Improving Persian Text Classification Using Persian Thesaurus

Hamid Parvin, Behrouz Minaei-Bidgoli, and Atousa Dahbashi

School of Computer Engineering,
Iran University of Science and Technology (IUST), Tehran, Iran
{parvin,b_minaei,dahbashi}@iust.ac.ir

Abstract. This paper proposes an innovative approach to improve the performance of Persian text classification. The proposed method uses a thesaurus as a helpful knowledge to obtain the real frequencies of words in the corpus. Three types of relationships are considered in our thesaurus. This is the first attempt to use a Persian thesaurus in the field of Persian information retrieval. Experimental results show a significant improvement in the case of employing Persian thesaurus rather common methods.

Keywords: Persian Text, Persian Thesaurus, Semantic-Based Text Classification.

1 Introduction

In the current century Information Technology is considered as one of the most important fields (if not the most important field) among the researchers. Because the information is growing in a significant rapid way, its appropriate management and usage are inevitable. Indeed proper responding to the user queries is crucial in the Information Technology [1]. One of the most challenging problems in the field of Information Technology is how to do text retrieval and how to employ efficient algorithm on the mass of information.

In this direction, usage of keywords is very promising way for researchers to handle the job. A very important desire for researchers is to find the best representative keywords in the field of information retrieval. One of the most straightforward ways is based on frequency based keywords. Although this method is a very handful solution, the between word relationships are ignored there. It means while two synonym words are counted by the algorithm as two different words, it is better for the algorithm to count them as a single word and for its frequency to be equal to sum of frequencies of those two words.

To response queries of users relevantly, indexing is necessary. In general each context is consisted of two main parts: (a) external part and (b) body part. In library indexing based on first part is *descriptive cataloging* and based on second part is *subject cataloging*. Indexing needs the cognition of context. If indexing is done by computer, this will be automated indexing [2].

In text retrieval systems, indexing can be produced completely automatically. Research on creating or improving indexing methods and the automatic search for

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 391–398, 2011.

information in texts for different languages has always been hot. The most sensitive and difficult step in the process during indexing should be automatically selection of the words that are used for index construction. In practice, indexing based on all words contained in the context has very high overhead. It is worthy to mention that indexing based on all the words is unnecessary.

In information processing, many systems were established. These systems are categorized in five main groups, which include: (I) Management Information System, (II) Data Base Management System, (III) Decision Support System, (VI) Question Answering System and (V) Information Retrieval System.

Text retrieval systems belong to information retrieval systems. Since a lot of similarities between information retrieval systems and database management systems, somebody may confuse the two systems with each other.

While data processing operations are performed on documents and duty of the systems is to store documents, provide and create access to documents or their representatives. In text retrieval systems, data input is natural language text (full text or selections, or abstract full text) [3]. In information retrieval systems, output in response to a search query is in the form of a set of references. These references show information about system user favorite items to them [4]. The duty of a database management system is the storage, the preservation and the retrieval system in a system, i.e. the information in this system is not natural language text; it is in the form of certain data elements that are stored in tables.

This paper has been to use existing relationships between words to help build a suitable technique for automatic thesaurus-based indexing in the Persian language.

Rest of this paper is organized as follows. Section 2 is related works. In section 3, we explain the proposed method. Section 4 demonstrates results of our proposed method against traditional comparatively. Finally, we conclude in section 5.

2 Related Work

In 1999, Turney showed that keyword extraction is one of the most important factors accelerate and facilitate information retrieval applications, but until then there is no attempt to improve quality of extracted keywords [5].

Then simultaneously in 1999, Frank et al. who worked in the field of artificial intelligence, while they were presenting machine processing algorithm, they tried to improve the quality of extracted keywords. Their work was based on Simple Bayes algorithm. Their system is named "KEA". In this method, although the quality of the extracted words significantly increased, linguistic issues were not considered [6]. The general process for extracting keywords was introduced by Liu in 2005. They elected the first candidate keywords, and then assigned a weight to each word and finally extracted the keywords with the highest weights [7]. Franz in 2002 combined statistical analysis and linguistic analysis [8]. He believed that without considering information about linguistic knowledge, statistical analysis considers disadvantageous and non-key words [8].

In direction of previous researches in 2005, to solve the drawbacks of the extraction of disadvantageous and non-key words, Freitas et al. modeled the process of keyword extraction as a classification problem [9]. Zhang et al. used a decision tree as

classifier to recognize the keywords [10]. Halt used n-gram in the context of information retrieval [11]. In the first attempt, Deegan used thesaurus in 2004 [12] to improve information retrieval efficacy. After that Hyun tried to specialized thesaurus for a special query [13]. There are some successive works, tried to improve information retrieval efficacy after them [14]-[16].

Some of the work done in the field of Persian language is as follows [17]-[21]. While there are many methods in the Persian language, there is a lack of employing thesaurus in the Persian so far.

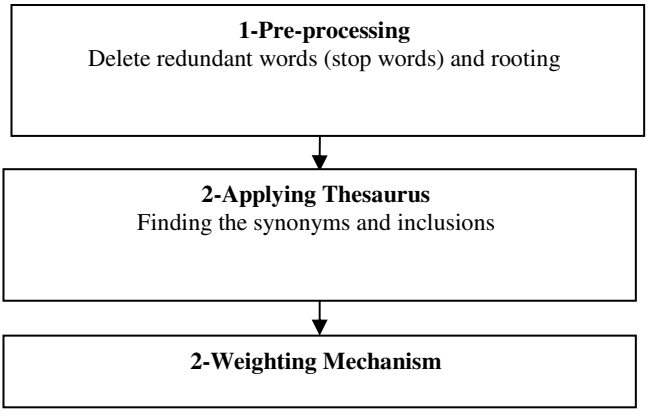

Fig. 1. Proposed Indexing framework

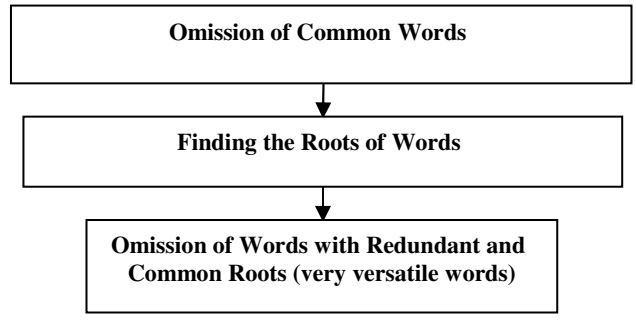

Fig. 2. Pre-processing phase of proposed framework

3 Proposed Framework

Fig. 1 depicts the proposed framework. The first step of the Fig. 1 is expanded in the Fig. 2. As seen in Fig. 2, in preprocessing step, Persian texts are refined to extract useful texts along with keywords to be ready for indexing stage. Indeed the pre-processing phase of proposed framework consists of three sub-parts. First the

common words like prepositions are omitted. Then the root of each word is found. Third the common roots, like "*be*", are omitted.

In the Fig. 3 assume that the word1, word2 and word3 are synonyms. Using thesaurus these three words are converted to first word, i.e. the frequency of word1 is considered as 3.

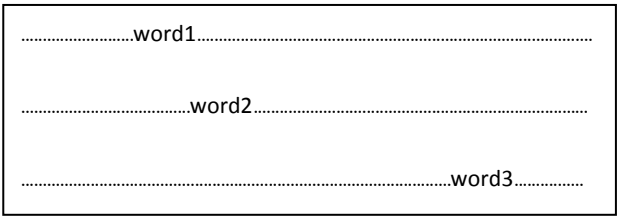

Fig. 3. A typical text with three words that are synonyms

So a table is produced from Fig. 3 that the frequencies of words are like Table 1.

Table 1. Table with frequencies of words of Fig. 3

word	#	
. . .		
word1	3	head
word2	3	
word3	3	
. . .		

So in the table of words frequencies synonym relationship is considered by a weight equals to 1, i.e. emerging the synonym of a word is equal to emerging that word. Another relationship that is taken into consideration is inclusion. For example a word an *animal* includes a *wolf*. So if in a text first *animal* is emerged, emerging a word *wolf* is equal to emerging animal with weight α, where α is less than one and vice versa. It means if an inclusion word has been emerged so far, emerging an included word is to emerge the included word by weight one, and including word by a weight below one. For example consider text of Fig. 4. Assume that *word5* is a special kind of *word4* and *word4* is special kind of *word3*. As before, *word1*, *word2* and *word3* are synonyms.

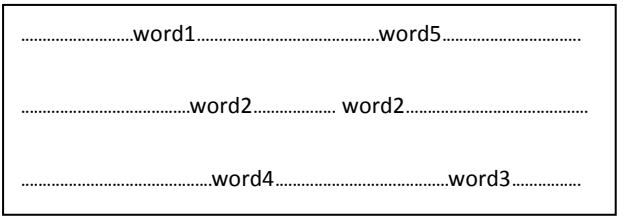

Fig. 4. A typical text with three words that are synonyms

Now a table is produced from Fig. 3 that the frequencies of words are like Table 2. For simplicity assume that α is 1/4.

Table 2. Table with frequencies of words of Fig. 4

word	#	
. . .	. . .	. . .
word1	4+1/4+1/4*1/4	head
word2	4+1/4+1/4*1/4	
word3	4+1/4+1/4*1/4	
word4	1+4*1/4+1/4	head
word5	1+1/4+4*1/4*1/4	head
. . .	. . .	. . .

In the table 2 *word1* is the head for three words, *word1*, *word2* and *word3*. Because the words, *word1*, *word2* and *word3*, are emerged 4 times, their frequencies are considered 4 at least. Besides due to emerging the *word4* that is a special kind of *word3*, a 1/4 is added to their frequencies. Due to emerging the *word5* that is a special kind of *word4*, a 1/4*1/4 is added to their frequencies. From another side, the frequency of the *word4* is at least 1, due to its appearance. Because of four appearances of the *word1*, 4 times 1/4 is added by its one appearance. Besides because of one appearance of *word5* another 1/4 is added to its frequency. This scenario is valid for *word5*. It means that one appearance of word5, plus 1/4 due to appearance of *word4* plus 4 appearances of *word1* that has inclusion relationship with length 2, i.e. 4*1/4*1/4, is considered as frequency of *word5*.

4 Experimental Results

In order to test the proposed method five different categories has been collected from Hamshahri [22] newspaper. The detail of the dataset is presented in the Table 3.

Table 3. Details of used dataset

Row	Topic	# of articles	Average # of words
1	Sport	146	204
2	Economic	154	199
3	Rural	171	123
4	Adventure	89	160
5	Foreign	130	177

After refinement of dataset, the average number of words in each category is reduced as the Table 4.

Table 4. Dataset after refinement

Row	Topic	Average # of words	Average # of words after refinement phase
1	Sport	204	149
2	Economic	199	135
3	Rural	123	76
4	Adventure	160	115
5	Foreign	177	124

After applying refinement phase, we produce a feature space as illustrated in the Table 5.

Table 5. Dataset after refinement

	Head Word1	Head Word2	Head Word1	.....	Head Wordn
Article1					
Article2					
00000					
Articlem					

In Table 5, parameter n is the number of all Head Word which is a head word in an article at least. The entity *j*th column of *i*th row in Table 5 is equal to frequency value of head word *j* in the *i*th article.

By filling the Table 5 values by using thesaurus and without using thesaurus we obtain two different datasets. By 4-fold cross validation and 1-neareast neighbour classifier, we reach the results in the Table 6.

Table 6. Accuracy of 1NN classifier with and without thesaurus

	Without thesaurus	With thesaurus
Accuracy of classification	68.3%	78.4%

5 Conclusion and Future Works

In this paper, we have proposed a new method to improve the performance of Persian text classification. The proposed method uses a Persian thesaurus to reinforce the frequencies of words. With a simple classifier, it is shown that using thesaurus can improve the classification of Persian texts. We consider two relationships: synonyms and inclusion. We use a hierarchical inclusion weighting, and linear synonym weighting.

As a future work, one can turn to research on the different weighting methods.

Acknowledgments. This research is supported by Iran Communication Research Center, Tehran, Iran.

References

1. American Society of Indexers, Frequently Asked Questions Indexing. Index review in Books, Ireland (1994), http://www.asindexing.org/site/indfaq.shtml
2. Maron, M.E.: Automatic indexing: an experimental enquiry. Journal of the ACM 8, 404–417 (1961)
3. Montgomery, C.A.: Linguistics and information science. Journal of the American Society for Information Science 23, 195–219 (1972)
4. Brooks, H.M.: Expert Systems and Intelligent Information Retrieval. Information Processing and Management 23(4), 367–382 (1987)
5. Turney, P.D.: Learning Algorithms for Keyphrase Extraction. Information Retrieval 2(4), 306–336 (1999)
6. Frank, E.: Domain-Based Extraction of Technical Keyphrases. In: 6th International Joint Conference on Artificial Intelligence, India (1999)
7. Liu, Y., Ciliax, B.J., Borges, K., Dasigi, V., Ram, A., Navathe, S.B., Ingledine, R.: Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering. In: 4th IEEE Computational Systems Bioinformatics Conference (CSB 2004), Stanford (2005)
8. Frantzi, K., Ananiadou, S., Mima, H.: Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. Digital Libraries 3(2), 115–130 (2002)
9. Freitas, N., Kaestner, A.: Automatic text summarization using a machine learning approach. In: 16th Brazilian Symposium on Artificial Intelligence (SBIA), Brazil (2005)

10. Zhang, Y., Heywood, N.Z., Milios, E.: World Wide Web Site Summarization Web Intelligence and Agent Systems. Technical Report, CS-2002-8 (2006)
11. Hult, A.: Improved automatic keyword extraction given more linguistic knowledge. In: 8th Conference on Empirical Methods in Natural Language Processing (EMNLP), Japan (2003)
12. Deegan, M.: Keyword Extraction with Thesauri and Content Analysis, http://www.rlg.org/en/page.php?Page_ID=17068
13. Hyun, D.: Automatic Keyword Extraction Using Category Correlation of Data, Heidelberg, pp. 224–230 (2006)
14. Witten, W., Medley, I.H.: Thesaurus based automatic keyphrase indexing. In: 6th ACM/IEEE-CS JCDL 2006 (Joint Conference on Digital Libraries) (2006)
15. Klein, M., Steenbergen, W.V.: Thesaurus-based Retrieval of Case Law. In: 19th International JURIX Conference, Paris (2006)
16. Martinez, J.L.: Automatic Keyword Extraction for News Finder, Heidelberg, pp. 405–427 (2008)
17. Shahabi, A.M.: Abstract construction in Persian literature. In: Second International Conference on Cognitive Science, Tehran, p. 56 (1381) (in Persian)
18. Bahar, M.T.: Persian Grammar, ch. IV, p. 111 (1342) (in Persian)
19. Khalouei, M.: Indexing Machine. Journal Books 6(3) (in Persian)
20. Karimi, Z., Shamsfard, M.: Automatic summarization systems Persian literature. In: 12th International Conference of Computer Society of Iran (1385) (in Persian)
21. Yousefi, A.: Principles and methods for computerized indexing. Journal Books 9(2) (in Persian)
22. Hamshahri newspaper, http://www.hamshahrionline.ir