International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I6P115

Volume 68 Issue 6, 87-93, May 2020

Original Article

Text Summarization using K-Means, Tanimoto Distance & Jaccard Similarity

Annu Sharma[1], Nandini Sharma[2]

[1]*Mtech Student, Department of Computer Science and Engineering, SRCEM at Palwal, Haryana, India.*

[2]*Assitant Professor, Department of Computer Science and Engineering, SRCEM at Palwal, Haryana, India.*

Received Date: 09 May 2020
Revised Date: 25 June 2020
Accepted Date: 29 June 2020

Abstract - *Text Summarization is a reduction procedure of content, text, passage source into the tiny or short text nevertheless still preserve and retain the crucial and significant information enclosed. This scheme confers the Summarization of the information like reviews, blogs, news from the web pages based on the content and context for the specific category or class using machine learning techniques like K-Means, Tanimoto Distance Jaccard Similarity and word frequency weighting. The aim of contemplation is to recapitulate, minimize and summarize the reviews, blogs and news web pages automatically to abridge the procedure of discovering a middle of reviews, blogs and news information. The analysis was completed by measuring the accurateness of the précis and summary by precision and recall calculation. From the analysis consequences, it was established that the précis or summary produces accuracy rate of the precise summary is approx 80% for and concise summary is approx 73% for English language reviews, blogs and news available online. The proposed scheme depicts that by assimilation of two or more techniques using machine learning were relatively successful and effectual in intriguing the essence of equivalent reviews, blogs and news that taken manually by humans as a précis or summary.*

Keywords *- Automatic Text Summarization, Machine Learning, K-Means Clustering, Tanimoto Distance, Jaccard Similarity.*

I. INTRODUCTION

Information is data that has been processed into a form that is useful for the recipient and tangible [1], in the form of values that can be understood in current and future decisions. With the information, every human being will get a variety of things they want, such as science, news, or just entertainment. Information in the form of data in the form of facts must meet at least three criteria, including accurate, timely, and relevant. Accurate means precise; in this case, the information must be in accordance with the facts that occur. Being punctual in terms of information must be there when needed. While relevant in this case means that the information sought must be useful. One of the information media which is still widely used by the general public is written information media. From this article, information can be conveyed to the public through printed media such as newspapers, magazines, or bulletins [5]. In addition, there is another media that can represent digital information which is commonly called electronic media, such as the web, in which there is the information needed by many people. The contents of the web are also diverse and what is popular today is the news web. This news web facility can provide easy access to the news because the news is always up to date and presented from time to time with an easy and simple display. News that is presented practically and can be accessed from anywhere and at any time is a supporting factor for the selection of news web media as a news source. This news website is needed by workers who need up-to-date information. In this case, most users are office workers who deal with information such as stock workers, taxes, and much more [3].

Web news viewer has advantages and disadvantages in their use. One of the advantages is that the public can access news that is being discussed hotly at any time. Besides, all the information they need can be found through this media. While the drawback is that sometimes the sentence used in the delivery of information is too long and complicated, so information seekers cannot be straight to the point in finding the information needed. In the case of office workers, as exemplified by the weaknesses of the web, this news is very influential because, in the world of work, all things are demanded efficiency, especially time [2-5]. Therefore we need a text summarizing technology to simplify long and convoluted news information into simple news information but do not lose the main essence of the news. Automatic text summarization technology, commonly called Automatic Text Summarization, is one solution to overcome this problem. Automatic Text Summarization Technology is a technique where a whole text of a news document is automatically summarized by a computer and produces a short text that does not lose the essence of the information contained in the news document.

According to Bari [6], Text summarization is a process in which subjectivity in a document is taken as the final result. Most research on document summarization focuses on the quality of summaries and utilities that are

assessed by a group of experts. Broadly speaking, there are two types in making a summary that takes the most important part of the original text, namely abstractive and extractive.

The extractive summation is essentially taking a few sentences from the original document. While abstractive summarizing rearranges information into new sentences that are not contained in the original document [7], there are two characters in making a summary, namely summarizing accurately and summarizing quickly. Summarizing accurately generally relies on the final result and is inefficient with time. Usually used for summarizing a single document whose summary results must be exact on the essence of the document. At the same time, summarizing quickly generally relies on time efficiency with a moderate level of accuracy. Usually used for a multi-document summarizing, the summary results must be fast, even though the documents must be summarized in large numbers [8-15]. In this thesis research, the researcher wants to build a summarization application using an extractive type approach using the K-Means Clustering[16] and Tanimoto Distance Jaccard Similarity Method[17,18] by sorting all sentences in a document text and entering it in a detailed data table (Terms Frequency-Inverse Document Frequency) or commonly called TF-IDF [19]. The similarity is a technique for matching text documents. Basically, the algorithm works by checking the similarity of the text of an existing document with the new document you want to find.

According to Lavin [19], TF-IDF works by determining the relative frequency of words in a particular document and comparing it with the proportion of words from all documents. This application is expected to minimize the complexity of sentences in a paragraph and find the essence of information without having to read the entire contents of the document. In addition, this application must be able to quickly summarize a collection of news from various websites because, in just one hour, the news that enters the virtual world can reach hundreds or even thousands, so it is necessary to summarize news that is efficient with time so that important news is not missed by everyone. Summarizing news, of course, requires a source where

II. LITERATURE SURVEY

Koit, Mare. (2014).[21] The actual data mining task is an automatic or semi-automatic analysis of large amounts of data to extract previously unknown and interesting patterns, such as groups of data records, which is why technology can generate new opportunities for businesses, providing the following capabilities [20]. But it is also important to understand that in addition to structured data, those other that come from known sources of information and are therefore easy to measure and analyze through traditional systems, we begin to be able and want to handle unstructured data Tascón.

(Anchalia, P. P., Koundinya, A. K., & Srinath, N. K.[22]. For the year 2013, in their article, they announced the importance of K-Means Clustering, which is a method used to classify structured or unstructured data sets. This is common and effective for classifying data due to its simplicity and ability to handle large data sets. The number of clusters and the initial set of centroids is accepted as parameters. The distance of each element in the data sets is calculated with each of the centroids of the respective cluster. The element is then assigned to the cluster with which the article distance is the shortest. The centroid of the group to which the item has been assigned is recalculated. One of the most important and commonly used methods of Grouping the elements of a data set using K-Means Clustering is calculating the distance between the point and the chosen mean. This distance is usually the Euclidean Distance, although there are other existing distance calculation definitions. This is the most common metric for point comparison.

(Kalimoldayev 2019 [23]). In 2019, he analyzed his study on the best-known structures for the treatment of big data, the structure of a data set, this type of information is beginning to be an undeniable requirement for the survival of many companies and organizations. As a consequence, terms such as big data, Mapreduce, Hadoop or cloud computing have emerged in recent years. Thus, the demand for so-called "data scientists" is growing exponentially. In his article, he presents an informative introduction to all these terms and analyzes the best-known structures for the treatment of big data.

(Aliguliyev, R. M. 2010 [24]). They talk about the Grouping of text documents, which is a central problem in text mining that is defined as the division of a set of documents into groups according to their main themes or content. Document grouping has many purposes, including information retrieval, summary generation, automatic topic extraction, browsing document collections, organizing information in digital libraries, and detection topics. Cluster with which the article distance is the shortest. The centroid of the group to which the item has been assigned is recalculated. One of the most important and commonly used methods Grouping the elements of a data set using K-Means Clustering is calculating the distance between the point and the chosen mean. This distance is usually the Euclidean Distance, although there are other existing distance calculation definitions. This is the most common metric for point comparison.

(Kalimoldayev 2019 [25]). In 2019, he analyzed his study on the best-known structures for the treatment of big data, the structure of a data set, this type of information is beginning to be an undeniable requirement for the survival of many companies and organizations. As a consequence, terms such as big data, Mapreduce, Hadoop or cloud computing have emerged in recent years. Thus, the demand for so-called "data scientists" is growing exponentially. In his article, he presents an informative introduction to all these terms and analyzes the best-known structures for the treatment of big data.

(Aliguliyev, R. M. 2010 [26]). They talk about the Grouping of text documents, which is a central problem in text mining that is defined as the division of a set of documents into groups according to their main themes or

content. Document grouping has many purposes, including information retrieval, summary generation, automatic topic extraction, browsing document collections, organizing information in digital libraries, and detection topics.

III. PROPOSED WORK

In this section, the discussion will focus on elements related to automatic text summarization research such as down-loaders/web-scrapper, extractions, summaries, as well as everything needed in the text mining process and creates conclusions that are appropriate to the text content. General Description of the System Text summarizing application that will be made is a system that makes a summary of the text that is composed of several sentences and will summarize into several sentences which are the contents of the text. Because the text to be summarized is an online news document, to download site pages using a simple web crawler, while for summarizing site pages will use the pre-processing, clustering and Tanimoto similarity method followed by weighting the frequency of occurrence of words so that the results of summation are more accurate. Therefore the process flow diagram below is depicted for ready reference.

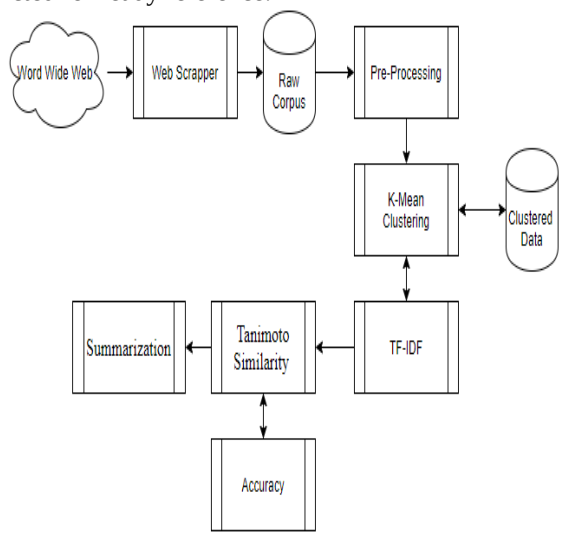

Fig. 1 Proposed Techniques or Proposed Design and Methodology

The above diagram is depicted in the algorithms manner for further elaboration and understanding.

A. Web Scrapper Algorithm

Step1. Initiate seed URLs
Step 2. Add the URL to the frontier
Step 3. Select the URL from the frontier
Step 4. Taking web pages related to URL
Step 5. Parsing the downloaded page and extracting all URLs.
Step 6. Add all the links that you haven't visited in the URL list (frontier).
Step 7. Return to the second step and repeat until frontier empty (URL does not exist).

B. K-Means Clustering

a) *K-Mean Algorithm*

K-Means is one method of non-hierarchical data grouping that attempts to partition existing data in the form of one group or more. This method partitioned the data into clusters or groups so that data have the same characteristics grouped into one same cluster and data have different characteristics grouped into other groups. In conducting clusters using K-Means, there are several things that must be done other queues:

Step1. Determine the number of Cluster K.
Step2. Initialising the K centre point of this cluster can be done by random method and used as the initial cluster centre point.
Step 3. Locate all data or objects to the closest cluster. The proximity of two objects is determined based on the distance of the two objects. To calculate the distance of all data to each cluster centre point, the Euclidean distance theory is formulated and place each data or object to the nearest cluster. The proximity of two objects is determined by distance. The distance used in the k -Means algorithm is Euclidean Distance (d).

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

x = x1, x2 ,. . . , xn, and y = y1, y2,. . . , yn is the number of n attribute (column) between 2 records.

Step 4. Recalculate the cluster centre with the current cluster membership. The cluster centre is the mean (mean) of all data or objects in a particular cluster.
Pseudo Code
Input: D = {d1, d2, d3, d4 .... dn}
Output: feature set
Method:

Step 1.
Declare Variants: j, X, Y, min, cluster, d, sumXY(), isStillMoving
isStillMoving = True
Step 2.
if totalData <= numCluster Then
only last data is put here because it is designed to be interactive
Data(0, total Data) = total Data ' cluster No = total data
Centroid(1 , total Data) = Data(1 , total Data) ' X
Centroid(2 , total Data) = Data(2 , total Data) ' Y
Else
'Calculate minimum distance to assign the new data min = 10 ^ 10 'big number
X = Data(1, totalData)
Y = Data(2, totalData)
For i = 1 To numCluster
d = dist(X, Y, Centroid(1, i), Centroid(2, i))
If d < min Then
min = d

```
                cluster = i
                End If
            Next i
        Data(0, total Data) = cluster
```

Step 3.

```
        Do While isStillMoving
        this loop will surely convergent calculate new centroids
    1 =X, 2=Y, 3=count number of data
    sumXY(1 To 3, 1 To numCluster)
    For i = 1 To totalData
    sumXY(1, Data(0, i)) = Data(1, i) + sumXY(1, Data(0, i))
    sumXY(2, Data(0, i)) = Data(2, i) + sumXY(2, Data(0, i))
    Data(0, i))
    sumXY(3, Data(0, i)) = 1 + sumXY(3, Data(0, i))
    Next i
    For i = 1 To numCluster
    Centroid(1, i) = sumXY(1, i) / sumXY(3, i)
    Centroid(2, i) = sumXY(2, i) / sumXY(3, i)
    Next i
    'assign all data to the new centroids
    isStillMoving = False
    For i = 1 To total Data
    min = 10 ^ 10 'big number
    X = Data(1, i)
    Y = Data(2, i)
```

Step 4.

```
Recalculate the distance using Euclidean Method.
    For j in 1 To numCluster
    d in dist(X, Y, Centroid(1, j), Centroid(2, j))
    If d < min Then
    min = d
    cluster = j
    End If
    Next j
    If Data(0, i) <> cluster Then
    Data(0, i) = cluster
    isStillMoving = True
    End If
    Next i
    Loop
   End If.
```

C. Term Frequency Invert Document Frequency

Pseudocode:

Step 1.

```
        Node Start
        Initialize TF, IDF, DF, c,
        TF Term Frequency;
        IDF inverse Document Frequency;
        DF Document Frequency;
        C threshold;
```

Step 2.

```
        For each document in the test, corpus do
        Remove tags, punctuation, other language text and non-alphanumeric text
        Perform case folding
        Perform Bag of Words
        Remove stop words
        End for
```

Step 3.

```
        For each remaining word in the dataset, do
        Perform Porter Stemmer and store in a vector (Word List)
        End for
```

Step 4.

```
        For each word in the Word List, do
        Calculate Modified TF/IDF and store the result in a weight matrix
        End for
```

Step 5.

```
        For each element in the weight matrix
        Set the threshold 'c.'
        Calculate Document Frequency (DF) for each term
        If DF < c, then
        Remove the term along with its weight from the weight matrix
        End if
        End for
```

D. Tanimoto Distance and Jaccard Similarity

Step 1.

```
Define Tanimoto Score (vec1, vec2, weights=None):
Remarks:
Return the Tanimoto score between vec1 and vec2.
Arguments: vec1, vec2: The two vectors to find the Tanimoto coefficient for. It MUST be of the same length
Kwargs or Keyword Arguments:
Weights: If given, this must be an iteration of the same length as vec1 and vec2. If k^th element of weights, weights[k] = wk, it means that vec1[k] and vec2[k] can take up values of either 0 or wk. If not given, or when None, a value of (1, 1, ... up to ten(vec1) elements) is assumed(i.e,
the Jaccard index of the binary vectors vec1 and vec2 is returned in this case).
```

Step 2.

```
        N = len(vec1)
           If weights is None:
              valid_ranges = [(0.0, 1.0) for i in vec1]
           else:
              valid_ranges = [(0.0, w) for w in weights]
        assert N == len(vec2) == len(valid_ranges)
        v1v2, v1v1, v2v2 = 0., 0., 0.
         For i in xrange(N):
    if vec1[i] not in valid_ranges[i] or vec2[i] not in valid_ranges[i]
       raise ValueError
    v1v2 += vec1[i] * vec2[i]
    v1v1 += vec1[i] * vec1[i]
    v2v2 += vec2[i] * vec2[i]
```

return v1v2 / (v1v1 + v2v2 – v1v2)

If the overhead of the validation is not needed(the common case, maybe), the implementation gets simpler:

Step 3.

Tanimoto_Score(vec1, vec2)
Return the Tanimoto score between vec1 and vec2.
Args: vec1, vec2: The two vectors to find the Tanimoto coefficient for. MUST be of the same length
No validation is performed except same length checks. It is assumed that the caller passes properly weighted data.

```
N = len(vec1)
assert N == len(vec2)
v1v2, v1v1, v2v2 = 0., 0., 0.
  For i in xrange(N)
  v1v2 += vec1[i] * vec2[i]
  v1v1 += vec1[i] * vec1[i]
  v2v2 += vec2[i] * vec2[i]
 return v1v2 / (v1v1 + v2v2 – v1v2)
```

Both these implementations return a value between 0 and 1, with a higher value indicating more similarity;

IV. IMPLEMENTATION AND RESULTS

The process of summarizing is divided into 3 main processes, namely the process of obtaining and extracting the site pages, document clustering, and the process of summarizing or Summarization.

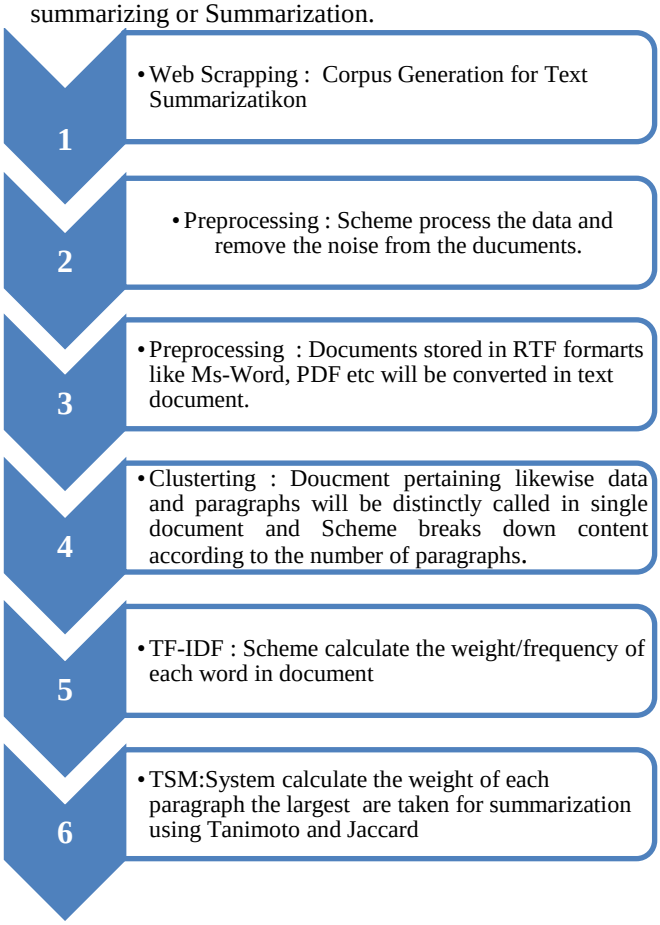

Fig. 2 Work Flow of Scheme Step Wise

A. Analogy or Automation Of Scheme

- The user enters the main URL of the site to be visited, and the news text that is on the site will be downloaded by the web scrapper.
- The system regulates that all news/blog/review URLs on the site are detected and forwarded to the download process.
- The system downloads the URL list.
- The system extracts downloaded news/blog/review site pages to get the title and content as corpus.
- The corpus will be pre-processed (noise removal, transformation) and converted into raw text.
- The corpus will be evaluated in clusters, and redundant documents will be accumulated in a single cluster with a threshold.
- The system takes content and is broken down according to the number of paragraphs. The first paragraph becomes a sample (training data) for the process of similarity and takes another paragraph whose contents are similar to the first paragraph using.
- After selecting one paragraph that has similarities to the first paragraph, the next process is that the two paragraphs are extracted per sentence to weigh the frequency of occurrence of words TF-IDF.
- After getting the weight of each sentence, the sentence with the highest weight will be the result of Summarization followed by a sentence that has a weight below as summarized data or information using Tanimoto and Jaccard.

B. Scheme Implantation

Graphical user interface implementation, often referred to as GUI, is an implementation of the design of the appearance of programs that can be enjoyed by the user. The implementation of the user interface must be made in accordance with the design while maintaining convenience in operating the program (user friendly). Graphical user interface application summarizing text using the above-mentioned method includes displaying the application form that is displayed using Visual Studio, which is supported by .Net Framework. The following is the appearance of the interface design of the news web summary application using the clustering and Tanimoto distance similarity method. The implementation of the interface design is as under:-

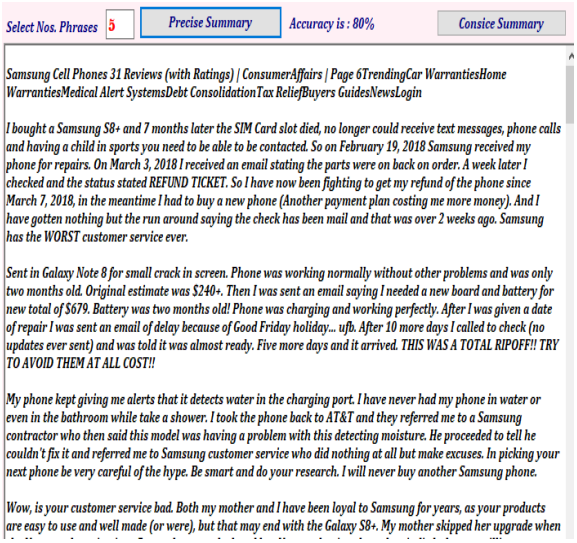

Fig. 3 Precise Summarization using Proposed Scheme with Accuracy of 80%

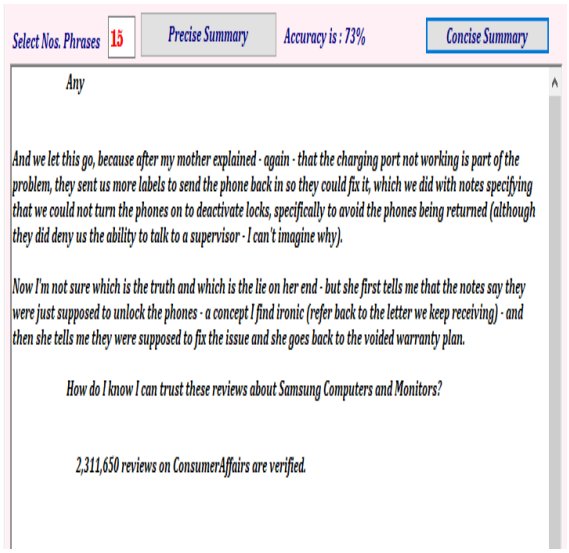

Fig. 4 Concise Summarization using Proposed Scheme with Accuracy of 73%

The recall value can be easily obtained by means of the system will produce recommendations for all the words in the article. But this is not desirable because the word annotation of the article becomes unclear. Therefore, an accuracy test is done by calculating the precision and recall to get the f-measure value of each calculated review content. The recall is a term used for invoked documents that is relevant to the statement (query) entered by the user in an information retrieval system. The recall is actually difficult to measure because the number of all relevant documents in the repository is very large. Therefore precision is usually one of the measures used to assess its effectiveness. F-measure is one of the evaluation calculations in information retrieval that combines recall and precision. The value of recall and precision in a situation can have different weights. The f-measure results will compare the quality level of the cluster at the time of calculation until the best f-measure value is obtained. Precision and recall measurements are very dependent on the ideal summary length and also the length of the summary to be evaluated. Accuracy will decrease as the summary length decreases. The precision and recall assessment requires a manual summary which will later be used as a comparison and become a query that will be the recommendation of choosing words in each paragraph of the article to be processed in order to be selected as a summary of the paragraphs. Consequently, while getting the precise Summarization using the scheme, the accuracy is different from getting Summarization using the concise Summarization.

V. CONCLUSION AND FUTURE SCOPE

A. Conclusion

From the results of implementation and trials that researchers have done can, it was concluded that the Tanimoto Distance Jaccard Similarity Algorithm is an algorithm that can be used as a summary of web reviews/blogs/news for the English language. This is proven by all good reviews/news websites of the English language, which can be summarized even though the corpus is not well-formed or lacking or grammar and punctuations. The level of accuracy of the precise Summarization performed by the application is quite far compared to the concise Summarization that is done in English-language text is different (precise 80% and concise 73%). Therefore, the addition of sentence weighting is very helpful in the summarizing process because if only with similarity, the results are still not good. Although by weighting the sentence summary results are relatively few but deep the measurement of the f-measure is still quite a result even though the length of the summary manuals and application summaries are relatively appropriate and understandable**.**

B. Future Scope

Of course, there are still many shortcomings in the research of summarized applications this. Therefore the authors suggest several things for the ingredients further development, including:

- Develop a more interesting summarization application, both in terms of appearance and system flow.
- However, adding training data to the similarity process will be able to affect the results of the concise accuracy performed by the system and manuals by experts.
- Develop special algorithms for extracting from web documents to web content.
- Add a stop word collection to filter the contents of the document on the web.
- Add a steaming process to increase the level of accuracy classification.
- Using antonyms and synonyms for feature extraction using a dictionary.
- Perform Summarization on mammoth repositories.

REFERENCES

[1] Nicholson, John. Information Retrieval(2019). 10.4324/9780367809416-63.

[2] Dawson, Catherine. Information retrieval. (2019). 10.4324/9781351044677-24.

[3] Boughanem, Mohand & Akermi, Imen & Pasi, Gabriella & Abdulahhad, KaramInformation Retrieval and Artificial Intelligence. (2020).. 10.1007/978-3-030-06170-8_5.

[4] Mainenti, David. Information retrieval: retaining its relevance. 2019).

[5] Banerjee, Swapna Information Retrieval. (2017).

[6] Bari, Poonam & Nihlani, Pracheta & Dev, Martand & Choudhary, Samruddhi. Automatic Text Summarizer. (2019).

[7] Simske, Steven & Lins, Rafael. Automatic Text Summarization and Classification. DocEng '18: Proceedings of the ACM Symposium on Document Engineering 1(2) (2018). 10.1145/3209280.3232791.

[8] Al-Taani, Ahmad. Automatic text summarization approaches. (2017). 93-94. 10.1109/ICTUS.2017.8285983.

[9] Chettri, Roshna & Kr, Udit. Automatic Text Summarization. International Journal of Computer Applications. 161(2017) 5-7. 10.5120/ijca2017912326.

[10] Torres-Moreno, Juan-Manuel. Automatic Text Summarization. (2014). 10.1002/9781119004752.ch3.

[11] Patil, Annapurna & Dalmia, Shivam & Ansari, Syed & Aul, Tanay & Bhatnagar, Varun. Automatic text summarizer(2014)1530-1534. 10.1109/ICACCI.2014.6968629.

[12] Prakash, B. & Sanjeev, K. & Prakash, Ramesh & Chandrasekaran, K. & Rathnamma, M. & Ramana, V.. Review of Techniques for Automatic Text Summarization (2020). 10.1007/978-981-15-1480-7_47.

[13] Bhole, Varsha. Automatic Text Summarization. (2014).

[14] Mathews, Lincy & Sathiyamoorthy, E.. Intricacies of an Automatic Text Summarizer. International Journal of Engineering and Technology. 5 (2013) 2871-2878.

[15] Soumya, S. & S Kumar, Geethu & Naseem, Rasia & Mohan, Saumya. Automatic Text Summarization. (2011). 10.1007/978-3-642-25734-6_140.

[16] Zhou, Hong. K-Means Clustering. (2020). 10.1007/978-1-4842-5982-5_3.

[17] Berenger, Francois & Yamanishi, Yoshihiro. Combining a Bisector tree with the Tanimoto Distance for Similarity Searches and Beyond. (2018). 10.13140/RG.2.2.15044.53121.

[18] Yan, Ziqi & Wu, Qiong & Ren, Meng & Liu, Jiqiang & Liu, Shaowu & Qiu, Shuo. Locally Private Jaccard Similarity Estimation. Concurrency and Computation: Practice and Experience. (2018). 10.1002/cpe.4889.

[19] Lavin, Matthew. Analyzing Documents with TF-IDF. The Programming Historian. (2019). 10.46430/phen0082.

[20] Chowdhary, K.. Natural Language Processing. (2020). 10.1007/978-81-322-3972-7_19.

[21] Koit, Mare. (2014). (Semi-)automatic analysis of dialogues. ICAART 2014 - Proceedings of the 6th International Conference on Agents and Artificial Intelligence. 1 (2014) 445-452.

[22] Tascón, M. (2013). Introduction: Big Data. Past, present and future. Telos: Communication notebooks and innovation, 95 (2013) 47-50.

[23] Anchalia, Prajesh & Koundinya, Anjan & Nk, Srinath. MapReduce Design of K-Means Clustering Algorithm. 1(5)(2013). 10.1109/ICISA.2013.6579448.

[24] Kalimoldayev, Maksat & Siládi, Vladimír & Satymbekov, Maksat & Naizabayeva, Lyazat. Solving mean-shift Clustering Using MapReduce Hadoop. (2017).

[25] Ramiz M. Aliguliyev, A new sentence similarity measure and sentence based extractive technique for automatic text summarization, Expert Systems with Applications, 36(4) (2009) 7764-7772, https://doi.org/10.1016/j.eswa.2008.11.022.

[26] Piegorsch, Walter. Confusion Matrix. (2020). 10.1002 / 9781118445112. Stat 08244

[27] B.Srinivasa Rao, S.Vellusamy Raddy, A Hard K-Means Clustering Techniques for Information Retrieval from Search Engine SSRG International Journal of Computer Science and Engineering 4(2) (2017).