Received: May 15, 2017

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Keyword Extraction Approach for Single Document Extractive Summarization Based on Topic Centrality

Veera Venkata Murali Krishna Ravinuthala[1]* Satyananda Reddy Chinnam[2]

[1] *G.V.P College of Engineering, Visakhapatnam, India*
[2] *Andhra University, Visakhapatnam, India*
* Corresponding author's Email: rvvmuralikrishna@gvpce.ac.in

Abstract: Graph based keyword extraction approaches represent the text document as a graph with the words as vertices. An edge between two vertices exists only if the corresponding words are related. The effectiveness of the graph representation is dependent on how the relationship is defined between two words. Early approaches used word order, word co-occurrence and syntactic relationships. Recent graph representation approaches are using lexical association measures for defining relationship between two words. The existing relationships can only help in identifying the words which form the topics in a document. But the topics in a document essentially form a theme in the document. Theme is the central idea conveyed though the topics in a document. In this paper a new relationship is defined between the words to capture the words which convey the theme of the document. This relationship is defined using lexical association among the words in the document. Based on the relationship a graph representation of text is defined and used for extracting the words that build the theme of the document. These words are used for selecting important sentences in the document. Experiments on DUC 2002 data set indicated that the proposed keyword extraction approach improves the quality of Extractive summarization.

Keywords: Extractive summarization, Keyword extraction, Topical word, Topic centrality.

1. Introduction

Extractive summarization is the process of selecting important sentences based on the keywords in the text [1]. Keywords are the basic text units which associate with each other to create topics. Some keywords contribute towards the theme of the document by being central to all the topics. The presence of these keywords in an extractive summary would help the readers understand the theme of the document. The earlier keyword extraction approaches focused on extracting topic representative keywords using statistical methods [2-3] and linguistic methods [4]. Recent approaches are using a combination of statistical and linguistic methods to build an intermediate representation of text. This intermediate representation is analyzed to extract keywords from the text [5]. In these approaches the quality of keyword extraction is dependent on the effectiveness of the text representation. Graph based text representations offer more flexibility and are proved to be effective in keyword extraction [6]. Directed graphs and undirected graphs are two types of graph representations for text. Directed graph based text representations define edge direction based on Word order relationship [7] and syntactic relationship. Undirected graph representations mostly use word co-occurrence relationship. The experiments with directed graph representations did not produce better results for keyword extraction when compared to undirected graph based text representations. Edge direction plays an important role in assessing the importance of a vertex in the directed graph. Vertex ranking algorithms are using the number of incoming edges as a parameter for assessing the vertex relevance [8, 9]. But the existing definitions for edge direction are not allocating more incoming edges for important keywords. Because these definitions do not present a topic based relation among the vertices. Word order and syntactic relationships do not carry information about the topic of the document. But topics are formed based on the association of keywords in the document. All other words in the document combine with the keywords only to add more meaning/information to the keywords. For example consider an article describing the life of famous personality. The person's name is main keyword which associates with other keywords in the article. Consider his name as Mr. X. So all the keywords are attributing some information to Mr. X. The keywords can be Mr. X's educational qualifications, Professional designations or family member's names. Keywords can again be described by few other words in the article. But these associations are not considered in word order and syntactic relations. Word co-occurrence relations capture topic based association among the keywords but do not represent the flow of information. When two keywords associate with each other, the flow of information is necessarily from a less important keyword to a more important keyword. This concept leads to topic based edge direction between two vertices.

In this paper, a graph based text representation technique is proposed to identify the keywords that describe the theme of the document. This graph representation is referred as Topic Association Graph (TAG). In TAG, the vertices are individual words which generate the topics in the source document. These words are referred as topical words. The proposed work defines a novel relationship among the topical words based on the flow of information. Since the relationship is defined among the topical words, it represents topic association in the source document. The proposed keyword extraction algorithm extracts the words which establish association among the topics.

The present work is an extension to the work proposed in [10]. The proposed work adopts the vertex selection process described in [10] but defines its own technique for connecting the vertices. A new approach for vertex connectivity is proposed for identifying topic centrality of keywords. The next section introduces previous graph based keyword extraction approaches.

2. Related work

Keygraph [11] is a pioneering work in identifying topic central words in the document. In Keygraph, text is represented as an undirected graph with frequent words as vertices. Vertices are related based on the word co-occurrence frequencies in the document. Keyword extraction approach finds the words which associate maximal connected sub graphs in the graph representation. Matsuo et al. [12] proposed the concept of Small World to identify keywords in a document. Small World is modeled using word co-occurrence information in the document. The importance of words in the document is measured as its contribution to the Small World. Based on the path lengths the significance of words is evaluated. Matsuo et al. [13] have proposed Keyworld, as an extension of Small World concept by utilizing inverse document frequency.

All the above approaches have utilized word co-occurrence information for building an undirected graph representation of text. Undirected graph representation neither preserve word order information nor the flow of information. So the vertex ranking mechanisms are complex and take time. Faster and more accurate vertex ranking mechanisms such as degree centrality measures cannot be directly applied on undirected graphs. Mihalcea et al. [7] presented a seminal work on graph based keyword extraction. A formal definition is given for graph based text representation. Based on the concept of PageRank [8], a keyword extraction technique, TextRank, is proposed and tested on undirected and directed graph representations of text. TextRank has given better results with undirected graphs constructed using nouns and adjectives in the document as vertices. The problem with PageRank [8] is that the process has to be repeated till convergence is achieved. For large graphs it is tend to be slow.

Vertex centrality measures are found to be effective in keyphrase extraction [14] and keyword extraction [6]. In [15], selectivity-based keyword extraction (SBKE) is proposed based on the average weight distribution on the edges of a node. Vertex selectivity is calculated as the ratio of vertex strength and vertex degree. Goyal et al. [16] proposed a context based keyword extraction approach based on the lexical association derived from a large text corpus. Using lexical association an undirected graph is constructed and analyzed for keyword extraction. Experimental results have shown that when vertex connections are based on lexical association, a graph based text representation facilitates extraction of topic representative keywords in the document. But the overhead in this process is that Lexical association has to be calculated using a large related corpus before any document can be summarized. So the summarization process described by Goyal et al. [16] cannot be applied to the documents for which related corpus is not available In [17], Thematic text graph representation is proposed for identifying keywords representing the theme of the document. Vertex connections in thematic text graph are based on the word co-occurrence frequency. The direction of the edge is based on the inverse sentence frequencies of words corresponding to the vertices. The limitation in thematic text graph is in the computation of the edge strength between two vertices. Edge strength computation considers the inverse sentence frequency of only one of the participating vertices. The vertex which is pointing to another vertex is considered as source vertex and its inverse sentence frequency determines the strength of the edge. Since both the vertices are words in the document and appear randomly across the lines in the document, the association can be determined by considering the inverse sentence frequency of both the words.

3. Proposed Keyword extraction approach

The earlier graph based approaches focused on extracting topical words in the document [10] [16]. Among the recent approaches, Murali et al. [17] focused on the identifying theme carrying keywords in the document by defining a new graph representation which is referred as thematic text graph. Thematic text graph is designed such that thematic words get more incoming edges. So the Indegree strength of a vertex can be used to differentiate between thematic words and other words in a document. In this paper, theme is referred as topic centrality and a graph representation is proposed to find topic central words in the text. The proposed work extends the keyword extraction approach presented in [10]. The next section describes the concept of topic centrality of words in a document.

3.1 Topic centrality of keywords

A topic is a distribution over a fixed vocabulary. In other words a topic is represented by several related keywords which associate (co-occur) with each other. These words can be referred as topical words [16]. The discourse of a text document contains one or more topics and hence there exist clusters of topical words. The proposed keyword extraction approach aims to find the topical words which are central to most of the clusters. A topical word is said to possess topic centrality if it maintains close association with the topical words of other clusters. So the main task in this approach is to find lexical association in the document. Lexical association refers to the association between the words in a text or a corpus [18]. It can be broadly classified in to two categories, statistical association and semantic association, considering a single source text document. Semantic association reflects conceptual relationship between the words whereas statistical association is based on distributional patterns of the words in the text. The process followed in this paper for computing lexical association is based on statistical association. Word co-occurrence statistics are computed from the source text to find the association between any two words in the document. The following process describes the computation of statistical association between any two words in a given document.

Consider a sentence as a collection of words.

$$s \leftarrow \{w_0, w_1 \dots w_{m-1}\} \tag{1}$$

Let the set *S* consist of all sentences in the document *D*. For k sentences in the document *D*, set *S* can be defined as follows

$$S \leftarrow \{s_0, s_1 \dots s_p, \dots s_{k-1}\} \tag{2}$$

Co-occurrence of two words $w_i\ and\ w_j$ in a given sentence s_p is defined as follows

$$Co - occurrence_{s_p}(w_i, w_j) = \begin{cases} 1 & for\ w_i \in s_p\ and\ w_j \in s_p \\ 0 & otherwise \end{cases} \tag{3}$$

$Co - occurrenceFrequency(w_i, w_j)$ gives the number of sentences in which $w_i\ and\ w_j$ appear together irrespective of their order in the sentences.

$$Co - occurrenceFrequency(w_i, w_j) = \sum_{p=0}^{k-1} Co - occurrence_{s_p}(w_i, w_j) \tag{4}$$

For *Z* number of words in a document, the average co-occurrence frequency in *D* can be computed as follows

$$AverageCo - occurrenceFrequency_D = \frac{\sum_{i=0;j=0;i\neq j}^{z-1} \mathrm{Co{-}occurrenceFrequency}(w_i,w_j)}{\frac{\mathrm{z}\times(\mathrm{z}-1)}{2}} \tag{5}$$

Let *ISF* (w_i) gives the number of sentences in the document *D* that contain the word w_i, then lexical association between two words can be defines as

$$lexicalassociation(w_i, w_j) = \frac{Co{-}occurrenceFrequency(w_i,w_j)}{ISF(w_i)+ISF(w_j)-CO{-}OCCURRENCES(w_i, w_j)} \tag{6}$$

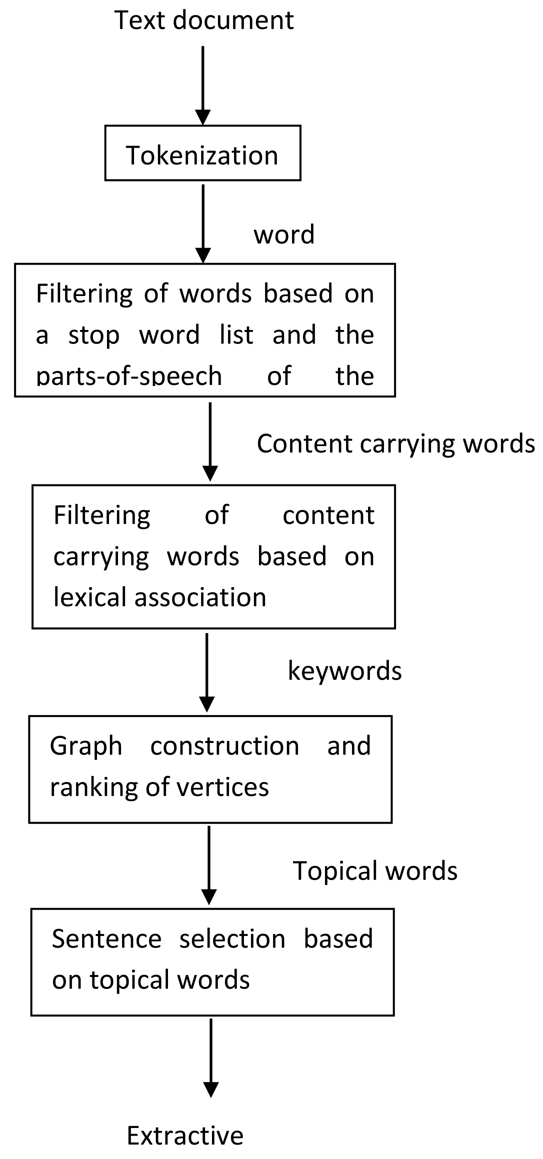
Text document
Tokenization
word
Filtering of words based on a stop word list and the parts-of-speech of the
Content carrying words
Filtering of content carrying words based on lexical association
keywords
Graph construction and ranking of vertices
Topical words
Sentence selection based on topical words
Extractive

Figure.1 Topical word extraction process for summarization

In our earlier work [10], lexical association is used in identifying topical words in a text document. Fig.1. describes the topical word extraction and summarization process presented in [10]. The proposed work adopts keyword extraction process described in [10]. In the preprocessing stage, stop words and the words whose parts-of-speech [19-20] does not fall in the set {Noun, Adjective, Verb, Adverb}, are removed from the document. For the remaining words in the document lexical association is computed. Two words are closely associated to each other if the lexical association between them is greater than the average lexical association value computed from all pairs of words in that document. Association Frequency (AF) of a word is the number words in the document with which it maintains close association. In this paper a directed graph representation of text is defined for representing the association among topical words. The next section describes the construction of graph representation for text document.

3.2 Graph representation of text

Consider $G(V, E)$ as a directed graph where V and E represent set of vertices and set of edges respectively. Each vertex $v \in V$ corresponds to a word (w) in the document (D) and

$$v = \{\, w \mid AF(w) > Avg \ and \ pos(w) \epsilon\, T \,, w \notin S \,\}$$

Where $Avg = \frac{\sum_{i=0}^{N} AF(w_i)}{N}$

$pos(w)$ gives the parts-of-speech of the word w, T ={*Noun, Verb, Adverb, Adjective*} and set P represents collection of stop words. An edge $e_{ij} \in E$ is drawn between two vertices v_i and v_j if they represent closely associated pair of words in the document. The direction of the edge is chosen based on the Association Frequency of the corresponding words. The following pseudo code describes the logic behind edge direction.

```
for $w_i \in v_i$ and $w_j \in v_j$
    if AF ($w_i$) > =AF ($w_j$)   then      $v_j \rightarrow v_i$
    else if AF ($w_i$) < AF ($w_j$) then    $v_i \rightarrow v_j$
```

The proposed approach to vertex connectivity is based on the hypothesis that topic central keywords have more association than the remaining words.

The traditional directed graph based text representations use word order relationship for connecting vertices. The application of a graph based ranking algorithm would result in the generation of topical words in the document [10]. But the objective of the present work is to extract words based on topic centrality. The Association Frequency of a topical word indicates its association with the words in the document whereas Topic centrality of a topical word is a measure of its association with other topical words in the document. A topic central word associates with other topical words to connect the topics in the document. The proposed graph representation is referred as Topic association graph. The vertex connections in Topic association graph are designed to increase the incoming edges for topic central words. It is assumed that a topical word with lower Association Frequency points to the topical word with higher Association Frequency. Let us refer to this relationship as topic association and the corresponding graph as topic association graph. For a graph based illustration of word order and topic association relationships. Consider the following document containing four sentences.

S1: Sita is wife of Rama.
S2: Sita accompanied Rama to the forest.
S3: Ravana abducted Sita from Rama.
S4: Rama killed Ravana and rescued Sita.

Table 1. Closely associated pairs of words in document D

Word pair	Lexical association
Sita, wife	0.25
Sita, rama	1
Sita, accompanied	0.25
Sita, forest	0.25
Sita, Ravana	0.5
Sita, abducted	0.25
Sita, killed	0.25
Sita, rescued	0.25
Wife, rama	0.25
Rama, accompanied	0.25
Rama, forest	0.25
Rama, Ravana	0.5
Rama, abducted	0.25
Rama, killed	0.25
Rama, rescued	0.25
Accompanied, forest	1
Ravana, abducted	0.5
Ravana, killed	0.5
Ravana, rescued	0.5
Killed, rescued	1

Table 2. Words with Association Frequency above the average

Keyword Code	Word	Association Frequency
K1	Sita	8
K2	Rama	8
K3	Ravana	5
K4	Killed	4
K5	Rescued	4

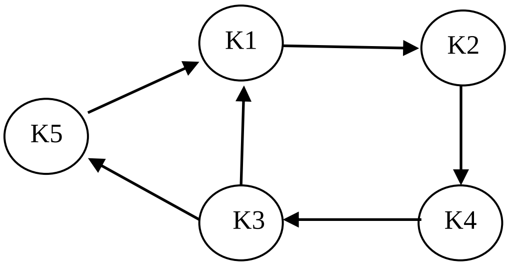

Figure.2 Directed graph based on word order relationship

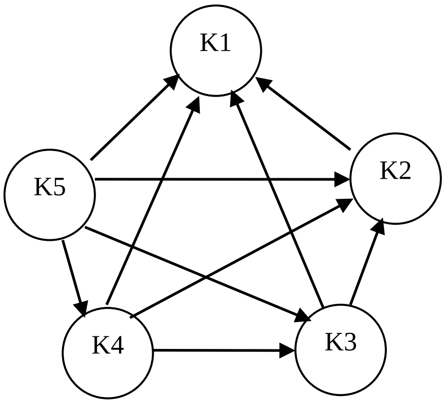

Figure.3 Topic association graph

Based on the Eq.(6), the average lexical association computed over all unique word pairs in document D is 0.23611. Table 1 shows the word pairs having lexical association above the average. Table 2. Presents the words in the document that are likely to be the topical words in the document. Fig.2. shows a graph representation considering the words in table 2 as vertices and their relationship based on their order of appearance in the sentences. Let us refer to this graph as word order graph. Fig. 3 shows a graph representation considering the words in the table as vertices and their relationship based on topic association.

3.3 Keyword extraction using topic association graph and sentence selection

The topic central words are spread all over the text and associate with the topical words. So topic central words have more associations than other words. By design, Topic association graph facilitates more incoming edges for the Topic central words. This property enables the use of degree centrality measures for the calculation of vertex strength. In this paper vertex strength is measured using In-degree and In-degree Strength [21] centrality measures. The topic association graph in Fig. 3 is modified to include edge weights based on the lexical association between the vertices. Eq. (7) describes the calculation of in-degree strength of a vertex.

Table 3. ROUGE-1 scores of proposed summarization system

Summarization System	ROUGE-1		
	Precision (P)	Recall (R)	F-measure (F)
ES-KWI	0.51257	0.61380	0.55838
ES-KWIS	0.51430	0.61643	0.56050

Table 4. ROUGE-2 scores of proposed summarization system

Centrality measure	ROUGE-2		
	Precision (P)	Recall (R)	F-Measure (F)
ES-KWI	0.40122	0.48129	0.43741
ES-KWIS	0.40323	0.48410	0.43977

$$Weight\ (V_i) = \sum_{V_j \in In(V_i)} weight(e_{ji}) \qquad (7)$$

Weight of e_{ji} is the lexical association value between the words corresponding to Vertex V_i and V_j. After calculating the weights of the vertices (keywords), the significance of a sentence in the source text document can be computed based on the number of keywords it possesses. If w_0 to w_n are the weights of keywords in the sentence, then the significance of the sentence is measured as follows

$$sentence\ weight = \sum_{i=0}^{n} w_i \qquad (8)$$

The sentences in the document are ranked according to their weights in decreasing order. Based on the user's input (say N sentences), top N sentences can be retrieved as an extractive summary of the document.

4 Results

The focus of this paper is to improve the quality of single document extractive summarization. So the keywords extracted from the topic association graph are utilized in single document extractive summarization. Two vertex centrality methods are used for keyword extraction. First one based on In-degree centrality and the second one using In-degree strength centrality. So two sets of keywords are extracted and each set is separately utilized for summarization. Let us refer to the Extractive summarization system based on Indegree centrality as ES-KWI and the summarization based on vertex in-degree strength as ES-KWIS. DUC 2002 data set [22] is used for the evaluation of the proposed summarization process. DUC2002 dataset contains 533 unique text documents and 8316 corresponding human written summaries. Since it is tedious work to match 533 system generated summaries with 8316 human written summaries, ROUGE tool [23] is used in the evaluation process. Using ROUGE tool, Unigram matching score (ROUGE-1) and Bi-gram matching score (ROUGE-2) are computed. ROUGE-1 and ROUGE-2 are experimentally proven to be close to human judgment in the relevance calculation of summaries [23]. Tables 3 and 4 present the performance details of ES-KWI and ES-KWIS over DUC2002 dataset.

5 Discussion

Based on Eq (1), lexical association is computed on the source documents and their corresponding summaries of DUC 2002 dataset. Lexical association is computed only on the words having parts-of-speech as Noun/Verb/Adverb/Adjective and does not come under stop word category. It is observed that the lexical association is increasing in the summaries with the increase in the size of their source text document. While there is not much difference among the lexical association values computed over source documents. Fig.4. shows the lexical association values computed separately over source text documents and summaries of DUC2002.

Table 5. Average lexical association computed on the documents of DUC 2002 dataset

	Source Text Documents	Manual Summaries
Average Lexical Association	0.05779	0.40207

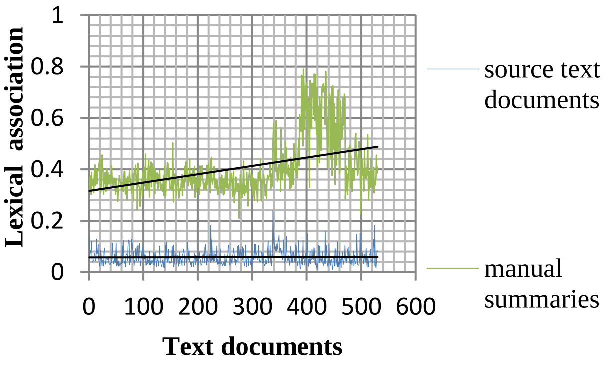

Figure 4. Average lexical association in source text documents and summaries

Table 6. Comparison with existing summarization systems

SUMMARIZATION SYSTEM	AVERAGE RECALL	
	ROUGE-1	ROUGE-2
UniformLink+bern +neB [16]	0.46432	0.20701
Directed graph based summarization system [10]	0.48645	0.39927
ThemeRank based summarization [17]	0.60563	0.46838
ES-KWI	0.61380	0.48129
ES-KWIS	0.61643	0.48410

Table 7. Comparison with the best keyword extraction system in SemEval-2010

	Proposed keyword extraction system (IS-KW)			HUMB		
Number of keywords	P	R	F	P	R	F
5	31.7	33.7	32.6	21.2	27.4	23.9
10	21.5	34.1	26.3	15.4	39.8	22.2
15	17.0	38.4	23.5	12.1	47.0	19.3

Table 5 presents the average lexical association computed over 533 source text documents and 8316 summaries of DUC 2002 dataset. The results show that lexical association value is more in the summaries than the source text documents. This results shows that lexical association is an important parameter for text summarization. Since the proposed summarization approach uses lexical association for text summarization, it should produce good summaries. The text summarization systems, ES-KWI and ES-KWIS, are compared with the existing text summarization systems in table 6. It is observed that ES-KWI and ES-KWIS have given better performance indicating the efficiency of the proposed topic association graph. Topic association graph is a better representation than the Thematic text graph because of the underlying definition of a directed edge. The authors of Thematic text graph [17] assumed that keywords with more inverse sentence frequency will have more associations and hence they are more likely to be topic central words. But keywords with more inverse sentence frequency necessarily need not have more associations. For example, if a keyword appears in 100 sentences of a document and all the 100 sentences are constructed from four keywords, then the keyword association is only with four words. It is possible that another keyword which is appearing in 10 sentences can be associated with more keywords. Topic Association Graph has a better definition for edge strength when compared to that of Thematic text graph. Because topic association graph considers the ISF of both the vertices (words) for calculating the edge strength between them.

The performance of ES-KWIS indicates that vertex in-degree strength measure is relatively better than vertex in-degree measure for keyword weighting. For the evaluation of the vertex in-degree strength based keyword weighting process (IS-KW), the authors used SemEval-2010 test dataset. SemEval-2010 test data set includes 100 text documents and manually assigned keywords for each of the document in stemmed format. Since author and reader assigned keywords contained some keyphrases, the authors have tokenized them in to individual words for comparison with our system. Top 5, 10 and 15 keywords generated by the proposed application are compared with the author assigned keywords for each document in the test dataset. Table 7 presents a comparison of the IS-KW and the best keyword extraction system in SemEval-2010 [24] based on the author assigned keywords. IS-KW has given better performance which has further strengthened the accuracy of the ES-KWIS summarization system. It can be said that topic association graph with lexical association based edge weights is a better representation for the topic connectivity in a document.

6 Conclusions and future scope

In this paper a topic association graph is defined and utilized for representing the text document. The application of in-degree and in-degree strength centrality measures on topic association graph has facilitated an improved keyword extraction mechanism. Based on the experiments conducted on DUC2002 summarization dataset, it is observed that in-degree strength measure is slightly better than in-degree measure for keyword weighting. Comparison with author generated keywords of SemEval-2010 test dataset indicated the efficiency of the proposed keyword extraction process. Since the text documents of DUC 2002 dataset are newswire articles, most of the significant sentences are found to be the first few lines of text document. But in scientific documents the first few lines of the text introduce the topic and the portions of text in the middle actually describe the topic. So it would be interesting to test the proposed keyword extraction process, IS-KW, on scientific documents. It would also be interesting to test the proposed keyword extraction technique on IS-KW on very large text documents. IS-KW can be applied to Information Retrieval and Information Filtering tasks.

References

[1] H. P. Edmundson, “New methods in automatic extracting”, *Journal of the ACM*, Vol. 16, No. 2, pp. 264-285, 1969.

[2] M. Andrade and A. Valencia, “Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families”, *Bioinformatics*, Vol. 14, No. 7, pp. 600–607, 1998.

[3] E. Hovy, and C.Y. Lin, “Automated text summarization in summarist”, *Advances in Automatic Text Summarization*, MIT Press, pp. 81-94, 1999.

[4] A. Hulth, “Improved automatic keyword extraction given more linguistic knowledge”, In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Morristown, NJ, USA, Vol. 10, pp. 216–223, 2003.

[5] A. Nenkova and K. McKeown, "Automatic Summarization", *Foundations and Trends® in Information Retrieval*, Vol. 5, No. 2–3, pp 103-233, 2011.

[6] S. Lahiri, S. R. Choudhury, and C. Caragea, “Keyword and keyphrase extraction using centrality measures on collocation networks”, *arXiv preprint*, arXiv:1401.6571, 2014.

[7] R. Mihalcea and P. Tarau, “Textrank Bringing order into texts”, In: *Proc. of EMNLP*, 2004.

[8] S. Brin and L. Page, “The anatomy of a large-scale hyper textual Web search engine”, *Computer Networks and ISDN Systems*, Vol. 30, pp.1–7, 1998.

[9] J.M. Kleinberg, “Authoritative sources in a hyperlinked environment”, *Journal of the ACM*, Vol.46, No.2, pp. 604–632, 1999.

[10] V.V.M.K. Ravinuthala and Ch.S. Reddy, “Extractive Text Summarization Using Lexical Association and Graph Based Text Analysis”, *Advances in Intelligent Systems and Computing*, Vol. 410, pp. 261-272, 2015.

[11] Y. Ohsawa, N.E. Benson and M. Yachida, “KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor”, In: *Proc. of the Advanced Digital Library Conference*, pp. 12-18, 1998.

[12] Y. Matsuo, Y. Ohsawa, and M. Ishizuka , “A document as a small world”, *New Frontiers in Artificial Intelligence*, Springer Berlin Heidelberg, pp. 444-448, 2001.

[13] Y. Matsuo, Y. Ohsawa, and M. Ishizuka, “Keyworld: Extracting keywords from documents small world”, *Discovery Science*, Springer Berlin Heidelberg, pp. 271-281. 2001.

[14] Z. Xie, “Centrality Measures in Text Mining: Prediction of Noun Phrases that Appear in Abstracts”, In: *Proc. of 43rd Annual Meeting of the Association for Computational Linguistics*, ACL, University of Michigan, USA, ,pp. 103-108,2005

[15] S. Beliga, A. Mestrovic, and S. Martincic-ipsic, “Toward Selectivity Based Keyword Extraction for Croatian News”, *arXiv preprint*, arXiv:1407.4723, 2014.

[16] P. Goyal , L. Behera, and T.M. McGinnity, “A Context-Based Word Indexing Model for Document Summarization”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.25, pp.1693-1705, 2013.

[17] V.V.M.K. Ravinuthala and Ch.S. Reddy, “Thematic Text Graph: A Text Representation Technique for Keyword Weighting in Extractive Summarization System”, *International Journal of Information Engineering and Electronic Business*, Vol. 8, pp 18-25, 2016.

[18] P. Pecina,” Lexical association measures and collocation extraction”, *Language Resources and Evaluation*, Springer, Netherlands, Vol .44, pp. 137-158, 2010.

[19] K. Toutanova and C.D. Manning, “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger”, In: *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63-70, 2000.

[20] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”, In: *Proc. of HLTNAACL*, pp. 252-259,2003.

[21] S. Beliga, A. Mestrovic, and S. Martincic-ipsic, “An overview of graph-based keyword extraction methods and approaches”, *Journal of Information and Organizational Sciences*, Vol. 39, pp. 1-20, 2015.

[22] P. Over and W. Liggett, “Introduction to DUC: An Intrinsic evaluation of Generic News Text Summarization Systems”, In: *Proc. of DUC workshop on Text Summarization*, 2002.

[23] Ch. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", In: *Proc. of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*, Barcelona, Spain, 2004.

[24] S.N. Kim, O. Medelyan, M.Y. Kan, and T. Baldwin, "Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles", In: *Proc. of the 5th International Workshop on Semantic Evaluation*, pp. 21-26, 2010.