Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print), ISSN:2521-3520 (online) Volume 34, Issue 1, 2023 DOI: http://doi.org/10.23851/mjs.v34i1.1195

Research Article **Open Access**

Text Summarizing and Clustering Using Data Mining Technique

Zainab Abdul Wahid Salman

Department of Computer, College of Science, Mustansiriyah University, Baghdad, IRAQ.

Contact: drzsalman@uomustansiriyah.edu.iq

Article Info

Received
25/08/2022

Accepted
04/12/2022

Published
30/03/2023

ABSTRACT

Text summarization is an important research topic in information technology because of the large volume of texts and the large amount of data on the Internet and social media. As a result, summarizing the text has gained significant importance, requiring finding highly efficient ways to extract knowledge in various fields. Thus, there was a need for methods of summarizing texts for one document or multiple documents. Furthermore, the summarization methods aim to obtain the main content of the set of documents simultaneously to reduce redundant information. This paper proposes an efficient method to summarize texts that depends on the word association algorithm to separate and merge sentences after summarizing them. As well as the use of data mining technology in redistributing information according to the (K-Mean) algorithm and the use of (Term Frequency Inverse Document Frequency TF-IDF) technology for measuring the properties of summarized texts. The experimental results found that the summarization ratios are good by deleting unimportant words. Also, extracting characteristics for texts was useful in grouping similar texts into clusters, which makes this method possible to be combined with other methods in artificial intelligence, such as fuzzy logic or evolutionary algorithms, in increasing summarization rates and accelerating cluster operations.

KEYWORDS: Information systems; texts summary; large data; learning machine; K-, TF-IDF means.

INTRODUCTION

Recently, text summarization has become an essential issue in the information systems field after there was a problem with the excessive cost of storing large data files. Therefore, it was sufficient to store and schedule document summaries instead of storing full texts and searching in new tables instead of storing the entire file [1]. This summarization process is represented by the content produced through a set of operations on the texts to find texts of less size and meaning close to the original text [2]. As well as collecting files of similar information and extracting only the vital information to be added to the summary. When the user searches for information via the query, the search will provide many files that match the result with the relevant content in the query; in return, the user will waste time searching for related content [3].

This problem grows exponentially in proportion to the increase in the flow of information and the increase in the number of texts and from different information sources, as summarizing the text is a way to retrieve information from several documents, where the output will be a generally processed text document with an extension required accurate content whenever the user inquires, depending on the nature of the text Representation in documents, for which the abstract can be categorized as a summary [4]. Although the method of extracting a sentence is not the usual way for humans to create summaries of documents, some sentences in documents represent some aspect of their contents to some extent. Speed is also essential when integrating a web-based summary [5]. Therefore, extraction-based summarization is still helpful on the web. A multi-document extractive process can be precisely formulated as extracting significant text units from multiple related documents, eliminating redundancy, and rearranging the units to produce an efficient summary. An alternative approach to ensure good coverage and avoid redundancy is a clustering-based approach that groups similar text

units (paragraphs, sentences) into multiple groups to identify common information themes and selects text units one by one from clustering to the final summary [6]. Each block consists of similar text units representing a subtopic (theme). Domain independence and language dependence are the major features of the aggregation-based approach to multi-document text summarization. This paper presents a multi-document text summarization system, which aggregates sentences using a similarity-based transmission algorithm to select multiple subtopics (topics) from a relevant document input set and selects representative sentences from appropriate combinations to form the summary [7].

There are large groups of texts in various fields in digital information systems, and it is always necessary to obtain technology that helps retrieve information as quickly and accurately as possible—for example, working in search engines to retrieve information. However, it requires reducing the search field and searching only valuable areas. Therefore, texts and information are collected by summarizing them and then grouping documents that are similar or similar in content into separate packages, which facilitates the process of searching and retrieval.

The current research aims to solve the problems of Retrieving information by using two methods: studying the grouping of similar texts in separate tables and relying on summarizing them in documents and forming groups with similar characteristics. The second is using the frequency term - inverse document frequency (TF-IDF) to measure the degree of congruence between texts.

RELATED WORK

There are several papers related to the topic of summarizing and scheduling texts in diverse ways, a group of which has been selected as previous works, as follows:

- Fabio Bif Goulartea et al. in 2019, researchers introduced an automated text evaluation process that relies on fuzzy logic and a variety of extractive features to find the most critical information in the evaluated texts. The summaries produced for these texts are compared with the reference summaries created by experts in the field. Unlike other proposals in the literature, this method summarizes by checking for correlation to reduce dimensionality and, thus, the number of vague rules used to summarize the text. Thus, the proposed approach of summarizing text with a small number of ambiguous grammars could benefit the development and use of future expert systems capable of evaluating writing automatically [8].
- Sanchez-Gomez and et.al. in 2020. Presented a multi-document extractive study that summarizes a method aiming to obtain the main content of a set of documents to reduce redundant information simultaneously. This can be addressed from an optimization point of view. There is a lack of multi-objective methods applied in this context through the Multipurpose of the Artificial Bee Colony (MOABC) algorithm for this task. Experiments were conducted using datasets from the Document Understanding Conference (DUC). Model performance was evaluated using the Recall-Oriented Understudy (ROUGE) scales, typical in this cognitive domain [9].
- Rasim M. Alguliyev et.al. in 2019. Suggested a method for aggregation and optimization through the evolutionary algorithm of text summarization. To summarize the text, a two-stage sentence selection model was proposed based on Clustering and optimization techniques called COSUM. For the first stage, the set of sentences is compiled using the k-mean method to find out all the topics in the text. As for the second stage, to select the distinguished sentences from the groups, an improvement model is proposed. This model optimizes an objective function expressed as a harmonic average of objective functions that enforce the coverage and diversity of the sentences chosen in the abstract. Furthermore, to provide the readability of the summary, this form also controls the length of sentences specified in the candidate's summary. Finally, to solve the optimization problem, an adaptive differential evolution algorithm with a new mutation strategy was developed.
- R. Janani and S. Vijayarani et. al. in 2018. Suggested a brand-spectral clustering technique using particle swarm optimization (SCPSO). The starting population is randomized while considering global and local optimization functions. This project aims to combine spectral clustering with swarm optimization to manage the enormous volume of text documents. The benchmark database compares the proposed algorithm SCPSO against other current methods. Furthermore, comparisons are made between the Spherical K-means, Expectation Maximization

ISSN: 1814-635X (print), ISSN:2521-3520 (online) Volume 34, Issue 1, 2023 DOI: http://doi.org/10.23851/mjs.v34i1.1195

Method (EM), and the traditional PSO Algorithm with the proposed algorithm SCPSO [11].

- C. Kruengkrai and C. Jaruskulchai. Suggested a workable method for selecting the key phrases from the original text to create a summary. Our strategy is designed to take advantage of sentences' local and universal characteristics. The global property can be considered the relationships between each sentence in the document, whereas the local property can be thought of as groups of significant words within each sentence. These two characteristics are combined to create a single metric that captures the informativeness of sentences. Our method outperforms a commercial text summarizer, according to experimental results [12].

MATERIALS AND METHODS

In this research, an important method has been proposed for text summarization. Organizing the results into groups with similar characteristics facilitates the process of searching and retrieving the stored information as intermediate tables are created that contain the information that is summarized and searched instead of searching in the raw data, which requires a long time and ample storage space. Figure 1 shows the general scheme of the proposed method.

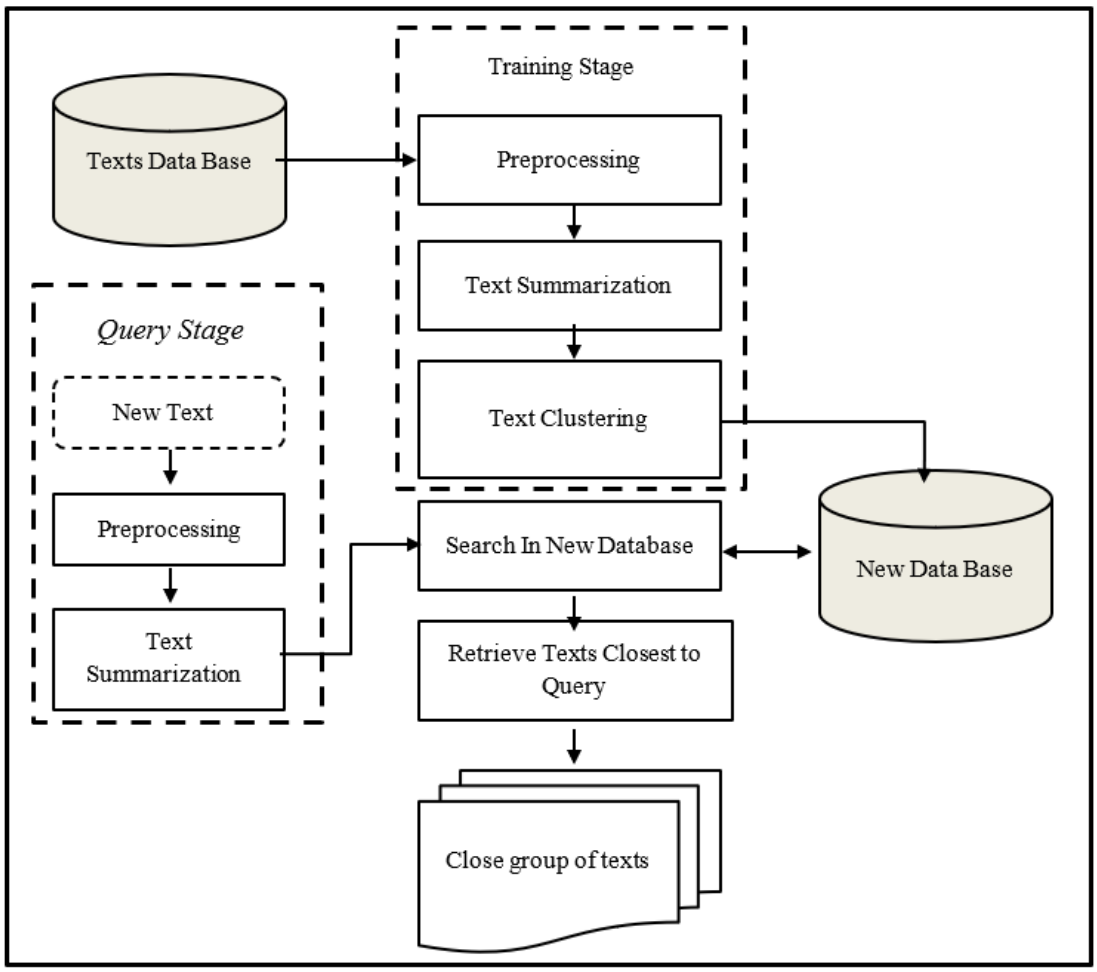

Figure 1. general scheme of the proposed method.

1. Preprocessing

The first step in the initial processing is the tokenization process, which includes cutting the text into smaller units known as the tokens according to separating signs such as spaces, sentence endpoints, separators, annotations, and parentheses. As a result, we have a set of parts, as shown in Table 1.

Table 1. An example of tokenization process.

Text	Cybersecurity is critical to the national infrastructure, federal and local government, military, industry, and personal privacy.
Tokens	“Cybersecurity”, “is”, “critical”, “to”, “the”, “national”, “infrastructure”, “federal”, “and” “local”, “government”, “military”, “industry”, “and”, “personal”, “privacy.”

The next step is to remove stop words like conjunctions. Each word in the document is compared to the list of stop words below and replaced with a blank space. As follow*:*

"a", "about", "above", "after", "again", "against", "all", "am", "an", "and", "any", "are", "aren't", "as", "at", "be", "because", "been", "before", "being", "below", "between", "both", "but", "by", "can't", "cannot", "could", "couldn't", "did", "didn't", "do", "does", "doesn't", "doing", "don't", "down", "during", "each", "few", "you've", "your", "yours", "yourself", "yourselves"…….

After that, the case of unification takes place, which is called Case Folding. In this step, the case of all symbols is converted to small letters. The last step in the initial processing is returning words to their roots, called stemming. In this step, the prefix and suffix additions to words are removed. Stemming is a technique used to extract the primary form of words by removing suffixes from them. It is like cutting the branches of a tree down to its roots. For example, the root word (eating, eats, eaten) is (eat).

Text Summarization

The text summarization process is based on a two-step word association algorithm. First, sort the correlation by descending order, calculate the correlation convergence rate, model the data network elements using a graph, and define (K-Nearest Neighbor), which means the relationship with nearby texts in words that are useful in the assembly process.

Secondly, splitting and merging take place according to the aggregation factor within one text and between other texts, as shown in Figure 2.

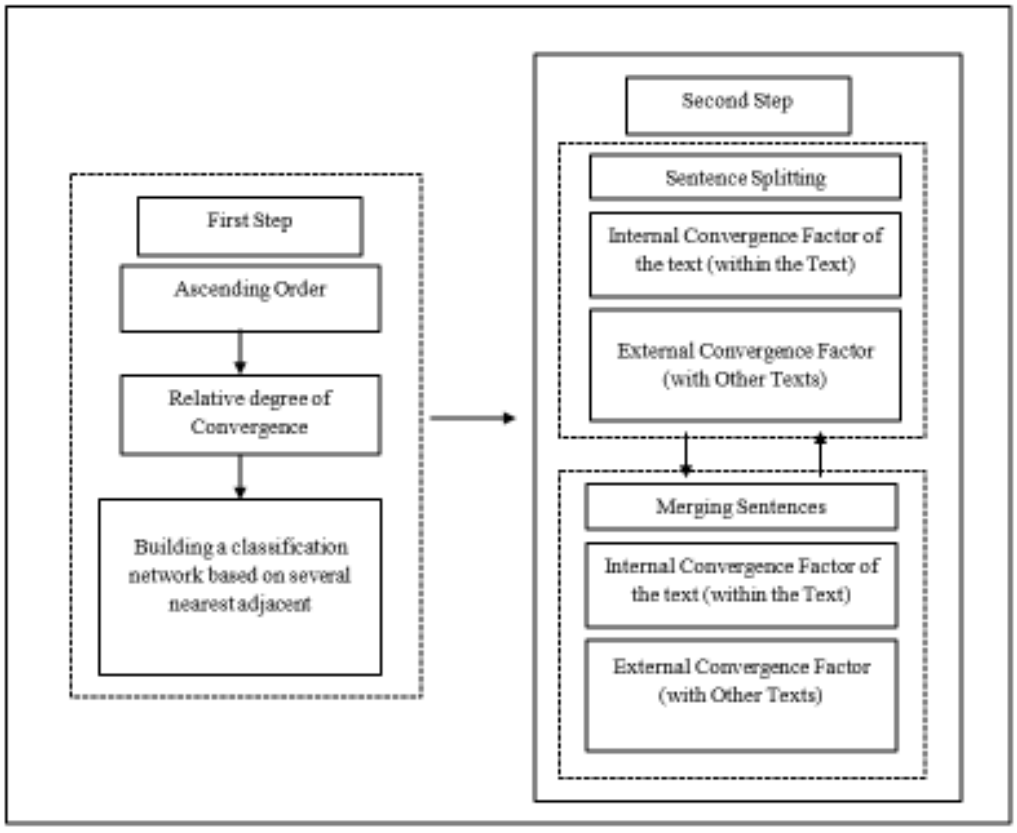

Figure 2. Text summary chart.

ISSN: 1814-635X (print), ISSN:2521-3520 (online) Volume 34, Issue 1, 2023 DOI: http://doi.org/10.23851/mjs.v34i1.1195

Text Clustering

The Clustering process relies on a scale of semantic similarity of abstracted texts called TF-IDF (Term Frequency and Inverse Document Frequency), a measure used in Information Retrieval (IR) and Machine Learning. TF-IDF can be divided into two parts: TF (Term Frequency) and (IDF) (Inverse Document Frequency). The first (TF) is called the repetition of the term by looking at the frequency of a particular term in the whole text and giving a weight to each term that is a scale of its importance in the document and depends on the number of times the word appears in the document (prime count). The frequency of the term is adjusted according to the length of the document. Each appearance is given a value of (1), and no appearance (0) is divided by the number of words in the document). The term (IDF) reflects the degree of prevalence of the word (its weight) in all existing documents. There may be a repetition of a particular word that is large, but in general, its weight is less in other documents, and this must be taken into consideration.

Query Stage

To retrieve a set of texts based on the text entered in the proposed system, the same initial processing steps are applied to the entered text and then summarized to obtain the summarized text to be dealt with for retrieval. Search will be conducted in the new database where the size is less than the original; thus, there is a speed in the search, and when the total of the closest texts is reached, and then the original texts associated with the summarized texts are retrieved.

Database Search

The text that has been processed and summarized and its conformity with the summaries in the new database built in the training phase are measured, and then the number of texts to be retrieved is determined (k). Retrieval will only be for the closest ones in the search process.

EXPERIMENTAL RESULTS

On the practical side, a set of texts related to articles presented in various fields such as computers, physics, chemistry, life sciences and mathematics was taken, with five articles for each specialty and one article for each specialization to examine the proposed method. Summarizing and grouping techniques have been worked out for grouping documents by summary. The results developed by the proposed method were compared using performance and health measures such as accuracy, recall and measurement. According to the results, the new method outperforms the other methods and reduces repetition due to aggregation. In the future, it is possible to improve the system by adding the sentence simplification technique to produce the abstract. These techniques are used to simplify complex and substantial sentences. This approach can also be extended to a multilingual platform. It is also possible to paraphrase to give the abstract more valuable properties.

We note in Table 2 that the summarization percentages are different according to each document's text, ranging from 41% to 79%. These percentages depend on the nature of the entered texts. For example, the percentage may be more significant if the texts contain a repetition of the existing words. Table 3 represents the distribution of documents after performing the grouping process according to clustering, the total words remaining in each group, and the words omitted from each group.

We note in the above table that the distribution of documents is not equal, depending on the measurement of similarity between the elements of one group and the degree of difference from other groups. This table is used in the retrieval process, as only one group is searched without searching in all groups, reducing the time required for retrieval.

Table 2. represents the number of documents used and the number of words summarized in each document.

Document Number	Number of Words	Words Summarized	Summarization Percentage	Document Number	Number of Words	Words Summarized	Summarization Percentage
1	211	141	41.54%	14	326	213	44.54%
2	291	182	68.80%	15	407	226	78.42%
3	361	262	65.84%	16	392	212	65.34%
4	238	106	54.08%	17	141	62	47.58%

5	223	160	67.96%	18	195	81	79.84%
6	366	287	72.01%	19	202	133	46.15%
7	314	235	58.44%	20	129	103	75.84%
8	476	361	51.59%	21	284	193	66.82%
9	418	301	74.84%	22	104	48	67.52%
10	372	177	71.75%	23	252	130	72.58%
11	314	212	55.53%	24	116	63	74.73%
12	459	343	62.54%	*25*	234	161	43.97%
13	154	90	54.31%				

Table 3. Distribution of documents after the process of grouping similar documents (clusters of groups).

Clusters of Groups	Number of Documents	Total Number of Words	Number After Summarization	Insignificant words
1	7	1879	1171	708
2	4	1196	751	445
3	5	1337	851	486
4	6	1606	1042	564
5	3	961	667	294
Total	*25*	6979	4482	2497

CONCLUSIONS

Organizing information and managing knowledge is one of the vital goals in information technology due to the urgent need to increase the speed of access to the required information in digital data warehouses. Therefore, there is an increasing need to find advanced word processing methods to become practical and not lengthy. Text summarization methods are considered a means of obtaining the main content of a collection of documents. At the same time, redundant information is reduced. In this work, data investigation was used in redistributing information according to the (K-Mean) algorithm and the (Term Frequency Inverse Document Frequency TF-IDF) technique was used in the measuring process of the properties of the summarized texts as well as the technique of sorting (split) and merging sentences (merge) to access the summarized texts. Satisfactory results are obtained with good summary ratios and deletion of words considered to be of little importance. It is essential to employ data investigation techniques in the clustering process. It is possible to benefit from this method and combine it with other methods in the process of summarizing to obtain better results in future research, as this method can be combined and work as a primary or secondary method.

Disclosure and conflict of interest: The authors declare that they have no conflicts of interest.

REFERENCE

[1] Mocnik, Franz-Benjamin. "Putting geographical information science in place-towards theories of platial information and platial information systems." Progress in Human Geography (2022). https://doi.org/10.1177/03091325221074023

[2] Zhang, Rui, Cairang Jia, and Jian Wang. "Text emotion classification system based on multifractal methods." Chaos, Solitons & Fractals 156 (2022). https://doi.org/10.1016/j.chaos.2022.111867

[3] Salloum, Said A., et al. "Using text mining techniques for extracting information from research articles." Intelligent natural language processing: Trends and Applications. Springer, Cham, 2018. 373-397. https://doi.org/10.1007/978-3-319-67056-0_18

[4] El-Kassas, Wafaa S., et al. "Automatic text summarization: A comprehensive survey." Expert Systems with Applications 165 (2021). https://doi.org/10.1016/j.eswa.2020.113679

[5] Wang, Danqing, et al. "Heterogeneous graph neural networks for extractive document summarisation." arXiv preprint arXiv:2004.12393 (2020).

[6] Jung, Chihoon, et al. "Knowledge Base Driven Automatic Text Summarisation using Multi-objective Optimization." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 12.8 (2021): 836-849. https://doi.org/10.14569/IJACSA.2021.0120895

[7] Memon, Muhammad Qasim, et al. "An ensemble clustering approach for topic discovery using implicit text segmentation." Journal of Information Science 47.4 (2021): 431-457. https://doi.org/10.1177/0165551520911590

[8] Goularte, Fábio Bif, et al. "A text summarization method based on fuzzy rules and applicable to automated assessment." Expert Systems with Applications 115 (2019): 264-275. https://doi.org/10.1016/j.eswa.2018.07.047

[9] Sanchez-Gomez, Jesus M., Miguel A. Vega-Rodríguez, and Carlos J. Pérez. "Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach." Knowledge-Based Systems 159 (2018): 1-8. https://doi.org/10.1016/j.knosys.2017.11.029

[10] Alguliyev, Rasim M., et al. "COSUM: Text summarization based on clustering and optimization." Expert Systems 36.1 (2019). https://doi.org/10.1111/exsy.12340

[11] R. Janani and S. Vijayarani "Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization" Expert Systems with Applications (IF 8.665) Pub Date: 2019-05-24, https://doi.org/10.1016/j.eswa.2019.05.030

[12] C. Kruengkrai and C. Jaruskulchai, "Generic text summarization using local and global properties of sentences," Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003), 2003, pp. 201-206. https://doi.org/10.1109/WI.2003.1241194

How to Cite

Z. A.-W. Salman, “Text Summarizing and Clustering Using Data Mining Technique”, *Al-Mustansiriyah Journal of Science*, vol. 34, no. 1, pp. 58–64, Mar. 2023.