# IMPROVING MYANMAR WEB SEARCH USING UNICODE

*Thin Zar Win, Khin Marlar Tun*
University Of Computer Studies, Yangon, Myanmar.
Email: thinzarwin07@gmail.com, marlartun@gmail.com

## ABSTRACT

As the web is increasingly hosting web pages indifferent languages, it is essential to be able to search for information stored in a specific language. A wide number of languages are spoken by human beings in the world and most of the people prefer to have information in their own language. Although there are so many search engines that exist on Internet, none of these can't fully support for searching with Myanmar language font. In this paper, we proposed a framework for Myanmar Unicode web searching. Our system first gathers the HTML documents of Myanmar Unicode Web Pages and then indexes the documents, searches and ranks the documents according to the users' given keywords. Finally, based on ranking scores, the most relevant documents are presented to the users. There are various kinds of Myanmar Unicode font using in Myanmar web pages. But there is no standard among these fonts and they use different code sequence. To handle these different kinds of Myanmar Unicode font, we also proposed unigram based font mapping table.

***Index Terms***— Myanmar web searching, Web-based information retrieval, Myanmar Unicode web search

## 1. INTRODUCTION

The world is increasingly using the Internet as a tool for different purposes like accessing information, communication, buying or selling of goods and services, etc. The size of the web has been increased rapidly during this year. If the information is so huge, it is impossible to get the relevant web page unless the URL of that page is known in prior. And it also becomes increasingly difficult and time consuming for the users to find the relevant information. As a result, we need a Search Engine to which we give the keywords and we get the relevant pages that we are looking for. Yahoo, Info-seek, Excite, AltaVista, Google are some search engines that offer extensive coverage by trying to index on the entire web. It is only confined to some Europeans languages.

As we all know, a wide number of languages are spoken by human beings in the world and most of the people prefer to have information in their own language. A person who is not particularly familiar with English should able to search with his native language. Although, there are so many search engines that exist on Internet, none of these can't fully support for searching with Myanmar language font. For these reasons, our proposed system intends to assist web searching for Myanmar Unicode language documents.

The rest of this paper is organized as follows. Section 2 introduces Myanmar Language. Section 3 reviews related work.. Section 4 presents overview of our proposed system architecture. Section 5 describes the experimental results of the system. Finally, concludes the paper in section 6.

## 2. UNICODE AND MYANMAR LANGUAGE

Myanmar language (formerly known as Burmese), is a member of Sino-Tibetan language family, according to the international language family trees. The Myanmar writing system derives from a Brahmi-related script borrowed from South India in about the eighth century for the Mon language. This system employs a letter to represent each syllable and consists of 33 symbols for consonants, 11 vowel symbols and various symbols to represent vowel sounds, tone marks, specified symbols and punctuation marks. Basic Myanmar consonants are shown in Table (1). It is the official language of Myanmar, where 32 million people speak it as their first language. Some people in China and India also speak Burmese.

Our Burmese Language is also a syllabic writing system that differs from English and many other western languages which are based on alphabetic .And Burmese Script is written from left to right and there are no spaces between words, although informal writing often contains spaces after each phrase. ASCII based fonts have been used for the Myanmar language data processing before Unicode. So, there were a lot of implementations and there is no standard among font encoding. After Unicode was invented, it is an international standard for all language. Myanmar Unicode was approved starting from Unicode 3.x. The range for Myanmar Unicode is from (U+1000 to U+109F).In Unicode 4.x, Unicode consortium defined standards for Myanmar Unicode Encoding Standards and Canonical Order. Nowadays, Myanmar Unicode fonts are widely used in Myanmar language Web Pages and documents. The latest Unicode version is Unicode 5.1 and

our proposed system intends to assist searching for any Myanmar Unicode font .

**Table 1: Basic Myanmar Consonants**

| | | | | |
|---|---|---|---|---|
| က | ခ | ဂ | ဃ | င |
| စ | ဆ | ဇ | ဈ | ည(ဉ) |
| ဋ | ဌ | ဍ | ဎ | ဏ |
| တ | ထ | ဒ | ဓ | န |
| ပ | ဖ | ဗ | ဘ | မ |
| ယ | ရ | လ | ဝ | သ |
| | ဟ | ဠ | အ | |

## 3. RELATED WORK

In [1] presented the linguistic processing technologies that incorporated in Infocious, a Web Search system designed to help users find information more easily by resolving ambiguities in natural language text. "Tumba", a new Internet search engine for the Portuguese Web was introduced in [2]. It uses a new repository architecture and implements innovative ranking and presentation algorithms.

The design and development of a search engine for the Indonesian Web have been proposed in [3]. And for Indian language, a search engine called WebKhoj [4] was implemented. It is capable of searching multi-script and multi-encoded Indian language content on the web. The design of a Tamil search engine [8] was proposed for Tamil documents on the web. The highlights of this search engine are its capability to handle multiple fonts, and the use of morphological analyzer. It also discussed the issues such as the crawler, the database storage architecture, the searcher and the functional modules of the search engine.

Another web-based Search engine for Indian languages is also proposed in [9]. It allows full-text indexing and searching of a database of HTML documents written in Brahmi-based Indian Language and English. Gatherer, Indexer and Search processor are the basic components of this search engine. It is capable of searching all phonetically equivalent words and different forms of keyword of any Indian language.

The architecture and algorithms of an Intranet search engine was proposed in [11] called D'alaxy. It employs a number of techniques to improve the search quality including relevancy ranking, personalization features, supporting Boolean operator, proximity information, spelling checker and thesaurus to aid the user. D'Galaxy is created for organizing intranet search within their university, MMU network. The challenges of web search and information retrieval in indexing the World Wide Web, the user behaviour, and the ranking factors was mainly discussed in [13].

A lot of research works on web information retrieval and web search engine have done for the variety of languages. In this paper, we propose a new framework for Myanmar Unicode web searching. And also there is a variety of Myanmar Unicode fonts used in the Internet and they are different in code sequence they define according to font developer. To overcome these difficulties, we presented unigram font mapping table, and as results of this, we can find any different kinds of Myanmar Unicode web pages.

## 4. SYSTEM OVERVIEW

As the number of non-English searches on the Web increases, our proposed system intends to improve in Myanmar language web search. Our system consists of two mainly steps, the first one is preprocessing and the second one is retrieving step. The general structure of our proposed framework is shown in Figure (1).

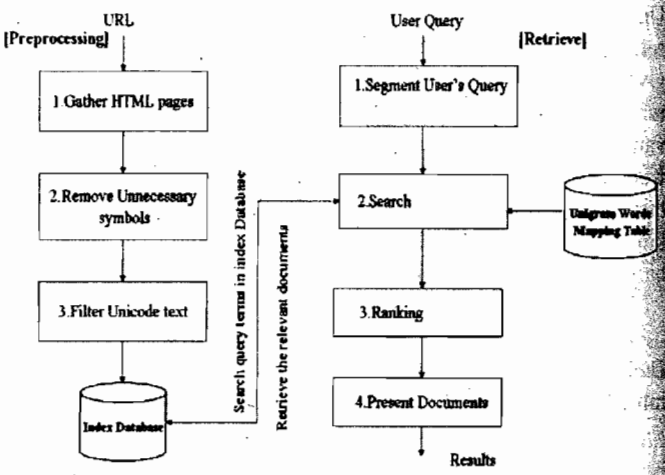


Figure.1 General Structure of the Proposed System

### 4.1. Prepocessing

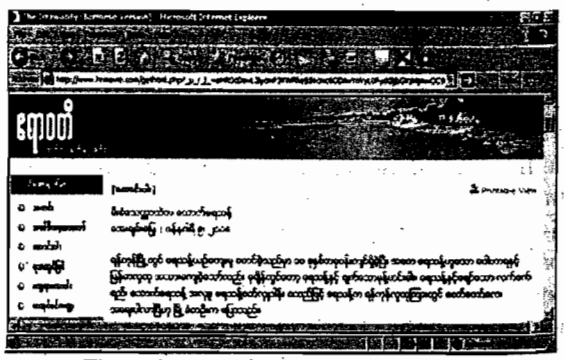

Figure.2 Myanmar Unicode Web Page

Firstly, our system takes a URL as input and works as a regular web user, sends HTTP request to the remote server for collecting the web documents. We store the entire web

site name and it's URL in our site table in index database. And then gathers HTML sources of Myanmar web pages from the web. After that it parses the whole page to remove script, style tags and all unnecessary symbols from these HTML documents and finds the link actually <a> tags. Web content is being written in different languages so we filter the Myanmar Unicode character and store these with related actual page links in index database.

English language is easy to separate the words because of having space between words. But there is no white space between Myanmar words, so to define word boundary for our language is a difficult task. Unlike English, Myanmar scripts are based on syllable and these syllabic words are composed of a combination of characters so we need to define word boundaries for each mono syllable words first.

The segmentation of syllabic words for Myanmar needs to identify the expressions or rules for the base character, vowel, consonants, medial etc. So, the content of the documents and users' query are tokenized as separated Myanmar words by using our own segmentation algorithm based on the predefined regular expression rules.

## 4.2. Retrieving

*4.2.1 Segment Query*

At retrieving phase, user query is accepted with Myanmar2 font which is based on Unicode version 5.1 and it is a standard font among Myanmar Unicode fonts. When the query enters into the system, we first tokenize the query as the mono syllable word according to our predefined regular expression rules for Myanmar scripts. And then we produce query into n-gram terms for both exact and partial match.

As n-gram approach is a language independent approach, most Asian languages used it effectively. If this approach is applied in Myanmar language, such as Myanmar word "စီးပွားရေးဂျာနယ်" –"Economic Journal" in English meaning, into tri-grams words, then , three tri-gram words "စီးပွားရေး" ၊ "ပွားရေးဂျာ" ၊ "ရေးဂျာနယ်" are obtained. The first tri-gram word "စီးပွားရေး" is meaningful word that means "Economic" in English , but the last two words have no meaning in Myanmar. To be effective search, we only produce the query into n-gram terms when the value of n is greater than or equal to 3 based on users' query, otherwise we directly search users' query.

*4.2.2 Searching*

After getting n-gram terms, we find these query terms in index database. And then record the word count for each document if the search term is matched. Finally, merge the result documents that contain all of the n-gram terms and retrieve other fonts relative to Myanmar2 font from Unigram font Mapping tables and search again.

*4.2.3 Unigram Mapping Table*

There are various kinds of Myanmar Unicode font using in web pages but they use different code sequence. To handle these different kinds of Myanmar Unicode font, we build unigram based font mapping table to the related searching. There are nearly 2000 unigram words are stored in our mapping table. In this mapping table, we use Myanmar2 font as our based font and store other Myanmar Unicode fonts related to Myanmar2 font. For this reason, our system can search any kinds of Myanmar Unicode font easily.

*4.2.4 Ranking*

Generally, most web search engine use two different kinds of ranking factors: query-dependent and query-independent factors. Query –dependent are all ranking factors that are specific to a given query, while query-independent factors are attached to the documents, regardless of a given query. The first one is also content-based ranking and these are measures like word documents frequency, the position of the query terms within the document or inverted document frequency. They also focus on finding the most relevant documents to a given query mainly by comparing queries and documents.

The second group of measures is used to determine the quality of a given document regardless of a certain query. The most popular of these factors is Page-Rank, which is a measure of link popularity used by the search engine Google. Such measures are necessary because there is a wide range from low quality to high quality documents on the web. The reason for this lies mainly in the link structure of the web. And every link is counted as a vote for the linked page. Link-based ranking only performs better in navigational, but not in informational queries. So, we use content based ranking to present relevant results to users based on word frequency according to users query.

*4.2.4 Present documents*

According to ranking scores the documents are presented to the users. The user interface of our system includes both querying facilities as well as display of query results in the same language in which the information is stored. The results returned by the system are including the Web links and their brief descriptions according to users' queries. The users can easily locate the links they want from among the results, and they may simply click the link to destination. The user interface of our system is presented in figure (3).

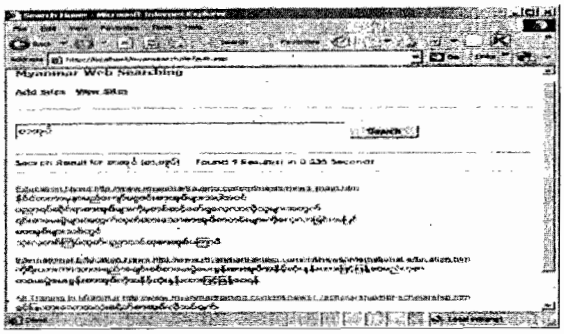

Figure.3. User interface of our system

## 5. EXPERIMENTAL RESULTS

There are many factors that together determine the quality of a Web search engine and web search tool. But the quality of information retrieval systems in general and search engines in particular is measured only with retrieval test. These takes into account standard measures like recall and precision and omit other factors that are not relevant in traditional information retrieval. The performance evaluation for our system is based on the precision and recall measure according to users' search keyword. In order to evaluate the performance of our system, we need to measure how far down the ranked list of results will a user need to look to find some or all the relevant documents.

## 6. CONCLUSION

As the Internet is becoming popular, the number of documents written in Myanmar language is also increasing day by day.Thus, information searching and retrieving on the Web have become more and more important. In this paper, we have presented a framework for searching Myanmar Language Web. Because of our proposed unigram mapping table, we can search every kinds of Myanmar Unicode web site easily by adding new fonts into our Unigram font mapping table. This paper is outgoing activities of our current research work and we have to prove our system's performance later. But, we believe that the framework proposed in this paper can be applied efficiently and effectively in Myanmar Unicode web searching.

## 7. REFERENCES

[1].A. ntoulas, G. Chao,J. Cho "The Infocious Web Search Engine: Improving Web Searching through Linguistic analysis", In Proceedings of the World Wide Web (WWW) Conference, 2005, Chiba.

[2]. M. J. Silva "The Case for a Portuguese Web Search Engine", IADIS International Conference WWW Internet 2003.

[3]. V. Berlian Vega SN, S. Bressan: Indexing the Indonesian Web: Language Identification and Miscellaneous Issues. WWW Posters 2001.

[4].P. Pingali, J. Jagarlamudi, V. Varma, "WebKhoj: Indian language IR from Multiple Character Encodings", *Electronic Edition* (ACM DL) BibTeX, (WWW 2006: 801-809)

[5].A N. Langville and C. D. Meyer, "Information Retrieval and Web Search.", *The Handbook of Linear Algebra*. CRC Press, 2006.

[6]. Martin Hosken, "Representing Myanmar In Unicode", Details and Examples.

[7].Hla Hla Htay, K.N.Murthy, "Myanmar Word Segmentation", ICCA 2007.

[8].J.Deepa Devi, Ranjani Parthasarathi, T.V. Geetha, "Tamil Search Engine", Resource Centre for Indian Technology Solutions-Tamil, School of Computer Science and Engineering, Anna University.

[9].Manoj Kumar malviya, "A Web-based Search Engine for Indian languages",Dept. of Computer Science & Engineering, Indian Institue of Technology, Kanpur, March ,1999,http://www.cse.iitk.ac.in/research/mtech1997/9711112.ps.gz

[10]. A. Natrajan, A. L. Powell and J. C. French, "Using N-grams to Process Hindi Queries with Transliteration Variations", ACM, 1997.

[11]. S. C. Haw, Y. W. Loh, S.K. Lua, "D'Galaxy:An Information Retrieval For Intranet Search", (IJCIM) Vol.14 No.1 (January-April 2006)

[12]. D. Gayo-Avello, D. Alvarez-Gutierrez, J. Gayo-Avello ,"Application of Variable Length N-gram Vectors to Monolingual and Bilingual Information Retrieval", In. 5th Workshop of the Cross-Language Evaluation Forum(CLEF) pages 73–82, 2004.

[13] Lewandowski, Dirk (2005) Web searching, search engines and Information Retrieval. *Information Services and Use* 18(3).