I.J. Information Engineering and Electronic Business, **2016, 4, 18-25**
Published Online July 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2016.04.03

Thematic Text Graph: A Text Representation Technique for Keyword Weighting in Extractive Summarization System

Murali Krishna V.V. Ravinuthala
G.V.P College of engineering (A), Visakhapatnam, India
E-mail: rvvmuralikrishna@gmail.com

Satyananda Reddy Ch.
Andhra University, Visakhapatnam, India
E-mail: satyanandau@yahoo.com

Abstract—Keyword extraction approaches based on directed graph representation of text mostly use word positions in the sentences. A preceding word points to a succeeding word or vice versa in a window of N consecutive words in the text. The accuracy of this approach is dependent on the number of active voice and passive voice sentences in the given text. Edge direction can only be applied by considering the entire text as a single unit leaving no importance for the sentences in the document. Otherwise words at the initial or ending positions in each sentence will get less connections/recommendations. In this paper we propose a directed graph representation technique (Thematic text graph) in which weighted edges are drawn between the words based on the theme of the document. Keyword weights are identified from the Thematic text graph using an existing centrality measure and the resulting weights are used for computing the importance of sentences in the document. Experiments conducted on the benchmark data sets SemEval-2010 and DUC 2002 data sets shown that the proposed keyword weighting model is effective and facilitates an improvement in the quality of system generated extractive summaries.

***Index Terms*—**Extractive summarization, keyword weighting, directed graph, Thematic text graph, ThemeRank.

I. Introduction

Automatic text summarization can be classified in to two types, Extractive summarization and Abstractive summarization [1]. Extractive summarization aims to pick up important sentences in the text document whereas Abstractive summarization process produces a short text that conveys the meaning of the original text. Keyword extraction and weighting methods influence the quality of system generated summaries. The existing studies on keyword extraction and weighting predominantly use vector space model [2] and graph models [3] for text representation. Vector space models represent text as a collection of independent words with some corresponding weights. Graph models have become popular as they preserve syntactical information as well as the semantic and statistical relationship among the words in the document.

Previous studies on graph based text representation have used word co-occurrence information, semantic similarity, syntactic relations and distance measures to draw edges between the vertices [4]. Graph based ranking algorithms utilize these edges to assign weight to each of the vertices. The ranking algorithms which are inspired by PageRank [5] and HITS [6], assign more weight to a word (node) with more number of recommendations (incoming edges). So the directions of edges in the graph play a crucial role in keyword weighting.

Distance measures use placement of words in sentences/entire text for defining the direction of an edge. For a window size of 2 words, all the neighboring words in a sentence/text are connected and the direction of edge follows the natural direction of the text. Hence a preceding word points to a succeeding word. This kind of representation will have two different forms for the same sentence written in active and passive voices leading to contrasting conclusions which is not acceptable. For a window size of 2 words, the subject word and the Object word in a sentence are not directly connected to each other. The strength of their relationship is determined by the number of edges in between them or based on the neighbors. In this paper a directed graph representation of text is proposed. This representation uses a topic based keyword selection procedure to represent the vertices in the graph. The graph is designed to capture the associations between the topics in the document and the edge direction indicates the flow of theme instead of the natural flow of text in the document. This paper focuses on using the proposed representation for finding keywords weights and using them for extractive summarization of the given document.

The rest of the paper is organized as follows: Section 2 describes the related work on graph based text representations and keyword weighting, Section 3 presents proposed summarization methodology, details of the dataset and evaluation measures used to test the proposed methodology and the results obtained are given in Section 4, comparison of proposed methodology with existing methodologies is done in Section 5 and Section 6 concludes the overall work presented in this paper.

II. Related Work

Keyword extraction is an important task in many Text mining and Natural language processing applications. Automatic text summarization (ATS) is one such application that uses keywords to select important sentences in the document. Usage of Graph based keyword extraction mechanisms in ATS applications has increased due to the flexibility they offer in processing unstructured text.

Ohsawa et al. [7] proposed KeyGraph, for keyword extraction based on the undirected graph representation of text. KeyGraph builds an undirected graph using term frequency and word co-occurrence information in the text. KeyGraph aims to segment the graph such that each sub graph contains keywords related to the sub topics in the document. The words which bridge relationship between the various topics in the document are chosen as Keywords. Matsuo et al. [8] proposed the concept of contractor and shortcut to identify important terms in a text document. The proposed concept treats the text document as a small world and identifies the importance of a word based on its contribution to the small world. The small world is represented as a word co-occurrence graph with edges representing the co-occurrence relationship between the words. Path lengths are considered as a measure to identify contractor nodes and shortcut nodes. Matsuo et al. [9] have extended their earlier work [8] by multiplying the contribution of a word with its inverse document frequency.

Mihalcea et al. [10] proposed a graph representation of text based on distance between the words in the text. Two nodes in the graph are connected to each other, if the keywords they represent appear within the window of N consecutive words in the text. N value can be any number in between 2 to 10. In their proposed work, TextRank, experiments were conducted on both directed and undirected graph representations of text document. Directed graph representation of text is based on the positions of words in the text. A directed edge can be drawn from node1 to node2, if the word represented by node1 precedes the word represented by node2 and if these words fall within a window of N words. TextRank has given better results when an undirected graph is constructed with nouns and adjectives as vertices for a window size of 2.

Node centrality measures are found to be effective in keywords weighting [11][12]. Lahiri et al. [13] made an extensive study on the usefulness of graph based centrality measures for keyword extraction. Experiments are carried on the possible combinations of directed, undirected graphs with and without weights. The central idea behind graph construction is based on TextRank model. Experimental results have shown that centrality based measures and PageRank [5] performed better than existing algorithms.

Beliga et al. [14] proposed a graph based keyword extraction technique known as selectivity-based keyword extraction (SBKE). SBKE examines the average weight distribution on the edges of a node. Vertex selectivity is calculated as the ratio of vertex strength and vertex degree. For a directed graph, a vertex has two strengths, in-degree and out-degree. Vertex in-strength is sum of weights of all incoming edges. Out-strength is the sum of weights of all outgoing edges. Pawan Goyal et al. [15] proposed a context based keyword extraction model based on lexical association in a large corpus. Text is represented as an undirected graph and keyword weights are computed using a PageRank based vertex ranking algorithm. Basing on the sentence similarity values important sentences are selected.

Murali Krishna et al. [16] proposed a graph based keyword extractive technique using directed graph representation of text document. A graph is constructed with the keywords exhibiting high lexical association as vertices and by relating the vertices based on their co-occurrence in the individual sentences of the document. The direction of edge is influenced by the word order relationship. PageRank [5] algorithm is applied on this graph to obtain the weights of the vertices. The proposed keyword weighting mechanism is applied to the extractive summarization task and found to be effective.

III. Text Representation Using Directed Graphs

Most of the existing directed graph based text representations use distance measure for connecting the nodes in a graph. According to distance measure, the words within a window of N consecutive words are related to each other and the order of their appearance decides the direction of edge between the corresponding nodes. Consider a sample sentence “Rama killed Ravana” for illustration. If Node1, Node2 and Node3 in the graph represent the words “Rama”, “killed” and ”Ravana” respectively, then the following relations hold for a window size of 2.

$$Node1 \rightarrow Node2, Node2 \rightarrow Node3 \quad (1)$$

If the same sentence is written in passive voice as “Ravana is killed by Rama”, then the relations will be as follows

$$Node3 \rightarrow Node2, Node2 \rightarrow Node1 \quad (2)$$

Since directed edges represent recommendations in between the nodes, there is a difference in the meaning conveyed by the sets of relations in (1) and (2). Another problem with Word order based edge direction is that it assumes relation between two unrelated words. If entire text in the document is considered as a single unit for applying word co-occurrence window of size 2, then the last word in a sentence and the first word in the following sentence are assumed to be related. This does not happen always particularly when the sentences are from two different paragraphs.

A. Thematic text graph

Theme is the central idea conveyed in a document. It can be explored by associating the topics/concepts in that document. So a Thematic text graph should be able to capture the topics as well as their associations in the source text document. The proposed Thematic text graph uses parts-of-speech information [17][18], word co-occurrence(wc) information and inverse sentence frequency(isf) of the words in the document. The following definitions are followed in our experimentation.

Word co-occurrence: A pair of words are said to co-occur once if they appear in a sentence of the document and they need not be neighbors.

Inverse sentence frequency (isf): It is the number of sentences in the document containing a given word.

Thematic text graph construction is based on the following assumptions

1. Words in the document with parts-of-speech as Noun/Verb/Adverb/Adjective are useful words. Clusters of frequently co-occurring pairs of useful words describe the topics/concepts covered in the document.
2. A keyword contributes to a topic in the document by co-occurring with many useful words in that topic. So their inverse sentence frequency is more than the useful words. A directed edge from a useful word towards a keyword indicates that the useful word is adding more information to the topic through the keywords.
3. The association of topics in the document can be captured using the keywords which are central to all the topics in the document. These keywords can be referred as ThemeWords and they co-occur with many useful words and keywords in the document.

Steps for constructing Thematic text graph

Begin

1. Extract Nouns, Verbs, Adverbs and Adjectives from the document D.
2. Pair words obtained in step1 such that each word pairs with all other words. For N words there can be (N*(N-1))/2 pairs of words. Count co-occurrence of words in each pair among the sentences in the document.
3. Using pair wise co-occurrence information obtained in step2, select pairs of words with frequency of co-occurrence above the average co-occurrence frequency of all pairs of words which co-occur at least once. Let us refer to these pairs of words as Frequently Co-occurring Word Pairs (FCWP).
4. Let G (V, E) be a directed graph for the document D, such that vertices/nodes in set V represent the words in the document which form FCWP. An edge in set E represents a directed edge between the nodes representing a pair of words in FCWP. The number of edges in the set E will be equal to the number of word pairs in FCWP. The direction of edge is based on the inverse sentence frequency (isf) of the words and is defined as follows

$$\begin{aligned}&\mathit{If}\ \left(W_i, W_j\right) \in \mathit{FCWP\ and\ isf}\ \left(W_i\right)\ >=\ \mathit{isf}\ \left(W_j\right)\\&\mathit{then}\ v_j \rightarrow v_i\ \mathit{for}\ W_i \in v_i\ \mathit{and}\ W_j \in v_j\\&\mathit{else}\ \ v_i \rightarrow v_j\ \mathit{for}\ W_i \in v_i\ \mathit{and}\ W_j \in v_j\end{aligned}$$

5. The weight of a directed edge(e_{ij}) from the node v_i to the node v_j is defined as

$$\mathit{Weight}\ \left(e_{ij}\right)\ =\ \mathit{WC}\ \left(W_i, W_j\right)\ /\ \mathit{isf}\ \left(W_i\right) \quad (3)$$

Where $\mathit{WC}\ \left(W_i, W_j\right)$ represents frequency of co-occurrences between W_i and W_j and isf(w_i) represents inverse sentence frequency of the word W_i .

End

B. How Thematic text graph works

We are considering three cases to describe the construction of Thematic text graph

Case 1: Consider a document *D1*, with two sentences as follows

S1: Sita is wife of Rama.
S2: Sita accompanied Rama to the forest.

“Sita”, “wife”, “Rama”, “accompanied”, “forest” are useful words in the given document *D1*. From the sentences in *D1* one can observe that there is only one frequently co-occurring word pair (Sita, Rama). So the Thematic text graph (as shown in Fig.1.) contains two nodes representing “Sita” and “Rama” respectively for the document *D1*. Since the words “Sita” and “Rama” have same inverse sentence frequency, the order of their appearance in the text is considered to draw a directed edge between them. The assumption in this case is, a word appearing first in the text will have more weight. So the node representing the word “Rama” will recommend the node representing “Sita”.

Case 2: Consider a document *D2* with three sentences

S1: Rama is the king of Ayodhya
S2: Sita is wife of Rama.
S3: Sita accompanied Rama to the forest.

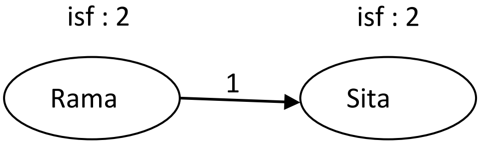

Fig.1. Thematic text graph for document *D1*

“Rama”, “king”, “Ayodhya”, “Sita”, “wife”, “accompanied”, “forest” are useful words in the document *D2*. Thematic text graph for *D2* will contain two nodes as there is only one frequently co-occurring word pair (Sita, Rama). In Fig. 2 we can observe that node representing “Sita” points to the node representing “Rama”, as the word “Rama” has more inverse sentence frequency than the word “Sita”.

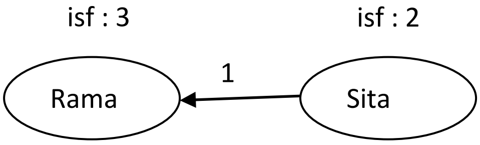

Fig.2. Thematic text graph for document *D2*

Case 3: Consider a document *D3* with four sentences as follows

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. Co-occurring word pairs in document *D3*

Pair Of words	Co-occurrence frequency
Sita, Rama	4
Ravana, Sita	2
Rama, Ravana	2

“Sita”, “wife”, “Rama”, “accompanied”, “forest”, “Ravana”, “abducted”, “killed”, “rescued” are useful words in the document *D3.* Thematic text graph for *D3* will have three nodes as there are three words “Sita”, ”Rama” and “Ravana” co-occurring with each other as shown in table 1 and three edges for representing three co-occurring pairs of words. Thematic text graph for document *D3* is shown in Fig.3. Since the word “Sita” has more inverse sentence frequency than the word “Ravana”, an edge can be drawn from the node representing “Ravana” to the node representing” Sita”. Similarly a directed edge can be drawn from the node: Ravana to node: Rama.

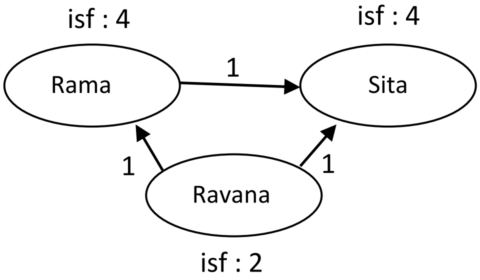

Fig.3. Thematic text graph for document *D3*

C. ThemeRank: Keyword weighting based on Thematic text graph

Studies indicated that centrality based measures in graph based keyword extraction perform better than Graph based ranking algorithms [11][13][19][20][21]. Centrality measures are very fast and easy to compute on single-document graphs. Among the centrality based measures, degree and strength measures are proved to be simple, fast and effective [4]. The design of Thematic text graph is based on the association between the keywords belonging to various topics in the document. According to the third assumption, ThemeWords will have more associations than the keywords as they are central to all the sub topics. Proposed design models these associations as directed incoming edges towards the vertices represented by ThemeWords. So the contribution of a keyword towards the theme of the document can be measured using the in-degree strength of its corresponding vertex in the graph. This approach is referred as ThemeRank and is computed as follows.

$$Weight\left(V_i\right)=\sum_{V_j\in In\left(V_i\right)} weight\left(e_{ji}\right) \tag{4}$$

Where e_{ji} represents directed edge from node v_j to node v_i

IV. Implementation and Results

The objective of our proposed text representation is to aid in improving the quality of single document Extractive summarization. To test our objective, summary generation process has been modeled based on the weights of the keywords obtained using ThemeRank. This section describes the experiments carried out to test the accuracy of keyword weighting based on ThemeRank and the quality of generated extractive summaries. Experimentation is carried out with the applications developed using Java programming language on windows platform. Section A and B describe the standard evaluation tools and data sets used for testing the developed applications.

A. *ThemeRank Evaluation Results*

ThemeRank is evaluated with the test dataset provided for keyword/keyphrase extraction in the Semantic Evaluation Workshop (SemEval-2010) [22]. Test dataset contained 100 scientific documents along with the keywords in stemmed format. Keywords are manually assigned by the author and the reader of the scientific articles. Since author and reader assigned keywords contained some keyphrases, we have tokenized them in to single word keywords for easy comparison with our system. The evaluation metrics used in SemEval-2010 are Precision, Recall and F-measure. Let the symbol SK denote the set of system generated keywords and the symbol MK denote the set of manually generated keywords. Equations (5), (6) and (7) are used for computing Precision, Recall and F-measure.

$$Precision = \frac{|SK \cap MK|}{|SK|} \qquad (5)$$

$$Recall = \frac{|SK \cap MK|}{|MK|} \qquad (6)$$

$$F-measure = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (7)$$

ThemeRank is applied on the first document in the dataset i.e, *c-1.txt.final* to extract the top 5 keywords. Table 2 shows the top 5 keywords along with their inverse sentence frequency and weight. From Table 3, it can be observed that three of the words generated by ThemeRank match with the author generated top 5 keywords. Fig.4 shows that, out of the 309 frequently co-occurring words in c-1.txt.final, only 109 words are assigned a non zero weight by the ThemeRank. The remaining words are assigned a weight of zero.

Table 2. Top 5 keywords retrieved by ThemeRank from the document c-1.txt.final

Rank	Keyword	Inverse Sentence frequency of the keyword	Keyword weight
1	Uddi	85	40
2	Registry	81	33
3	Servic	82	27
4	Discoveri	24	15
5	Dht	62	15

Table 3. Top 5 keywords extracted by ThemeRank and Author from the document c-1.txt.final

Rank	ThemeRank	Author
1	Uddi	Uddi
2	Registry	Dht
3	Servic	Web
4	Discoveri	Servic
5	Dht	Grid

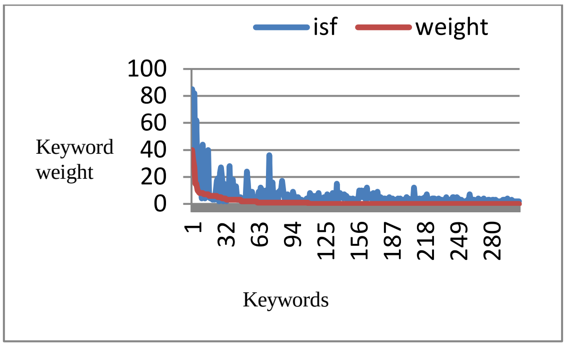

Fig.4. Details of keywords in c-1.txt.final

Table 4 shows the Precision, Recall and F-measure of ThemeRank over SemEval-2010 dataset. ThemeRank has given consistent performance withrespect to the precision, but the Recall value dropped drastically over Reader assigned keywords. This is due to the increase in number of keyphrases in reader assigned keywords/ keyphrases. In Tables 5 and 6, ThemeRank is compared with the keyword/keyphrase extraction systems participated in SemEval-2010. Considering only Precision measure over Reader assigned keywords, ThemeRank stands at 4th position in overall rankings among all participating systems in SemEval-2010. Withrespect to obtained Recall values on Author assigned keywords, ThemeRank stands at 1st position among all the participating systems.

Table 4. Evaluation of top 5 keywords generated by ThemeRank using SemEval-2010 dataset

Keywords assigned by	Precision	Recall	F-measure
Author	30%	28.3 %	30.1%
Reader	30%	7.89%	12.4%
Author and Reader	30%	7.09%	11.4%

Table 5. Comparison with participating systems in SemEval-2010 based on Reader assigned keywords

Keyword extraction systems	Precision
SEERLAB[22]	31 %
WINGNUS[22]	30.6 %
HUMB[22]	30.4 %
THEMERANK	30 %

Table 6. Comparison with participating systems in SemEval-2010 based on Author assigned keywords

Keyword extraction systems	Precision
HUMB[25]	21.2 %
MAUI[25]	20.4 %
KP-MINER[25]	19 %
THEMERANK	30 %

B. *Extractive Summarization Using ThemeRank*

The weight of the sentence in the source text document is computed as the sum of the weights of its constituent keywords [23]. Sentences in the document are sorted in decreasing order. Based on the requirement on the summary size, the topmost sentences can be presented as the extractive summary for that document. We have tested our ThemeRank based summarization (TRS) approach using DUC2002 dataset [24]. DUC2002 dataset consists of 533 unique text documents and 8316 manually written summaries. The performance of the application is indicated using ROUGE-1 and ROUGE-2 scores [25]. During the initial testing phase, TRS is applied on the top three largest files (LA101590-0066, LA042789-0025, LA011990-0091) in the DUC2002 dataset. Tables 7 and 8 show the ROUGE scores obtained for the three documents. Since the results were competitive with the existing standards, we have applied TRS on all the 533 documents in DUC2002 dataset and the results are shown in Table 9.

Table 7. ROUGE-1 scores for 100 words summary generated by TRS

Document	**No. Of Lines**	**Average Recall**	**Average Precision**	**Average F-measure**
LA101590-0066	123	0.64621	0.57164	0.60664
LA042789-0025	163	0.45053	0.38462	0.41497
LA011990-0091	186	0.52827	0.44719	0.48436

Table 8. ROUGE-2 scores for 100 words summary generated by TRS

Document	**No. Of Lines**	**Average Recall**	**Average Precision**	**Average F-measure**
LA101590-0066	123	0.51897	0.45849	0.48686
LA042789-0025	163	0.23682	0.20183	0.21793
LA011990-0091	186	0.42899	0.36250	0.39295

Table 9. Average ROUGE scores for 100 words summaries generated on DUC2002 dataset

Quality Measure	**Average Recall**	**Average Precision**	**Average F-measure**
ROUGE-1	0.60563	0.50509	0.55055
ROUGE-2	0.46838	0.39000	0.42541

V. Result Analysis

Tables 5 and 6 show that ThemeRank can compete with the existing keyword extraction techniques in terms of Precision. The competitiveness in terms of Precision indicates that ThemeRank is a reliable keyword extraction approach. The proposed text representation facilitates more recommendations for the keywords with high *isf* value. However Fig. 4 shows that keyword weights are not biased towards inverse sentence frequency of the words. Even if a keyword gets more recommendations due to its high *isf* value, it is actually the strength of the incoming edges that determines the weight of the keyword. In Table 2, it can be observed that the keywords “dht” and “discoveri” have the same weight despite a huge difference in *isf* value. The results shown in Table 10 indicate that the proposed summarization system performed better than the recent graph based summarization systems. It is due to the identification of Theme conveying keywords in the document. ThemeWords reflect the overall opinion conveyed in the document whereas normal keywords highlight the individual topics/concepts in the document. So an extractive summary based on ThemeWords will be more coherent. An example summary presented in Fig. 6 indicates that ThemeRank has given more weight to the Theme carrying words, as it contains semantically coherent important sentences in the document. The performance of TRS over DUC2002 dataset has indicated that the proposed text representation is able to capture the theme of the document. During the experimentation it is observed that very few words in DUC text documents are frequently co-occurring with each other. On an average basis 29 words are forming Frequently Co-occurring Word Pairs in a 28 lines text document. So there will be less number of vertices in the Thematic text graph in comparison with the existing graph representations. The analysis concludes that Thematic text graph is a compact text representation for fast and enhanced topic analysis.

Table 10. Comparison with graph based summarization systems

SUMMARIZATION SYSTEM	**AVERAGE RECALL**	
	ROUGE-1	**ROUGE-2**
UniformLink+bern+neB[15]	0.46432	0.20701
Directed graph based summarization system [16]	0.48645	0.39927
Proposed summarization system	0.60563	0.46838

[1] leonard bernstein, the renaissance man of music who excelled as pianist, composer, conductor and teacher and was, as well, the flamboyant ringmaster of his own nonstop circus, died sunday in his manhattan apartment. [2] bernstein was the first american-born conductor to lead a major symphony orchestra, often joining his new york philharmonic in playing his own pieces, while conducting from the piano. [3] bernstein, known and beloved by the world as "lenny," died at 6:15 p.m. in the presence of his son, alexander, and physician, kevin m. cahill, who said the cause of death was complications of progressive lung failure. [4] cahill said progressive emphysema complicated by a pleural tumor and a series of lung infections had left bernstein too weak to continue working. [5] invited to a dance recital in new york, bernstein sat in the balcony next to a man he did not recognize. [6] on his 25th birthday, aug. 25, 1943, bernstein was told by the koussevitzkys that he should visit artur rodzinsky, newly appointed music director of the new york philharmonic, at his stockbridge, conn., farm .

Fig.5. 6 lines summary generated for the DUC2002 document LA101590-0066

VI. Conclusion

This paper proposes Thematic text graph, a directed and weighted graph representation for unstructured text. The vertices in the graph represent the keywords related to the topics in the document. A new definition is given to the relationship between the vertices to capture the flow of the theme in the document. This relationship is modeled as a directed and weighted edge between two vertices. The edge weights represent the association strength between subtopics and hence the in-degree strength of a vertex indicates its centrality among the topics in the document. We propose ThemeRank to find the weights of the vertices in a Thematic text graph. ThemeRank has given encouraging results when tested with SemEval-2010 dataset. The performance of ThemeRank based summarization over DUC2002 dataset proved that ThemeRank is a competitive technique for keyword weighting. ThemeRank is language independent and executes in less time due to the simple structure of Thematic text graph. On an average basis, ThemeRank has extracted 12 keywords per 28 lines in DUC2002 documents. Further work will carry experiments on the applicability of ThemeRank for textbook index generation and chapter summarization.

References

[1] S. Gholamrezazadeh, M. A. Salehi and B. G.: "*A Comprehensive Survey on Text Summarization Systems*", 2nd International Conference on Computer Science and its Applications, 2009, pp.1,6, 10-12.

[2] G.Salton, A.Wong and C.S. Yang, "*A vector space model for automatic indexing"*, Vol. 18, 1975, pp.613–620.

[3] R. Mihalcea,D. Radev,*Graph-based Natural Language Processing and Information Retrieval,* Cambridge University Press , 2011

[4] S.Beliga, A.Mestrovic, S. Martincic-ipsic, ,"An overview of graph-based keyword extraction methods and approaches", Journal of information and organizational sciences,Volume 39,2015, pp 1-20.

[5] S. Brin and L. Page., "*The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems"*, Vol. 30, 1998, pp.1–7.

[6] J.M. Kleinberg. "*Authoritative sources in a hyperlinked environment*". *Journal of the ACM*, 46(5), 1999, 604–632.

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

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

[9] Matsuo, Yutaka, Y. Ohsawa, and M. Ishizuka. "*Keyworld: Extracting keywords from document s small world*." Discovery Science. Springer Berlin Heidelberg, 2001, pp271-281.

[10] R. Mihalcea and P. Tarau , "*Textrank: Bringing order into texts*", In Lin, D., & Wu,D. (Eds.), Proceedings of EMNLP,2004, pp. 404.

[11] 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, 2005,pp 103-108

[12] C. Huang, Y. Tian, Z. Zhou, C.X. Ling, T. Huang "*Keyphrase extraction using semantic networks structure analysis*" in IEEE Int. Conf. on Data Mining, 2006, pp.275-284.

[13] Lahiri, Shibamouli, Sagnik Ray Choudhury, and Cornelia Caragea. "*Keyword and keyphrase extraction using centrality measures on collocation networks.*" arXiv preprint, 2014,arXiv:1401.6571

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

[15] 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, 2013, pp.1693-1705.

[16] Krishna, RVV Murali, and Ch Satyananda Reddy. "*Extractive Text Summarization Using Lexical Association and Graph Based Text Analysis*." Computational Intelligence in Data Mining—Volume 1. Springer India, 2016. 261-272.

[17] K. Toutanova, D. Klein, C. Manning, and Y. Singer: "*Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLTNAACL(2003)*", pp. 252-259.

[18] K. Toutanova, C. D.Manning: "*Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)"*, (2000), pp. 63-70.

[19] M. Litvak, M. Last, "*Graph-based keyword extraction for single-document summarization*" in ACM Workshop on Multi-source Multilingual Information Extraction and Summarization, 2008, pp.17-24.

[20] M. Litvak, M. Last, H. Aizenman, I. Gobits, A. Kandel, "*DegExt — A Language-Independent Graph-Based Keyphrase Extractor*" in Advances in Intelligent and Soft Computing V. 86, 2011, pp 121-130.

[21] F. Boudin, "*A comparison of centrality measures for graph-based keyphrase extraction*". in International Joint Conference on Natural Language Processing (IJCNLP), 2013, pp. 834-838.

[22] Kim SN, Medelyan O, Kan MY, Baldwin T. "*Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles*". In Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 21-26

[23] Edmundson, Harold P. "New methods in automatic extracting." *Journal of the ACM (JACM)* 16.2, 1969, 264-285.

[24] P. Over, W. Liggett: Introduction to DUC: "*An Intrinsic evaluation of Generic News Text Summarization Systems*", Proc. DUC workshop Text Summarization., 2002.

[25] C.-Y. Lin and E. H. Hovy, "*Automatic evaluation of summaries using N-gram co-occurrencestatistics, in Proc*". 2003 Language Technology Conference (HLT-NAACL), 2003, pp. 71–78.

Authors' Profiles

Mr. R.V.V. Murali Krishna is a faculty of information technology in G.V.P College of engineering (Autonomous), Visakhapatnam. He is currently working on Document Summarization.

Dr. Ch. Satyananda Reddy is an Associate Professor in CS&SE department, Andhra University, Visakhapatnam. His research interests include Software Engineering, Software Cost Estimation, Software Metrics and Document Summarization.

How to cite this paper: Murali Krishna V.V. Ravinuthala, Satyananda Reddy Ch.,"Thematic Text Graph: A Text Representation Technique for Keyword Weighting in Extractive Summarization System", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.8, No.4, pp.18-25, 2016. DOI: 10.5815/ijieeb.2016.04.03