Suggesting new words to extract keywords from title and abstract

Hadeel Qasem Gheni, Ahmed Mohammed Hussein, Wed Kadhim Oleiwi
Department of Computer, College of Science for Women, University of Babylon, Iraq

Article Info

Article history:

Received Dec 22, 2018
Revised Apr 22, 2019
Accepted May 6, 2019

Keywords:

Information retrieval
Keyword extraction
Natural language processing

ABSTRACT

When talking about the fundamentals of writing research papers, we find that keywords are still present in most research papers, but that does not mean that they exist in all of them, we can find papers that do not contain keywords. Keywords are those words or phrases that accurately reflect the content of the research paper. Keywords are an exact abbreviation of what the research carries in its content. The right keywords may increase the chance of finding the article or research paper and chances of reaching more people who should reach them. The importance of keywords and the essence of the research and address is mainly to attract these highly specialized and highly influential writers in their fields and who specialize in reading what holds the appropriate characteristics but they do not read and cannot read everything. In this paper, we extract new keywords by suggesting a set of words, these words were suggested according to the many mentioned in the researches with multiple disciplines in the field of computer. In our system, we take a number of words (as many as specified in the program) that come before the proposed words and consider it as new keywords. This system proved to be effective in finding keywords that correspond to some extent with the keywords developed by the author in his research.

Corresponding Author:

Hadeel Qasem Gheni
Department of Computer,
Babylon University,
Hilla, Babil, Iraq.
Email: hadeelqasem84@gmail.com

1. INTRODUCTION

Due to the great development in an online publishing [1] and availability of electronic books and websites [2] the retrieval of information can achieved quickly and easily using search engines [3] where the vast amount of information is accessible to people through the Internet [4]. This information is available as database, document, or multimedia format and access to this information governed by the availability of an appropriate search engine [5]. Newly, many documents are electronically available and you can easily choose any document you want to read or to know the relationship between the documents by extracting the suitable keywords [6]. Keywords are words used by the users of search engine to get what they want from information and research, so they are terms that given to the words that indicate the content of the subject [7]. Keywords have been everywhere in our daily lives, from searching for the information we need on the web via search engines to online ads that match the content we're currently browsing [8]. Unfortunately, many documents do not contain keywords [9]. People do not have enough time to read the entire research; the best for them is a brief reading of it like the abstract instead of the entire text [10]. Keywords can be an intensive summary of the document, developed the retrieval of information, or become an entry for the document set [11].

Keywords are important and meaningful words in the document, which give an accurate overview of their content and reflect the author's intention to write [12]. Hence, these words are all for the writer of the article and should focus on the selection and identification to increase the interest of visitors to read articles and compete with other articles in all search engines. In the extraction of pertinent information, keywords have a very significant role where the summary of the entire document content can extracted via a few keywords [13]. It may consist of one or two words that have a general meaning [14], therefore called short keywords, or may consist of short sentences that have an average length and be more specialized, or may be long sentences.

Keywords extraction is to extract several prominent words from a particular text and use these words to represent the entire text [15], it is an automatic process to define a group of terms, which represent the information in a document that have been discussed [16]. Extracting of keywords is an essential task and it is an important research trend in text mining, information retrieval, and natural language processing [17], it let us to represent text document in an intensive way [18]. Keywords usually extracted by extracting the relevant higher frequency words from others, with emphasis on important words [19]. The algorithm of keyword Extraction has great capabilities in summarizing the entire document [20].

Extraction of keywords is the basis of information retrieval process and numerous of techniques have been proposed to address this problem [21], [10] suggest a form that extracts keywords from the title and abstract by constructing a list of words arranged in descending order depending on the number of their appearance in abstract and title. [22] Embodiments can include word analysis in an individual document by stop words, delimiters, or both to identify candidate keywords. Then, for each word, word scores are count in the candidate keywords based on the frequency function. [23] Tested a set of central metrics on words and networks compile the noun sentences and analysis of the achievement on four standard databases. [24] Build approach with the scale of a new network - node selectivity, driven by the research of a centralized approach based on the graph. Node selectivity is define as the distribution of average weight on single node connections; they extract nodes (keyword candidates) based on the value of selectivity. In addition, they expand the extracted nodes into word groups with the highest in / out selective values. [25] Suggesting an algorithm depending on an end-to-end neural keywords extraction by using the network of Siamese LSTM, removing the need to engineer manual features.

2. PROPOSED METHOD

In general, the basic idea is to propose a set of words expected to be mention or always come in the title and abstract, several areas of computer applications were consider and identify words that are extremely used in research, this word are clarified in Table 1.

Table 1. The suggested keywords

Suggested Keywords					
Networks	processing	extraction	detection	expansion	method
tool	Segmentation	way	model	system	steps
Approach	localization	Scheme	Technique	diagram	Flowchart
Framework	Compression	Cryptography	Processor	stitching	texts
Circuit	steganography	Security	Structure	Paradigm	features
Algorithm	discrimination	privacy	exposure	chains	Wireless
environment	Communications	technology	applications	patterns	analysis

The proposed system consists of several steps: (1) Tokenization, (2) Stop word removing, (3) Search about the suggested words in the title and abstract and take (N) words that comes before it, (4) Find the similarity between the extracted keywords and keywords mentioned by the author in the research. Figure 1 illustrates the mechanism of work of the proposed system.

The first step is to divide the paragraphs of title and abstract into single words by the tokenization process, then, delete the stop words in the second step and convert all words to lower case to minimize the comparison time. After that, we will search in the title and abstract paragraphs about the proposed words, and if they found, the words that come before them will be taken in amount of (N) specified in the program. (N) Represents the number of words will be truncate from the sentence that comes before the words that suggested. Then analyzing the results and find the similarity between the extracted keywords and the keywords developed by the author. The proposed system shown in Figure 1. Figure 2 explain our algorithm that demonstrate the work of the proposed system.

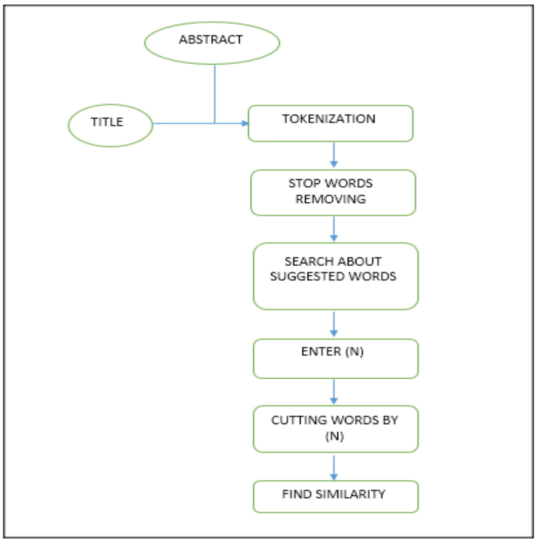
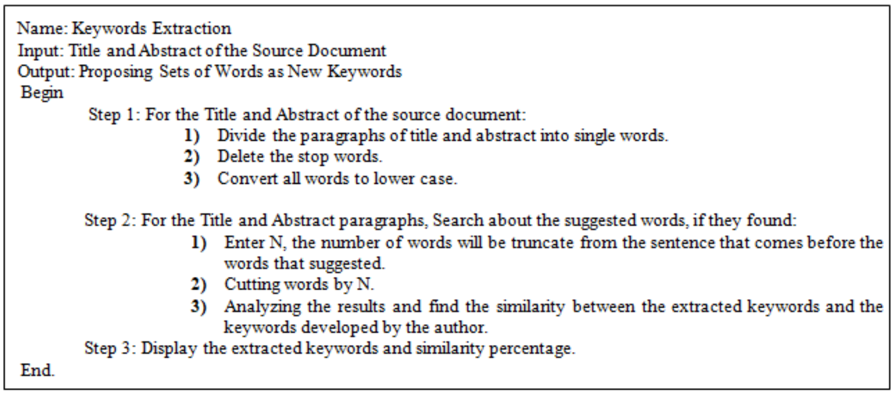


Figure 1. The proposed system



```
Name: Keywords Extraction
Input: Title and Abstract of the Source Document
Output: Proposing Sets of Words as New Keywords
Begin
    Step 1: For the Title and Abstract of the source document:
        1) Divide the paragraphs of title and abstract into single words.
        2) Delete the stop words.
        3) Convert all words to lower case.

    Step 2: For the Title and Abstract paragraphs, Search about the suggested words, if they found:
        1) Enter N, the number of words will be truncate from the sentence that comes before the
           words that suggested.
        2) Cutting words by N.
        3) Analyzing the results and find the similarity between the extracted keywords and the
           keywords developed by the author.
    Step 3: Display the extracted keywords and similarity percentage.
End.
```

Figure 2. The algorithm

3. RESULTS AND ANALYSIS

Database built to store researches that will tested by our system. This database consists of three fields, namely: the first field to store the title of research, the second field to store the abstract of research and the third field is to store the keywords written by the author of the research. About 100 scientific researches were stored in the database for testing; these researches randomly selected from publications. When searching for and finding the words we want, we will determine the number of words to be truncate before the proposed words, and this is what (N) expressed.

Three values were tested for (N) and these values are when (N) is one, two and three. When the value of (N) is one, it means truncating one word before the proposed word that found in the title or abstract, and when the value of (N) is two, it means truncating two words, and therefore when the value of (N) is three, it means truncating three words. When truncates the words by (N), these words will be a sentence with the proposed word according to the sequence in the paragraph and thus we will have a short sentence representing the new keyword.

The results showed that there is 78% matching between the extracted keywords and keywords developed by the author in his research when choosing the value of (N) equal to one, where the word taken by (N) and the proposed word are formed a short sentence of two words.

While when selecting the value of (N) equal to two, there is 54% matching between the extracted keywords and keywords developed by the author, where the words taken by (N) and the proposed word are formed a short sentence of three words. Finally, when selecting the value of (N) equal to three, there is 23% matching between the extracted keywords and keywords developed by the author, where the words taken by (N) and the proposed word are formed a sentence of four words. Table 2 shows the matching ratio of (N) values. The results obtained in Table 3 show that there is a significant correlation between the keywords that were drawn with the keywords developed by the author in his research.

Table 2. Matching ratios

	Extracted Keywords Where		
Author Keywords	N=1	N=2	N=3
	78%	54%	23%

Table 3. Matching results

Total Keywords	Total Matches	Partial Matches	Failures
552	53	331	168
Ratio	9.6 %	59.9 %	30.4 %

4. CONCLUSION

In general, as shown in Tables 1-3, 100 research papers containing keywords were tested. The results showed that there were 82 studies found a match between its keywords and the proposed keywords and that there were 18 studies that failed to match. More specifically, 552 keywords were tested and found 53 total matches, 331 partial matches and 168 match failures. As statistical numbers, for 100 abstract and 100 title there are 82% match and there are 18% failure. For 552 keywords there are 9.6% total match, which is mean 53 study. There are 59.9% partial match, which is mean 331 study. There are 30.4% failure match, which is mean 168 study.

REFERENCES

[1] C. Jiang, *et al*., "Implicit semantics based metadata extraction and matching of scholarly documents," *Journal of Database Management (JDM,)* vol. 29, no. 2, pp. 1-22, 2018.

[2] E.S. Al-Shamery and H.Q. Gheni, "Plagiarism detection using semantic analysis," *Indian Journal of Science and Technology*, vol. 9, no. 1, 2016.

[3] S. Niwattanakul, *et al*, "Using of Jaccard coefficient for keywords similarity," *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1-6, 2013.

[4] A M. Hussein, "Paraphrasing: Solution to the Problem of Plagiarism," *Asian Journal of Information Technology*, vol. 15-16, pp. 2743-2746, 2016.

[5] M. Habibi, 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.

[6] 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, no. 01, pp. 157-169, 2004.

[7] G. K. Palshikar, "Keyword extraction from a single document using centrality measures," *International Conference on Pattern Recognition and Machine Intelligence*, Springer, Berlin, Heidelberg, 2007.

[8] F. Rousseau and V. Michalis, "Main core retention on graph-of-words for single-document keyword extraction," *European Conference on Information Retrieval*. Springer, Cham, 2015.

[9] C. Zhang, "Automatic keyword extraction from documents using conditional random fields", *Journal of Computational Information Systems*, vol. 4-3, pp. 1169-1180, 2008.

[10] Y. HaCohen-Kerner, "Automatic extraction of keywords from abstracts," *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, Berlin, Heidelberg, 2003.

[11] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," *Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics*, 2003.

[12] L. Yang, *et al*., "A new network model for extracting text keywords," *Scientometrics*, vol. 116, no.1, pp. 339-361, 2018.

[13] Sh. Luthra, *et al*, "A Statistical Approach of Keyword Extraction for Efficient Retrieval," *International Journal of Computer Applications*, vol. 168-7, 2017.

[14] R. C. Anju, *et al*., "Keyphrase and Relation Extraction from Scientific Publications," *Advances in Machine Learning and Data Science, Springer,* Singapore, pp. 113-120, 2018.

[15] X. Wan, *et al*., "Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction," *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007.

[16] A. Ruhul, and M. Chakraborty, "Algorithm for Bengali Keyword Extraction," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE*, 2018.

[17] A. Tixier, *et al*, "A graph degeneracy-based approach to keyword extraction," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[18] A. Onan, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232-247, 2016.

[19] R. Naidu, *et al*., "Text summarization with automatic keyword extraction in telugu e-newspapers," *Smart Computing and Informatics, Springer*, Singapore, pp. 555-564, 2018.
[20] X. Wu, *et al*., "A visual attention-based keyword extraction for document classification," *Multimedia Tools and Applications*, vol. 77, no. 19, pp.25355-25367, 2018.
[21] Y. Chen, *et al*., "Single document keyword extraction via quantifying higher-order structural features of word co-occurrence graph," *Computer Speech & Language*, 2019.
[22] S. Rose, *et al*, "Rapid automatic keyword extraction for information retrieval and analysis," U.S. Patent No. 8,131,735. 6 Mar 2012.
[23] Sh. Lahiri, "Keyword and keyphrase extraction using centrality measures on collocation networks," arXiv preprint arXiv: 1401.6571, 2014.
[24] S. Beliga, "Toward selectivity based keyword extraction for Croatian news," arXiv preprint arXiv: 1407.4723, 2014.
[25] J. Villmow, *et al*., "Automatic keyphrase extraction using recurrent neural networks," *International Conference on Machine Learning and Data Mining in Pattern Recognition, Springer*, Cham, 2018.

BIOGRAPHIES OF AUTHORS

Hadeel Qasem Gheni obtained a Bachelor's Degree in Computer Science from the University of Babylon - Faculty of Science for Women's - Computer Department in 2006, and then got a Master's Degree in Artificial Intelligence from the University of Babylon - Faculty of Information Technology - Software Department in 2016. Do as assistence lecturer at the University of Babylon/College of Science for Women/Computer Department Since 2006 until now.

Ahmed Mohammed Hussein, Bachelor of Computer Science in 2004 from the University of Babylon/Faculty of Science, Master of IT/university of Tenaga/Malaysia in 2012, and do as lecturer at the University of Babylon/College of Science for Women/Computer Department Since 2006 till now.

Wed Kadhim Oleiwi got a Bachelor Degree of Computer Science from the University of Babylon/College of Science/Department of Computer in 2005, then earned a Master's Degree from the University of Babylon/College of Science/Department of Computer in 2012 in the field of Artificial Intelligence, and do as lecturer at the University of Babylon/College of Science for Women/Computer Department Since 2007 till now.