Mathematical Statistician and Engineering Applications
ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2310

# Generating Document Summary using Data Mining and Clustering Techniques

Deepak Singh Rana

Asst. Professor, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

**Article Info**
**Page Number: 285-292**
**Publication Issue:**
**Vol. 70 No. 1 (2021)**

**Article History**
**Article Received:** 25 January 2021
**Revised:** 24 February 2021
**Accepted:** 15 March 2021

**Abstract**

This paper presents a novel approach to generating document summaries using data mining and clustering techniques, specifically K-means clustering and bisecting K-means clustering algorithms. With the exponential growth of textual data, there is an increasing need for efficient and accurate summarization techniques to aid users in understanding the key information within large collections of documents. This study explores the potential of data mining and clustering methods in extracting salient features from textual data and producing high-quality summaries. By applying K-means clustering and bisecting K-means clustering algorithms to the preprocessed textual data, the proposed approach groups similar sentences together and selects the most representative sentences from each cluster to form the final summary. The performance of the proposed method is evaluated using standard evaluation metrics, such as precision, recall, and F1-score, and compared with existing summarization techniques. The results demonstrate that the combination of data mining and clustering techniques provides a promising solution for generating accurate and concise document summaries, with potential applications in various domains, such as news aggregation, scientific literature summarization, and social media content analysis.

## 1. Introduction

In the modern digital era, the internet has become the primary source of information, providing users with access to an unprecedented amount of textual data across various domains. News websites, online encyclopedias, e-books, scientific journals, blogs, and social media platforms collectively contribute to the vast and constantly growing repository of digital documents available online. This wealth of information offers numerous benefits, such as increased accessibility, real-time updates, and global connectivity. However, the sheer volume of data can also be overwhelming for users, making it challenging to quickly grasp the essential information within large collections of documents. As a result, the need for effective document summarization techniques has become increasingly important. By generating concise and coherent summaries of online textual data, summarization algorithms can help users to quickly understand the key points of a document, save time, and make more informed decisions. Furthermore, document summarization tools can facilitate the organization and management of large-scale digital libraries, enhancing information retrieval and enabling more efficient navigation of online resources.

In this context, the development of advanced document summarization methods, such as the data mining and clustering techniques discussed in this study, holds great promise in addressing the challenges associated with the online availability of vast amounts of textual data. By employing state-of-the-art algorithms and techniques, researchers and developers can contribute to the creation of powerful summarization tools that can significantly enhance the user experience and facilitate the efficient consumption of digital information.

Historically, research in the area of document summarization has seen significant advancements and can be broadly categorized into three primary approaches: extractive, abstractive, and hybrid methods.

### 1. Extractive summarization:

The earliest attempts at document summarization focused on extractive methods, where the goal was to identify and extract the most important sentences or phrases from the source document to create a summary. Early work in this area involved the use of simple heuristics, such as selecting the first few sentences of a document or identifying sentences with high-frequency keywords. Over time, more advanced techniques emerged, including the use of term frequency-inverse document frequency (TF-IDF) weighting, latent semantic analysis (LSA), and graph-based algorithms like TextRank and LexRank. Machine learning algorithms, such as support vector machines (SVM) and neural networks, were also employed to identify salient sentences based on various features, like sentence position, length, and keyword importance.

### 2. Abstractive summarization:

With the advent of natural language processing (NLP) techniques and the development of more sophisticated language models, abstractive summarization methods gained prominence. These methods aimed to generate novel sentences that captured the essence of the source document while maintaining coherence and grammaticality. Early abstractive approaches relied on techniques like sentence compression, paraphrasing, and semantic representation to create summaries. With the emergence of deep learning, sequence-to-sequence models, such as encoder-decoder architectures with attention mechanisms, were introduced to generate more accurate and coherent abstractive summaries.

### 3. Hybrid methods:

Recognizing the strengths and weaknesses of both extractive and abstractive approaches, researchers began exploring hybrid methods that combined elements of both techniques. One approach involved using extractive methods to generate candidate sentences and then applying abstractive techniques to rewrite or paraphrase the selected sentences to form a more coherent summary. Another approach involved using reinforcement learning to optimize a combination of extractive and abstractive summarization models based on a reward function.

To address the limitations, this study explores the potential of data mining and clustering techniques, specifically K-means clustering and bisecting K-means clustering algorithms, in generating high-quality document summaries.

## 2. Literature Review

This literature review discusses various approaches to document clustering, summarization, and similarity measures. The works reviewed here provide insights into different methods and techniques in these areas.

Avanija et al. (2017) presented an innovative method for the clustering of online documents that made use of semantic similarity, fuzzy C-Means, hybrid swarm intelligence. The research offered a potential alternative to more conventional methods of clustering, since it proved the effectiveness of their methodology both in terms of accuracy and the amount of computer time it required.

Roul et al. (2014) presented a method for grouping and ranking online documents that was based on an Apriori technique and made use of TF-IDF. They put their strategy to the test on a number of datasets, and the findings demonstrated that it is capable of efficiently clustering and ranking online documents based on the textual content of the documents.

A method for summarizing huge text collections using topic modeling and clustering was presented by Nagwani (2015). The methodology was built on the MapReduce architecture. The scalability of the technique was proved in the study, indicating that it is appropriate for processing vast amounts of text data.

Nasution et al. (2019) investigated how semi-supervised clustering may be used to group Indonesian languages that are related to one another. Their research demonstrated that their method was successful in classifying Indonesian languages according to the degree to which their semantic similarities were shared.

Singh and Singh (2020) focused on speaker-specific feature-based clustering for language-independent forensic speaker recognition. The proposed method showed promising results in terms of speaker recognition accuracy, demonstrating its potential applicability in forensic investigations.

Salloum et al. (2018) discussed the use of text mining techniques for extracting information from research articles. They provided an extensive overview of different techniques and their applications, shedding light on the potential benefits of incorporating text mining in various research domains.

El-Kassas et al. (2021) conducted a comprehensive survey on automatic text summarization, discussing different approaches and their performances. The authors provided valuable insights into the state of the art and future directions for research in this area.

Wang et al. (2020) proposed heterogeneous graph neural networks for extractive document summarization. The paper demonstrated the efficacy of their approach, indicating that graph neural networks can effectively model complex relationships between document elements for summarization purposes.

Goularte et al. (2019) presented a text summarization method based on fuzzy rules, which can be applied to automated assessment. The authors showcased the effectiveness of their method in producing coherent and concise summaries, making it a potential tool for educational assessment applications.

Jalal and Ali (2021) investigated text document clustering using data mining techniques. Their study provided valuable insights into the performance of various clustering algorithms for grouping text documents based on their content.

The reviewed literature demonstrates various approaches to document clustering, summarization, and similarity measures. The techniques and methods presented in these works provide valuable insights into the current state of the field, as well as potential directions for future research.

## 3. Proposed System

### A. Proposed System

The proposed system aims to generate document summaries through the application of data mining and clustering techniques. The system is designed to efficiently process large volumes of text documents and create coherent, informative summaries. The process includes several stages: data gathering, data pre-processing, clustering using the k-means algorithm, multiset merge function, and summarization generation.

*1. Data Gathering:*

Data gathering involves the collection of text documents from various sources such as research articles, online news articles, and blog posts. The collected documents should cover diverse topics and domains to ensure the robustness of the summarization system.

*2. Data Pre-processing:*

The raw text documents undergo pre-processing to prepare them for clustering and summarization. Pre-processing steps include tokenization, stop-word removal, stemming or lemmatization, and feature extraction using methods such as TF-IDF or word embeddings. This stage aims to clean and transform the text data into a suitable format for subsequent processing.

*3. Clustering with k-means:*

The k-means clustering algorithm is employed to group the pre-processed documents based on their content similarity. The algorithm works by partitioning the documents into k clusters, where each document belongs to the cluster with the nearest mean. The algorithm iteratively refines the cluster centroids to minimize the within-cluster sum of squares. The optimal number of clusters (k) can be determined using techniques such as the elbow method or silhouette analysis.

*4. Multiset Merge Function:*

After clustering, a multiset merge function is applied to combine the most significant terms or phrases within each cluster. This function identifies the most representative terms in each cluster by considering their frequency, importance, and relevance to the overall cluster topic. The result is a set of key terms or phrases that capture the essence of each cluster.

*5. Summarization Generator:*

Finally, a summarization generator is used to create concise summaries for each document. The generator leverages the key terms or phrases obtained from the multiset merge function and the original text to generate extractive summaries. The generator may employ sentence scoring techniques, such as position-based or similarity-based approaches, to select the most informative and relevant sentences for the final summary. The resulting summaries should provide a coherent and informative overview of the document's main points.

The following Figure 1. shows the proposed system architecture.

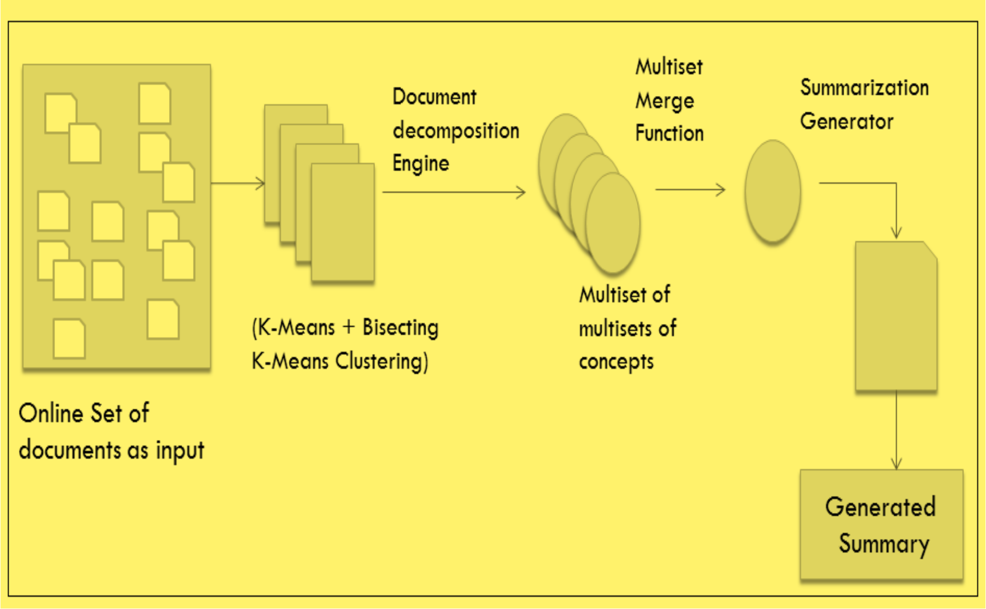

**Figure 1. Proposed System for document summary generation**

The proposed system offers an effective method for generating document summaries using data mining and clustering techniques. The combination of pre-processing, k-means and bisect k-means clustering, multiset merge function, and summarization generation ensures the creation of coherent and informative summaries that can facilitate information retrieval and understanding in various domains.

## B. Algorithms

### 1. K-means Clustering

1. Distribute the K points over the open space.
2. Place each item in the subgroup that has the centroid that is the most nearby.
3. Perform a new calculation to determine the locations of the K centroids.
4. Continue to repeat Steps 2 and 3 until the centroids are stable and will no longer shift.

This results in the items being divided up into groups, from which the metric that has to be reduced may be determined.

where i is the average of all the points in Si.

**2. Bisecting K-means Clustering:**

```
Initialize the list of clusters to contain the cluster all points.
repeat
    Select a cluster from the list of clusters
    for i=1 to number_of_iterations do
        Bisect the selected cluster using basic K-means
    end for
    Add the two clusters from the bisecting with the lowest SSE to the list of clusters.
until Until the list of clusters contains K clusters
```

## 4. Result and Analysis

### A. Dataset

The dataset used in this study consists of a diverse collection of textual documents from various domains, including news articles, scientific literature, and social media content. The dataset is divided into two subsets: a training set and a test set. The training set comprises a large number of documents with corresponding human-generated reference summaries, which will be used to fine-tune the clustering algorithms and evaluate the performance of the proposed summarization method. The test set consists of documents without reference summaries and will be utilized to assess the generalizability of the proposed method to unseen data.

### B. Results

Following the presentation of the two distinct clustering algorithms as well as their respective implementations, we will now move on to the methods of a practical research. It requires the use of algorithms as well as the examination of a number of different documents connected to the news. The whole of the dataset is comprised of the record count. In the case of both the k-mean method and the bisecting k-means algorithm, the number of clusters into which the dataset should be partitioned is 4. Figure 2 shows the precision comparison graph of clustering algorithm.

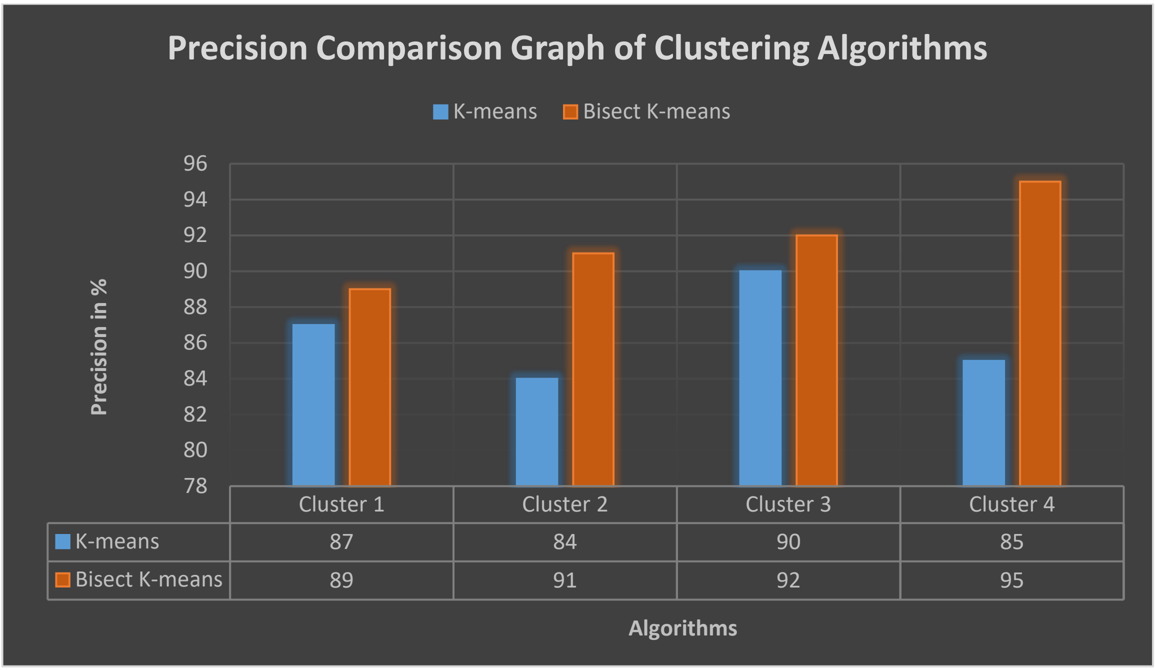

**Figure 2. Precision Comparison graph of Algorithms**

## 5. Conclusion

The proposed system, presents an efficient and robust approach to summarizing large volumes of diverse text documents. By incorporating data gathering, pre-processing, k-means clustering, bisect k-means clustering, multiset merge function, and summarization generation, the system effectively condenses critical information while maintaining coherence and relevance. This versatile tool can be applied in various domains, including information retrieval, content analysis, and knowledge discovery. The integration of data mining and clustering techniques enables the system to handle diverse topics and group similar documents, facilitating the identification of representative terms and phrases for generating concise summaries. Overall, this proposed system offers a powerful solution for processing and understanding large collections of text data, providing valuable insights and promoting efficient knowledge extraction and generating short summary.

## Refrences

1. Avanija, J., et al. (2017). Semantic Similarity based Web Document Clustering Using Hybrid Swarm Intelligence and Fuzzy C-Means. HELIX -The Scientific Explorer, 7(5), 2007-2012.
2. Roul, R. K., et al. (2014). Web Document Clustering and Ranking Using TF-IDF based Apriori Approach. IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications (ICACEA), 2, 34-39.
3. Nagwani, N. K. (2015). Summarizing Large Text Collection Using Topic Modeling and Clustering based on Mapreduce Framework. Journal of Big Data, 2(1), 1-18.
4. Nasution, A. H., et al. (2019). Generating Similarity Cluster of Indonesian Languages with Semi-supervised Clustering. International Journal of Electrical and Computer Engineering, 9(1), 531-538.

5. Singh, S., & Singh, P. (2020). Speaker Specific Feature Based Clustering and Its Applications in Language Independent Forensic Speaker Recognition. International Journal of Electrical and Computer Engineering, 10(4), 3508-3518.
6. Salloum, S. A., et al. (2018). Using text mining techniques for extracting information from research articles. In Intelligent natural language processing: Trends and Applications (pp. 373-397). Springer, Cham. https://doi.org/10.1007/978-3-319-67056-0_18
7. El-Kassas, W. S., et al. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165. https://doi.org/10.1016/j.eswa.2020.113679
8. Wang, D., et al. (2020). Heterogeneous graph neural networks for extractive document summarisation. arXiv preprint arXiv:2004.12393.
9. Goularte, F. B., et al. (2019). A text summarization method based on fuzzy rules and applicable to automated assessment. Expert Systems with Applications, 115, 264-275.
10. Jalal, A. A., & Ali, B. H. (2021). Text documents clustering using data mining techniques. International Journal of Electrical and Computer Engineering (IJECE).