International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 6, December 2020, pp. 5909~5916
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i6.pp5909-5916

Improving keyword extraction in multilingual texts

Bahareh Hashemzadeh[1], Majid Abdolrazzagh-Nezhad[2]
[1]Department of Computer and Information Technology, Faculty of Engineering, Torbat-E Heydariyeh University, Iran
[2]Department of Computer Engineering, Faculty of Engineering, Bozorgmehr University of Qaenat, Iran

Article Info

Article history:

Received Jun 18, 2019
Revised Apr 29, 2020
Accepted May 12, 2020

Keywords:

Data retrieval
Graph-based algorithm
Keyword extraction
Language independent
Text mining
TF-IDF algorithm

ABSTRACT

The accuracy of keyword extraction is a leading factor in information retrieval systems and marketing. In the real world, text is produced in a variety of languages, and the ability to extract keywords based on information from different languages improves the accuracy of keyword extraction. In this paper, the available information of all languages is applied to improve a traditional keyword extraction algorithm from a multilingual text. The proposed keywork extraction procedure is an unsupervise algorithm and designed based on selecting a word as a keyword of a given text, if in addition to that language holds a high rank based on the keywords criteria in other languages, as well. To achieve to this aim, the average TF-IDF of the candidate words were calculated for the same and the other languages. Then the words with the higher averages TF-IDF were chosen as the extracted keywords. The obtained results indicat that the algorithms' accuracis of the multilingual texts in term frequency-inverse document frequency (TF-IDF) algorithm, graph-based algorithm, and the improved proposed algorithm are 80, 60.65, and 91.3%, respectively.

Corresponding Author:

Majid Abdolrazzagh-Nezhad,
Department of Computer Engineering,
Bozorgmehr University of Qaenat,
9761986844 Bozorgmehr University of Qaenat, Abolmafakher St, Qaen, South Khorasan, Iran.
Email: abdolrazzagh@buqaen.ac.ir

1. INTRODUCTION

Designing data retrieval systems of large databases is one of the research areas for the application of information technology in the information business. We faced an increasing demand for types of data retrieval systems able to cross the interlingual boundaries, while text data expands in different languages and on the web [1-6]. Therefore, by developing the volume of electronic data in various languages, the data retrieval, independent of document languages, has gained importance. The extraction of effective keywords is a time-consuming and human-processing task. Recently, automatic keyword extraction, especially keyword extraction in different languages, introduced an interesting topic for text mining and data retrieval [7-9].

The fields of text mining and information retrieval and especially their implementation on the database is of particular importance. The first step is to identify and extract keywords from the texts in the fields. One of the main challenges to extract keywords is existing very diverse languages for contextual information and depending the available keyword extraction methods on the language's type and its verbal structure. The multilingual keywords extraction is the current research problem and the research object is considered based on designing an unsupervised language-independent algorithm to the extraction. So, it is done by focusing on the property of repeating keywords in each text and their intensifying in other texts by utilizing the TF-IDF algorithm.

The rest of the current paper is organized as follows: Section 2 reviews the state-of-the-art keywords extraction methods. The problem of keywords extraction descrids in Section 3. The proposed language independent keywords extraction algorithm and its experimental results are discussed in Section 4 and Section 5. Finally, a conclusion and recommendations are described in section 6.

2. LITERATURE REVIEW

Several methods were proposed so far for the identification and extraction of keywords, all of which could be classified into two groups of supervised and unsupervised methods [10-12]. In the following, we discuss shortly about the proposed methods to realize the probable research challenges. The first group is the supervised methods. In this group, there is a training data set, by learning of which a model is designed and by incorporating this model on new document the phrases will divided into two classes of key and non-key phrases.

The supervised method of word extraction is considered as a clustering problem, which should be trained like a genetic algorithm [13, 14]. In Bayes linear algorithm, which is called a keyphrase extraction algorithm (KEA) and proposed by [15], TF-IDF and keyphrase relative distance from the beginning of the text are two algorithm inputs [16]. They also used a binary clustering algorithm that its input features include some references to the text. Decision tree of [17], conditional random field of [18], and a type of KEA in [19] are among other types of supervised word extraction. The functionality of this method is highly dependent on training data and lack of such high quality data could cause an efficiency drop in the system of keyphrase extraction. In this method, the designed model is specific to a domain and works based on the domain of usage.

Another approach to extracting keywords is through unsupervised methods. In these methods, word extraction is dealt with as a ranking issue [20], the most important of which is the TF-IDF. In this method, the relation between the number of a word repetition within a text is calculated according to the number of its repletion in other texts [21]. Graph-based methods are also among the unsupervised methods [22]. The works of [22-24] are examples of graph-based methods for word extraction. In unsupervised methods, there is no need for training data and the most important contextual phrases could be extracted by using the ranking strategies. Unlike the supervised methods, the unsupervised methods are applicable for each text to any domain type independent of domain of usage. By the qualitative analysis and comparison of the proposed methods several advantages and disadvantages were found, which could be noted as follows.

The first advantage of the unsupervised methods is their applications in constructing models of any text type and domain. No efficiency drops in case of existence of poor quality data, independently of training data, lower time consumption for keyword extraction, compared with the supervised methods, useful functionality for high-volume data, and high accuracy are among the advantages of the unsupervised methods. In contrast to these advantages, low compatibility is the most tangible shortcoming of these methods. As mentioned previously, there are some disadvantages/advantages of the supervised methods, among which we could refer to the existence of training data with the quality of regular data categorization. However, one of the significant shortcomings of this method is that it is dependent on the training data and lack high-quality data could lead to an efficiency drop of the keyword extraction system, the constructed model is for one domain only, and it acts based on the domain of usage. Providing training data is a time-consuming and laborious task. Moreover, evaluations which are made based on frequency are not applying for high-volume data. One of the challenges of such a method is that providing training data is time-consuming and if such data are not available, the algorithm faces problems and has low efficiency, but it is not the case in the unsupervised method [1, 3]. Hence, we employ this method for the proposed algorithm.

Despite the simplicity, TF-IDF algorithm is one of the effective methods for keyword extraction [16, 25]. The practical simplicity and efficiency of this algorithm has attracted a considerable attention. A logarithm is proposed for word extraction in the present study to improve TF-IDF. This method is based on TF-IDF, but uses the information of each text in several languages to enhance keyword extraction based on TF-IDF. To implement such an objective, we concentrated on the repetition of words in the context and deleted the conjunctions, prepositions, and verbs. Further, we used simultaneous multilingual information for a certain text, to improve its usage. This process is elucidated in details in the following.

3. PROBLEM DISCRIPTION OF KEYWORDS EXTRACTION

Data retrieval is used extensively in the everyday life of people. Enhancing efficiency and improving performance is of great importance for the designers of data retrieval systems. As mentioned previously, one way to increase the productivity of data retrieval systems is through the use of statistical plans. In these plans, a frequency is set of keywords, based on which words with the highest frequency are selected as keywords.

In aim of the present study was to propose an algorithm, which has the required features, including non-supervisory, language-independent, simplicity, and high speed for processing considerable amount of data. By using the proposed algorithm along with the TF-IDF, which is a statistical, simple, language independent and non-supervisory algorithm, by relying on a sequence of calls with Unicode format, and by designing an online database keyword could be extracted independently of language in large databases.

By assessing the applications of data retrieval and text mining, we could realize that existing keywords within a text play a significant role and facilitate the process in this field. For example, by finding important words in the news and by detecting sentences with more important words, we could extract that sentence in the abstract and better comprehend the text. Since important words are often in headings and important sections, by realizing the structure of a text and by extracting keywords out of these parts, we could get access to these words with a minimum of time. Feed or RSS is used for reading news, which make a news extract available in a structured way by XML format. News reading and saving template are Unicode. For extracting keywords of news texts, we need websites with proper and authentic feed addresses. Hence, we select those feeds, which provide appropriate information. These feeds, however, are selected for every language. After calling information from feeds, they will be saved in a database. Some words are available with high frequency in all texts with no contextual value, like pronouns, adverbs, prepositions, conjunctions, and some frequent verbs. These elements are called public words. By omitting the public words in statistical text mining, we have less calculations and higher efficiency. Words take an equal weight based on their frequency in the document. Actually, this weighting system shows how much a word is important for a document. This fact has no functionality in data retrieval. The weight of a word in a text increases by the number of repetitions in that text, but it is controlled by the number of words in the text. This method is an unsupervised one, which is applied to a simple text. In contrast to the supervised methods, this method does not need the training dataset, in that proving an appropriate training data is a time-consuming and not an easy task and in case the data lack the desired quality, they reduce the efficiency of the supervised keyword extraction system.

4. THE PROPOSED ALGORITHM

Figure 1 presents the oerall structure of the proposed algorithm in seven steps, which its detail is discussed as follows:

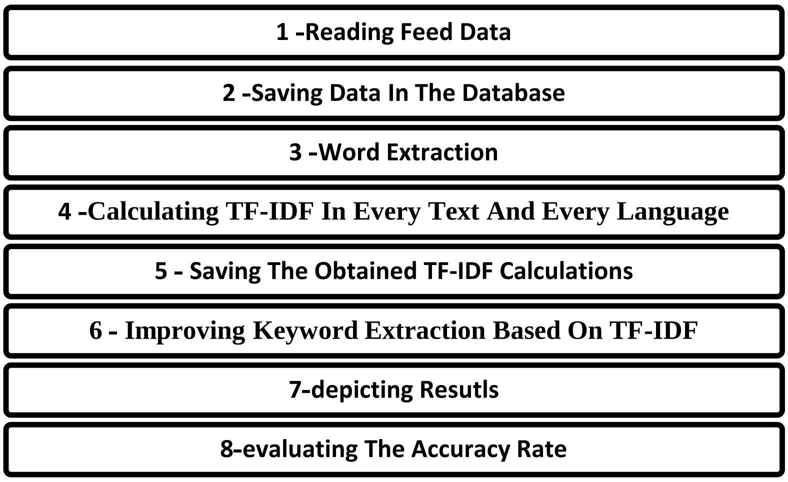

1 -Reading Feed Data
2 -Saving Data In The Database
3 -Word Extraction
4 -Calculating TF-IDF In Every Text And Every Language
5 - Saving The Obtained TF-IDF Calculations
6 - Improving Keyword Extraction Based On TF-IDF
7-depicting Resutls
8-evaluating The Accuracy Rate

Figure 1. The overall structure of the proposed algorithm

Step 1 (selecting feeds and retrieval): in order to gain access to various documents of different languages, we tried to select the appropriate feeds. Data retrieval of each document, like title or body is carried out in this step. Since our algorithm is language independent, information is read by the unicode format. Step 2 (saving document information in the large database): the read information is stored in the database, separately. Data are stored in the Unicode format. This format covers most of the languages. Step 3 (word extraction): all words are extracted from the text and omitted in the step related to this action. Every language has a list of repetitive words, which should be deleted from the extracted words.

Step 4 (TF-IDF calculations): TF-IDF calculations are carried out in the step for every text and language and finally the calculated TF-IDF of each text in a different language is used for improving the keywords. In this method, each word has a frequency-based weight in the document. Actually, such weighting system shows how much a word is important for a particular document. This process is used frequently in data retrieval. The weight of a word is increased by the increase of its repetition in a certain text, but is controlled by the number of words in the context, because if the text is lengthy some words would be repeated, naturally, though they do not have any significance in the meaning. Term frequency is a criterion for the range of common and repetitive words in a text, which is calculated as follows:

$$TF(f,d) = 0.5 + 0.5 \times \frac{f(t,d)}{max\{f(w,d): w \in d\}} \quad (1)$$

where in the numerator, d is the number of words in the selected text. w is the most frequent words in the selected text.

IDF (inverse document Frequency) is a criterion for the range of the most frequent and repetitive words. This criterion is achieved by dividing the total number of texts in the number of texts including the common word. For example: suppose that there are 1000 texts in the whole databases. If there is a certain word in all of them (like, is) the result of an algorithm is 1000 divided by 1000, which is 0, that is, this word is among the common words and must be taken the coefficient of 0. However, if the repetition is occurred in 500 texts, the result is 1 and takes the coefficient of 1. The more the repetition of a word, the less is the IDF weight. In case a word has no repetition and dominator becomes 0, we put +1 in dominator, which is calculated through second formula:

$$IDF(t,D) = log(\frac{D}{1+\{d \in D : t \in d\}}) \quad (2)$$

where, D is the number of existing texts in the numerator and the number of texts bearing the word in the dominator. The TF-IDF is calculated through formula (3) as follows:

$$TF\_IDF(t,d,D) = TF(t,d) * IDF(t,D) \quad (3)$$

Step 5 (saving calculations in the database): the performed calculations are saved in the database by TF-IDF algorithm. Step 6 (improving the extraction of the proposed TF-IDF): in the conventional TF-IDF, in a text in a certain language, words with the highest frequency of TF-IDF are considered as keywords in that text with the same language. However, in the proposed method, words are called keywords if their averages TF-IDF are high for that text with the same language and other languages. Therefore, the average TF-IDF is considered for a text with the same language and other languages and instead of using TF-IDF of a text in a language, its average TF-IDF is used in available languages. This simple, but useful method could improve the extraction of keywords, significantly. In this paper, average and maximum TF-IDF method for a text in different languages is also tested, the result of which outweighed the conventional one. However, the method, which calculates the average TF-IDF has the highest accuracy.

Step 7 (depicting results): this step shows those keywords, which were extracted by TF-IDF improved algorithm. Step 8 (evaluating the accuracy rate): in this step, the keyword extraction accuracy of the algorithm is calculated through the following formula:

$$Accuracy\ rate = \frac{No.of\ correct\ extracted\ words}{total\ no.of\ words\ extracted\ as\ keywords} \times 100 \quad (4)$$

where, the number of correct extracted keywords are those words, which are common between actual keywords and the extracted one by the algorithm. The dominator is also the total number of extracted words by the algorithm as a keyword.

The pseudo-code of the proposed keywords extraction algorithm is presented in Figure 2. The algorithm is unsupervised and could be run on the simple text. It means that unlike supervised keyword extraction algorithms, there is no need for appropriate training data sets. As known as, providing appropriate training data is time consuming and difficult. If the data is not of good quality, it will lead to a decline in the efficiency of the supervised keyword extraction algorithms.

```
Begin
If data have ASCI format, change them to Unicode format.
Read information from feeds with Unicode by Get RSS data function.
Store the information in a database.
Generate Ignore array based on prepositions, conjunctions, adverbs and verbs.
Read all words by GetWord function and save them in Key_word array.
Remove the words of Ignore array from Key_word array.
Calculate Equ. (3) for Key_word array by running TF-IDF algorithm.
Calculate the average of TF-IDF for Key_word array in different languages.
Save any words of Key_word array as keywords if their averages TF-IDF are high for the same and the other languages.
Calculate Equ. (4) for the identified keywords.
End
```

Figure 2. The pseudo-code of the proposed algorithm

5. EXPERIMENTAL RESULTS

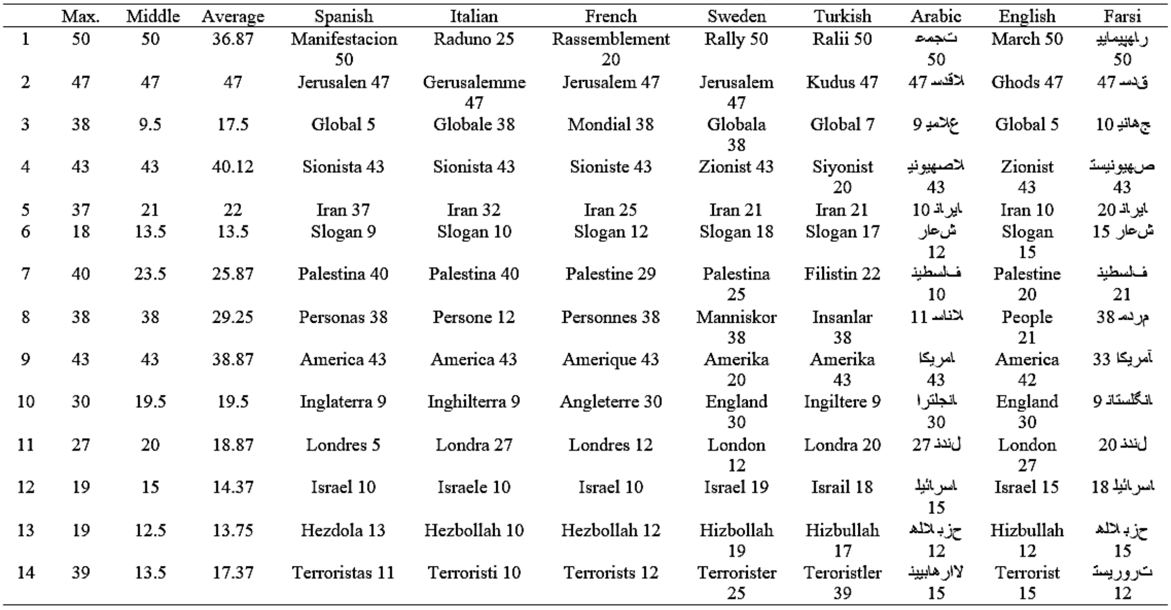
The proposed algorithm was programed in SQL Server 2012 and Visual Studio 2013 and simulations were performed on the Intel Core i5, 64 B, CPU 2.50 GHz and RAM 21 GB. The database used for evaluating the efficiency and performance of the proposed keyword extraction algorithm has been an online dataset containing 200 news collected from BBC website in various languages. Each news is in eight languages. The reason for using such a dataset was to provide updated information, which are processed at the same time. The proposed method is assessed by counting the number of matching between extracted keywords by the proposed method and given keywords.

5.1. The results of the proposed algorithm

An algorithm is designed in this study, which is language independent and has a simple structure. In contract to language-dependent algorithms (like [26]), which are using the Persian roots for keyword extraction, this algorithm is simply functional for large databases in every language. In the TF-IDF algorithm, high-frequency words in a text, but in all languages (TF-IDF mean in all languages) were selected as keywords and the accuracy of the algorithm, considering the text in various languages is improved. It is noteworthy that in a text, non-keywords, including verbs and prepositions are repeated, considerably, so, we set all non-keywords a side at the very beginning. The proposed algorithm is applicable to all multilingual websites and here the results were shown just on BBC News Website. The database used is comprised of 200 news collected from BBC Website in eight languages (a total of 1600 news). As can be seen in Table 1, words with relatively high TF-IDF (here TF-IDF more than 20) were considered, while in the conventional TF-IDF algorithm, in every language, those independent words with highest TF-IDF value is counted as keywords. As can be seen in Table 2, in Persian language, the word “America” is detected as a keyword (in thickened Table 2 mistakenly, while in English language, three words of “America, England, and London” (in thickened Table 2 were mistakenly detected as a keyword. In other languages, two or three keywords were also known as keywords, mistakenly.

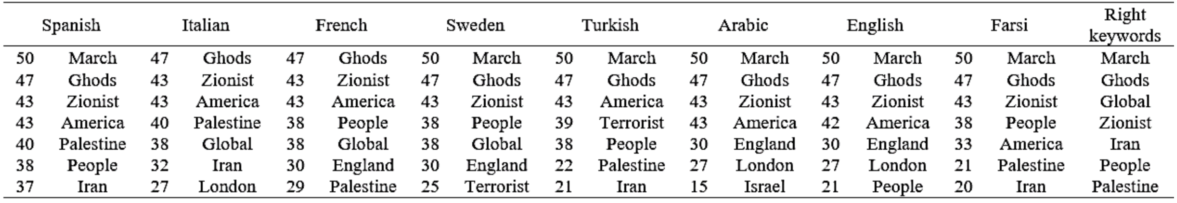
Table 3 illustrates the proposed algorithm results for the selected text. As can be seen in the table, the mean TF-IDF is calculated in eight languages (the proposed algorithm) for each word depicted in Table 1 and seven keywords were selected. The selected keywords in this method are considered for all eight languages, such that for all languages in this text, keywords in the mean method, which are shown in Table 3 include Quds, Zionist, America, demonstration, people, Palestine, and Iran, in which America is detected mistakenly as a keyword for all languages. However, as we mentioned in Table 2, in the conventional TF-IDF method the number of wrong detected keywords is different and more than one word for most languages. If we evaluate the accuracy of mean TF-IDF algorithm (the proposed one) and that of the conventional algorithm, the conventional algorithm (which is shown in Table 2) 6 of 7 Persian words, 4 of 7 English words, and 3 of 7 Arabic words, as well as other words in other languages were detected, correctly. In total, in 8 languages and among 56-7*8 correct keywords, 39 were detected correctly and the accuracy of the algorithm is 0.69=39/56, while in the mean TF-IDF method, 6 of 7 words were detected correctly for all languages and the accuracy of the algorithm is 0.85=6/7. This is the case of the mean and maximum method.

Table 1. The TF-IDF value for words that are most likely to be among the key words for the selected text



	Max.	Middle	Average	Spanish	Italian	French	Sweden	Turkish	Arabic	English	Farsi
1	50	50	36.87	Manifestacion 50	Raduno 25	Rassemblement 20	Rally 50	Ralii 50	تجمع 50	March 50	راهپیمایی 50
2	47	47	47	Jerusalen 47	Gerusalemme 47	Jerusalem 47	Jerusalem 47	Kudus 47	القدس 47	Ghods 47	قدس 47
3	38	9.5	17.5	Global 5	Globale 38	Mondial 38	Globala 38	Global 7	عالمي 9	Global 5	جهانی 10
4	43	43	40.12	Sionista 43	Sionista 43	Sioniste 43	Zionist 43	Siyonist 20	الصهيوني 43	Zionist 43	صهیونیست 43
5	37	21	22	Iran 37	Iran 32	Iran 25	Iran 21	Iran 21	ايران 10	Iran 10	ایران 20
6	18	13.5	13.5	Slogan 9	Slogan 10	Slogan 12	Slogan 18	Slogan 17	شعار 12	Slogan 15	شعار 15
7	40	23.5	25.87	Palestina 40	Palestina 40	Palestine 29	Palestina 25	Filistin 22	فلسطين 10	Palestine 20	فلسطین 21
8	38	38	29.25	Personas 38	Persone 12	Personnes 38	Manniskor 38	Insanlar 38	الناس 11	People 21	مردم 38
9	43	43	38.87	America 43	America 43	Amerique 43	Amerika 20	Amerika 43	امريكا 43	America 42	آمریکا 33
10	30	19.5	19.5	Inglaterra 9	Inghilterra 9	Angleterre 30	England 30	Ingiltere 9	انجلترا 30	England 30	انگلستان 9
11	27	20	18.87	Londres 5	Londra 27	Londres 12	London 12	Londra 20	لندن 27	London 27	لندن 20
12	19	15	14.37	Israel 10	Israele 10	Israel 10	Israel 19	Israil 18	اسرائيل 15	Israel 15	اسرائیل 18
13	19	12.5	13.75	Hezdola 13	Hezbollah 10	Hezbollah 12	Hizbollah 19	Hizbullah 17	حزب الله 12	Hizbullah 12	حزب الله 15
14	39	13.5	17.37	Terroristas 11	Terroristi 10	Terrorists 12	Terrorister 25	Teroristler 39	الارهابيين 15	Terrorist 15	تروریست 12

Table 2. The results of typical TFIDF algorithms, thick words are mistaken for keywords



Spanish		Italian		French		Sweden		Turkish		Arabic		English		Farsi		Right keywords
50	March	47	Ghods	47	Ghods	50	March	50	March	50	March	50	March	50	March	March
47	Ghods	43	Zionist	43	Zionist	47	Ghods	47	Ghods	47	Ghods	47	Ghods	47	Ghods	Ghods
43	Zionist	43	America	43	America	43	Zionist	43	America	43	Zionist	43	Zionist	43	Zionist	Global
43	America	40	Palestine	38	People	38	People	39	Terrorist	43	America	42	America	38	People	Zionist
40	Palestine	38	Global	38	Global	38	Global	38	People	30	England	30	England	33	America	Iran
38	People	32	Iran	30	England	30	England	22	Palestine	27	London	27	London	21	Palestine	People
37	Iran	27	London	29	Palestine	25	Terrorist	21	Iran	15	Israel	21	People	20	Iran	Palestine

Table 3. Results of the proposed algorithm, TFIDF improved thick words are words that are mistakenly identified as keywords

Maximum Method (Selected Keywords for Any 8 Languages)		Medium Method (Selected Keywords for All 8 Languages)		Average method (Selected keywords for all 8 languages)		Right keywords
50	March	50	March	47	Ghods	March
47	Ghods	47	Ghods	40.12	Zionist	Ghods
43	Zionist	43	Zionist	38.87	America	Global
43	America	43	America	36.87	March	Zionist
40	Palestine	38	People	29.25	People	Iran
29	Terrorist	23.5	Palestine	25.87	Palestine	People
38	People	21	Iran	22	Iran	Palestine

5.2. The comparison of the obtained results with the other related algorithms

To evaluate the efficiency and performance, the rate of accuracy of the proposed algorithm is compared with that of the other methods. The algorithm was tested with 200 texts in eight languages, which are shown in Table 1, and 1200 correct keywords were achieved. The rate of accuracy of the conventional TF-IDF algorithm for 1014 correct words and 1672 obtained keywords is 60.6%, while the proposed algorithm, namely the mean. TF-IDF, for 1164 correct words of 1275 words, the rate is 91.3%. In the proposed algorithm with the median method, 1092 correct words of 1456 words indicate the rate of 75%. Moreover, if we calculate the accuracy rate for the maximum method, 1021 correct words of 1531 words by the accuracy rate of 66.6% is obtained. The rate of accuracy for graph-based algorithm [27] for these data is 80%. Concerning the obtained rates, mean with the accuracy rate of 91.3% is the best method. Table 4 shows the summary of results on BBC data. This suggests that the proposed algorithm not only extracted the keywords language independent, but has achieved a considerably better results. Table 5 shows comparison the algorithm with other related algorithms.

Table 4. Suggested algorithm accuracy rates and keywords extraction algorithms on BBC data

Algorithm	TF-IDF Maximum suggestion	TF-IDF Suggestion middleware	TF-IDF Suggested average	Graph [27]	TF-IDF normal
Accuracy rate	66.6%	75%	**91.3%**	80%	60.6%

Table 5. Comparing the algorithm with other related algorithms

Algoritm	Accuracy
The Proposed Algorithm	91.3%
Graph[27]	80%
Kp[28]	47.7%
MSF [29]	60%
GATE[30]	64.4%
Habibi[1]	75%
Single-Document[31]	83.2%

6. CONCLUSION

Data retrieval is widely applied in everyday life. Increasing the efficiency and performance of information retrieval systems is very important for their designers. We realized based on investigating the applications of the data retrieval and text mining that the keywords of a text are important and facilitate the oriantations of the processes. For example, by finding the keywords in the news or some sentences with more keywords, we could summarize or comprehend the text more easily. To achieve to this aim, an unsupervised keywords extraction algorithm is proposed based on improving the TF-IDF algorithm for multi-language texts. In the proposed algorithm, the average TF-IDF of the candidate words were calculated for the same and the other languages. Then the words with the higher averages TF-IDF were chosen as the extracted keywords. A database, which was collected 200 news from BBC website in various languages, was considered to evaluate the efficiency of the proposed algorithm. The experimental results show that the selected keywords are more similar to the mentioned keywords by the website and this confirms the reliability of the algorithm. The overall accuracy rate of the algorithm is 91.3% that it is higher than the state-of-the-art keyword extraction algorithms. We would like to introduce three strategies as our future works, to improve the proposed algorithm in application, complexity and time. Finding complex keywords could be added to the algorithm, real-time and on-line behaviour could be created by focusing on parallel processing and normalizing the feeds' addresses could be considered to facilitate access.

REFERENCES

[1] M. Habibi and A. Popescu-Belis, "Keyword extraction and clustering for document recommendation in conversations," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 23, no. 4, pp. 746-759, 2015.

[2] M. Savić, et al., "A language-independent approach to the extraction of dependencies between source code entities," *Information and Software Technology*, vol. 56, no. 10, pp. 1268-1288, 2014.

[3] S. Siddiqi and A. Sharan, "Keyword and keyphrase extraction techniques: a literature review," *International Journal of Computer Applications*, vol. 109, no. 2, pp. 18-23, 2015.

[4] T. S. Chung, et al., "A survey of flash translation layer," *Journal of Systems Architecture*, vol. 55, no. 5-6, pp. 332-343, 2009.

[5] N. I. Abdulkhaleq, et al., "Improving the data recovery for short length LT codes," *International Journal of Electrical & Computer Engineering*, vol. 10, no. 2, pp. 1972-1979, 2020.

[6] N. N. Kulkarni and S. A. Jain, "Checking integrity of data and recovery in the cloud environment," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 2, pp. 626-633, 2019.

[7] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48-57, 2014.

[8] V. Jain and S. V. A. V. Prasad, "Ontology based information retrieval model in semantic web: a review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 8, pp. 837-842, 2014.

[9] K. Kim, et al., "Language independent semantic kernels for short-text classification," *Expert Systems with Applications*, vol. 41, no. 2, pp. 735-743, 2014.

[10] D. Deshwal, et al., "Feature Extraction Methods in Language Identification: A Survey," *Wireless Personal Communications*, vol. 107, pp. 2071-2103, 2019.

[11] S. K. Bharti and K. S. Babu, "Automatic keyword extraction for text summarization: A survey," *arXiv preprint arXiv:1704.03242*, 2017.

[12] E. Ferrara, et al., "Web data extraction, applications and techniques: A survey," *Knowledge-based systems*, vol. 70, pp. 301-323, 2014.

[13] P. Turney, "Learning to extract keyphrases from text," *National Research Council Canada*, 2002.

[14] S. S. Hong, et al., "The feature selection method based on genetic algorithm for efficient of text clustering and text classification," *International Journal of Advances in Soft Computing and its Applications*, vol. 7, no. 1, pp. 22-40, 2015.

[15] E. Frank, et al., "Domain-specific keyphrase extraction," in *16th International joint conference on artificial intelligence (IJCAI 99)*, vol. 2, pp. 668-673, 1999.

[16] C. Caragea, et al., "Citation-enhanced keyphrase extraction from research papers: A supervised approach," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1435-1446, 2014.

[17] G. Ercan and I. Cicekli, "Using lexical chains for keyword extraction," *Information Processing & Management*, vol. 43, no. 6, pp. 1705-1714, 2007.

[18] F. Fkih and M. N. Omri, "Complex terminology extraction model from unstructured web text based linguistic and statistical knowledge," *International Journal of Information Retrieval Research*, vol. 2, no. 3, pp. 1-18, 2013.

[19] G. Figueroa, et al., "RankUp: Enhancing graph-based keyphrase extraction methods with error-feedback propagation," *Computer Speech & Language*, vol. 47, pp. 112-131, 2018.

[20] S. Lahiri, et al., "Keyword and keyphrase extraction using centrality measures on collocation networks," *arXiv preprint arXiv:1401.6571*, 2014.

[21] P. Tonella, et al., "Using keyword extraction for web site clustering," in *Fifth IEEE International Workshop on Web Site Evolution, 2003. Theme: Architecture. Proceedings*, pp. 41-48, 2003.

[22] S. K. Biswas, et al., "A graph based keyword extraction model using collective node weight," *Expert Systems with Applications*, vol. 97, pp. 51-59, 2018.

[23] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404-411, 2004.

[24] S. Duari and V. Bhatnagar, "sCAKE: Semantic Connectivity Aware Keyword Extraction," *Journal of Information Sciences*, vol. 477, pp. 100-117, 2019.

[25] K. S. Hasan and V. Ng, "Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 365-373, 2010.

[26] R. Farhad, et al., "Improved Clustering Persian Text Based on Keyword Using Linguistic and Thesaurus Knowledge," *Signal and Data Processing*, vol. 13, no. 1, pp. 87-100, 2016.

[27] A. R. Nabhan and K. Shaalan, "Keyword identification using text graphlet patterns," in *International Conference on Applications of Natural Language to Information Systems*, pp. 152-161, 2016.

[28] C. B. Ali, et al., "A two-level keyphrase extraction approach," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 390-401, 2015.

[29] D. Y. Lee, et al., "A New Extraction Algorithm for Hierarchical Keyword Using Text Social Network," in *Information Science and Applications (ICISA) 2016*, pp. 903-912, 2016.

[30] P. Nesi, et al., "A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction From Web Pages and Documents," in *21st International Conference on Distributed Multimedia Systems (DMS 2015)*, pp. 1-7, 2015.

[31] F. Rousseau and M. Vazirgiannis, "Main core retention on graph-of-words for single-document keyword extraction," in *European Conference on Information Retrieval*, pp. 382-393, 2015.

BIOGRAPHIES OF AUTHORS

Bahare Hashemzade is lecturing and researching at Electrical and Computer Engineering, University of Torbat Heydarieh from 2016. She has graduated at M.Sc. of Information Science from Birjand Uniersity in 2015. Her interest fields are information technology, obfuscation and data mining.

Majid Abdolrazzagh-Nezhad is lecturing and researching as assistant professor at the computer engineering department of the University of Bozorgmehr Qaenat since 2013 and dean the department since 2016. He was dean of the faculty of computer science between 2013 to 2016. Abdolrazzagh-Nezhad Supports Master and PhD students of Islamic Azad University of Birjand since 2016. He has graduated at PhD in Computer Science from Information Science and Technology, Faculty of the National University of Malaysia (UKM) in 2013. Also, he received his master degree of Operation Research from University of Sistan and Blochstan in 2007, and his bachelor's degree from University of Birjand in 2004. His interest fields are artificial intelligent, optimization, data mining, scheduling and uncertain systems. He is a young Professionals member of the Institute of Electrical and Electronics Engineering (IEEE) and reviewer of valid journals such as Information Sciences, Applied Soft Computing, Soft Computing, International Journal of Production Research, IEEE Transaction on Industrial Electronics and IEEE Transaction on Industrial Informatics.