Journal of Computer and Communications, 2016, 4, 79-89
Published Online August 2016 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2016.410009

Performance Analysis of Optimized Content Extraction for Cyrillic Mongolian Learning Text Materials in the Database

Bat-Erdene Nyandag[1], Ru Li[1], G. Indruska[2]

[1]School of Computer Sciences, Inner Mongolia University, China
[2]IT Consultant, Destination Consulting Co., Hattisaar Hub, Putalisadak, Kathmandu, Nepal
Email: 1918611467@qq.com, baterdene@muls.edu.mn

Received 31 May 2016; accepted 28 August 2016; published 31 August 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

This paper had developed and tested optimized content extraction algorithm using NLP method, TFIDF method for word of weight, VSM for information search, cosine method for similar quality calculation from learning document at the distance learning system database. This test covered following things: 1) to parse word structure at the distance learning system database documents and Cyrillic Mongolian language documents at the section, to form new documents by algorithm for identifying word stem; 2) to test optimized content extraction from text material based on e-test results (key word, correct answer, base form with affix and new form formed by word stem without affix) at distance learning system, also to search key word by automatically selecting using word extraction algorithm; 3) to test Boolean and probabilistic retrieval method through extended vector space retrieval method. This chapter covers: to process document content extraction retrieval algorithm, to propose recommendations query through word stem, not depending on word position based on Cyrillic Mongolian language documents distinction.

Keywords

Cyrillic Mongolian Language, Content Extraction Formatting, Learning Text Materials Style

1. Introduction

Basic training material and data distinction:

How to cite this paper: Nyandag, B.-E., Li, R. and Indruska, G. (2016) Performance Analysis of Optimized Content Extraction for Cyrillic Mongolian Learning Text Materials in the Database. *Journal of Computer and Communications*, **4**, 79-89.
http://dx.doi.org/10.4236/jcc.2016.410009

Problems related to natural language always followed and studied any kind of research work in Mongolia. New Mongolian or Cyrillic Mongolian language is official language of Mongolia. All levels of academic education are operated in Mongolian as a natural language completely.

Cyrillic Mongolian language included a kind of agglutinative language and words were depended on rules for word generating and inflecting. A word generating is based on attaching suffix and affix to word stem [1] that rules completely different from other languages. For Cyrillic Mongolian language, a word generating is based on attaching suffix and affix to word stem [2]. For example: general structure type is as following: "word + root morphemes + affix morphemes".

Root morphemes: indicating main idea and not possible to parsing word structure [3]. For example: "хүн", "ном", "үзэг" etc.

Affix morphemes: It is named all possible morphemes to attached root morphemes. Affix morphemes are divided into two categories as a generating suffix and attaching affix. A generating suffix is inflecting word and generating new word. For example: "Хүн + лэг", "Ном + хон", "Үзэг + дэл" etc.

Attaching affix indicated the relationship between two words. For example: "Хүн + ээс", "Ном + ын", "Үзэг + ээр" etc.

Word stem: main part of the word and inflecting by any affix [1].

It has to face several challenges during calculating text documents that depend on the features of Cyrillic Mongolian. In order to rule for attaching word affix, word stem should be described as a "word stem + affix 1 + affix 2 +⋯+ affix N".

Morphological position at Cyrillic Mongolian language has distinctions from other languages. It is the biggest problem for calculating. For example: A comparative example of the most widely used language is shown in **Table 1** and **Table 2**.

There are several opportunities to express this meaning except this example in Cyrillic Mongolian language according to **Table 1** and **Table 2**. For Cyrillic Mongolian language, principal noun members are free to change their positions. In that case, **Table 1** or **Table 2** shows that sentence meaning will not change.

Principal noun members don't have constant position. Therefore, it has been facing challenge to search full sentence. It shows that word sequence at the sentence or full sentence search are not optimized method.

Table 1. Cyrillic Mongolian morphological position at the sentence is compared to English and Chinese.

Languages	Morphological position at the sentence					
	(1)	(2)	(3)	(4)	(5)	(6)
English	She (1)	to gave (2)	him (3)	the (4)	book (5)	(6)
Chinese	她 (1)	给 (2)	他 (3)	这本 (4)	书 (5)	(6)
Cyrillic Mongolian	Тэр (1)	өгсөн (2)	түүнд (3)	энэ (4)	номыг (5)	(6)

Table 2. Cyrillic Mongolian morphological position at the sentence.

Cyrillic Mongolian			Morphological position at the sentence			
positions	(1)	(2)	(3)	(4)	(5)	(6)
position 1	Тэр (1)	Өгсөн (2)	Түүнд (3)	энэ (4)	Номыг (5)	6
position 2	Энэ (4)	Номыг (5)	тэр (1)	түүнд (3)	Өгсөн (2)	6
position 3	Тэр (1)	Түүнд (3)	энэ (4)	Номыг (5)	Өгсөн (2)	6
position 4	Түүнд (3)	энэ (4)	Номыг (5)	тэр (1)	Өгсөн (2)	6
position 5	Энэ (4)	Номыг (5)	Түүнд (3)	тэр (1)	Өгсөн (2)	6

2. Methodological Sequence

Main purpose of this research work was not dedicated to language processing for characteristic of Cyrillic Mongolian language. This research work focused on optimized content extraction from training document at the distance learning training system database. Collection of documents should be written on Cyrillic Mongolian language. It will be reached successful result for optimized content extraction from any kind of documents, when we decided the several characteristics. Therefore we have suggested the following methods. It should be include the following:

1) To preprocessing for documents on distance learning system database.
2) To parse word structure, separate word affix and identify word stem using inflecting word method in NLP at Cyrillic Mongolian language.
3) To search new sentence through word stem without word affix from words or the first words with affix including question sentence, right answer, key words at e-test.
4) To extract key words from Cyrillic Mongolian language documents.
5) To search words through vector space search method based on statistic.
6) To select words through TF-IDF method from query result.
7) To calculate similar quality using Cosine method.
8) To test and process search algorithm document optimized content extraction from Cyrillic Mongolian language.

2.1. Text Segmentation

Text segmentation is the process of dividing written text into meaningful units, such as words, sentences, or topics [4]. Word segmentation is the problem of dividing a string of written language into its component words [5].

Sentence segmentation is the problem of dividing a string of written language into its component sentences [6]. However even in Cyrillic Mongolian language, this problem is not trivial due to the use of the full stop character for abbreviations, which may or may not also terminate a sentence. For example: “H” is not its own sentence in “Topic written by N. Bat-Erdene”. When processing plain text, tables of abbreviations that contain periods can help prevent incorrect assignment of sentence boundaries. As with word segmentation, not all written languages contain punctuation characters which are useful for approximating sentence boundaries.

2.2. To Separate Section of the Text

Researchers who are working with text information are required to break line of the text for comparing the quality and optimization [7]. For our research work, it is required to separate one or several lines from learning material based on search result.

In other words, it is need to break line by one or several sentence for detailed search results. Also need to compare between content and separate section. For example: It is required to calculate optimizations how to meet for search purpose which are sentence or section found from text material.

Therefore, we need to distinguish word, sentence and part of the painting lines among text using the painting algorithm. Words appropriate to search among the text are painted and the after painted sentence or part should be break. This type of painting algorithm is an effective way to avoid next calculation.

3. Research on Information Retrieval Mechanism and Its Retrieval Methods

Information search system is able to display particular word roots and its inflecting form from database when defined the word roots.

In linguistics, a root word holds the most basic meanings of any word and uninflected form. The first study of the roots recognition algorithms made in 1968. For example: word stem should be input in the roots recognition algorithms such as “Book, ном” is found “Books, номнууд”, “person, хүн” is found “хүний”, “хүнээс”, “хүнтэй” and “fish, загас” is found “fishing, загасчлах”, “fished, загасчилсан”, and “Fisher, загасчин”.

Advanced search is a search of information which is expanded the request by user. The following techniques widely used:

1) To search by synonymic.

2) To search word roots and its inflected word within search request.
3) To correct all mistakes and to search correct request by automatically or to suggest this search.
4) To search for each of the elements in the original request.

3.1. Information Retrieval Mechanism

Information retrieval process will be starts request entered at system by user. However, user request is not determined only object. Information retrieval systems are usually ranked multiple objects according to requests level. The information retrieval system can measure the following parameters:

1) precision: the fraction of relevant objects that are retrieved.
2) recall: the fraction of relevant objects that are retrieved.
3) error: the fraction of relevant objects that are not retrieved.
4) Harmonic mean.

The general information search models are classified as following methods [8]:

1) IR model based on set theory (Set Theoretic models). It is included Boolean model, model based on Mohy set and extended Boolean model.
2) IR model based on algebraic theory (Algebraic models). It is included vector space model, indexing model and neural network model.
3) IR model based on probabilistic statistics (Probabilistic models). It is included regression model, probabilistic model, IR model for language model and class network model.

In addition, it is also include list for machine learning based on statistics. Boolean methods are used in all branches of science. In other words, search methods based on Boolean models and it has been developed and expanded with other technologies.

3.2. Vector Space Model

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. In this model; texts are vector, which consist of t space terms. It is usually calculate constant weigh of specific single word [9]. Therefore a text document should be consisting of combination of specific t space terms and it is also express main idea and content of text.

The similarity between documents can be found by computing the Cosine similarity between their vector representations [10]. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. If angle between vectors are less, similarity between documents should more. Doc-Term Matrix of text (Equation (1)):

$$A_{m*n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{nn} \end{pmatrix}. \tag{1}$$

Matrix A_{m*n} should be consisting of n number of text and m number of index. Matrix column is a vector for text and matrix row is a vector for single word.

Two vectors can be determined between vectors position at space. There are various functions for similarity calculation. Angle functions between two vectors are used commonly. The similarity of desired text and revealed text is calculated the following Equation (2).

$$\text{cosine}\left(d_j, q\right) = \frac{d_j * q}{\left|d_j\right| * \left|q\right|} = \frac{\sum_{i=1}^{t} w_{ij} * w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} * \sqrt{\sum_{i=1}^{t} w_{iq}^2}}. \tag{2}$$

In the Cosine similarity calculation using following formula, text and every search at t dimensional space are used as a point of numerical value. Thus, each character can be a measure of t dimensional space. The first point at space with specific relatively and point of numerical value are created vectors. In that case, Cosine similarity

is angle which is created by vector at calculation space. If angles between vectors are less, content similarities should be more. It is possible to revealed same two texts. In that case, the value of Cosine similarity between two vectors would be 1.

3.3. TF-IDF Algorithm for Specific Weigh Calculation

Identification of main content using word weight is implemented by computer. Specific *TF-IDF* is made combination of frequency for frequency and reverse relevant [9]. *TF* (Term Frequency) [11] word frequency is a number of indication for a word at text file. When word indicated from same text file, it is more indicated difference between other files. *IDF* (Inverse Document Frequency) reverse document frequency is indicated that some words are very few at text file. If specific words are focused on more than others, it is represented relatively high frequency word from small size documents. If we consider t_i as a middle word in specific text, their quality can express following Equation (3).

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

where, n_{ij} is a number of indication words at d_j text. But divided is a sum of all the detected word in the d_j text.

Reverse text frequency is a number of commonly used words. *IDF* of a given word is a number of general subtract the number of text including word. It can be found the following Equation (4):

$$idf_i = \log \frac{|D|}{\left|\left\{ j : t_i \in d_j \right\}\right|}. \tag{4}$$

Here: $|D|$ is a number of every text at language material database, $\left|\left\{ j : t \in d_j \right\}\right|$ is number of texts including t_i (where, number of text $n_{ij} \neq 0$) if this word wouldn't find material database, number of deletion from text is equal to 0.

4. Search Algorithm for Optimized Content Extraction of Learning Text Documents

When the optimized content extraction is search from learning text documents, certain answer found from search system through user's desired requirements and it suggested back to user. This action should be performed by computer dynamic, to break suitable lines through compare similarity of between query word and documents and to suggest brief parts.

The following algorithms developed and tested based on several methods of search engine studies. This model consists of 4 main components, such as to process wrong answer form based on result of e-test, to recognize the word stem based on parsing word structure, keyword extraction, and the search process.

1) The processing wrong answer form is to prepare search key data for each text set which is including answer key word, correct answer, text of question sentence.

2) In the first step of recognizing the word stem based on parsing word structure, correct answer of wrong answered question and question sentence are combined each other. The parsing word structure at text set, recognizing word stem, forming new set with word stem are next step. The following 2 kinds of set will be formed in the result of those steps. a) The first basic set consists of exam question and correct answer. For example: Байгаль орчинд үзүүлэх эерэг болон сөрөг нөлөөллийг тодорхойлох үнэлгээ аль нь вэ?: Хүрээлэн буй орчны үнэлгээ. Which assessment can assist of positive or negative impact of environment?: Environmental assessment b) The new set with word stem also consist of exam question and correct answer. For example: Байгаль орчин үзүүлэх эерэг сөрөг нөлөөлөл тодорхойлох үнэлгээ?: Хүрээлэн орчны үнэлгээ. Assessment assist positive, negative impact environment?: Environmental assessment. Possible form of search can be shift next search action.

3) In the second step, processed two kinds of set can used and it can extract five key words from each set using Stop Word method and statistical method (In Stop Word method, word and symbol without meaning should be deleted from set. For example: Байгаль орчин үзүүлэх эерэг болон сөрөг нөлөөлөл тодорхойлох үнэлгээ аль нь вэ?: болон, аль, нь, вэ,? etc. Which assessment can assist of positive or negative impact of environment?: or, which, can,? etc. Nominated key words are selected by statistical method). In this step, search should be use each keywords and it should be compared three form of search data result which is processed part 1.

4) The search can be based on 1 and 3 form from text set by VSM method at search action part. The search should be covered two kinds of forms at database documents. Two kinds of forms are processing basic documents at database and forming new document consisting of word stem. For example: a) basic document, b) new document consisting of word stem.

The search action is calculating search result and finding suitable content. For example: The action should be made the following principle such as making search, selecting content, breaking line of selected text border, indexing selected parts, ranking content by the highest rank, suggesting through search sequence, and showing to users. The suggested search system architecture diagram is shown in **Figure 1**.

5. Experimental Data Selection and Analysis

For the experiment in this chapter, making search and content extraction at are similar to data monitoring method. In e-learning system, the training covered the overall average 4 - 8 basic course and professional courses. There are select 3 basic courses and 2 professional courses selected from learning text materials by search method and an optimization can calculate by mathematical method.

5.1. Experimental Data Selection

The text materials for e-subject such as “Tourism regional planning” (T1), “Marketing” (T2), “Management” (T3), “Psychology” (T4) and “Education science” (T5) are selected and tested which is taught bachelor course at the National University of Mongolia, Mongolian University of Science and Technology, Mongolian University of Life Science and University of the Humanities.

Those learning materials are dominated by theory form, contents are similar mutually. It is main reason to select those subjects.

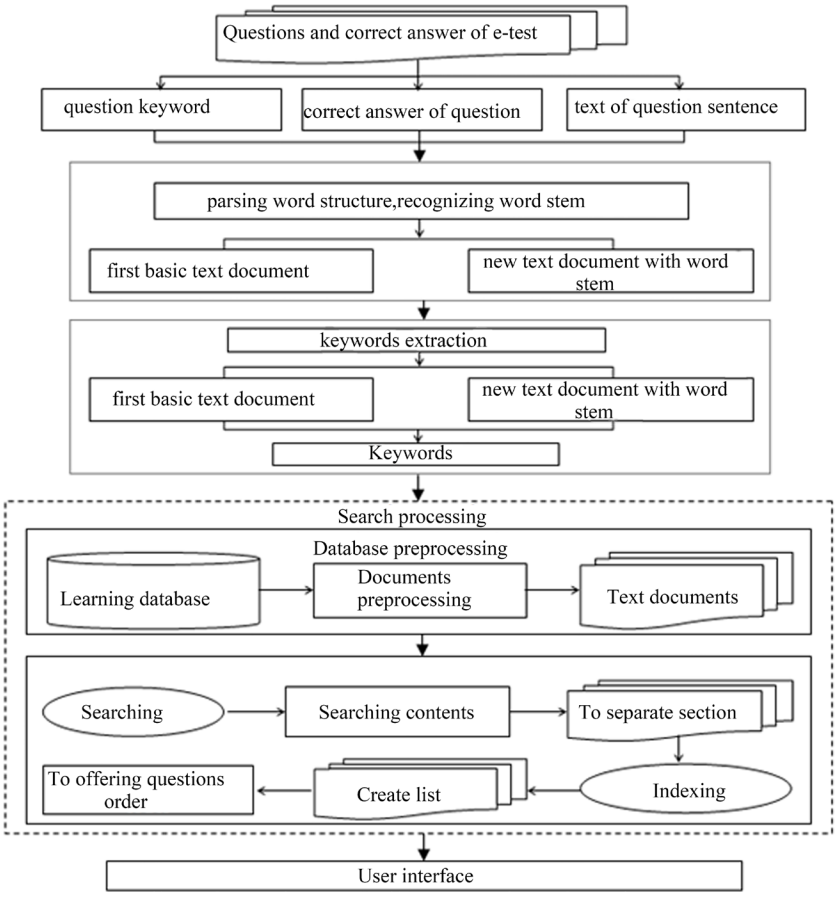

Figure 1. Search system architecture diagram.

5.2. Experimental Results and Analysis

It was preprocessed experimental text according to requirements of text document statistical analysis and calculated basic statistical information. It is including: numbers of symbol, word and sentence for each experimental text material are shown in **Table 3**. In those numerical indicators, average length of word and sentence each text material are calculated by statistical method.

300 question sets at e-text database are made experiment for each text material using VSM search method. It is selected a question form from experiment and introduced experimental results and analysis at learning material by detailed. For example: Question 1: Which assessment can assist of positive or negative impact of environment?: For this question set, experimental search form would be the following form.

Question key word: Байгаль, нөлөөлөл, үнэлгээ.

Correct answer: c. Хүрээлэн буй орчны үнэлгээ.

Question (word basic form and basic document): Байгаль орчинд үзүүлэх эерэг болон сөрөг нөлөөллийг тодорхойлох үнэлгээ аль нь вэ?

Word parsing structure (by word stem and new document): Байгаль орчин үзүүлэх эерэг болон сөрөг нөлөө тодорхой үнэлгээ аль нь вэ?

Experimental results are shown in **Tables 3-13**.

The average length of word and sentence at experimental text material are calculated by statistic and results are shown in **Table 4**. It is defined WFR and ALM. ALW is comparison of sentence length at text material and number of all words.

It is possible to make conclusion based on results, T2 and T3 can be the most understandable learning material. The sentence lengths are same and sentence words are few are indicated that content are understandable.

Search results using key words by hand and mechanically are shown in **Table 5**. If query word and reveled text are same, it called Exact Match during the making basic database search. In the experimental result, T1 text material is meet the search condition through key word statistic and probability distribution information. It have too much work to do if training teacher set off limitation border of key word as an experimental basic search data. But it is one way of experiment. It is possible to show as a list which is ranked by indicated probabilities. In other word, results are show the following principle “if it is not match, selects next”. According to this principle, at first it will be suggest T1 search result and if it is not match it will be suggest next as the second T5, the third T4, the fourth T3 and the fifth T2 etc.

Table 3. Statistic of experimental text materials.

Text documents	Count of characters	Count of words	Count of sentence
T1	66,510	10,151	554
T2	127,107	20,579	1918
T3	107,291	16,581	930
T4	130,302	20,634	1329
T5	94,645	14,416	700

Table 4. Average length of word and sentence (WFR and ALM).

Text documents	Average length of sentence	Average length of word
T1	18.32	6.55
T2	10.73	6.17
T3	16.09	6.47
T4	15.52	6.31
T5	20.59	6.56

Table 5. Search results using keywords by hand.

Search words (keywords by hand)	Frequency of words									
	T1		T2		T3		T4		T5	
	fre	Pd	fre	Pd	fre	Pd	fre	Pd	fre	Pd
Keyword 1	21	0.00207	2	0.00010	0	0.00000	6	0.00029	16	0.00111
Keyword 2	1	0.00010	1	0.00005	3	0.00018	0	0.00000	0	0.00000
Keyword 3	3	0.00030	0	0.00000	2	0.00012	2	0.00010	4	0.00028
Totals	25	0.00246	3	0.00015	5	0.00030	8	0.00039	20	0.00139

a. fre—frequency of word; b. Pd—probability distribution.

Table 6. The statistical result to search by correct answer word.

Search word (correct answer word)	Frequency of words									
	T1		T2		T3		T4		T5	
	fre	Pd	fre	Pd	fre	Pd	fre	Pd	fre	Pd
Хүрээлэн	10	0.00099	0	0	4	0.00024	12	0.00058	4	0.00028
Буй	17	0.00167	13	0.00063	26	0.00157	33	0.00160	21	0.00146
Орчны	19	0.00187	4	0.00019	18	0.00109	17	0.00082	8	0.00055
үнэлгээ	21	0.00207	0	0	15	0.00090	6	0.00029	4	0.00028
Total	67	0.00660	17	0.00083	63	0.00380	68	0.00330	37	0.00257

Experimental search results using correct answer words are shown in **Table 6**. Search word indicated frequency for each text material and distribution probabilities are calculated. When the correct answer word increased by one word, indicated word frequency and distribution probability are completely changed and compared with results of **Table 5**. For example: It will be made following sequence such as T4, T1, T3, T5 and T2 based on word frequency at search result. It will be made following sequence such as T1, T3, T4, T5 and T2 based on word distribution probability at search result. However, the word frequency ranked the highest in the T4, it will be ranked the third by distribution probability. Also the word frequency ranked the second in the T1, it will be ranked the highest by distribution probability.

Experimental search results using sentence main word are shown in **Table 7**. It performed to search by question sentence main words with given form. Inner composition and text similarities at each text document are calculated by Cosine calculation use Equation (3). The following results are shown.

Text set of T1 text material are formed angle equal to 0.73 and it meet to purpose of experiment for theory. Text set of T4 text material are formed angle equal to 0.71. However, their numerical difference was 0.71 and 0.73, but their difference was 0.045 percent.

If made little mistake, it easy to face with difficulties during learning machine. Experiment is made to compare between computer search result and hand method. T1 learning materials meet to search content at this experiment.

But it was become doubtful, when query words of experimental texts are indicated too much. The following experiments are made to solve these issues. It is including the following works. The parsing word structure removes word affix, composing new sets by word stem, and then extracts optimized content. The searching process will not consider word position. The results are shown in **Table 9**.

Experimental search results using sentence word stem are shown in **Table 9**.

Inner composition and text similarities at each text document are calculated by Cosine calculation use Equation (3).

Text set of T1 text material are formed angle equal to 0.84 and it meet to purpose of experiment for theory. For other text materials, all numerical value was increased but it was not meet to search purpose.

Table 7. The statistical result to search by sentence main word (Doc-Term Matrix of basic document).

Search word (answer sentence main word)	Frequency of words					q
	T1	T2	T3	T4	T5	
Байгаль	21	0	0	9	15	3
орчинд	6	5	34	13	23	5
үзүүлэх	4	2	7	5	10	5
эерэг	7	0	1	3	0	3
болон	31	28	39	33	46	5
сөрөг	4	0	4	5	0	3
нөлөөллийг	2	2	5	1	1	5
тодорхойлох	8	2	3	5	6	5
үнэлгээ	7	3	7	5	9	5
аль нь вэ	0	0	0	0	0	0
Total words	90	42	100	79	110	

Table 8. The Cosine calculation results.

Text material	Inner composition	Cosine
T1	386	0.73
T2	210	0.55
T3	490	0.69
T4	361	0.71
T5	520	0.70

Table 9. The statistical result to search by sentence word stem (Doc-Term Matrix of new document).

Search word (sentence word stem)	Frequency of words					q
	T1	T2	T3	T4	T5	
Байгаль	21	0	0	9	15	3
орчин	19	13	53	14	29	4
үзүүлэх	23	2	25	23	19	4
эерэг	7	0	1	7	0	3
болон	62	46	52	51	53	4
сөрөг	4	0	4	6	0	3
нөлөөлөл	13	2	34	62	2	4
тодорхойлох	63	5	3	41	45	4
үнэлгээ	22	4	15	6	21	4
аль нь вэ	0	0	0	0	0	0
Нийтүг	234	72	187	219	184	

Table 10. The cosine calculation results.

Text documents	Inner composition	Cosine
T1	1106	0.84
T2	360	0.56
T3	925	0.80
T4	1051	0.83
T5	890	0.81

Table 11. Automatically selected keywords.

Search word (word stem)	Frequency of words				
	frequency of single word	TF	IDF	Weight of word	Rank of weight
Байгаль	21	0.03	14.86	0.457	**IV**
орчин	19	0.01	14.10	0.098	**V**
эерэг	7	0.03	43.38	1.298	**II**
сөрөг	4	0.04	95.76	3.614	**I**
үнэлгээ	22	0.03	21.64	0.646	**III**

Table 12. The statistical result to search by automatically extracted key word.

Search word (word stem)	Frequency of words					q
	T1	T2	T3	T4	T5	
Байгаль	21	0	0	9	15	3
Орчин	19	5	53	14	29	4
эерэг	7	0	1	7	0	3
сөрөг	4	0	4	5	0	3
үнэлгээ	22	3	15	6	21	4
Totals	73	8	73	37	65	

Table 13. The cosine calculation results.

Text documents	Inner composition	Cosine
T1	301	0.93
T2	40	0.78
T3	355	0.73
T4	151	0.92
T5	295	0.87

Five key words are extracted from Cyrillic Mongolian language documents through keyword extraction algorithm which is processed in using natural language processing technology. The nominated key words are shown in **Table 11**.

The search result relevance is calculated based on query words as a $t_1, t_2, \cdots, t_n$ and these results are shown in **Table 11**.

In this text material, the weight of following words such as positive "сөрөг", negative "эерэг", assessment "үнэлгээ", natural "байгаль" and environment "орчин" are indicated the highest. Therefore it is possible to search the highest five values. The experimental results using extracted key word are shown in **Table 12**.

Inner composition and text similarities at each text document are calculated by Cosine calculation use Equation (3). These results are shown in **Table 13**.

Text set of T1 text material are formed and increased angle equal to 0.93 and it meet to purpose of experiment for theory.

6. Conclusions

In this chapter, we introduced optimized content extraction algorithm in text documents from distance learning system database and its result of application.

Optimized content is extracted by NLP, TF-IDF and cosine methods. The following methods have been done to achieve the research goal. We also approved to search by word stem and word position without consideration.

The text material pre-processing is important to determine the results of statistic accurately. These text materials had average word length of 6.41 and average sentence length of 16.25. This study performance was similar to other studies on the Cyrillic Mongolian language.

In the experiment result with making search using word stem of question sentence, word frequency and distribution probability are become more sophisticated and cosine is increased until 0.84 percent.

Those indicators meet the purpose for search. Five key words are extracted from Cyrillic Mongolian language documents through key word extraction algorithm using natural language processing technology. The search was using automatically extracted key words. Text material similarities are 0.93.

It was proved through experiment that our developed optimized content extraction search algorithm from Cyrillic Mongolian language was very useful.

References

[1] Altangerel, A. and Tsend, G. (2008) N-Gram Analysis of a Mongolian Text. *IFOST* 2008 *Proceedings*, Tomsk, 23-29 June 2008, 258-259.

[2] Byambadorj, D. (2006) Mongolian Language Forms Studies. Mongolia University of Education Printing, Ulaanbaatar City.

[3] Luvsansharav, Sh. (1999) Modern Mongolian Language Structure "Mongolian Language Words and Terms". Sanaa and Tungaa Printing, Ulaanbaatar City.

[4] Romero, A. and Nino, F. (2007) Keyword Extraction Using an Artificial Immune System. *Proceedings of the* 9*th Annual Conference on Genetic and Evolutionary Computation*, London, England, 7-11 July 2007, 181-182. http://dx.doi.org/10.1145/1276958.1276995

[5] Yang, W. (2002) Chinese Keyword Extraction Based on Max-Duplicated Strings of the Documents. *Proceedings of the* 25*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, 8-11 August 2002, 439-440. http://dx.doi.org/10.1145/564376.564483

[6] Grefenstette. G. and Tapanainen, P. (1994) What Is a Word, What Is a Sentence? Problems of Tokenization. *The* 3*rd International Conference on Computational Lexicography*, Budapest, 7-10 July 1994, 79-87.

[7] Bat-Erdene, N. (2014) Trends in E-Education. 2014 *International Conference on Embedded Systems and Applications*, Ulaanbaatar, 17-19 August 2014, 331-335.

[8] Al-Kabi, M.N., *et al.* (2013) Keyword Extraction Based on Word Co-Occurrence Statistical Information for Arabic Text. *ABHATH AL-YARMOUK*: "*Basic Sciences and Engineering*", **22**, 75-95.

[9] Meier, U., Ciresan, D.C., Gambardella, L.M. and Schmidhuber, J. (2011) Better Digit Recognition with a Committee of Simple Neural Nets. 2011 *International Conference on Document Analysis and Recognition*, Beijing, 18-21 September 2011, 1250-1254. http://dx.doi.org/10.1109/icdar.2011.252

[10] Sparck Jones, K. (1972) A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, **28**, No. 1.

[11] Berger, A., *et al.* (2000) Bridging the Lexical Chasm: Statistical Approaches to Answer Finding, *Proceedings of the International Research and Development in Information Retrieval*, New York, 24-28 July 2000, 192-199. http://dx.doi.org/10.1145/345508.345576

Scientific Research Publishing

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: http://papersubmission.scirp.org/