Proposing a Keyword Extraction Scheme based on Standard Deviation, Frequency and Conceptual Relation of the Words

Shadi Masaeli
Dept. of computer science and engineering, School of electrical and computer engineering, Shiraz University, Shiraz, Iran

Reza Boostani
Dept. of computer science and engineering, School of electrical and computer engineering, Shiraz University, Shiraz, Iran

Seyed Mostafa Fakhrahmad[*]
Dept. of computer science and engineering, School of electrical and computer engineering, Shiraz University, Shiraz, Iran

Betsabeh Tanoori
Dept. of computer science and engineering, School of electrical and computer engineering, Shiraz University, Shiraz, Iran

***Abstract*—At each text there are a few keywords which provide important information about the content of that text. Since this limited set of words (keywords) is supposed to describe the total concept of a text (e.g. article, book), the correct choosing of keywords for a text plays an important role in the right representing of that text. Despite several efforts in this field, none of the so far published methods is accurate enough to elicit representative words for retrieving a vast variety of different texts. In this study, an unsupervised scheme is proposed which is independent on domain, language, structure and length of a text. The proposed method uses the words' frequency in conjunction with standard deviation of occurred location of words in text along with considering the conceptual relation of words. In the next stage, a secondary score is given to those selected keywords by the statistical criterion of TFISF in order to improve the basis method of TFIDF. Moreover, the proposed hybrid method does not remove the stopwords since they might be a part of bigram keywords while the similar approaches remove all stopwords at their first stage. Experimental results on the known SEMEVAL dataset imply the superiority of the proposed method in comparison with state-of-the-art schemes in terms of F-score and accuracy. Therefore, the introduced hybrid method can be considered as an alternative scheme for accurate keyword extraction.**

Keywords—Keyword extraction; key-phrase extraction; TFISF; standard deviation; frequency

I. Introduction

Since a tremendous of texts in the form of book, scientific paper, news, technical reports are daily added to the internet, researchers in big databases develop automatic methods for analyzing the texts and finding semantic relations between them. To search a desired topic among a huge number of texts, brute force search does not work and to diminish the size of each text into some representative words, the idea of keyword extraction is emerged [1]. Keywords are one or a short consequence of limited words that represent a text [2,3]. In some research fields like natural language processing (NLP), there is a serious need to investigate a huge number of texts. Therefore, by providing a set of keywords as indicators of each text, this investigation can be remarkably eased, especially when a text is searched according to its keywords [4].

In addition, there are some text processing applications that their methods contain a similar trend to the keyword extraction techniques. These applications include: automatic text summarization [5,6], information retrieval [7], text classification [8], fast and accurate searching of texts in web [9] and automatic indexing [10,3]. Keyword extraction methods are the basis schemes for all of these applications and the right choosing of essential keywords is the main purpose of this research. Reading keywords of a text can help the query to choose or reject of that text. In fact, keyword checking can be an effective way to find a relative text where the searched keywords are enough similar to the keywords of the corresponding text. Therefore keyword extraction can help the query in learning how to arrange correct keywords in the search engines to find his favorite document quickly.

Keyword extraction is usually performed in two stages. At the first stage, the text is preprocessed using heuristic rules and some words are selected as the keyword candidates. One of the most applicable heuristic rules is removing the stopwords since these words are repeatedly scattered throughout the text and the conventional keyword extraction methods select them as the best candidates while they carry no I nformation about the text's concept. Some methods use a list of stopwords and remove every similar stopword in the text. In some other methods, those words with high frequencies throughout the text are removed.

In the second stage, the methods are divided into two categories including supervised and unsupervised schemes. In both categories a dataset with labels (keywords) is required in order to assess the results. One of the problems in supervised approaches is requiring a training phase while in unsupervised

[*] The corresponding author

schemes the train is carried out through the stages of algorithms and there is no need to a separate train phase. In addition, supervised methods have usually more computational complexity rather than the unsupervised ones. One of the basic methods employed in both categories is the statistical methods which benefits from the simplicity and low computational complexity. One of the famous statistical methods is TFIDF (Term Frequency-Inverse Document Frequency) which provides suitable results in different applications [11]. Nevertheless, TDIDF suffers from high dependency to the length and the size of the corpus; consequently, researchers made a lot of attempts to overcome its drawback by combining it with other basic methods [12].

II. Related Work

In this regard, Witten et al. [13] developed a supervised method by combining TFIDF [14] with the first occurrence location factor in the text and called this method as KEA and applied it to 75 journal papers. In this algorithm, root location is adopted as the preprocessing method and in the learning phase, Naïve Bayes classifier is employed and finally the recall measure is utilized for the assessment. The accuracy of this method is assessed based on Human Judi and is highly dependent to the size of the train set. Nevertheless, this method could be assessed by a more robust measure. This method is also employed for text abstraction, searching through the web and classification of different texts is utilized.

Hulth [15] suggested a supervised method for the keyword extraction just from the abstract [16] parts of scientific papers using INSPEC dataset. He combined the statistical measure with the linguistic knowledge in order to remove the dependency of the frequent words to both the length of document and number of documents. Nevertheless, it still suffers from the dependency to the structure of the text. Their performance in terms of F-measure provides 33.9, while its main deficiency is to estimate the correct label for the dataset in the training phase. To compensate this defect, it is possible to use the knowledge of finding the relevance among the words and assign label to them accordingly. They removed the stopwords according to their high frequency. They also incorporate the concept of each sentence containing a candidate keyword for giving a score to each keyword. This is done by a graph based method for giving a score to the candidate keywords according to the type of the graph and the semantic role of that keyword in the text. Finally the keywords were sorted according to their scores. They could get a considerable improvement in terms of F-measure by incorporating the sentence information for the keyword extraction. Nevertheless, this scheme highly biased the keyword to the importance of the sentence and if they used TFIDF, their results could be more robust to the sentences.

In another attempt, Zahedi et al. [17] for improving the better accessibility to the web content using keywords, focused on the search engines for ranking the keywords. They used the retrieved information results as the training data and for improving their supervised method, genetic algorithm is deployed to optimize the features. The preprocessing of the text contained two steps where in the first stage, the unification is done and then the stopwords were eliminated. The key point in this algorithm is to determine a threshold such that 60% of predefined keywords by the authors of that text considered as the candidate keywords. In this scheme TFIDF is used and assessed by F-score (40.82) on Farsi websites. Their results outperformed the other compared methods on their dataset, though their method suffered from the problem of supervised approaches and also imposes lots of computational burden.

Sharon et al. [18] applied a different method in construction of a semantic graph for each text in an unsupervised manner. They used word net for determining the conceptual links using the GEPHI tool for graph demonstration. Each node takes a score according to the sorted rank of keywords using human judgment. They applied their method to two different datasets and demonstrated keywords are those having a high conceptual value despite TFIDF that considers keywords as high frequency words in a text. Unlike TFIDF, this approach does not need a lot of texts.

Lu et al. adopted an unsupervised scheme for keyword extraction that uses literature references within the framework of word co-occurrence and topic distribution. They applied their method to ACM digital library and assessed their method by F-score resulted in 0.276. It is has been demonstrated that using references could improve the performance of each method compared to the situation that the references is not incorporated.

Das et al. [12] developed an unsupervised method based on collocation and fuzzy set theory for keyword extraction to handle the ambiguity of high occurrence rate of words in a text [19]. This method does not need any corpus and can be applied on a single document. Moreover, this method could solve the dependency of TFIDF to the size of documents. This method is able to elicit both keywords and key-phrases. They applied their method to the electronic documents which contains scientific articles on different issues which are accessible by different sources like Wikipedia. For the exactness of the elicited keywords, the precision is determined up to 95% accuracy. This high precision rate can be originated from applying a filtering in the preprocessing stage.

Siddiqi et al. [20] utilized an unsupervised scheme for keyword extraction which I independent to the length of corpus, its domain and type of language. Their method helped TFIDF to consider both the frequency of words and spatial distribution of the words. In this method, for each word a spatial sequence is generated which demonstrate the places of that word in a document. By this distribution, stopwords can be detected since they are regularly distributed throughout the text while key words do not have a specified spatial pattern. Each word is given a weight according to its spatial distribution and this weighting is determined such that the stopwords are automatically takes very low value. After eliciting the keywords according to the mentioned weight mechanism the precision is determined. They applied this method to an Indian book and achieved 0.8 precision. Since this method uses a statistical criterion, it needs a big document to be able to correctly find the keywords.

Yang et al. [21] suggested an unsupervised graph based scheme for keyword extraction. They pay attention to each sentence during keyword extraction. In their method, the importance of each word within a sentence is measured and these words are ranked according to their importance in an iterative manner throughout a document. The ranking of keywords is determined according to their ranking of their corresponding nodes in the graph of that document. One advantage of this method is its applicability on just one document which overcomes to the TDIDF drawback and decreases its dependency to the size of corpus. They have applied their scheme to the WEB TEXT 13702 dataset and compared their result with the text ranking algorithm and demonstrated that they have got a better F-score (25.2%) in comparison with text ranking method.

In this study, we propose a new method based on multi-factorial features, which can quickly be extracted from each single document. In fact, the proposed method is an extension of the method introduced by Siddiqi et al. [20] by being added a language model for generating better results. We have applied the proposed method to the SEMVAL dataset and compared our method to the base method.

The rest of this paper is organized as follows. Section 2 is devoted to introduction of the proposed method along with the base method [20]. Section 3 presents the experimental results. Finally, section 4 concludes the paper and gives an outline to clarify the horizon of this study in the future.

III. Methods and Materials

In this section, the method proposed by Siddiqi et al. [20] and the scheme proposed here are presented. Afterward, the employed datasets in this study are introduced and their features are explained. In accordance with the limitation of supervised methods, most of the conventional algorithms in this field are unsupervised whose goal is the independency to the type of language, topic, length and structure of the document while preserving the accuracy. Eliciting of the correct representative keywords remarkably help a search engine to find its required document in big data bases at a glance.

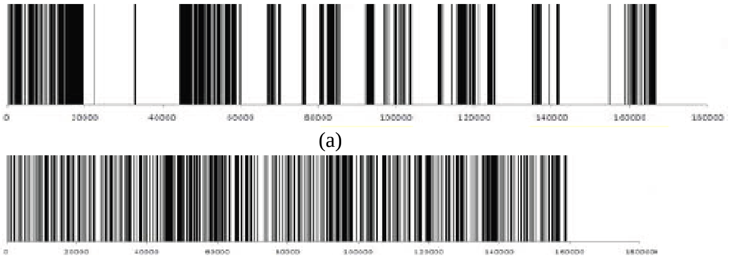
One of the research approaches which provide logical results, published by Siddiqi et al. [20]. The used the concept of standard deviation over a difference sequence for each word. Their method was dependent to the length of data since statistical criterion for limited numbers of occurrences cannot provide an acceptable index. This method is able to significantly differentiate keywords and stopwords, as shown in Fig. 1 (a) and (b), respectively.

In order to compare and demonstrate the spatial occurrences of a real keyword and stopword, two diagrams in the form of barcode are shown in Figure 1. The horizontal axis in each diagram presents the spatial length of the document, while the vertical axis shows the occurrence of that word. As we see in this figure, the occurrence place of a keyword is much irregular than that of a stopword. This difference rises from this fact that each sentence needs one of more stopwords and therefore its spatial distribution is uniform while a single keyword which carries a part of main concept of a document cannot be appeared in all sentences since there are some other keywords and each part of text is concentrated on a certain concept. Consequently, standard deviation of this barcode graph can be a good indicator to distinguish a stopword from a keyword.

In the Siddiqi's method [20], keywords and key-phrases are extracted in two phases by two algorithms, each of which has similar steps. The candidate keywords are chosen at the first step, then the chosen keywords are assessed and the correct ones are determined. Both the keyword extraction algorithm (SDFKWE) and the modified version of SDFKPE (MSDFKPE) are proposed by Siddiqi et al. [20] use only the standard deviation criterion for extracting the keywords. The results are then used for extracting the key-phrases in the second algorithm. In this study, we have attempted to improve their method by adding additional factors. In the following sections, the proposed method and its steps are illustrated.

A. *Keyword extraction phase*

The first step is to preprocess the input text by eliminating the irrelevant words. In this regard, the sentences are separated according to the separating signs, i.e., punctuations. Then, all of the grammatical signs within each sentence are removed and the text is tokenized for separating single words. The whole text is then stored in the form of sentences captured in arrays. After that, all unique words along with their synonyms whose frequency is lower than a threshold are removed. The most profit of the proposed scheme is that it does not need to remove stopwords at this stage. The second step is carried out for diagnosing the precision of the elicited keywords by calculating two scores for obtaining the Final-Score.



(a)

(b)

Fig. 1. Comparison of the spatial distribution of a) A keyword, b) A stopword throughout a document [20]

Score1:

In the second phase, for each word, two different scores are determined and the final decision is made according to these two scores. The first score is related to the occurrence sequence of each word and its synonyms [22] and the words with the same root and then standard deviation of spatial distribution of each word is determined. To do this, if a word with its synonyms and the same root are placed in the $X_1, X_2, \ldots, X_N$ locations, the differential sequence of occurrence locations as the standard deviation of spatial filter will be determined as equation 1.

$$S_W = \{\ (X_2 - X_1), (X_3 - X_2), (X_4 - X_3), \ldots (X_N - X_{N-1})\ \} \quad (1)$$

The mean of these locations for each word is denoted as μ and is determined as equation 2.

$$\mu = \frac{(X_2 - X_1) + (X_3 - X_2) + \cdots + (X_N - X_{N-1})}{N} = \frac{(X_N - X_1)}{N} \quad (2)$$

After that, by the equation 3 we can determine the standard deviation of difference sequence for each word.

$$\sigma = \sqrt{\frac{\sum((X_{i+1} - X_i) - \mu)^2}{N}} \quad (3)$$

The above formula is adopted from the method proposed by Siddiqi et al. [16]. In order to diminish the dependency of the standard deviation to the repeated number of a word, this squared deviation is divided by the number of repetition of that word. Then the difference deviation values of all words are sorted in a descending order. The top ranked words are selected as the primary candidate set of keywords. In order to finalize the keywords, the second score, here called Score2, needs to be calculated.

SCORE2:

Consider a text about the computer; it is obvious that in such a text the keyword *computer* is repeatedly occurred throughout the text. Hence, this word is not likely to be removed in the former processing stage. In addition, this word takes a high value in Score1. As this general word at all texts about the computer would have the same status; it would be considered as a general word and cannot be a good keyword. For this reason, in addition to the criterion of the word frequency, the number of sentences containing that keyword over the whole number of sentences should also be considered so as to avoid choosing very general words. In fact, a real keyword should be occurred in specific part of a text, not to be appeared in all sentences. General words act like stopwords since they are occurred in all sentences and must be removed. In order to incorporate the sentences containing the candidate keywords, in this work, we have used the TFISF (Term Frequency-Inverse Sentence Frequency) statistical criterion. TFISF is somehow similar to TFIDF in order to normalize the frequency of each word based on the number of corresponding sentences. The TFISF criterion is determined as follows:

Consider the text D with the array of N sentences denoted as D = {S_1, S_2, ..., S_N} where each S_i includes a set of words in the form of w_{1i},w_{2i}, ..., w_{ni}. TFISF is composed of two main elements, i.e., TF and ISF. TF_{ti} is roughly defined as the number of time that the keyword t has occurred in the i^{th} sentence. ISF for a typical keyword t is defined as equation 4.

$$\mathrm{ISF_t} = \log(\frac{\mathrm{N}}{\mathrm{N_t}}) \quad (4)$$

Where N is the whole number of sentences in text D and Nt stands for the number of sentences in text D that contain the word t. Finally, the TFISF for the word t in the ith sentence is obtained from the equation 5.

$$\mathrm{ISF} \times \mathrm{TF} = \mathrm{TFISF} \quad (5)$$

To determine score2 using the concept of TFISF over the retrieved keywords up to the rank of 120 with score1 is performed and the final keyword is determined by forming the equation 6.

$$\text{Final-Score}(W_i) = \text{Alpha} \times \text{score1}(W_i) + (1\text{-Alpha}) \times \text{score2}(W_i) \quad (6)$$

Regarding the achieved value of a word in the score1 and score2 observed that there was a difference in the range of these two scores; consequently, before combining these two scores each of them is normalized in the interval of [0,1]. Next in order to make these two scores comparable the value of ALPHA within the range of [1, … , 0.1, 0.05, 0] is considered to regularize these two score and the summation of this regularization will determine the final score. It is obvious that this regularization parameters should be determined to the cross validation phase.

In order to provide an efficient algorithm for extracting the keyword, we proposed Improved-Standard Deviation Based Keyword Extraction algorithm (Improved-SDFKWE) in which in addition to considering the statistical factors like frequency and standard deviation, in the SDFKWE algorithm, introduced by Siddiqi et al. [20], considering the semantic links between the sentences overall the text.

1) *Proposed method: improved –SDFKWE algorithm*

In this study, an improvement is performed on the efficient algorithm of SDFKWE and equipped with the following factors:

- Generating the unique word list as candidate keyword and arrays of synonym and the same root words for each word separately.
- Determine the frequency of each word with onsidering the frequency of that word and its synonyms and words which have the same root.
- Eliminating the word with the frequency bellow than the threshold.
- Constructing a different sequence of occurrence of the each word in consecutive sentences with considering the location of that word and the corresponding the synonyms words and the words which have the same root.
- Determining the score1 with finding the standard deviation of the sequence SW and normalizing of that with the mean of this sequence and determining score2 with calculating the TFISF for each word by considering its synonyms and those which have the same root and finally determining the Final-Score of each word in the text.
- Now the lists of keywords are ranked according to the Final-Score in an descending order and selecting those which have a higher rank according to a threshold.

B. Key-phrase Extraction Phase

In this algorithm for extracting key-phrase containing two words, denoted as bigram, is considered and the phrases which exceed two words are eliminated. Key-Phrase extraction is performing in two similar steps in which the routines for the functional each word is determined as followed.

In the first step the input text was preprocessed to generate the candidate bigrams in the text. We should first determine these bigrams according to the discriminative sign at each sentence for each unique bigram. At each sentence like Q which contains N words in the form of W_1 W_2 W_3 W_4 … W_N unique bigrams in the form of W_1 W_2 ,W_2 W_3 ,W_3 W_4 ,….. , W_{N-1} W_N are considered and finally an array of consecutive sentences of bigrams are stores afterward the number of occurrence times of each bigram in the text determined and those bigrams which their number of repetition exceed than a threshold are considered and those which are lower than this threshold are eliminated.

The second step is diagnosing the precision of elicited bigrams by calculating two different scores for creating the Final-Score.

The first step to calculate the Final-Score is to determine the score1 of its constructing bigrams elements (Score1: word#1, Score1: word#2) and the next step is to determine the score2 of the bigram's TFISF.

SCORE1:

This score is determined by the summation of score1 of each component in a bigram phrase in the keyword extraction phase as shown in equation 7.

$$\text{SCORE1}(W_i\ W_{i+1}) = \text{SCORE1}(W_i) + \text{SCORE1}(W_{i+1}) \quad (7)$$

Where SCORE1(W_i) is calculated from equation 3.

Then the candidate bigrams according to the score1 are listed in a descending order.

SCORE2:

This score for those statistical samples up to the rank of 120th achieved from the sorting of bigrams resolving score1 is determined. To do this, the TFISF criterion is employed here by this difference that TFISF determination instead of a single word, employ a bigram there for TFISF. In this phase, the number of occurrence time of bigrams in a sentence is determined and compared to the occurrence times of that bigram in all over sentences of the text.

The Final-Score is the summation of score1 and score2 as shown in equation 8, in regularization manner for the bigram in this phase and this score is sorted in a descending order as similar to the base algorithm and finally the top ranks bigrams are selected as the key-phrase candidate.

$$\text{Final-Score}(W_i\ W_{i+1}) = \text{Alpha}\times\text{score1}(W_iW_{i+1}) + (1\text{-Alpha})\times\text{score2}(W_iW_{i+1}) \quad (8)$$

In order to provide an efficient algorithm for extracting the Key-Phrase, we proposed Improved-Modified Standard Deviation Based Key-Phrase Extraction algorithm (Improved-MSDFKPE) in which in addition to considering the statistical factors like frequency and standard deviation, in the MSDFKPE algorithm, introduced by Siddiqi et al. [20], considering the semantic links between the sentences overall the text.

1) Proposed method: Improved –MSDFKPE algorithm

To ease the understanding of the proposed method, it has been clarified in terms of pseudo-code.

- Generating the list of unique bigrams as the key-phrase candidates.
- Determining the frequency of each unique bigrams and eliminating the other bigrams which their occurrence is lower than the predefined threshold.
- Determining score1 for each bigram is determined by the summation of score1 and score2 of that bigrams using the statistical criterion of TFISF and finally the score is descending in a regularized manner.
- Ranking of candidate bigrams according to the Final – Score and choosing the bigrams as the key-phrases.

The execution time of this algorithm with a certain value of Alpha is approximately suitable; therefore, it should be considered similar to the improved-MSDFKPE algorithm that for achieving the best precision, this regularization parameter should be found in an iterative manner.

C. Datasets and Tools

In this section, empirical results of applying the proposed method along with SDFKWE method on two different datasets are presented. In order to find the synonyms and common-root words of each keyword, the Wordnet software is used. Wordnet is written by JAVA at MIT University[1].

As we mentioned before, SEMEVAL and the book titled "on the origin of species" were used as the datasets to evaluate the methods. The mentioned book contains 14 chapters and 222 pages[2].

SEMEVAL is one of the known standard datasets for assessing the keyword extraction and phrase extraction methods[3]. This dataset contains 284 documents in four fields to cover several topics in which by providing train and test sets, in addition to the unsupervised methods, it is possible to apply them to the supervised methods. The labels of this dataset for evaluating each of the supervised and supervised schemes are introduced in three categories. This information contain the labels (keywords) that the authors selected for their documents, the labels that the readers assign to each document and labels which are the combination of labels are assigned by the authors and readers for each document.

[1] JWI 2.4.0-2015-[online] Avalable in: http://projects.csail.mit.edu/jwi. [Accessed 20 September 2016].

[2] Download ebook for the origin of species – 2016- [online] Available in: https://www.goodreads.com/ebooks/download/22463.The_Origin_of_Species. [Accessed 12 September 2016].

[3] GirHub-2016-[online]Avalable in: https://github.com/snkim/AutomaticKeyphraseExtractionR.M.[Accessed 10 may 2016].

IV. Experimental Results

In the preprocessing stage, for executing over each of the above mentioned datasets, the words of each sentence are separated and all the capital letters are converted to lower case letters. The advantage of the proposed scheme is that in its first stage, the stopwords are not eliminated since a part of a bigram keyword can be a stopword [12]. For instance the keyword of "Sun" along with a stopword "the" can make a bigram keyword of "The Sun". Therefore, preserving the stopwords in the first stage of the algorithm is one of the key points while the other conventional methods try to remove all stopwords at the first stage.

In order to preparing the first dataset (the book), we did not apply our method to each chapter separately while we concatenate all chapters and produce a long document. In contrast, the SEMEVAL data set, our proposed method and SDFKWE were separately applied to each document separately.

A. Experiment on Basic Dataset

1) Results Provided by Improved - SDFKWE algorithm

Siddiqi et al. [20] applied their method to this dataset and used precision criterion for evaluating their method. They elicited keywords and compared them by the indexed words of that book, as the real labels. For comparing the proposed method to the introduced rivals, the proposed method is applied to this book and their extracted keywords are compared to those labels which are selected by Siddiqi et al. [20]. The comparison results of the proposed method to SDFKWE are illustrated in terms of precision, Score-1 and the final regularized Score in Table 1.

TABLE. I. Results of the Improved-SDFKWE over Different Number of Top Ranked Words in Terms of Precision Evaluation Measure

method	Top-10	Top-20	Top-30	Top-40
Improved - SDFKWE: Score1	0.2400	0.4000	0.4857	0.5250
Improved - SDFKWE: Final-Score	1	1	0.9667	0.9500
SDFKWE	0.7000	0.6500	0.6000	0.6000
Improvement rate	0.30	0.35	0.37	0.35

As it can be seen in Table 1, precision of the final regularized Score of the proposed method in comparison with the Score-1 gets better and by comparing to SDFKWE, our results are enhanced up to 30%. Keywords in a text are not necessarily those which are highly repeated in all its sentences. As we mentioned before, to find the real occurrence rate of a word, its synonyms and its common-root words are selected. Therefore, is a word is repeated with a low rate, it cannot be considered as a keyword. The threshold for considering a word as a key word is empirically set to 10. The highest accuracy for the Improved-SDFKWE in Table 1, with considering Alpha=0.55, is achieved. In addition, the threshold for the base algorithm, proposed by Siddiqi et al. [16] was set to 10.

2) Results Provided by Improved- MSDFKPE

In this phase, according to the improvement method proposed by Siddiqi et al. [20], the correctness of the bigrams as keywords is assessed. The results of their method and the proposed method in this datasets are brought in Table 2.

TABLE. II. Results Of Different Improved -MSDFKPE over Different Number of Top Ranked Words in Terms of Determines Precision Evaluation Measure

method	Top-10	Top-20	Top-30	Top-40
Improved- MSDFKPE:SCORE1	0.40	0.30	0.26	0.27
Improved - MSDFKPE: Final-Score	0.60	0.55	0.43	0.40
MSDFKPE	0.30	0.30	0.23	0.17
Improvement rate	0.30	0.25	0.20	0.23

As we can see in Table 2, the accuracy of the proposed scheme in terms of final score is improved compared to Score-1. Moreover, our scheme compared to the SDFKWE, at each iteration of retrieved keywords is improved about 20 to 30%. The threshold for eliminating the words with low occurrence rate was set to 5 by Siddiqi et al. [20]. The parameters of our scheme for this comparison were set to 4.88 as the threshold and 0.46 for the regularization parameter of Alpha. The run time of the proposed improved SDFKWE was higher than that of the Improved- MSDFKPE. This difference is raised from calculating the Score-1.

B. Experiment on SEMEVAL Dataset

In order to compare the performance of the proposed algorithm to the conventional schemes, in the keyword extraction phrase, the SEMEVAL dataset is adopted. As far as the most documents in this dataset are with their correct corresponding keywords and key phrases, to customize this dataset for our application, just the single keywords and bigrams were selected and the precision of our method compared to the other ones are made according to single and bigram keywords and the longer key words were all eliminated.

Since the labels of this dataset are brought in the form of root using the porter stemmer, all of the achieved unique words in the preprocessing stage (phase 1), using this rooting algorithm are represented in form of root. Next, the proposed algorithm on each document of the SEMEVAL dataset is executed separately and finally to determine the precision of the algorithm all over the documents, for the top ranked words exceed than micro average, the precision and recall are determined and finally F-Score for the top ranked keywords are determined as follows:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}} \quad \textbf{(8)}$$

Where P and R denote for precision and recall, respectively. It is necessary to note that in the evaluation of keywords extraction performance is carried out in the form of exact match and as far as in the comparison Table 2, the results of algorithm are brought for 5, 10 and 15 top ranked keywords, the performance of proposed algorithm is determined accordingly. The performance of conventional supervised and unsupervised algorithms with respect to the base lines of SEMEVAL, are brought in Table 3 [23].

TABLE. III. PERFORMANCE OF SUBMITTED SYSTEM OVER THE COMBINED ASSIGNED KEYWORDS RANKED BY F-SCORE [23]

System	Rank	TOP-5			TOP-10			TOP-15		
		P	R	F	P	R	F	P	R	F
HUMB	1	39.0	13.3	19.8	32.0	21.8	26.0	27.2	27.8	27.5
WINGNUS	2	40.2	13.7	20.5	30.5	20.8	24.7	24.9	25.5	25.2
KP-Miner	3	36.0	12.3	18.3	28.6	19.5	23.2	24.9	25.5	25.2
SZTERGAK	4	34.2	11.7	17.4	28.5	19.4	23.1	24.8	25.4	25.1
ICL	5	34.4	11.7	17.5	29.2	19.9	23.7	24.6	25.2	24.9
SEERLAB	6	39.0	13.3	19.8	29.7	20.3	24.1	24.1	24.6	24.3
KX-FBK	7	34.2	11.7	17.4	27.0	18.4	21.9	23.6	24.2	23.9
DERIUNLP	8	27.4	9.4	13.9	23.0	15.7	18.7	22.0	22.5	22.3
Maui	9	35.0	11.9	17.8	25.2	17.2	20.4	20.3	20.8	20.6
DFKI	10	29.2	10.0	14.9	23.3	15.9	18.9	20.3	20.7	20.5
BUAP	11	13.6	4.6	6.9	17.6	12.0	14.3	19.0	19.4	19.2
SJTULTLAB	12	30.2	10.3	15.4	22.7	15.5	18.4	18.4	18.8	18.6
UNICE	13	27.4	9.4	13.9	22.4	15.3	18.2	18.3	18.8	18.5
UNPMC	14	18.0	6.1	9.2	19.0	13.0	15.4	18.1	18.6	18.3
JU-CSE	15	28.4	9.7	14.5	21.5	14.7	17.4	17.8	18.2	18.0
Likey	16	29.2	10.0	14.9	21.1	14.4	17.1	16.3	16.7	16.5
UvT	17	24.8	8.5	12.6	18.6	12.7	15.1	14.6	14.9	14.8
POLYU	18	15.6	5.3	7.9	14.6	10.0	11.8	13.9	14.2	14.0
UKP	19	9.4	3.2	4.8	5.9	4.0	4.8	5.3	5.4	5.3

As we in Table 3, the first and second rank methods are supervised while the third rank belongs to the KP-Miner which is an unsupervised. By a short look through this list, we can see that the performances of unsupervised algorithms are approximately similar to the supervised ones in the first 15 top ranked methods. The achieved F-Scores are not impressive since one of the weaknesses of the conventional keywords extraction is its subjectivity. Therefore, reaching to the F-Score 100% in this field is infeasible. On the other hand, in this task, the maximum number of extracted keywords is 15. It should be mentioned that if we set the number of keywords more than 15, the F-Score might be improved but for comparing to the other methods, we had to set the same number to them [23].

According to Table 3, among the unsupervised approaches, KP-Miner [24] provided the best rank; therefore, we have compared our method to KP-Miner on the SEMEVAL dataset and bring the comparative results in Table 4. The threshold of the top ranked keywords is set to 10 to provide the best result. To show the effectiveness of the parameters like the number of top ranked keywords, regularization parameter (Alpha), the proposed algorithm is executed by different values and in those parts that our method outperformed KP-Miner is bolded in Table 4.

The main goal of this research is presenting a method for the evaluation of the amount of relativity of keywords to a text. Regarding the subjectivity of this area, whatever the elicited keywords get near to opinion of reader-assigned, it can be claim that this method could provide higher accuracy. In contrast, whatever the extracted keywords are nearer to the authors' suggested keywords rather to the readers' assigned keywords, it shows the lack of enough accuracy in extracting the keywords. As it is demonstrated in Table 4, the accuracy of the proposed method is better than that of KP-Miner for the 5 top ranked keywords. In retrieving the keywords more than 5, although our method outperforms KP-Miner in terms of recall, its precision could not compete to that of KP-Miner. This is therefore the low number of bigrams in the SEMEVAL labels leading to diminish the F-Score.

TABLE. IV. PERFORMANCE OF KP-MINER & IMPROVED -MSDFKPE ALGORITHM ON P, R AND F INDEXES, GIVEN AS PERCENTAGES

Rank		TOP-5			TOP-10			TOP-15		
Assigned key-phrases	**Method**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Author	**KP-MINER**	19.0	24.6	21.4	13.4	34.6	19.3	10.7	41.6	17.1
	Improved-MSDFKPE	13.26	**27.85**	17.96	9.78	**41.09**	15.80	7.53	**47.48**	13.00
Reader	**KP-MINER**	28.2	11.7	16.5	22.0	18.3	20.0	19.3	24.1	21.5
	Improved-MSDFKPE	22.39	16.37	**18.91**	16.84	**24.64**	20.01	12.89	**28.29**	17.72

V. CONCLUSION AND FUTURE WORK

Keyword extraction for a text is very necessary since most of the search engines find the related documents according to the maximum matching between the searched items and keywords of each text. For this purpose, developing a method for providing high rate of true keywords is of interest. In this paper, a multi aspect algorithm was introduced which considers several factors. The proposed scheme uses the standard deviation of difference sequence of the occurred place for each word, in addition to preserving the synonyms and common-stem words of each keyword candidate while it does not eliminate the stopwords. This unsupervised scheme was executed on two datasets and could outperform state-of-the-art methods for detecting the 5 top ranked keywords. The accuracy of the algorithm is assessed by both reader-assigned and author-assigned approaches and the matching of the elicited keywords to the reader-assigned labels imply on the higher performance of the proposed scheme compared to the rivals.

As a future work, for applying the proposed method to a book with several chapters, using TFISF and TFIDF in keyword extraction may improve the accuracy of elicited keywords for each chapter. In addition, considering the semantic relation of unique words with the title of each chapter and considering it as an additional score could be enhance the results.

REFERENCES

[1] Stuart Rose, Dave Engel, Nick Cramer, Wendy Cowley, "Text Mining: Applications and Theory", Michael W. Berry and Jacob Kogan, USA,pp 1-20,2010.

[2] S. Jones and G. W. Paynter, ”Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications,” Journal of the American Society for Information Science and Technology, vol. 53, no. 8, pp. 653-677, 2002.

[3] K. Coursey, R. Mihalcea, and W. Moen, ”Automatic keyword extraction for learning object repositories,” in Proc. Conf. Amer. Soc. Inf. Sci. Technol., 2008.

[4] Iryna Oelze,First examiner:Prof.Dr. techn. Dipl.-Ing. Wolfgang Nejdl,Second examiner:Prof. Dr.Heribert Vollmer,Supervisor:MSc. Dipl.-Inf. Elena Demidova:" Automatic Keyword Extraction for Database Search",Hannover, den 27 Februar 2009.

[5] Noopur Srivastava, Bineet Kumar Gupta," An Algorithm for Summarization of Paragraph Up to One Third with the Help of Cue Words Comparison",International Journal of Advanced Computer Science and Applications, Vol. 5, No. 5, 2014.

[6] Aliguliyev," AUTOMATIC DOCUMENT SUMMARIZATION BY SENTENCE EXTRACTION",Journal Computational Technologies scholar ,Issue number 5, volume 12,2007.

[7] Umar Manzoor, Mohammed A. Balubaid, " Semantic Image Retrieval: An Ontology Based Approach", International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.4, 2015.

[8] Maryam Habibi, Andrei Popescu-Belis," Keyword Extraction and Clustering for Document Recommendation in Conversations ",IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, VOL. 23, NO. 4, APRIL 2015.

[9] Kohei Arai and Herman Tolle, “E-learning Document Search Method with Supplemental Keywords Derived from Keywords in Meta-Tag and Descriptions which are Included in the Header of the First Search Result” International Journal of Advanced Computer Science and Applications (IJACSA), 3(4), 2012.

[10] M. Andrade and A. Valencia, “ Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families”, Bioinformatics, Vol.14(7, pages 600-607),1998.

[11] Kazi Saidul Hasan,Vincent Ng," Automatic Keyphrase Extraction: A Survey of the State of the Art ", 52nd Annual Meeting of the Association for Computational Linguistics, pages 1262–1273, June 23-25 2014.

[12] Bidyut Das, Subhajit Pal, Suman Kr. Mondal, Dipankar Dalui, Saikat Kumar Shome, “Automatic Keyword Extraction From Any Text Document Using N-gram Rigid Collocation”, International Journal of Soft Computing and Engineering (IJSCE),ISSN: 2231-2307, Volume-3, Issue-2,may 2013

[13] Witten, G. Paynte, E. Frank, C. Gutwin, C. Nevill-Manning. KEA: practical automatic keyphrase extraction. In Proceedings of the 4th ACM Conference on Digital Library, 1999.

[14] 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, p. 3770–3774, 2005.

[15] Anette Hulth, “Improved Automatic Keyword Extraction Given More Linguistic Knowledge”, Proceeding EMNLP '03 Proceedings of the 2003 conference on Empirical methods in natural language processing, Pages 216-223.

[16] Y. HaCohen-Kerner, ”Automatic extraction of keywords from abstracts,” in Proc. 7th Int. Conf. Knowledge-Based Intell. Inf. Eng. Syst., vol. 2773, pp. 843-849, 2003

[17] H. H. Kian and M. Zahedi, “AN EFFICIENT APPROACH FOR KEYWORD SELECTION; IMPROVING ACCESSIBILITY OF WEB CONTENTS BY GENERAL SEARCH ENGINES ”, International Journal of Web & Semantic Technology (IJWesT) Vol.2, No.4, October 2011.

[18] Chandra Shekhar Yadav, Aditi Sharan, Manju Lata Joshi," Semantic Graph Based Approach for Text Mining", ICICT,2014.

[19] 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. 1, pp. 157-169, 2004.

[20] Sifatullah Siddiqi, Aditi Sharan," Keyword and Keyphrase Extraction from Single Hindi Document using Statistical Approach ", 2nd International Conference on SignaProcessing and Integrated Networks (SPIN),2015.

[21] Fan Yang, Yue-Sheng Zhu, Yu-Jia Ma," WS-Rank: Bringing Sentences into Graph for Keyword Extraction ",Springer International Publishing, Switzerland, APWeb 2016, Part II, LNCS 9932, pp. 474–477, 2016.

[22] Yanchun Lu, Ruixuan Li, Kunmei Wen, Zhengding Lu, " Automatic Keyword Extraction for Scientific Literatures Using References ", international innovative design and manufacturing, Montreal, Quebec, Canada, August 13-15 2014.

[23] Su Nam Kim, Olena Medelyan, Min-Yen Kan, Timothy Baldwin, "SemEval-2010 Task 5: Automatic Key-phrase Extraction from Scientific Articles, 5th International Workshop on Semantic Evaluation", Uppsala, Sweden, pages 21–26, July 2010.

[24] El-Beltagy, S. R. and A. Rafea, ‘KP-Miner: Participation in SemEval-2’. In: Proceedings.of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden, pp 190-193,2010.