AN EFFICIENT APPROACH FOR KEYWORD SELECTION; IMPROVING ACCESSIBILITY OF WEB CONTENTS BY GENERAL SEARCH ENGINES

H. H. Kian[1] and M. Zahedi[2]

School of Information Technology and Computer Engineering, Shahrood University of Technology, Shahrood, Iran

[1]Hodhodkian@shahroodut.ac.ir, [2]zahedi@shahroodut.ac.ir

ABSTRACT

General search engines often provide low precise results even for detailed queries. So there is a vital need to elicit useful information like keywords for search engines to provide acceptable results for user's search queries. Although many methods have been proposed to show how to extract keywords automatically, all attempt to get a better recall, precision and other criteria which describe how the method has done its job as an author. This paper presents a new automatic keyword extraction method which improves accessibility of web content by search engines. The proposed method defines some coefficients determining features efficiency and tries to optimize them by using a genetic algorithm. Furthermore, it evaluates candidate keywords by a function that utilizes the result of search engines. When comparing to the other methods, experiments demonstrate that by using the proposed method, a higher score is achieved from search engines without losing noticeable recall or precision.

KEYWORDS

Automatic keyword extraction, search engine, genetic algorithm, web contents accessibility, Farsi stopwords

1. INTRODUCTION

Nowadays, unlimited growth of data creation causes the production of great datasets like the web which contains a vast amount of information. Consequently, many methods are proposed for analysis of such a huge amount of information specially text information. Automatic keyword extraction is one of the most important processes for text data analysis. It is the task of extracting a small set of words, or key phrases from a document which is able to describe the main meaning of the document [1]. Keyword extraction is the basis of many text mining solutions and applications such as automatic indexing, automatic summarization, automatic classification, automatic clustering, automatic filtering, topic detection and tracking, information visualization, etc [2].

Keyword extraction methods can be divided into four categories: The most popular methods belong to the category of statistical approaches which need no explicit training data and use only statistical information of the words to identify the keywords in the document. For instance, R. Mihalcea and P. Tarau introduce a graph-based ranking model which proposes an unsupervised

DOI : 10.5121/ijwest.2011.2406

method for keyword and sentence extraction [3], Y. Matsuo and M. Ishizuka propose a method which extracts keywords based on word co-occurrences [4] and L.F. Chien uses PAT-tree [5].

Some other researchers employ linguistics approaches where the keywords are selected on their linguistic features. For example, G. Ercan and I. Cicekli use lexical analysis [6] and A. Hulth works with a syntactic analysis [1].

On the other hand, machine learning approaches use extracted keywords by authors to create a model and use the created model to extract keywords from new documents. Kea method is a keyword extraction algorithm based on Naive Bayes classifier [7]. It calculates three features for each word after pre-processing: the normalized term frequency (TF), the inverted document frequency (IDF) and the distance of first occurrence of the word from the beginning of the document. The candidate words from each document are combined and used to construct a Naive Bayes classifier. The final classifier decides that a given phrase is a keyword or not. Also [8] generates a decision tree to evaluate a given phrase as a keyword. It uses nine features such as the number of words in a phrase, the location of the first occurrence of a phrase in a document, the frequency of occurrence of a phrase within a document, the relative length of the whole phrase, being the whole phrase a proper noun, ending the whole phrase in a final adjective, and occurring a common verb in the whole phrase.

Also in [9] the keyword extraction method uses four features including the normalized term frequency (TF), the inverted document frequency (IDF), whether or not a phrase appears in the title or heading (THS), and the normalized paragraph distribution frequency (PDF). Using the introduced features, it uses a multilayer feed-forward neural network as classifier. Also, [10] introduces another approach which use support vector machine (SVM) and GenEx is another method introduced in [8] that uses genetic algorithm as a machine learning approaches.

Most of other approaches for automatic keyword extraction combine previous methods or use some heuristic knowledge such as the position, length, layout feature of the words, HTML tags around the words and etc [11].

However, existing keyword extraction methods, concern only about getting better recall, precision and other criteria which describe the performance of method in obtaining authors keyword list, but considering the undeniable role of search engines in today's information world and the matter that every search engine has its own special algorithm for web page indexing, a method is needed which get a good score on recall and precision criteria and also acquire a better rank and accessibility to the document by general search engines.

In this paper, a new method is introduced that utilizes the results of popular web search engines to optimize the keyword extractor function. This optimization leads to selecting keywords which are more important to search engines ranking algorithms and by using famous keyword extraction features; an acceptable recall/precision is achievable at the same time. In the research community, no previous study has investigated similar method, to the best of our knowledge.

The proposed method uses two feature sets (A and B) and two keyword evaluation functions. The first evaluation function, gives a score to each tokens by using feature set A. Since some of features in feature set A depend on multi document datasets, a multi level document dataset containing 4500 web pages is created, which 600 of them are keyword-assigned by experts. The second evaluation function named as SERE uses the features in feature set B on web search engines feedback results, based on single document keyword extraction methods.

To elicit significant keywords for a web page following phases are performed:

1. **Pre-processing:** in this phase, stopwords are removed at the first step, and then the document is tokenized.

2. **Keyword extraction and evaluation**: in this step, extracted terms are evaluated and sorted by score function which uses statistical features of terms, and then top terms of the extracted keyword list are used to construct a search query. The search query is submitted to popular search engines and the results obtained from the search engine are analyzed to rank query terms by a fitness function based on outputs of SERE function. SERE is also used as the base of a genetic algorithm fitness function to optimize the score function.

The paper is organized as follows: section 2 describes the pre-processing phase according to Farsi language structures. Section 3 presents keyword extraction and evaluation phase. Section 4 gives the used dataset information and experimental results. Section 5 includes the summary and conclusion.

2. PRE-PROCESSING

Keyword selection methods use a pre-processing phase which depends on the language of dataset and affects directly final results. The pre-processing phase contains two steps: unification and stopwords removing.

2.1. Unification

Some of Farsi words can be written in different ways like "کتابها" and "کتاب ها" which are both the same (means "books") and there is a need to unify shape of these words. We use 13 simple rules for this step. For example the "تر" postfix which is used to create the comparatives in Farsi can be part of a word or can be departed from it, after the unification step it is always appeared in joined form. Here is an example for this rule: "خوب تر" is converted to "خوبتر" (means better in English).

2.2. Removing Stopwords

In information retrieval researches, the words that do not carry any information about the contents of the document are known as stopwords. These words usually play grammatical roles in documents. Examples of stopwords in Farsi are "اگر" (if), "اما" (but) and etc. The removal of stopwords can increase the efficiency of the indexing process as stopwords can represent 30 to 50% of the tokens in large text collaboration [12].

There are few sources for Farsi stopwords [13, 14]. Therefore to create a stopwords list, the stopwords listed in [13] are used as an initial list of stopwords and the other stopwords occurring in our dataset are added to the list. At the end of this step, all the words occurring in the documents that belong to the stopwords list are removed and the remaining part is passed to the keyword extraction phase.

3. KEYWORD EXTRACTION AND EVALUATION

Figure 1 shows the process of the proposed method for keyword extraction. The method contains two sections: train section and test section.

3.1. Train Section

In this section two functions are used: score function and search engines results evaluation function (SERE).

This research uses unigram, bigram and trigram models. At the first step of this section, the features defined as feature set A are calculated for every extracted n-gram as parameters of score function. Feature set A is a set of effective keyword selection features defined and used in previous related works. After scoring, tokens are sorted by their score. To select a token as a candidate keyword, a threshold value is defined and tokens with a score more than it are assigned as candidate keywords. The threshold value is determined so that leads to appearance of 60% of author's selected keywords of training dataset in the candidate keywords list.

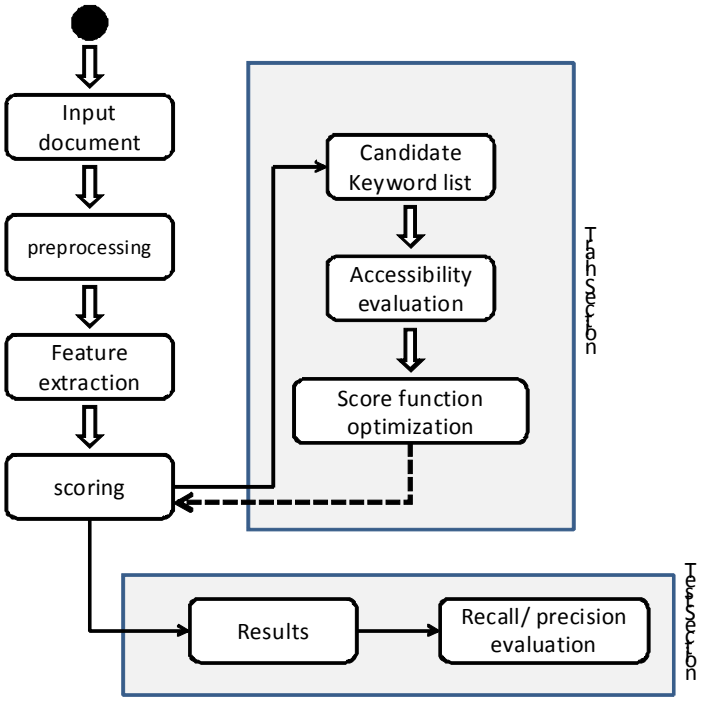

Figure 1. The keyword extraction method

After candidate list creation, the list is delivered to keyword evaluation module. This module uses popular search engines like Google, MSN, yahoo and etc by sending candidate keyword list to them and receiving the first result page as a feedback. The result page is analyzed by SERE function which uses the features from feature set B as the parameters and assigns a score to every candidate keyword list.

The method uses a genetic algorithm for module modification and result improvement. The genetic algorithm fitness function is based on SERE and the score function coefficients are the genes in genetic algorithm. The training phase process repeats until fitness function get to its maximum value.

3.2. Scoring Function

The scoring function is presented as equation 1. This function is used in both training section and test section where genetic algorithm improves α, β and γ coefficients in the training section.

$$\text{Scoring function } (w) = NPW^* (\alpha^*NPLen + \beta^*NSL + \gamma^*NTFIDF)/3 \qquad (1)$$

Initially α, β and γ get random values between 0.4 and 0.6. During training section each of them changes to maximize fitness function. This leads to a better score for keyphrases which are more important for web search engines and because of using the threshold value, there is not a noticeable recall decrement. It is important to notice that the number of parameters depends to the research condition and can be changed. Other scoring function parameters are defined as below:

Normalized Phrase Words (NPW)

NPW is the number of words in the selected phrases, normalized to the maximum number of words. The values of this feature can be 1, 1/2, or 1/3. The hypothesis is that key-phrases consisting of three words are better than key-phrases containing two words, and so on.

Normalized Phrase Length (NPLen)

NPLen is the length of the candidate phrase (in words), divided by the number of words in the sentence. This feature has a value of one, when the whole sentence is a key-phrase. The hypothesis is that it captures titles and subtitles of the document, which seems to contain key-phrases.

Normalized Sentence Location (NSL)

NSL measures location of the sentence containing the candidate phrase within the document. Equation 2 is used as a simple distribution function where "L" is location of the sentence within a document divided by total number of sentences in that document (m). The maximum value of NSL is 1 for first (L=0), and last sentences (L = m) in the document.

$$NSL= (2(L/m)-1) \wedge 2 \quad (2)$$

Normalized TFIDF

TFIDF weighting method is one of the major developments of term weighting in information retrieval [2]. Most modern term weighting algorithms are a new version of the family of TFIDF weighting algorithm. TFIDF is normalized and calculated with equation 3.

$$TFIDF\ (word) = (WF/MF)*\log(N+1/n+1) \quad (3)$$

- WF shows the frequency of word in this document
- MF presents the frequency of most repeated word in this document
- N is number of all documents
- n is number of documents containing the word

Since the value of the coefficient is between zero and one, it is divided by the number of coefficients in order to normalize the returned value.

3.3. SERE Function

This function is used in the training phase as the basis of the genetic algorithm fitness function. In a technical view, SERE is a single document keyword selection function (equation 4).

Since search engines results are web pages; it is possible to analyze key phrases importance with HTML tags. Parameters in parentheses are calculated using HTML tags based on the features that search engines report as important ones.

$$SERE = PRF*(TET+NLCT+NHCT+NIRT) \quad (4)$$

SERE parameters are defined as below:

The Phrase Relative Frequency (PRF)

RPF represents the frequency of the candidate phrase normalized by the most frequent phrase in the given document. PRF has a maximum value of one; when the candidate keyphrase is the most repeated one in a given document.

Term Existence in Title (TET)

TET value is zero if the term does not exist in the title tag, and else it is 1 divided by number of title words.

Number of Links Containing the Term (NLCT)

NLCT represents the frequency of links containing the term normalized by number of all links.

Number of Headings Containing the Term (NHCT)

NHCT represents the frequency of headings containing the term normalized by number of all headings.

Number of Images Related with the Term (NIRT)

There are two important attributes for <img> tag which define an image for browsers in HTML. The first attribute is "src" which contains image address and the second one is "alt" which describes the image by text. NIRT is the sum of frequency of src and alt tags containing the term normalized by dividing it by number of all <img> tags. Table 1 presents the HTML tags used for four recent parameters.

Table 1. HTML based features and their HTML tags.

Name	HTML tag
TET	<title>
NLCT	<a>
NHCT	<h1> , other headings such as <h2>,… can be used
NIRT	<img>

3.4. The Implemented Genetic Algorithm

A genetic algorithm may be viewed as a method for optimizing a string of bits, using techniques based on natural selection and evaluation. Set of bit strings is called a population of individuals. The initial population is usually randomly generated. New individuals (new bit strings) are created by randomly changing existing individuals (this operation is called mutation) and by combining substrings from parents to make new children (this operation is called crossover). Each individual is assigned a score (called its fitness) based on some measure of the quality of the bit string, with respect to a given task. Fitter individuals get to have more children than less fit individuals. While the genetic algorithm is running, new individuals tend to be fit increasingly, up to some asymptote. The fitness function shown in equation 5 is used for the improvement step. This function has no limitation on search engines number. The coefficients named as Se1.weight, Se2.weight … SeN.weight are defined to customize importance of search engines due to the application.

$$\text{Fitness function} = Se.1weight * SERE(\text{word in } Se1.result) + Se.2weight * SERE(\text{word in } Se2.result) + \cdots + SeN.weight * SERE(\text{word in } SeN.result) \quad (5)$$

Also, the number of resulted documents used for each search engine, can be selected arbitrary, e.g. three first results of the first search engine, two first results of the second search engine and so on.

Every coefficient is defined by one byte and this string of bites (totally 24 bits) is used as population of individuals for the implemented genetic algorithm. The genetic algorithm uses a simple crossover function which combines 4 bits of every coefficient to create a new population member.

4. EXPERIMENTAL RESULTS

To create the stopword list and to evaluate the extracted keywords, a multi level database is constructed by gathering Persian web pages from official and non-official web sites. After downloading the web sites; text content of each page is extracted and saved in the database. The mentioned database consists of the contents of 4500 Persian web pages. Then 600 web contents are selected and keyword-assigned by expert readers as a second level of dataset. The number of the annotated keywords for second level of dataset ranges from 5 to 10 and the average of annotated keywords is 7.48 per documents. Figure 2 presents the database creation process.

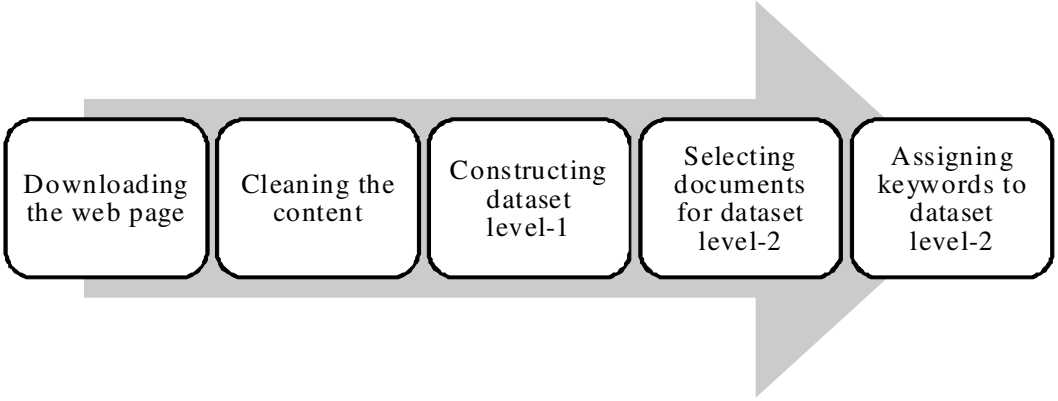

Figure 2. The database creation process

According to the human keyword selection process, there are two types of words or phrases, which are keywords and non-keywords assigned by humans. On the other hand, every automatic keyword extraction method divides words and phrases in two categories: the keywords and non-keywords extracted by system. Table 2 shows the contingence table on the results of keywords extraction method and keyword assignment manually.

Table 2. Contingence table on the results of automatic keyword extraction and keyword assignment manually

	Keywords assigned by human	**Non-keyword assigned by human**
Keywords extracted by system	A	b
Non-keywords extracted by system	C	d

Experiments on keyword extraction, generally evaluate methods efficiency by precision (P), recall (R) and F1-Measure criteria which are defined as follows:

$$P=a/(a+b) \quad (6)$$

$$R=a/(a+c) \quad (7)$$

$$F1(P,R)=2PR/(P+R) \quad (8)$$

The experimental results show effect of training set size on recall, precision, and F1-measure (Table 3). In order to perform the experiments, 100 documents are used as test set and ten pages from the first results of Google, Yahoo and MSN web search engines as fitness parameters.

Table 3. Effect of training set size on general measurement

Training set	**Average R**	**Average P**	**Average F1**
250	44.77	26.14	39.18
300	46.78	22.09	32.54
350	**49.14**	27.26	38.51
400	48.95	**32.21**	**40.82**
450	47.32	20.9	29.86
500	46.49	18.51	29.8

To avoid the effect of previous training step, the training phase is established from beginning for every training set size and also the documents are choose randomly.

Best results are obtained with 400 training set. It can be seen that after a while increasing training set, it causes lower precision. Actually system is over trained and reflects behaving as search engines algorithm. Because of recall relative threshold value, training set size has no significant effect on this criterion.

As it is shown in figure 3 and 4, γ interval is getting higher value, faster than other coefficients, during the training phase iteration. Since fitness function is an ascending function, it can be concluded that NTFIDF is the most important feature for search engines algorithms.

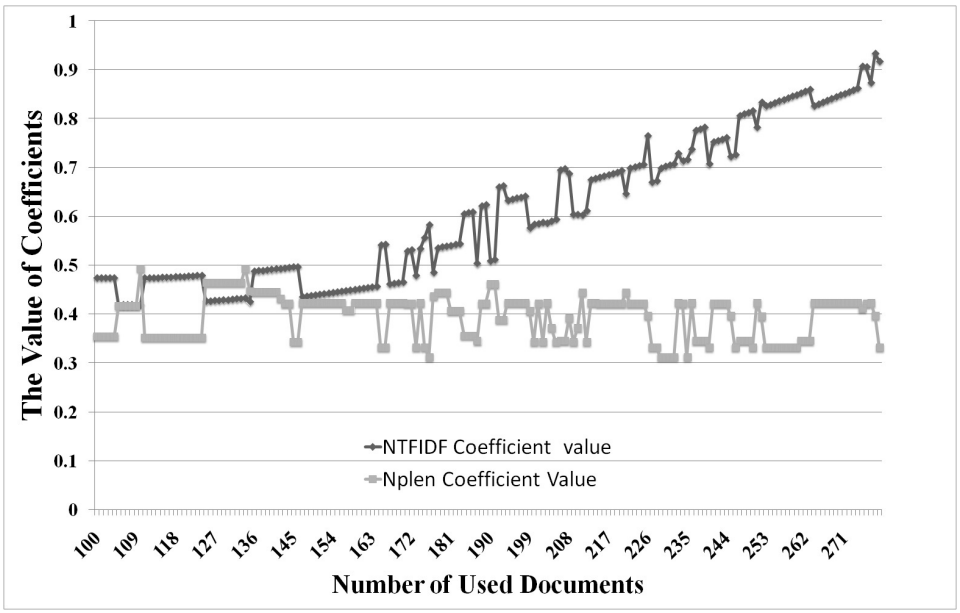

Figure 3. Coefficient of NTFIDF (γ) growth faster than NPLen coefficient (α) during training phase iteration

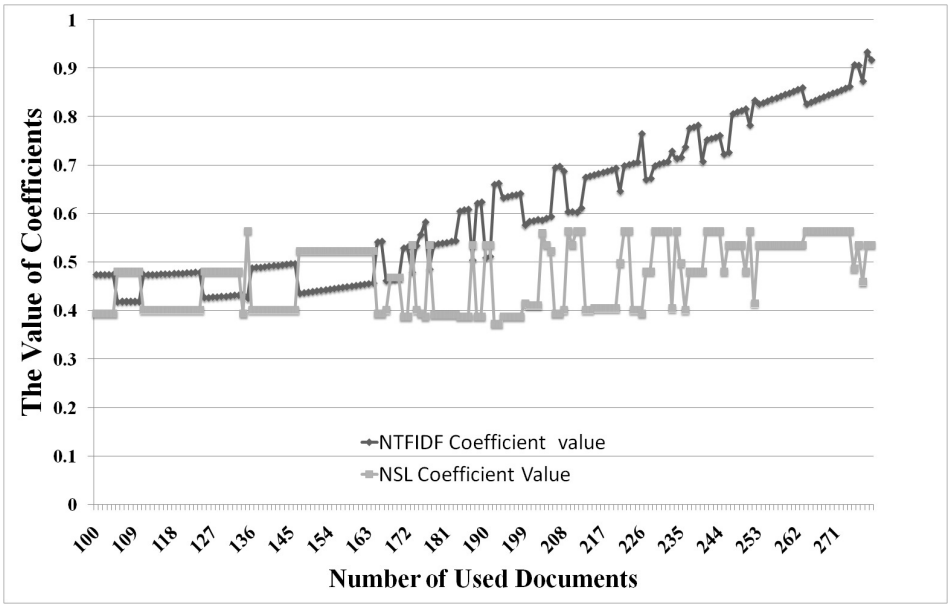

Figure 4. Coefficient of NTFIDF (γ) growth faster than NSL coefficient (β) during training phase iteration

5. CONCLUSION

In this paper, a new method is proposed for keyword extraction from Farsi websites using two score function, which one of them is used to estimate tokens importance and the other one evaluates results of first score function by search engine results and optimizes the first function too. Experimental results show that using this method has better recall and approximately the same precision as the best researches on Farsi documents until today. Also because of using search engine results for optimization of the keyword selection function, a better accessibility by

popular search engines is achievable. The method also is used to evaluate effect of keyword selection features on accessibility of documents by search engines. According to results shown in this paper, TFIDF is the most important feature in general search engines behaviour.

REFERENCES

[1] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 216-223.

[2] J. Kaur and V. Gupta, "Effective Approaches For Extraction Of Keywords," Journal of Computer Science, vol. 7, 2010, pp. 144-148.

[3] R. Mihalcea and P. Tarau, "textRank: bringing order into texts," *Proceedings of EMNLP 2004*, 2004, pp. 404-411

[4] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, 2004, p. 157–170.

[5] L.F. Chien, "PAT-tree-based keyword extraction for Chinese information retrieval," *ACM SIGIR Forum*, ACM, 1997, p. 50–58.

[6] G. Ercan and I. Cicekli, "Using lexical chains for keyword extraction," *Information Processing & Management*, vol. 43, Nov. 2007, p. 1705–1714.

[7] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin, and C.G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," *Proceedings of the fourth ACM conference on Digital libraries*, ACM, 1999, p. 254–255.

[8] P. D. Turney. Learning to Extract Keyphrases from Text. NRC Technical Report ERB-1057, National Research Council, Canada. 1999: 1-43.

[9] J. Wang, H. Peng, and J.S. Hu, "Automatic keyphrases extraction from document using Neural Network," *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, IEEE, 2005, p. 3770–3774.

[10] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," *Advances in Web-Age Information Management*, 2006, p. 85–96.

[11] C.T.Y. s l G. Salton, C. S. Yang, "A Theory of Term Importance in Automatic Text Analysis," Journal of the American society for Information Science, vol. 26, 1975, pp. 33-44.

[12] P. Schäuble, "Multimedia information retrieval : content-based information retrieval from large text and audio databases", Boston: MA:kluwer Academic publishers, 1997.

[13] M.R. Davarpanah, M. Sanji, and M. Aramideh, "Farsi lexical analysis and stop word list," *Library Hi Tech*, vol. 27, 2009, p. 435–449.

[14] K. Taghva, R. Beckley, and M. Sadeh, "A list of farsi stopwords," *Retrieved September*, vol. 7, 2003, p. 2006.