Text document clustering using statistical integrated graph based sentence sensitivity ranking algorithm

G Kannan* and R Nagarajan

1Department of Computer Science, Government Arts and Science College, Manalmedu, India
2Department of Computer and Information Science, Annamalai University, India

*E-mail: kannanaucse@gmail.com

Abstract. The proposed methodology employs a novel statistical integrated graph-based sentence sensitivity ranking algorithm for text document clustering. Clustering of documents is a task of grouping a document automatically into a list of meaningful clusters; in order for the documents inside a group to share the same topic. In this paper, first, a novel integrated graph-based methodology using the sentence sensitivity ranking is proposed to extract keyphrases from the documents. In the standard statistical approach, keyphrases are extracted on the basis of the sentence sensitivity ranking; and in the graph-based method, the candidate keyphrases are automatically created as graphs by applying the sentence sensitivity ranking. With the aid of the top listed keyphrases, the documents clustering are carried out by implementing the proposed sentence sensitivity ranking algorithm. The simulation results reveal that the proposed graph-based text document clustering using statistical integrated graph-based sentence sensitivity ranking algorithm obtained the best results for clustering the text documents.

Keywords. Graph-based approach, Keyphrase extraction, Statistical method, Sentence sensitivity ranking algorithm, Text document clustering.

1. Introduction
Over the last decade, significant advances in computer technology have provided more rational and sophisticated configuration systems, making it impossible to manually assign keyphrases. Burst data is growing day by day and must be maintained and analyzed for effective use or processing. The data is available in image, spatial, and text formats. Most textual data is represented in many ways, including text, graphs, and predicates. In newspapers, social media is used for posting and messaging and contains all the information about the company, so keyword extraction plays a big role in the text mining process.

Unstructured data services are growing significantly due to the large broadcast, widespread, availability, and storage space provided by the Internet. As a result, it is difficult to evaluate these details and collect interesting facts. Text mining is a well-adapted tool for this purpose. Machine-based extraction and processing of textual information quickly capture useful information and skills. Of these methods, this paper finds text clustering to be a matter of concern. Document clustering is the job of automatically grouping documents into a list of meaningful clusters so that the group documents can share the same topic. In other words, in order to satisfy human needs and desires in the search and analysis of information, this approach offers a heterogeneous, ordered, synthesised text of standard languages. The method of extracting keywords from a text document or webpage is known as
extracting keywords [1]. A keyphrase extraction method is also used to retrieve information from a broad database on the basis of a particular query. Many real-world keyword extraction apps are available, such as site search, text summarization, Twitter trends, etc. Clustering is one of the most popular practices of data mining and is extensively researched in the sense of the text in order to arrange a large number of text documents. It has a wide range of features, including grouping, visualisation and arrangement of text documents. Text document clustering is essential for data indexing, database recovery, management and exploitation of extensive text data on the network.

Due to the increasingly rising usage of the Internet, more than enough user-generated websites and applications have been created that arrive every day. Most of the websites and applications are short text, such as Twitter, Quor, and StackOverflow, etc. The user-generated content is scattered faster and shorter. The handling of a large number of short user-generated content has become increasingly important to the web application service provider [2], [3]. Their control focuses on the retrieval and quality of the information. Many researchers have recently researched the extraction of keyphrases from short text texts and are called microblogs. Recently, microblogs have attracted users to chat and communicate with others. There are so many places for microblogs, but Twitter is one of the most common sites for microblogs. A number of experiments have been carried out in recent text analysis studies, and the speed of work in the future is also gradually rising. Various techniques for information retrieval are available. One of these is the manual extraction of keyphrases where the manual level reading and extraction of keywords for text can be done. However, this is time-consuming and wasteful in terms of human capital usage. The other approach is to automatically retrieve keyphrases as documents are read by computers. When it comes to the automation process, it's a very difficult job to articulate a lot of sentences using a few words. Keyphrase extraction, on the other hand, is actually one of the most well-known research techniques. Keyphrase extraction plays a key role in various areas such as data retrieval, document grouping, text resuming, and in different fields of information processing.

The remaining of this paper is organized as follows: Section 2 reviews the literature related to graph-based keyphrase extraction and text document clustering. Section 3 presents the graph analytics and various centrality measures. The proposed text document clustering using the integrated graph-based method and its implementation procedures are given in Section 4. The comparison results and discussion of the proposed algorithm with the existing methods are given in section 5. Finally, in Section V, a conclusion is drawn.

2. Related work
The rapidly growing web-based culture adds large amounts of data on the Internet. The growing amount of unstructured and vast amounts of data on the Internet makes people more readable but makes it very difficult to access and gain knowledge from it. Keyphrase extraction is a popular subject in the field of text processing, and there are several ways to extract keyphrases [1]. Mihalcea et al. described two innovative unsupervised approaches to extracting keywords and sentences from a document [4]. They converted the graphed text into structured data. They introduced a TextRank graph-based ranking model for processing text in documents and showed how this approach could be used to advantage in natural language text. The results obtained from the TextRank graph-based ranking model have been successfully compared to existing published results. Islam et al. propose a new extension method for extracting keywords that use a random walk model, taking into account the position of terms in the document, and the information gain of terms covers the entire document set [5]. In addition, the authors used a random walk model to incorporate mutual information in terms and extract keywords from the document. They created a random walk model with TextRank. Previously, there were several types of random walks that were successfully applied and developed in various applications such as citation analysis, social networks, and weblink analysis. Their work in the post-processing stage is similar to that of Mihalcea et al. [4].
Search engines work based on keyword extraction, such as entering keyphrases in the search bar and displaying more suggestions in the search bar. After searching, Google will show you some links related to that keyphrase. Brin et al. presented Google with a large search engine that enables a wide
range of applications for the frameworks available in hypertext [6]. Google is modelled to effectively crawl and index the web, so your search results will be more satisfying than the methods already available. They addressed the question of how to develop a practically large network with additional available information in hypertext.

Wang et al. describe sentiment analysis on Twitter based on a new graphical model, and the analysis data is crawled from Twitter [7]. They adopted the literal meaning of hashtags as semi-supervised information in enhancing classification settings to improve the performance of the presented graph model.

Wang et al. proposed a keyword extraction method based on WordNet and PageRank [7]. In this approach, candidate words are first presented as an undirected graph with links to node-related information and node relationships. They applied PageRank to undirected graphs to perform word-sense disambiguation (WSD). Then the author removed the word graph and reapplied PageRank to the resulting graph for keyword extraction. The results of their study show that WordNet and PageRank-based keyword extraction techniques are more efficient and practically applicable.

Zhou et al. described an approach for ranking keyword results against structured data [8]. Their work is based on a schema graph-based method for keyword search, including candidate network generators and their evaluation stages. By using this idea, a ranking process was performed on the keywords that resulted in optimized words from the document.

Wen et al. discussed different classification methods for keyword extraction [9]. In their job, they ran classifiers to extract keywords from news articles. They created a candidate keyword graph-based on the TextRank by calculating the similarity between words as the node transition probability. Then they estimated the word scores by the iterative method and finally selected the top N keywords as the final keywords. To extend the work of keyword extraction from graph-based models [9], many other researchers have dragged this area to various aspects of the graph.

Li et al. proposed a graph-based ranking method by using Wikipedia as external information for extracting short text keywords [10]. They introduced Wikipedia to improve the quality of short text content to address the poor lack of knowledge of short text content. Comprehensive studies conducted in their study have shown that a graph-based ranking approach can improve indicators of F major, recall, and precision. They also said that the TextRank graph-based ranking method is suitable for keyword extraction of short text content by leveraging the information available on Wikipedia.

Litvak et al. has implemented two modern techniques for identifying keywords from text documents: unsupervised and supervised graph-based syntax techniques [11]. They also compared both of these methods. The unsupervised method used the HITS algorithm to extract keywords from text and web documents. The supervised method used a traditional spatial vector model to extract keywords from a text document. Spatial vector models are trained using keywords in the document summary collection. According to Osawa et al., Keyword extraction methods can be narrowly divided into statistical approaches, machine learning approaches, linguistic approaches and other approaches [12]. In addition, the author identified the extraction of keywords in graph-based data and categorized different graph forms. The majority of analyses are performed in co-occurrence diagrams since it is simple to measure and create two terms at the same time. Osawa et al. proposed a KeyGraph algorithm that extracts keywords by describing the important points described in a document, without relying on additional sources such as natural language processing tools or a corpus of large documents [12]. The algorithm presented focuses on graph segmentation and the presentation of a word cluster co-occurrence between words in a text. Despite the fact that the algorithm presented does not use the average frequency of terms in the document corpus, the results of their experiments showed that the extracted keywords were more accurate.

Cao et al. explained how to enhance graph-based keyword extraction. There, the author proposed an approach to calculate the importance of co-occurrence words in a document and modelled it as a graph to identify more relative keyphrases [13]. The authors also introduced word interrelationships in the document to improve performance while extracting the average number of keywords from the document.

There are many approaches available, but accurate automated text document clustering is still a major challenge in the fields of natural language processing, text mining, and information retrieval. In order
to resolve these, this paper proposes to cluster the text documents in a large corpus using a novel integrated graph-based technique with a traditional statistical approach and the term sensitivity ranking. First, a traditional statistical approach is used in which words are extracted on the basis of sentence sensitivity ranking; and a further graph-based procedure is used when the candidate words are automatically drawn up as graphs by applying sentence sensitivity ranking. Finally, keyphrases are extracted on the basis of statistical indices and centrality measures. For hundreds of journal papers, a systematic review of the proposed integrated graph-based methodology is carried out to validate its performance. The results for the proposed integrated graph-based approach using sentence sensitivity rating for text document clustering in a large corpus are discussed in the experimental results section.

3. Graph analytics
This section describes the basic centrality measures that are important for understanding the definition of a graph-based keyphrase extraction technique. Further detail on the centrality measures can be found in [14]–[16].

An undirected graph can be mathematical represented as G = (V, E) with V as a set of nodes (vertices) and E as a set of edges (links) [17]. In a weighted graph, each edge connects two nodes x and y with an associated weight \( w_{xy} \) (a real positive integer) [16]. The centrality measures are calculated to allot the ranking score on all nodes.

3.1. Degree centrality
The number of edges incident on a node \( V_i \) is referred to as degree centrality \( C_D(V_i) \) of the node \( V_i \) [15]. Let \( N(V_i) \) be the set of nodes that co-occurrence with the node \( V_i \) and \( |N(V_i)| \) be the number of nodes in set \( N(V_i) \). Hence, the degree centrality \( C_D(V_i) \) for a node \( V_i \) is expressed as,

\[
C_D(V_i) = \frac{|N(V_i)|}{|V|-1}
\]

(1)

3.2. Closeness centrality
Closeness centrality \( C_C(V_i) \) of a node \( V_i \) is defined as the inverse of farness between the node \( V_i \) and all the other nodes in a connected group [15]. Let \( \sigma(V_i, V_j) \) be the number of the shortest paths from nodes \( V_i \) to \( V_j \). Let \( \sigma(V_i, V_j | V_k) \) be the number of those short paths which pass over the node \( V_k \). Thus, the normalised closeness centrality \( C_C(V_i) \) of a node \( V_i \) is expressed as,

\[
C_C(V_i) = \left( \frac{|V'|}{|V|-1} \right)^2 \sum_{V_j \in V'} \frac{1}{\sum_{V_j \in V', \text{distance}(V_i, V_j)}}
\]

(2)

where, \( V' \) is a set of nodes in a connected group and \( |V'| \) is the number of nodes in that connected group.

3.3. Betweenness centrality
Betweenness centrality \( C_B(V_i) \) of a node \( V_i \) is defined as the number of times that the node \( V_i \) functions as a bridge along the shortest path between two other nodes [15]. Let \( \sigma(V_j, V_k) \) be the number of the shortest paths from nodes \( V_j \) to \( V_k \). Let \( \sigma(V_j, V_k | V_i) \) be the number of those short paths which pass over the node \( V_i \). Thus, the normalised betweenness centrality \( C_B(V_i) \) of a node \( V_i \) is expressed as,

\[
C_B(V_i) = \frac{\sum_{V_j \neq V_i \neq V_k \in V} \sigma(V_j, V_k | V_i)}{(|V|-1)(|V|-2)/2}
\]

(3)
4. Proposed text document clustering using the integrated graph-based method

This section describes a novel integrated graph-based method with the statistical approach for text document clustering that focuses on the journal and research articles. For each candidate keyphrase, besides measuring the centralities using the graph-based approach, the proposed integrated graph-based text document clustering approach also uses the sentence sensitivity to extract the candidate keyphrases. The proposed integrated graph-based text document clustering algorithm consists of two stages namely keyphrase graph construction and text document clustering which are described as follows,

4.1. Keyphrase graph construction

Let D be a target text document in a large corpus. The keyphrase graph is constructed as per the below procedures, in which the nodes denote the candidate keyphrase and the edges denote the direct relationship between the nodes.

Step 1: Read the target document first and split the paragraphs into separate sentences using the full stop delimiter.
Step 2: Words added to the graph are restricted with syntactic filters, which select only lexical units of a certain Part-of-Speech (nouns and adjectives). Sequences of adjacent words, restricted to nouns and adjectives only, are considered as candidate words. Use a stop words list (e.g., articles “a”, “an”, “the” in the stop words list) to remove the stop words in the document, because stop words hardly not having any information about the document.
Step 3: Two candidate words are considered redundant if they have the same stemmed form (e.g. “precisions” and “precision” are both stemmed to “precision”). This process is called it as stemming. Because this process identifies the root word, all candidate words are automatically pre-processed.
Step 4: After stemming, each candidate word in the document is treated as the node and occurrence between the candidate words is treated as edges. By using node and edge information, a keyphrase graph is constructed for the document.
Step 5: For each candidate word, keyphrase frequency (KF) and keyphrase sentence frequency (KSF) are computed. KF represents the total number of times a candidate word occurs in the stemmed document. Meanwhile, KSF denotes the total number of sentences where a selected candidate word presents.
Step 6: Rank the candidate keyphrases based on WF and it is defined as frequency ranking. If any of the candidate words have the same KF value, then the succeeded word among the candidate keyphrases is selected on the basis of the word having maximum KSF value. This is referred to as sentence sensitivity ranking.
Step 7: In this step, the top-ranked keyphrases are extracted with the aid of frequency and sentence sensitivity rankings. Using these keyphrases, the keyphrases graph is construction.

4.2. Text documents clustering using extracted keyphrases

In this stage, the extracted top-ranked keyphrases from all documents in a large corpus are used to cluster the documents into groups using Euclidean distance between the cluster centre and the documents.

Distance measure

Euclidean distance is a normal metric in the field of the text cluster used to quantify the distance (differences) between documents D and their centroids clusters C as expressed,

\[
\text{Min } \text{Dis}(D, C) = \sqrt{\sum_{j=1}^{d} (D_d - C_j)^2}
\]  

(4)
5. Experimental results

This section provides some experimental evaluation of the proposed graph-based text document clustering. Hundreds of text document clustering in a large corpus can only be evaluated based on the accuracy that helps predict or classify a particular document into the appropriate class. To evaluate the proposed statistical integrated graph-based sentence sensitivity ranking algorithm for text document clustering, the m-script coding is implemented in MATLAB using the Intel Core i7 processor, 500 GB hard disk, and 8 GB RAM system specification.

For an illustration of the proposed statistical integrated graph-based sentence sensitivity ranking algorithm, the documents considered for analysis are six journal papers (document ID: D1, D2, D3, D4, D5 and D6) in the area of text document clustering and power systems [1]–[3], [18]–[20]. As per the procedures given in section 4, the keyphrase graphs are constructed for all the six text documents. Figure 1 to 6 shows the candidate keyphrase graph for six text documents that are extracted using the statistical integrated graph-based sentence sensitivity ranking algorithm.

### Table 1. Top listed keyphrases in six documents with its keyphrase frequency (KF)

| S.No | Keyphrase       | Document ID |
|------|-----------------|-------------|
|      |                 | D1  | D2  | D3  | D4  | D5  | D6  |
| 1    | generation micro| 5   | 5   | 0   | 0   | 0   | 0   |
| 2    | micro grid      | 6   | 6   | 12  | 0   | 0   | 0   |
| 3    | eagle strategy  | 5   | 0   | 8   | 0   | 0   | 0   |
| 4    | strategy particle| 8  | 0   | 8   | 0   | 0   | 0   |
| 5    | particle swarm  | 9   | 0   | 8   | 0   | 0   | 0   |
| 6    | keyword extraction| 0 | 0   | 0   | 8   | 0   | 8   |
| 7    | extraction keyword| 0 | 0   | 0   | 1   | 0   | 2   |
| 8    | keyphrase extraction| 0 | 0   | 0   | 0   | 13  | 0   |
| 9    | term clustering  | 0   | 0   | 0   | 0   | 12  | 0   |
| 10   | clustering keyphrase| 0 | 0   | 0   | 0   | 12  | 0   |

Table 1 shows the top-listed keyphrases with its KF for all the six text documents. Figure 7 shows the document clustering output of the proposed graph-based text document clustering using graph-based sentence sensitivity ranking algorithm. The experimental study was conducted on six journal papers and the findings revealed that using the graph-based sentence sensitivity ranking algorithm, the presented graph-based text document clustering approach would efficiently cluster the journal documents. The proposed graph-based text document clustering model using graph-based sentence sensitivity ranking algorithm may therefore be better adapted for document clustering from the different databases or online records of journals.

In order to validate the accuracy and clustering speed of the proposed text document clustering method using statistical integrated graph-based sentence sensitivity ranking algorithm, a hundred documents in a corpus is considered for analysis. The corpus consists of fifty research articles from the area of computer science and fifty documents from the field of electrical engineering. The clustering result of hundred text document in a large corpus is depicted in Figure 8. The correctly classified documents in each class (computer science and electrical engineering), the overall accuracy of clustering the documents, and the clustering speed of the proposed statistical integrated graph-based sentence sensitivity ranking algorithm are compared with the statistical approach and graph-based method. The comparison results are tabulated in Table 2.
Figure 1. Keyphrase graph – D1

Figure 2. Keyphrase graph – D2

Figure 3. Keyphrase graph – D3
Figure 4. Keyphrase graph – D4

Figure 5. Keyphrase graph – D5

Figure 6. Keyphrase graph – D6
Figure 7. Graph-based text document clustering for six documents.

Figure 8. Graph-based text document clustering for a hundred documents.

Table 2. Comparison result of proposed integrated method with statistical approach and graph-based method

| Method               | Number of documents in a corpus | Correctly classified documents | Overall accuracy (%) | Clustering speed (sec) |
|----------------------|---------------------------------|--------------------------------|----------------------|------------------------|
| Statistical method   | 100                             | 42                             | 85                   | 29                     |
| Graph-based method   | 100                             | 44                             | 90                   | 17                     |
| Proposed integrated  | 100                             | 48                             | 96                   | 12                     |
From Figure 8 and Table 2, it is clearly seen that the overall text document clustering accuracy of the proposed integrated graph-based sentence sensitivity ranking algorithm is better than the existing methods. It also noted that the proposed integrated approach is the quickest method of text document clustering from a large corpus of documents.

6. Conclusions
In this paper, a novel integrated graph-based method with a statistical approach using frequency and keyphrase sensitivity ranking schemes for text documents clustering in a large corpus has been proposed. In addition to the keyphrase frequency count, the keyphrase frequency count of a keyphrase was considered to be the candidate nodes. The proposed integrated graph-based approach with a statistical ranking scheme utilises the advantages of both a graph-based approach and a statistical approach. Under this assumption, it is clear that the accuracy of the clustering of documents has been increased. In addition, a hundred different journal articles were used to validate the proposed novel graph-based model. Among these hundred documents, the experimental results of the six documents were presented. From an experimental analysis, the novel integrated graph-based approach using sentence sensitivity ranking has been shown to outperform other existing graph-based approaches and statistical methods. Hence, the proposed graph-based text document model can be best suits for document clustering from the various journals database or web documents.

7. References
[1] R. Nagarajan, S. Nair, P. Aruna, and N. Puviarasan, “Keyword extraction using graph based approach,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 10, no. 10, 2016.
[2] G. Kannan and R. Nagarajan, “Standard Statistical and Graph based Automatic Keyword Extraction,” Int. J. Innov. Technol. Explor. Eng., vol. 9, no. 2, pp. 5013–5019, Dec. 2019, doi: 10.35940/ijitee.B7601.129219.
[3] G. Kannan and R. Nagarajan, “Key Phrase Extraction by Term Clustering using Proposed Graph Based Ranking Model Method,” Int. J. Emerg. Trends Eng. Res., vol. 8, no. 9, pp. 5000–5005, 2020.
[4] R. Mihalcea and P. Tarau, “Textrank: Bringing order into text,” in Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.
[5] M. R. Islam and M. R. Islam, “An improved keyword extraction method using graph based random walk model,” in 2008 11th International Conference on Computer and Information Technology, 2008, pp. 225–229.
[6] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” 1998.
[7] J. Wang, J. Liu, and C. Wang, “Keyword extraction based on pagerank,” in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2007, pp. 857–864.
[8] J. Zhou, X. Yu, Y. Liu, and Z. Yu, “Ranking keyword search results with query logs,” in 2014 IEEE International Congress on Big Data, 2014, pp. 770–771.
[9] Y. Wen, H. Yuan, and P. Zhang, “Research on keyword extraction based on word2vec weighted textrank,” in 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 2016, pp. 2109–2113.
[10] W. Li and J. Zhao, “TextRank algorithm by exploiting Wikipedia for short text keywords extraction,” in 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), 2016, pp. 683–686.
[11] M. Litvak and M. Last, “Graph-based keyword extraction for single-document summarization,” in Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization, 2008, pp. 17–24.
[12] Y. Ohsawa, N. E. Benson, and M. Yachida, “KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor,” in Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL ’98, 1998, pp. 12–18.
[13] J. Cao, Z. Jiang, M. Huang, and K. Wang, “A way to improve graph-based keyword extraction,” in *2015 IEEE International Conference on Computer and Communications (ICCC)*, 2015, pp. 166–170.

[14] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, “An overview of graph-based keyword extraction methods and approaches,” *J. Inf. Organ. Sci.*, vol. 39, no. 1, pp. 1–20, 2015.

[15] F. Boudin, “A comparison of centrality measures for graph-based keyphrase extraction,” in *Proceedings of the sixth international joint conference on natural language processing*, 2013, pp. 834–838.

[16] S. Beliga, “Keyword extraction: a review of methods and approaches,” *Univ. Rij. Dep. Inform. Rij.*, pp. 1–9, 2014.

[17] R. Blanco and C. Lioma, “Graph-based term weighting for information retrieval,” *Inf. Retr.*, vol. 15, no. 1, pp. 54–92, 2012.

[18] K. Santhosh and R. Neela, “Optimization of Distributed Generation in Micro Grid using a Hybrid Metaheuristic Technique,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 9, pp. 5104–5110, 2020, doi: 10.30534/ijeter/2020/36892020.

[19] K. Santhosh and R. Neela, “Optimal Placement of Distribution Generation in Micro-Grid using Eagle Strategy with Particle Swarm Optimizer,” *Int. J. Pure Appl. Math.*, vol. 118, no. 18, pp. 3819–3825, 2018.

[20] S. Kasi and R. Neela, “Operation Cost Minimization of Micro Grid using Particle Swarm Optimizer and Eagle Strategy Micro Grid’s Operation Cost Minimization using PSO and ES,” in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Oct. 2020, pp. 1207–1212, doi: 10.1109/I-SMAC49090.2020.9243485.