable, for searching the entire Web and extensible. Olena Medelyan, Stefan Schulz, Jan Paetzold, Michael Poprar, Kornel Marko , [6] they used n-gram model for text categorization tool based on content of web pages using standard crawler Nutch and checked the domain of web pages with training documents collections. Dr Rajender Nath and Khyati Chopra [2] discussed about the Web Crawlers: Taxonomy, Issues & Challenges. They classified the crawlers according to coverage area, mobility, topic –domain and load distribution to Unfocused and Focused Crawler , Mobility Based Crawler , Domain specific crawler and Based on Load Intra and Distributed Crawler respectively. They also discussed issues of Crawlers. Crawler used in this paper is related to Domain (Specific) crawler because it does not need to collect the entire Web, but need to retrieve and collect only Myanmar Web pages. Finally, the relevance of the web page is determined by rule based the syllable percentage threshold.

3. VARIOUS TYPES OF CRAWLERS AND SOME OPEN SOURCE CRAWLERS

General web crawlers are designed to download as many resources as possible from a particular web site. Trupti V. Udapure1, Ravindra D. Kale, Rajesh C. Dharmik [3] are discussed four different types of web crawlers: **(1) Focused web crawler** : Focused Crawler is the crawler that tries to download the pages which are related to a specific and relevant of a topic that users interest**.(2)Incremental crawler:** In order to refresh the download pages, crawlers replaces the old documents with newly downloaded documents frequently based on the estimate of how often pages changes.**(3) Distributed crawler:** Different crawlers are working in distributed form in order to download the most coverage of the web, in which central crawler manages all other distributed crawlers.**(4) Parallel Crawler :** Many crawlers runs in parallel and a parallel crawler consists of multiple crawling process and it may be local or distributed at geographically distant location. In addition to another, some of the general open source web crawlers that are widely used today are also listed in table 1.

Table1. Some of the general open source Web crawlers.

Types of Crawlers	Definition
Web SPHINX , WebLech	Website-Specific or featured for HTML Information extraction.
Nutch, Crawler4j, JSpider, Heritrix	Highly configurable, extensible and customizable open source Web Spider.
WebEater, HttpTrack ,Web-Harvest	Web site retrieval and offline viewing.
JoBo ,Arachnid ,Java Web crawler	Simple Web spider.

4. MYANMAR SCRIPTS

Myanmar language is the official language of Myanmar, spoken as first language by two thirds of the population of 60 million and 10 million as a second language, particularly ethnic minorities in Myanmar. Myanmar script draws its source from Brahmi script which flourished in India from about 500 B.C.to over 300 AD. Myanmar Script like the Brahmi script is a system of writing constructed from consonants, vowels symbols related to the relevant consonants, consonant combination symbols, devowelizer and digits. Myanmar script is composed of 33 consonants, 12 basic vowels, 8 independent vowels, 11 consonant combination symbols and 27 devowelizer [9] and is written from left to right in horizontal line. Table 2 shows the characters of Myanmar script.

Table2. Some Myanmar Characters

Names	Respective characters
Digits	၀၊၁၊၂၊.....၉
Consonants	က၊ခ၊ဂ၊.....၊အ
Vowels	ာ၊ား၊ိ၊ီ၊.....၊ို၊ း
Independent Vowels	အ၊က္က၊ဣ၊.....၊ဪ
Devowelizer	က်၊ခ်၊ဂ်၊.....၊ဟ်
Consonant Combination	ျ၊ြ၊.....၊ွ

The combination of one or more character not more than eight character will become one syllable; combination of one or more syllable become one words and combination of one or more than one words becomes phases and theses phases are combined into sentences. Finally, a paragraph is formed by one sentence or more than one sentences. Figure 1 shows the structure of Myanmar sentence and figure2 shows structure of Myanmar syllable (ကြောင်) that is equivalent to 'cat' in English and contains 6 characters ဂ၊ ◌ျ ◌ာ၊ဝ၊ဂ၊◌်.

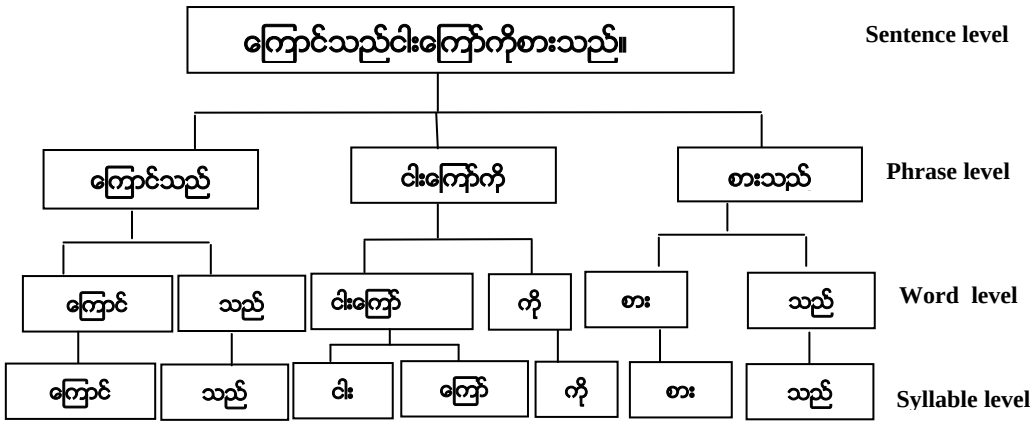

Figure1. Structure of Myanmar Sentence

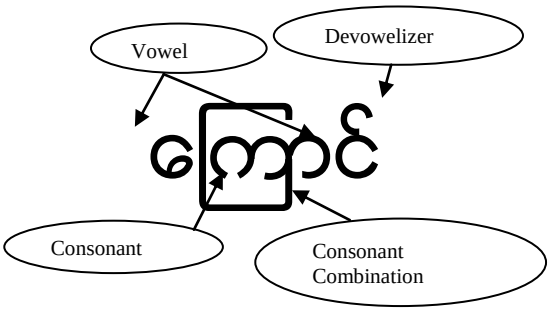

Figure2. Structure of Myanmar Syllable

4.1 Different fonts and encoding system for Myanmar Web sites.

The first generation of Myanmar encoding systems were ASCII code in which Latin English glyphs were replaced by the Myanmar script glyphs to render the Myanmar scripts which was no standardization of encoding characters. Firstly, Myanmar script was added to Unicode Consortium in 1999 as version 3.0 and improved Unicode5.1 in 2008 and Myanmar3, Padauk and Parabaik are in the range of U+1000 to U+109F. And then, various fonts such as Myazedi, Zawgyi_One have been created. Although Zawgyi_One is not Unicode standard, over 90% of Web sites use Zawgyi_One font, which are Although Unicode stores text in only one order and render correctly and Zawgyi–One can store text in several ways but superficially appear correct. Therefore, the proposed crawler converts all fonts to Zawgyi_One fonts and normalizes various writing style to one standard style. For example, user can write '့' ,'ိ'or 'ိ', '့' after writing consonant 'က' for syllable 'ကို' that is equivalent to 'Ko' in English. Table 3 shows different encoding sequences of Unicode and Zawgyi_One and Table 4 shows some examples of normalization of Zawgy_One character.

Table 3. Sequence style of using Unicode and Zawgyi_One for Myanmar Syllable

Fonts	Sequence Style
Unicode	က + ိ +ု =ကို
	က + ု + ိ =ကုိ
Zawgyi-One	က + ိ + ့=ကို
	က + ့ + ိ=ကို

Table 4. Normalization of Zawgyi_One character sequences.

Various forms of writing sequence	Normalize sequence
◌ိ◌ီ, ◌ီ◌ိ	◌ီ◌ိ
◌ေ◌ာ, ◌ာ◌ေ	◌ော
◌ေ◌ာ်, ◌ာ◌ေ်	◌ော်
◌ေ◌ာ်◌း, ◌ာ◌ေ်◌း, ◌ော်◌း	◌ော်◌း
◌ို◌ံ, ◌ိုံ◌, ◌ံ◌ို, ◌ုံ	◌ုံ
.....	....
◌ွ◌ဲ, ◌ဲ◌ွ, ◌ွဲ, ◌ဲွ	◌ွဲ

5. SYSTEM ARCHITECTURE FOR PROPOSED CRAWLER

The proposed crawler traverses identified famous Myanmar web sites seeds URLs systematically, it identifies all Urls containing in that page and adds them to the frontier, which contains the list of unvisited URLs. URLs from the frontier are visited one by one, fetch the web pages and parse the pages to parser to remove HTML tags in order to check Myanmar character. The proposed crawler normalizes various fonts to Zawgyi_One font since Zawgyi_One is mainly dominant font on Web pages. After normalization, the proposed crawler calculates the syllable threshold based on rule base syllable identification in order to judgment the relevant of the pages. It the web pages are relevant, store them in the pages repository in order to ready for indexer to extract the keywords of web pages. The process is repeated until the crawling process reach the specified depth of the crawler after starting from the specified seeds URLs **.**Figure3 shows the design of proposed crawler and the process flow of proposed crawler can be summarized in figure4.

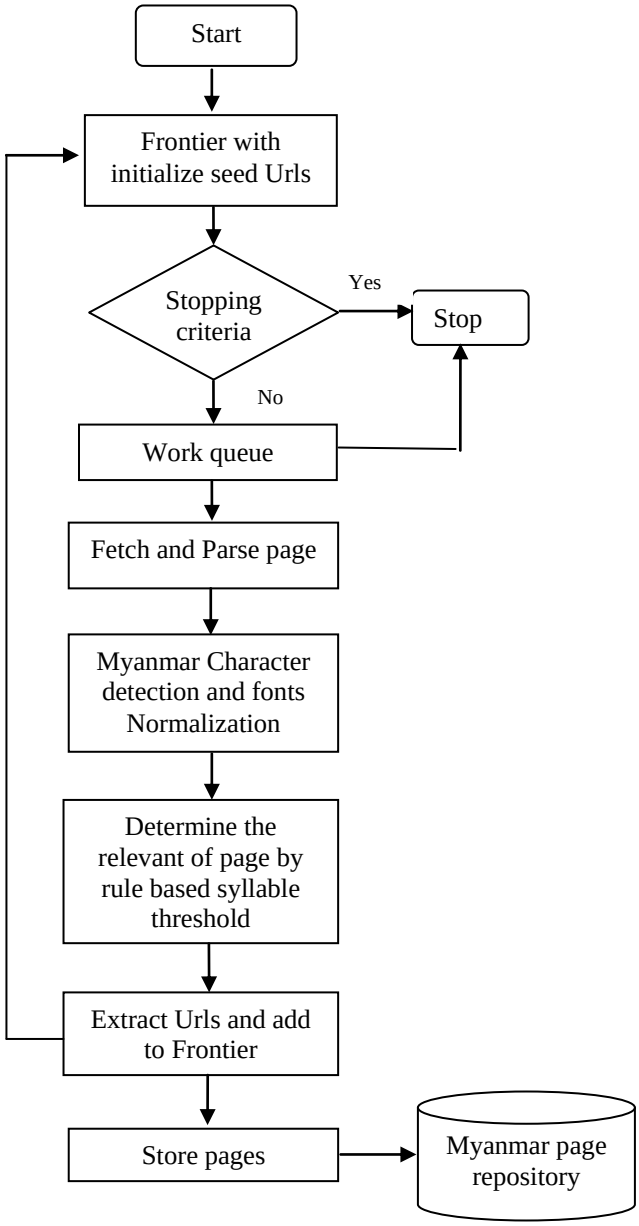

Figure3. The design of proposed crawler

1. Myanmar Languages web sites urls are put to the crawler as seed URLs.
2. The system check for stopping criteria.
3. If not reach the specified criteria, add URLs to the Work Queue.
4. Pick the URL up from the Work Queue.
5. Web pages are fetched and passed to parser in order to extract the content.
6. Myanmar characters are detected in the range between the decimal values of 4096 to 4255 defined by Unicode Consortium.
7. Various fonts are normalized to Zawgyi_One font.
8. The relevant of Myanmar Web pages are identified by proposed rule base syllable threshold.
9. Extract the Urls, add them to Frontier and pages are stored in repository.
7. Otherwise, discard the web pages.
Go to Step 2 and repeat when the specified
depth is reached.

Figure4. Process flow of proposed crawler**.**

5.1. Proposed rule based syllable segmentation

After detecting Myanmar characters and normalization to one standard font, the system segmented Myanmar sentences into syllable by proposed rule based syllable segmentation methods and calculate thresholds in order to identify the relevancy of Myanmar Pages since Myanmar Web pages are missed other languages .Proposed rule based syllable segmentation method is shown in figure 5. In here, the crawler are not considered the spelling checking of syllable since the proposed crawler only segmented content of web pages.

1. It we found one consonant and next character is not "္" or any consonant then take one syllable by combining the rest of characters until we found any consonants or 'ဍ' or 'ဋ'
2. If starting character is 'ဍ' or 'ဋ' and next character is consonants, take one syllable by combining the rest of character until we found another consonants or 'ဍ' or 'ဋ'
3. If first character is 'ဍ' and second character is 'ဋ'and next character is consonant , take one syllable by combining the rest of characters until we found another consonant or 'ဍ' or 'ဋ'

Figure5. Proposed rule based syllable segmentation method

Most of the Myanmar Web sites are not written only Myanmar Language and they are missed to other languages. For the combination of Myanmar and other language documents, Myanmar content exceed the predefined syllable threshold will be considered as relevant of Myanmar Web pages and stored them into page repository in order to further study of word segmentation and below the threshold will be discard as a non relevant pages to save storage space on disk. Threshold percentage is calculated the ratio of Myanmar Syllable count to the total numbers of Myanmar Syllable and other characters contained in that web pages. Figure 6 and 7 shows some example of web pages combined with other languages such as English Languages.

Figure 6 Greater threshold of Myanmar Syllable to other language

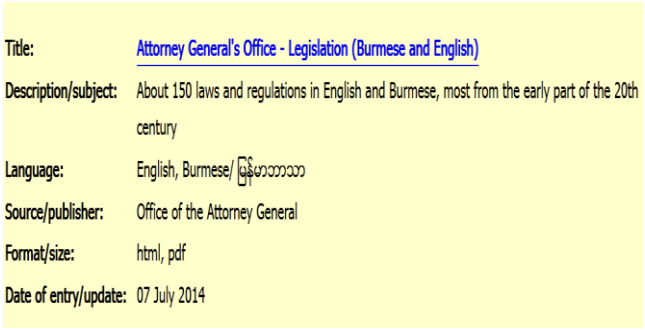

Figure7. Fewer threshold of Myanmar Syllable to other language

According to figure 6, the proposed crawler obtain 82% of Myanmar syllable percentage threshold to other language character and figure 7 obtain 2.5% of Myanmar syllable to other languages such as English language as shown in figure. The proposed crawler will regard figure 6 as relevant pages and stored in pages repository and figure 7 will be discard in order to reduce storage space in repository when syllable threshold is set to 3%. In this proposed crawler, users can easily define syllable percentage threshold depends on how much percentage of Myanmar language web pages to other language they want. It can easily to define and scalable.

6. EXPERIMENTAL RESULTS

6.1 Performance Evaluation

The evaluation methodology commonly and widely used in information retrieval is to calculate the precision. In the language specific crawling prospective, precision represents the ratio of the number of language relevant documents to the total number downloaded documents. Precision also called “harvest rate” in equation 1 is used for major performance metric for language specific crawler community.

$$\text{Precision (Harvest rate)} = \frac{\text{Language relevant pages}}{\text{Total download pages}} \quad (1)$$

6.2 Crawling experiment

In this section, the proposed crawler presents the result of experiment of crawler. The proposed crawler started with 11 Myanmar web site seeds URLs shown in Table 5, which are popular Myanmar Web sites. The crawler runs two times with 32 bit operating system, 4GB memory

with different internet downloads speed at day and night respectively. The first run of the crawler at 9: AM to 2: PM can download a set of only 8960 Html Myanmar documents with the depth of crawler is set to 7 and Myanmar syllable threshold to 4% .The second run of the crawler at 1: AM to 5: AM resulted in 12582 HTML documents with the depth of crawler is set to 10 and Myanmar syllable threshold to 3%. In total, 21812 documents were collected in this system and the results are shown in table 6. The result shown that fewer percent of syllable threshold can download more documents and greater percent of syllable threshold can download fewer documents.

Table 5. Myanmar web site seeds Urls

No	Urls	Description
1	http://www.president-office.gov.mm/	Information
2	http://my.wikipedia.org/wiki/.mm	Information
3	http://www.thithtolwin.com	News
4	http://www.7days.com	News
5	http://www.myanmarwebdesigner.com/blog/	Technology
6	http://winthumon.blogspot.com/2010/03/valueable-words.html	Literature
7	http://www.rfa.org/burmese/	News
8	http://hivinfo4mm.org/category/myanmar/	Health
9	http://www.myanmar-network.net/	Education
10	http://www.oppositeyes.info/	Politics
11	http://burmese.dvb.no/dvblive	News

Table 6. Different runs of crawler

	depth of crawler	syllable threshold	no of page collected
First run	**7**	**4**	**8960**
Second run	**10**	**3**	**12582**
Total			**21542**

It is a little difficult to calculate the precision of all download documents manually, the proposed crawler only calculates for first 1300 pages of each run. For the first run of crawler, by manually checking the relevancy of Myanmar pages ,1289 pages of 1300 are correctly download as Myanmar web pages and only 24 pages are download incorrectly and the precision was 98.15% . For the second run of the crawler, 1294 pages of 1300 are correctly download as Myanmar web pages and only 15 pages are download incorrectly and the precision was 98.84%. The experiments also evaluated that the proposed crawler outperformed n-gram based language identification which require sufficient training corpus for different fonts and encoding. The proposed crawler is not necessary training corpus and easily identify as Myanmar Language web site. Table 7 shows the average percentage of precision for proposed crawler and ngram-based crawler were 98.49 % and 96.6% respectively.

Table 7. Precision of the proposed crawler and n-gram based crawler

	Proposed crawler		N-gram Based Crawler	
	First run	Second run	First run	Second run
Correctly download as Myanmar pages	1289	1294	1192	1268
Incorrectly download as Myanmar pages	11	6	108	32
No of pages	1300	1300	1300	1300
Accuracy	99.15%	99.53%	91.69	97.54
Average Accuracy	99.34%		94.6%	

The crawler analyzed what kinds of top level domains are influenced on Myanmar Web sites. The average percentage of top level domain for Myanmar web sites in which the crawler downloaded are.com 83%, .mm 7% .org 5.2%, .net 3.24%, .info 0.92% and other for 0.56 respectively are shown in figure 8.

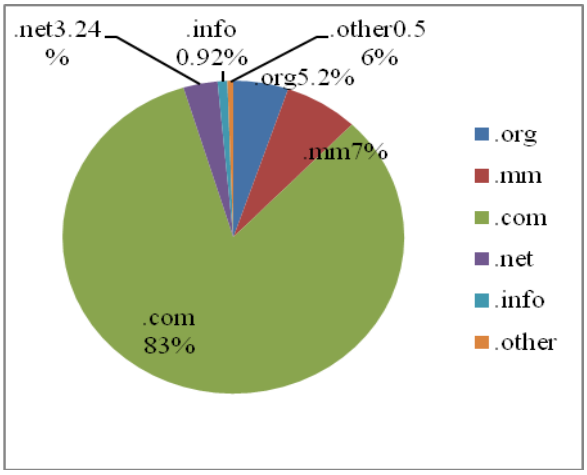

Figure8. Influence of different domains on Myanmar web sites.

The crawler also analyzed which fonts are mostly used for Myanmar web sites for each domain. Among them, Zawgyi_One is the widely used by web developer and Myanmar3 is the secondly used on Myanmar web site especially on governmental sites. Win Innwa is the thirdly used and the most rarely fonts is Padauk on Myanmar web sites according to result. Table 8 shows the fonts usage for each domain and Figure 9 shows bar chat representation for each font on each domain.

Table 8. Various fonts for each domain

	Zawgyi_One	Win Innwa	Myanmar3	Padauk	Total
.com	82.3	7.0	9.0	1.7	100%
.mm	76.5	2.0	20.0	1.5	100%
.org	87.4	4.0	7.6	1.0	100%
.net	86.0	4.0	8.3	0.7	100%
.info	92.7	4	3	0.3	100%
other	92.9	2	5	0.1	100%

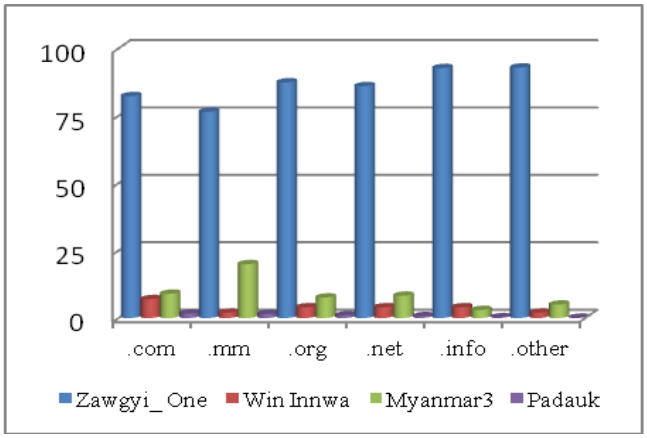

Figure.9 various fonts for each domain.

7. CONCLUSION

This system proposed language specific crawler in order to retrieve and download the Myanmar web pages for the supporting of web search engine for Myanmar Language. Myanmar characters of Web pages are detected and the relevance judgment of the web pages is determined by the proposed rule based syllable percentage threshold. This crawler can easily adjust the Myanmar syllable threshold in order to judgment the relevant of the pages. The proposed crawler can download various fonts written in web pages. This crawler also analyzes the various kinds of domain in Myanmar Language web sites and different fonts types for each domain. According to statistic, Zawgyi_One is the mostly influence in web pages and other fonts are fewer used on web pages. The proposed system is implemented in java language that is easy to install, develop and crawling speed is very high. The proposed crawler will improve the efficiency of language specific crawling for Myanmar Language in the future.

8. REFERENCES

[1] http://code.google.com/p/crawler4j/

[2] Dr Rajender Nath and Khyati Chopra, (2013)"Web Crawlers: Taxonomy, Issues & Challenges", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 4.

[3].Trupti V. Udapure, Ravindra D. Kale[2], Rajesh C.Dharmik[3], (2014)"Study of Web Crawler and its Different Types ", IOSR Journal of Computer Engineering. Volume 16, Issue 1, Ver VI ,PP01-05.

[4].Allan Heydon and Marc Najork , "Mercator: A Scalable, Extensible Web Crawler".

[5] Eda BayKan, Monka Henzinger, Ingmar Weber, (2008) "Web Pages Language identification based on URLs" .PVLDB '08 , Auchkand, New Zealand.

[6] Olena Medelyan, Stefan Schulz, Jan Paetzold, Michael Poprar, Kornel Marko "Language Specific and Topic Focused Web Crawling".

[7] Tomas OLVECKY (2005) "N-gram Based Statistics Aimed at Language Identification", M.Bielikova (Ed). IIT.SRC 2005, pp. 1-7.

[8] Takayuki Tamura, Kulwadee Somboonviwat and Masaru Kitsuregawa , (2007)"A Method for Language – Specific Web Crawling and Its Evaluation" , Systems and Computers in Japan , Vol.38, No.2.

[9]. Myanmar –English dictionary Department of the Myanmar Language Commission (2011).

Authors

Su Mon Khine received M.C.Sc and B.C. Sc, in Computer Science, from University of Computer Studies, Yangon. She is now PhD candidate in Information and Technology and currently doing research at University of Computer Studies, Yangon. Her research interest includes web crawling, information retrieval, natural language processing and data mining.

Dr. Yadana Thein is now working as an Associate Professor in University of Computer Studies, Yangon (UCSY) under Ministry of Science and Technology, Myanmar. She is particularly interested in Optical Character Recognition, Speech Processing and Networking. She published about 30 papers in workshops, conferences and journals. Currently, she teaches networking subject to under-graduate and post-graduate students. She supervises Master thesis and PhD research candidates in the areas of Image Processing.