Web Search Engine for Myanmar Language

Yadana Thein, Su Mon Khine
University of Computer Studies, Yangon
yadana@uscy.edu.mm,sumon5.8.1986@gmail.com

Abstract

Web search engines play an important role for information retrieval on World Wide Web. There are many search engines according to their languages, but very few research areas for Myanmar Language. Search engine that can handle the typical characteristics of Myanmar Language is needed. The proposed Myanmar search engine contains the ***crawler, indexer, and query processor*** *components. In crawling,* ***proposed new rule based syllable language identification*** *is used to determine the relevancy of Myanmar Web pages. There are two new methods as contributed in indexing.* ***Proposed new Myanmar grammar based word segmentation and proposed new Myanmar word Sorting*** *are used in indexing structure to provide the most relevant documents to the users in least possible time. Documents are ranked according to their degree of relevance to the query words segmented by Myanmar grammar rules and documents using the cosine similarity measure.*

Keywords: *Myanmar Web search engine, Myanmar word sorting, Myanmar word segmentation, word based indexing.*

1. Introduction

Web Search engines are designed to search for information on the World Wide Web. The search results are generally presented in a list of results to users. Generally, there are two types of search engines: crawler-based search engines and human-powered directories. Crawler-based search engines, such as Google, create their listings automatically. Crawlers are programs used by search engines to gather information about web pages, index the documents and return hundreds of thousands of relevant responses to users. A human-powered directory, such as Yahoo, depends on humans for its listings.

Today, Myanmar language is the primary language of instruction and English is the secondly language in Myanmar. In recent year, Internet users in Myanmar greatly rely on search engines to get information from the Web. General purpose search engines, Google, Yahoo and Bing, are mainly designed for English language and hence not suitable for Myanmar nations to search information on Internet. General search engines cannot response relevant information to the requesters for the Myanmar Language query. Thus, efficient retrieval of Myanmar language web pages is a challenging new research area. Therefore, the proposed system is developed an effective search engine that can handle for Myanmar language documents with retrieval effectiveness than general search engine. In addition to another objective is to optimize speed and performance in finding relevant documents for an input query using proposed index structure based on nature of Myanmar words. The proposed search engine consists of crawler, indexer, query processor and is intended to design that considers the typical characteristics of web contents in Myanmar Language.

This paper is organized into six sections. In the next section2, the literature reviews are discussed. Section3 explains about Myanmar scripts and encoding system on Web. Section 4 describes the proposed system. Experimental results are discussed in section5 and the proposed system will be concluded in section 6.

2. Related Work

In this section, the papers related to the proposed search engine are discussed. Monica Peshave and Kamyar Dezhgosha [2] described how a search engine works and demonstrated the working of web crawler developed in Java language. Hassen Redwan, Solomon Atnafun [3] discussed the approaches for effective IR for Amharic web contents. They developed three components for Amharic Search Engine. The language specific crawler is developed to collect only Amharic contents. They also addressed the unique features of the language by using stemming, alias injecting during indexing process to obtain better performance.

In this paper, Myanmar words based indexing structure for search engine is developed and Myanmar words sorting algorithm is used for B+ tree index structure for each Myanmar consonant to retrieve the

most relevant documents and to improve searching time in retrieving documents.

3. 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, likes the Brahmi script, is a system of writing constructed from consonants, vowels symbols related to the relevant consonants, consonant combination symbols; and devowelizer. Myanmar script is composed of 33 consonants, 12 basic vowels, 8 independent vowels, 11 consonant combination symbols and 27 devowelizer [1] and is written from left to right in horizontal line. Myanmar word contains one syllable or more than one syllable and one syllable contains only one character or not more than eight characters. Figure 1 shows the structure of Myanmar syllable .There are nine Part-of Speech classes for all Myanmar words described by Myanmar Language Commission [1].They are Noun, Pronoun, Adjectives, Verb, Adverb, Postpositional Marker, Particles and Interjection. The proposed system mainly considered Noun, Verb, Adjective, and some adverbs for meaningful keywords segmentation and Pronouns, Postpositional Marker and Particles are defined as stop -words.

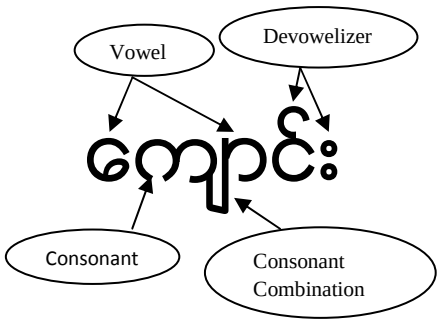

Figure 1. Structure of Myanmar Syllable

Table 1. Example of some Myanmar syllable

	Extended character						
consonants		ာ/ါ	ိ	.....	ျ	ျိ	ျောင်း
	က	ကာ	ကိ	.....	ကျ	ကျိ	ကျောင်း
	စ	စါ	စိ	.....	ရ	ရိ	ရောင်း
	.	.	.	.....	.	.	.
	.	.	.	.....	.	.	.
	အ	အာ	အိ	.....	---	----	----

In Table 1, syllable "ကား" (ကား) contains one basic character and two extended characters; syllable "ကျောင်း" (school) contains one basic character and six extended characters.

Example: က and ာ + း = ကား

က and ဇ +္+ာ+ ဝ +် + း = ကျောင်း

3.1 Myanmar Fonts and Encoding System on Web

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. Myanmar script was added to the Unicode Standard in September, 1999 with the release of version 3.0. Later, the release of the Unicode 5.1 Standard on 4 April 2008, three Unicode 5.1 compliant fonts have been available, including Myanmar3, Padauk and Parabaik and the range of Myanmar Unicode is from (U+1000 to U+109F). Various Myanmar font makers have created Burmese fonts including Win Innwa, CE Font, Myazedi, Zawgyi_One, Ponnya, Mandalay which are not Unicode compliant .It was extended in October, 2009 with the release of version 5.2.[5]. Although the meaning of Myanmar syllable is same, their encoding sequence is different in Unicode and Zawgyi_One. Although Zawgyi_One is not Unicode-compliant and does not meet the standards set by the Unicode Standard, it is currently and the most widely used solution for input in Myanmar –language websites. Unicode stores text in only one order and render correctly and Zawgyi_One can store text in several ways but superficially appear correct. Table 2 show different encoding sequences of Unicode and Zawgyi_One. Table 3 shows various types of font which follow the Unicode and does not follow the Unicode. The proposed system normalizes the different writing style of Zawgyi_One fonts to standard style in order to search efficiently for user.

Table 2.Unicode sequence style of Myanmar Syllable using Unicode and Zawgyi_One

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

Table3. Myanmar Fonts and their encoding

Encoding form	Types of fonts
Myanmar fonts which follow Unicode encoding for Window	Myanmar3
	Parabaik
	Padauk
	Thanlwin
	WinuniInnwa
	MyMyanmar
	Panglong
	Tharlon
Myanmar code which doesn't follow Unicode encoding	Zaygyi-One
	Ayar

4. Proposed Myanmar word based search engine design

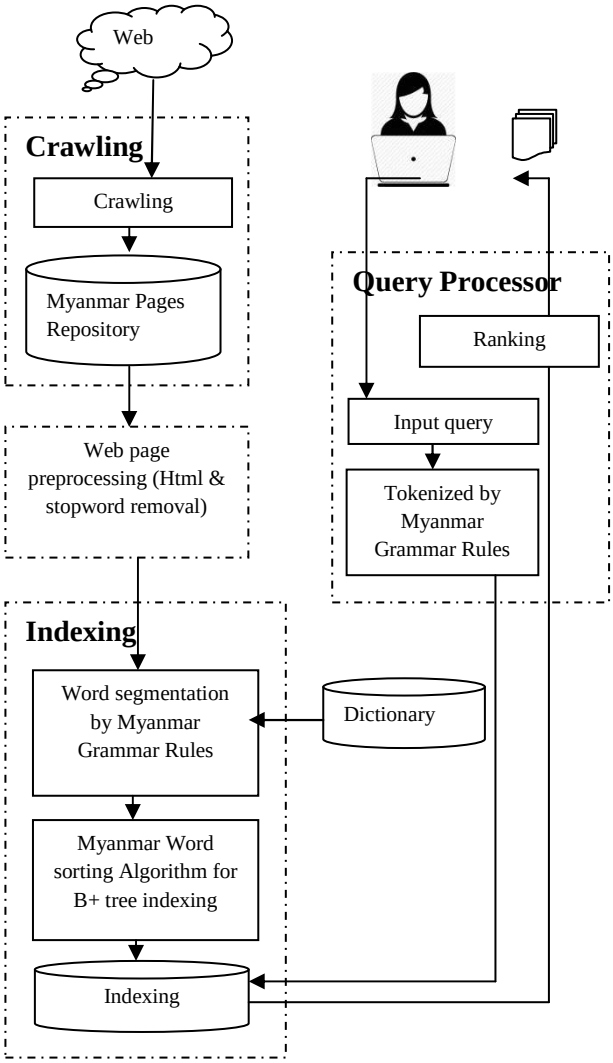

Figure 2. Proposed search engine design

The proposed system, Figure 2 consists of three parts, the crawler; indexer and query processor. The crawler only collects the Myanmar web documents for indexing. During the indexing process, Myanmar words for indexing are tokenized by proposed segmentation method. To optimize speed and performance in querying, inverted index and new Myanmar words sorting algorithm is used for B+ tree indexing structure. Finally, the query processor processes the user queries and documents are ranked according to their degree of relevance to the query words segmented by Myanmar grammar rules and documents using the cosine similarity measure.

4.1 Crawling Myanmar Web Pages

To collect Myanmar web pages, web crawler is an important component of web search engine. In this proposed system, web crawler is developed to download only the Myanmar Web pages. Language specific search engines are required to identify the language of a web page for further processing. Myanmar Web Page identification module discards the links and web pages for other languages. For the combination of Myanmar and other language documents, Myanmar content exceed the rule based syllable threshold will be considered as relevant pages for Myanmar Language and below the threshold will be discard. Relevant web pages are stored in the pages repository in order to ready for indexer to extract the keywords of web pages. The detail explanation of rule based syllable segmentation and crawling component are explained in [7].It can also segment Myanmar Parli and other words such as 'မေတ္တာ', 'တက္ကသိုလ်' and 'မဂ္ဂါ'.

4.2 Indexing

To optimize speed and performance, word based inverted index is constructed and B+tree index structure for each Myanmar consonant is used to access the inverted files. Word tokenization is an essential and important step in creating index terms and processing query for search engines especially for non-English documents. Therefore, Myanmar Word segmentation is needed to carry out additional task since Myanmar language does not use word boundaries like English. Myanmar words are formed syllables and syllable contains one or more characters. Firstly, rule based syllable segmentation is used for syllable segmentation and words are tokenized by proposed new Myanmar word segmentation algorithm.

4.2.1 Proposed Word Segmentation Method

Firstly, input sentence is segmented into syllables and remove the stop-words from the sentence such as preposition, conjunction, pronouns, particles that are repeatedly occurred in every document and that can degrade performance of web search system. Segmented phases are then segmented into words by finding the longest word in dictionary that matches the ending point of the phases. Figure 3 describes proposed words segmentation algorithm.We use Myanmar _English dictionary [1] for segmentation and we also added city names from Wikipedia to dictionary. We collected Myanmar proverb from [4] and add them to dictionary. Table4 shows some examples of stop-words.

Table 4. Some Myanmar stop-words

နာမ်စား (pronoun)	ကျွန်တော်၊ကျွန်မ၊ကျုပ်၊ထိုဟာ၊အဘယ်၊ဟောဒိ၊ ယင်းအရာ၊မိမိကိုယ်တိုင်၊မိမိဘာသာ၊ ကျွန်တော့်ကို၊ကျွန်တော့်အား၊ သူမအား...
ဝိဘတ် (post positional marker)	နှင့်အတူ၊နှင့်အညီ၊နှင့်အမျှ၊အနက်၊အထဲမှ၊ အထဲမှာ၊ဌာ၊အလိုဌာ၊အရာ၊အလျောက်၊ အားလျော်စွာ ၊များနှင့်အတူ ...
သမ္မန် (conjunction)	ဘဲနှင့်၊နှင့်တကြိုင်နက်၊နှင့်၊သောကြောင့်၊သော်လ ည်း၊သော်၊သဖြင့်၊သို့မဟုတ်၊သောအခါ၊လျှင်၊သာ မက၊သို့ရာတွင်၊ထိုအခါ၊ထို့ကြောင့်...
ပစ္စည်း (particle)	၊နိုင်တော့၊ပါလား၊ကြစို့၊စို့လေ၊သာလျှင်၊ကွဲသို့၊နှင့်ပ ါလား၊ကြပါစို့၊တည်းဟူသော ၊များသည် ၊များတွင် ၊တို့ ဖြင့် ၊ တို့ နှင့် ၊ တိတိ ၊ ကျစီ ၊စလုံး ....

We defined 1200 stopwords in our system. We also remove one syllable that are post positional marker of subjects (သည် ၊က၊မှာ), objects (ကို), departure (မှ), destination (သို့), usage (ဖြင့်၊နှင့်), reason(ကြောင့်), accept(အား), place (၌၊မှာ၊ဝယ်), time (၌,မှာ), possessive(၏) , aims(ဖို့) to get only meaningful words.

Some particles that demonstrate more than one things that come after noun, pronoun, verb, adverb "များကို", "တို့သည်", particles that modify counting of mathematical words "ကျစီ", "စလုံး" and some particles that deeply support noun such as "ကိုသာ", "များသာ", "ကိုချည်း" are considered as stop-words list for our system.

Example : ကလေး(n |နာမ်)***တို့ကို***(Particle|ပစ္စည်း)
ကျောင်းသား (နာမ်) ***များသာ***(Particle|ပစ္စည်း)
စာမေးပွဲ (n|နာမ်)***ကိုတော့*** (Particle|ပစ္စည်း)

In above examples, ***တို့ကို*** , ***များသာ***, ***ကိုတော့*** are also considered stopwords in this system.

Algorithm : WordSegment(Sentence S, Dictionary D, stopwords)
1. Input sentence is segmented into syllable .
2. L← Φ, p← Φ
2. **if** stopwords are contained in S, **then**
removed stopwords from S and retrieve phases P
3. **for** each phases p □ P **do**
4. **if** p is contained in D , **then**
Words ← Words U p;
if L ≠Φ and p= Φ **then**
p← reversed l and go to 4
5. **else if** p is not contained in D, **then**
p ← p-last syllable
L←L+last syllable and go to 4
6. **end if**
7. **end for**
8. return Words;

Figure 3. Proposed Word segmentation algorithm

Figure4 shows the output of the segmented words according to the above algorithm.

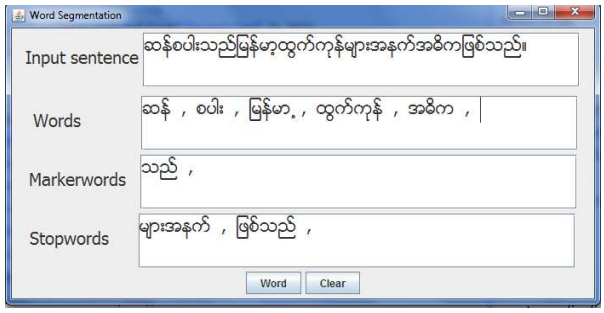

Figure4. Output of word segmentation

The proposed word segmentation algorithm can misunderstand in some words due to some particles such as 'အား', as in 'လူနာအားဆေးပေး': 'give medicine to patient' phrase. This proposed algorithm will deviate segment 'လူနာ: patient' 'အားဆေး: tonic',' ပေး: give ', instead of correctly segment 'လူနာ: patient ','ဆေး: medicine ',' ပေး: give '. The algorithm will be modify by adding some additional grammar rules and stemming to overcome such problems in later.

4.2.2. Myanmar Word sorting Algorithm for B+tree

The proposed system use B+tree index structure to access inverted file for speedup at searching time and document update efficiency. B+ tree define a **lexicographic order** over the data; the

complete value of a key is used to direct search. In this proposed system, Myanmar words are used as index terms and word sorting is important Myanmar character nature .Our algorithm sorts words by assigning weights to digits (D), consonants (C), vowels (V), devowelizer(DV) , and consonants combination (CC) symbols in ascending order. Firstly, input keywords are segmented into syllables and then we compute the weight of syllable and sorts Myanmar words according to the weight of each syllable .Table5 shows the assumed weight to D, C, V, DV, CC respectively and Figure 5 show our proposed word sorting algorithm.

Table5. Assume weight assignment of D, C, V, DV, CC

Names	Respective characters and their weight
consonants	ကခဂဃင ……..ဠအ 10,11,12,13,14,…….41,42
vowels	အ၊ အာ၊ အာ ……..အို၊ အိုး 43 ,44,45,……..,66
Devowleizer	က်၊ ၏၊ ၌ ……၊ ဟ် 67,68,……..,92
Consonant combination	ျ၊ ြ၊ ……..ၣ 93,94.,……,103

for B+ tree index structure. We proposed new Myanmar Word sorting algorithm based on

Algorithm MW_Sorting(W_1, W_2)
1. Assign weight to C, V, D, CC in ascending order .
2. MaxW ← Φ
3. S1 (s_1w_1,s_2w_1,….,s_nw_1)←syll-segment(W_1)
   S2 (s_1w_2,s_2w_2,….,s_nw_2)←syll-segment(W_2) // input words are segmented into syllable
4. **for** (k=0;k<min(W_1,W_2).length-1;k++) **do**
5. Result← Compare_Syll (s_kw_1, s_kw_2)
6. **if** result equal to 3 then MaxW←W1 break;
7. **if** result equal to 2 then MaxW← W2 break;
   if result equal to 1 **, then**
8. **if** k is last index ,**then**
9. **if** W1 length is equal to W2, **then** W1 is equal to W2
10. **else if** W1.lenght>W2.length , **then** MaxW←W1
11. **else** MaxW←W2
12. **end for**
13. **Return** MaxW

Function **Compare_Syll** (s_1w_1, s_1w_2)
1. [C, V, D, CC] ← s_1w_1 // separate each syllable into 4 parts
   [C, V, D, CC] ← s_1w_2
2. result ← compare_Val (s_1w_1 [C], s_1w_2 [C])
3. **if** result equal 1 , then
4. **for** (k= s_1w_1.lenght-1; k>0;k--) **do**
5. result = compare_Val (s_1w_1.[k], s_1w_2.[k])
6. **if** result ≠ 1 **then**
7. go to 10.
8. **end for**
9. **end if**
10. **return** result.

Function **compare_Val** (v_1,v_2)
1. if (v_1>v_2) then result ←3; // compare their weight
2. if (v_1<v_2) then result←2; //
3. else result=1.
5. return result

Figure 5. Word sorting algorithm

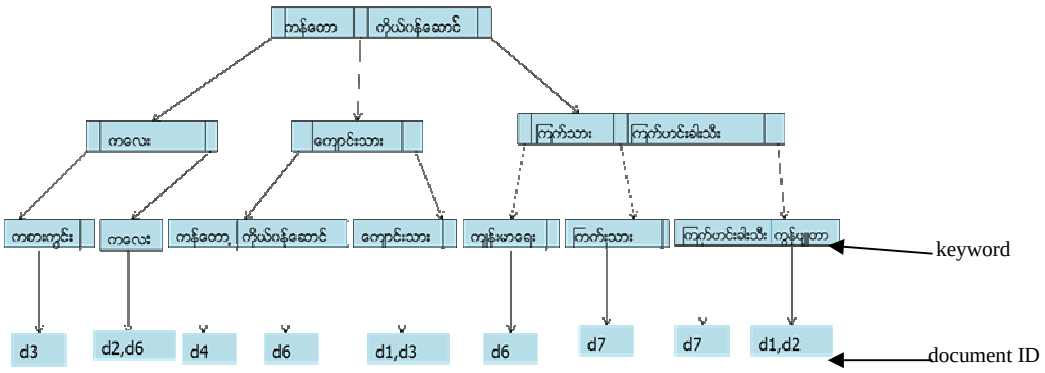

Figure 6. Sample index structure for Myanmar character word ' တ ' .

In this paper, we also explained how to implemented B+ tree index structure for Myanmar words in order to provide rapid response back to the web users and to improve the searching time in the indexed database. There are 33 consonants and we constructed 33 B+ tree for each Myanmar consonant. Figure 6 shows the access method of inverted file for Myanmar words that begin consonant 'တ' .

4.2.3 Document ranking

Query processor accepts the user query and tokenized the query by proposed segmentation method and search in the index database to rank according to the degree of relevance of the query. In this system, cosine similarity measure, equation (1) which

is the cosine of angle between the query vector q and document vector d_j, is used to calculate the similarity of the query q to each document collection D [6].

$$\cos ine(d_j, q) = \frac{(d_j \cdot q)}{|d_j||q|} = \frac{\sum_{i=1}^{|V|} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{|V|} w_{i,j}^2} \sqrt{\sum_{i=1}^{|V|} w_{i,q}^2}} \quad (1)$$

in which , w_{ij} is the final TF-IDF term weight , w_{iq} is the weight of term i in query q , tf_{ij} is normalized term frequency and idf_i is inverse document frequency

$$w_{ij}=tf_{ij}*idf_i \quad (2)$$

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{ij}, f_{ij}, ..., f_{ij}\}} \quad (3)$$

$$idf_i = \log\left(\frac{N}{d_{fi}}\right) \quad (4)$$

and the weight of term i in query q is :

$$w_{iq} = (0.5 + \frac{0.5 f_{iq}}{\max\{f_{1q}, f_{2q}, ..., f_{|V|q}\}}) * \log N / dfi \quad (5)$$

where,

f_{ij} is the raw frequency count of term t_i in document d_j, f_{iq} is the raw frequency count of term t_i in query q ,N is the total number of documents in the collection ,df_i is the number of documents that contains the term t_i , w_{iq} is the weight of term i in query q.

5. Experimental Result

This section describes the experimental results of crawling, indexing and query processing components.

5.1 Crawling experiment

In crawling experiment, the crawler started with 5 Myanmar web site seeds URLs which are popular Myanmar Web sites. The crawler run with 32 bit operating system, 4GB memory, from 9: AM to 12: PM, with the depth of crawler is 10 and Myanmar syllable threshold is set to 4 %. The crawler can download 1827 Html Myanmar documents.

5.2. Indexing and querying experiments

This section discusses for query processing time which is key for interactive application for one word query and phrase query that does not contain space between words using normal inverted file and proposed index structure for searching of keywords. In this system, the query processing times are computed among indexed keywords segmented from 500 crawled web pages. The following query words are used to test for the query processing time. Q1 to Q3 are single word queries and Q4 to Q5 are phrase queries that do not contain space between the words.

Q1: တက္ကသိုလ် (university)

Q2: ကျောင်းသား (student)

Q3: ရန်ကုန် (Yangon)

Q4:ကျန်းမာသောအစားအစာများ (healthy foods)

Q5:ကလေးအတွက်ကွန်ပျူတာ(Computers for children)

Q6:တောင်ပြုံးပွဲသို့သွားသောကားများ (Cars for 'Taung Pyone' festival

Table 6 shows query processing times for single words queries and Table 7 shows query processing times for phrase queries.

Table 6. Query processing times for single word query

	Query processing times	
Query	Normal Inverted index	Proposed Index structure
Q1	0.49sec	0.43sec
Q2	0.31sec	0.26sec
Q3	0.54sec	0.51sec

Table 7. Query processing times for phrase queries.

	Query processing times	
Phrase query	Normal Inverted index	Proposed Index structure
Q4	2.6sec	2.1sec
Q5	2.1sec	1.9sec
Q6	3.4sec	3.0 sec

According to the results shown in above Table 6 and 7 the proposed index structure outperformed than normal inverted index in query processing times.

The proposed system also compared these three phrase queries to other search engines, Google and Yahoo search engine. General search engine cannot handle for some Myanmar phrases that user types successively without space between words. The proposed search engine can perform well for that phrase queries since it segmented the phrase into words and search them in the index database. Table 8 shows the return results for three phrase queries.

Table 8. Experimental result with different phrases query

Phrase Queries	No of pages returned by Google	No of pages returned by Yahoo	No of pages returned by Proposed system
Q4	(0%)	(0%)	43 (8.6%)
Q5	(0%)	(0%)	20 (4%)
Q6	(0%)	(0%)	9(1.8%)

After that, the proposed search engine compared to other Myanmar Language search engine [8] in which, word based indexing including stop words are proposed. The retrieval performances of search engine is measured by precision and recall which is commonly measure in information retrieval. Precision is fraction of retrieved documents that are relevant and recall is the fraction of relevant docs that are retrieved. Table 9 describes the measurement results of other Myanmar Langue search engine in which word based indexing including stop words and proposed system that removes the stop words on query processing.

Table 9. Measurement results for word based indexing including stop words and proposed system.

Query	Word based indexing including stop words		Proposed system (Word based index that does not include stop words)	
	Precision	Recall	Precision	Recall
Q4	0.66	0.8	0.79	0.86
Q5	0.54	0.63	0.9	0.72
Q6	0.64	0.9	0.77	0.89

6. Conclusion

In this system, web search engine design for Myanmar Language which consists of crawler, indexer and query processor components are proposed. The proposed search engine can search both Unicode and non-Unicode web sites of different fonts and encoding on Web developed by different web developers since normalization of different writing style to Zawgyi-One fonts allows the users to search for various types of writing style. In this system, the crawler can download relevant Myanmar web pages by using rule based syllable threshold and discards the web pages below the threshold. It also proposed new Myanmar Word sorting algorithm and an efficient index structure for dynamic document collection by using B+ tree index structure. Constructing of one B+ tree index structure for each Myanmar consonants will optimize speed and performance in finding relevant documents for user queries. Finally, the proposed system can retrieve the most relevant documents due to the word based indexing rather than the syllable based indexing which will retrieve undesired result.

7. References

[1] Myanmar –English dictionary Department of the Myanmar Language Commissionı2011.

[2] Monica and Kamyar Dezhgosha "How search engine work and web crawler application".

[3] Hassen Redwan, Solomon Atnafun. "Design and implementation-Algorithms of Amharic Search engine for Amharic Search Engine System for Amharic Web Contents".

[4] Myanmar proverbs can be available at

[5] http://www.mmproverb.com/

[6] http://en.wikipedia.org/wiki/Burmese_alphabet

[7] Bing Liu "Web Data Mining, Exploring.

[8] Su Mon Khine ,Yadanar Thein ,"Myanmar web pages crawler", Fourth International Conference on NLP, Sydney, Australia.

[9] Pann Yu Mon, Chew Yew Choong, Yoshiki Mikami, “Myanmar Language Search Engine”, international Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.