PAKE: a supervised approach for Persian automatic keyword extraction using statistical features

Soghra Lazemi¹ · Hossein Ebrahimpour-Komleh¹ · Nasser Noroozi²

Received: 30 May 2019 / Accepted: 2 November 2019 / Published online: 8 November 2019
© Springer Nature Switzerland AG 2019

Abstract

Keywords are a collection of important words in a document that are the core topic of the discussion. This paper proposes a hybrid method for automatically extracting keywords from Persian documents and web pages. In the proposed method, firstly, based on linguistic knowledge, processing was performed at word and letter levels to optimize of the analysis. Then a new statistical features set is defined and extracted at the word level. At the final stage, keywords are determined using the SVM algorithm. Also, in this paper, due to the lack of a corpus for evaluating the methods of automatic extraction of Persian keywords, a large-scale corpus has been developed and introduced. The achieved F-measure for keywords and non-keywords are 99.89% and 99.99% respectively.

Keywords Automatic keyword extraction · Natural language processing · Persian

1 Introduction

Today, in the competitive world in which we face an information explosion, a significant part of usable data is presented as text data in the form of digital books, articles, web pages, e-mail, web blogs, comments on products etc. Searching and navigation among this huge volume of documents is very difficult, time-consuming and confusing, and it's almost impossible to find the intended materials without using text-mining techniques. Automatic extraction of keywords is a subset of text mining that facilitates the organization and retrieval of text data and helps the user understand and access information in a text (document) in a short time [1]. Keywords are a set of words (a word or set of words) is a text (document) that can be an indicator of its contents [2]. As a compressed representative of a text, keywords can be a useful and appropriate solution to the problem of curse of dimensionality [3]. Today, there are more than 130 trillion web pages on

the internet [4], of which nearly 1.7% are in Persian [5]. In other words, we have over 2.21 trillion web pages in Persian. Although one can determine the keywords of a web page by the meta tag in HTML, not all Persian web pages use this tag, or they do not use proper keywords for web pages. Since one of the main ways to display the core content of the text and search across web pages is the use of keywords, therefore, the use of automated keyword extraction systems is required. The importance of extraction of keywords from Persian texts is more pronounced due to the rapid increase of Persian electronic documents, the greater complexity of the grammar and writing in Persian language, and the little research performed in this field.

The rest of the paper is organized as follows: Sect. 2 surveys the conducted researches in the field of keyword extraction-especially in Persian. Section 3 describes our proposed method in detail. Section 4 presents

✉ Soghra Lazemi, soghra.lazemi@gmail.com; Hossein Ebrahimpour-Komleh, ebrahimpour.kashanu@gmail.com; Nasser Noroozi, Noroozi@kashanu.ac.ir | ¹Department of Computer Engineering, Faculty of Electrical and Computer Engineering, The University of Kashan, Kashan, Iran. ²Faculty of Mathematics, The University of Kashan, Kashan, Iran.

experiments and results. Finally, Sect. 5 makes conclusion on this paper and suggests some future work.

2 Literature review

So far, there has been a lot of research into automated keyword extraction. These researches have various features (language-dependent/independent, domain-dependent/independent, accuracy, speed) that can be examined from a variety of aspects. In this regard, some researchers have invented or domesticated a specific methodology; for example, for English language, [6] extract the keywords using the machine learning method in a supervised manner with the NB Classifier. Reference [7] use the machine learning method in a graph-based, completely unsupervised manner. Reference [8] have proposed a language-based approach which uses the syntactic analysis. For Arabic, [9] domesticate KEA for this language. Reference [10] extract keywords in English and Arabic by a statistical method, inspired by the nature of the documents and key phrases. For Chinese, [11] uses PAT-tree. Reference [12] introduce a method based on N-gram and Word co-occurrence, [13] propose a statistical method based on TF-IDF. Jonghera and Analouei [14] have proposed statistical, corpus-oriented methods for extracting keywords from Persian documents. In their proposed method, the statistical characteristics for the different words are calculated and the probable keywords are selected using fuzzy rules. Khozani and Bayat [15] have used statistical methods. The proposed method consists of two steps. In the first step, redundant words are excluded and the remaining words are weighed using the TF-IDF criterion. In the next step, using the n-gram method, a matrix is created for the words and based on the words' position, the weight of the words is updated. Then based on the obtained weight, the key sentences are determined and using the selected sentences, the keywords are extracted. Experiments have been carried out on 48 documents collected by the authors. Kian and Zahedi [16] have also used the statistical method to extract keywords. In their proposed method, after unification and deletion the stop words, the information about the neighbors of the word is extracted and based on the obtained score, the likelihood of a word being the keyword is calculated. Experiments have been conducted with 800 documents prepared by the authors.

3 Proposed method

Our proposed method generally runs in 2 phases. The first phase, called preprocessing. The input of this phase is a TXT file. In this phase, depending on the characteristics

of the Persian language, a series of changes are made to the file. We used PersianToolbox [17] to perform unifying, tokenizing, stemming and removing stop words tasks, and we papered and used a list of writing marks and numbers for removing writing marks and numbers. In second phase, to find keywords in Persian language texts, we first take a list of words in the text and store it in a database. Given that keywords may be a combination of several consecutive words, we put all the consecutive words up to a maximum of 5 words in the database as a compound word. Then we calculate the value of some of the features on each of the words in the database. These features that will determine whether a word is a keyword or not; they include:

Term Frequency determines how many times a word is repeated in the document and is calculated by Eq. 1 [18]:

$$TF(t_i, d) = \frac{f(t_i, d)}{N} \quad (1)$$

N is the total number of words in article d , $f(t_i, d)$ is the number of repetitions of the word, t_i , in article d , and $TF(t_i, d)$ is the repetition of the word t_i in article d .

Inverse Document Frequency Some words like verbs are repeated many times in all documents, but they are not keywords. In these cases, the IDF feature is used to indicate how many times a word has been repeated in "other" document [18]:

$$IDF(t_i) = \log \frac{D}{1 + |\{d \in D, TF(t_i, d) \neq 0\}|} \quad (2)$$

D is the total number of documents, and $|\{d \in D, TF(t_i, d) \neq 0\}|$ is the number of articles in which the t_i word has been repeated at least once.

Combination degree This feature, which is calculated only for compound words (for other words, it is considered equal to one), determines how many times each item of a compound word is repeated beyond that combination. This feature is derived by Eq. 3:

$$Deg(T) = \frac{f(T) \times |T|}{\sum_{i=1}^{|T|} f(t_i)} \quad (3)$$

T is a compound word composed of $|T|$ words $t_1, t_2, \dots, t_{|T|}$. $f(t_i)$ is the number of repetitions of the word t_i .

Head This feature determines how many times a word is repeated in the headings. To calculate this property, we first consider all the headings of the article as a new article called H , and then we use the Eq. 4 to calculate the value of this feature for the t_i word:

$$Head(t_i) = TF(t_i, H) \quad (4)$$

Entropy The keyword of an article is generally discussed from the beginning to the end of the article, while other words are only used in a particular section of the article, which is referred to as the “local keyword”. This feature determines how much a word has been distributed in the document. To calculate this feature, we divide the input article into at least 64 equal parts, and then use the Eq. 5:

$$Ent(t_i) = - \sum_{i=1}^n p(i) \times \log_2 p(i) \quad (5)$$

$p(i)$ is the number of repetitions of the word t_i in the i th section of the article, and n is the number of sections in which the input article is divided.

Transformation The keyword of an article usually does not change and remains constant from the beginning of the article to the end, while synonyms are used for the insignificant words to avoid uniformity. To calculate this feature, first we derive all the synonyms of the word t_i in Persian using FarsNet [19], T_j , and calculate their transformations by Eq. 6:

$$Trans(t_i, d) = \frac{1 + \sum_{j=1}^n f(T_{jti}, d)}{f(t_i, d)} \quad (6)$$

4 Experiments and results

As we have mentioned, we used 1570 papers in various areas as the corpus of the system. These articles were collected from two different sources. They include.

Articles and academic thesis Part of the articles in the corpus is actually articles published in Persian journals as well as academic dissertations. These articles have been obtained from the following sources:

1. Academic Jihad Scientific Information Database (<http://www.sid.ir/>).
2. The website of the libraries of some Iranian universities (which allows the user to download the thesis).

In these articles, the author generally presents the keywords after the abstract section. So you can easily extract the actual keywords. The format of articles provided by this source is PDF. Given that many PDF files have no standard structure and that their content is not readily readable, only articles have been selected from this source that have a standard PDF file that is convertible to text.

News and articles available on the Internet We collected news and articles from Persian websites using a HTTP crawler. Then we considered the content of the meta keywords tag from the HTML page as the keyword list of those articles. We chose websites for resource collection

in which, authors have listed the keywords of each article in the meta keywords tag. Some websites do not use the right keywords for each article, or they use a keyword for the entire website. For example, on the farsnews.com website, this rule is not complied with in all articles, and some articles do not contain proper keywords, while they are complied with in isna.ir. Also for integration and avoiding errors when constructing the model, we selected the articles from both sources that meet both of the following criteria:

1. Articles containing at least 500 and at most 40,000 words.
2. Each article must have at least 3 and at most 7 keywords to be selected.

The general specifications of the corpus used in are presented Table 1. For each of the file formats listed, a Handler was designed to convert those formats to TXT format.

In order to evaluate the performance of proposed method Precision, Recall, F1-measure and Accuracy are used as the evaluation metrics [20]. Unfortunately, due to the lack of annotated resources, Persian has attracted fewer researchers in the keyword extraction field, so we evaluate our proposed method by comparing the performance of our defined features. The experiments are configured in 2 stages: in the word level and in the document level. First Case: in this case, the words are utilized as instances and are classified as keyword or non keyword.

In the other word, classification process is performed in the word level. Total number of keywords and non keywords are 7084 and 3,698,752 respectively. Table 2 presents the predictive performance of the most frequent features (TF and IDF) and the introduced features in this paper on the mentioned measures. As it can be observed from Table 2, the highest scores are achieved by our proposed features in all measures. The results indicate that TF and IDF can extract non-keywords and for keyword extraction produce very poor results. As shown in Table 2, the combination of all features shows a large and significant improvement.

Second Case: in this case, the documents are utilized as instances. In this case, the keywords of 10 articles have been extracted incorrectly. We considered articles that

Table 1 Statistic about the corpus

Number of PDF files	432
Number of web pages in format HTML 4	26
Number of web pages in format HTML 5	1112
Average document length	235.62
Average number of document keywords	4.66

Table 2 Results on word level

Features	Recall	Precision	F-measure
TF + IDF	0.4392	0.04238	0.0779
TF + IDF + combination degree	0.5734	0.5570	0.5650
TF + IDF + combination degree + head	0.8396	0.7912	0.8146
TF + IDF + combination degree + head + entropy	0.9559	0.8933	0.9235
TF + IDF + combination degree + head + entropy + trans-formation	0.9995	0.9983	0.9989

the solution did not work properly on them. According to results, some of the keywords in these articles were words that were not repeated or emphasized directly and repeatedly in the article, but were extracted from the concept of text.

5 Discussion and conclusion

There is a dire need of Persian automatic keyword extraction approaches with regards to the fact that there are more than 2 trillion Persian web pages now and keywords of most of them are not determined correctly. This article presents a solution for the automatic extraction of keywords from Persian texts using SVM. To achieve the objective, a set of new statistical features is defined. In addition, due to the lack of a corpus for evaluation, a large-scale corpus has been developed. The results of the experiments show that the proposed strategy can detect the keywords of Persian documents with an accuracy of 99.3631%. As a future work for the discovery of keywords, it is suggested that the concept of sentences in the article is extracted. It should be noted that concept extraction is very time-consuming task and its use to disclose keywords is not appropriate in a large number of articles, therefore, it is recommended that optimizations are performed.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Lee D-Y, Kim K-R, Cho H-G (2016) A new extraction algorithm for hierarchical keyword using text social network. In: Kim KJ, Joukov N (eds) Information science and applications. Springer, Singapore, pp 903–912
- Matsuo Y, Ishizuka M (2004) Keyword extraction from a single document using word co-occurrence statistical information. *Int J Artif Intell Tools* 13(01):157–169
- Siddiqi S, Sharan A (2015) Keyword and keyphrase extraction techniques: a literature review. *Int J Comput Appl* 109(2):18–23
- Schwartz B (2016) Google's search knows about over 130 trillion pages. <https://searchengineland.com/googles-search-indexes-hits-130-trillion-pages-documents-263378>
- Usage of content languages for websites (2017) [cited 2017 13 November]. https://w3techs.com/technologies/overview/content_language/all
- Witten I, Paynter H, Frank GW, Gutwin EC, Nevill-Manning CG (2005) KEA: practical automatic keyphrase extraction. In: Proceedings of the 4th ACM conference on digital libraries. ACM, pp 129–152
- Mihalcea R, Tarau P (2004) TextRank: bringing order into text. In: Proceedings of the conference on empirical methods in natural language processing
- Ercan G, Cicekli I (2007) Using lexical chains for keyword extraction. *Inf Process Manag* 43(6):1705–1714
- Duwairi R, Hedaya M (2016) Automatic keyphrase extraction for Arabic news documents based on KEA system. *J Intell Fuzzy Syst* 30(4):2101–2110
- Awajan A (2015) Keyword extraction from Arabic documents using term equivalence classes. *ACM Trans Asian Low Resour Lang Inf Process* 14(2):1–7
- Chien LF (1997) PAT-tree-based keyword extraction for Chinese information retrieval. In: ACM SIGIR forum. ACM
- Jiao H, Liu Q, Jia HB (2007) Chinese keyword extraction based on N-gram and word co-occurrence. In: International conference on computational intelligence and security workshops. IEEE
- Seraji M (2015) Morphosyntactic corpora and tools for Persian, Acta Universitatis Upsaliensis
- Jonghera MM, Analouei M (2010) Keyword extraction Persian documents. In: 3th annual conference of computer society of Iran, Tehran
- Khazani SMH, Bayat H (2011) Specialization of keyword extraction approach to Persian texts. In: International conference of soft computing and pattern recognition. IEEE
- Kian H, Zahedi M (2013) Improving precision in automatic keyword extraction using attention attractive strings. *Arab J Sci Eng* 38(8):2063–2068
- Mohseni M, Ghofrani J, Faili H (2016) Persianp: a persian text processing toolbox. In: International conference on intelligent text processing and computational linguistics, pp 75–87
- Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. *Inf Process Manage* 24(5):513–523
- Shamsfard M, Hesabi A, Fadaei H et al (2010) Semi automatic development of farsnet; the persian wordnet. In: Proceedings of 5th global WordNet conference, Mumbai
- Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.