IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-10, Oct 2019 E-ISSN: 2347-2693

An Effective Clustering Approach for Text Summarization

Rajani S. Sajjan[1*], Meera G. Shinde[2]

[1,2]Dept.of Computer Science & Engineering, VVPIET, Solapur University, Solapur, India

**Corresponding Author: rajanisajjan78@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i10.191197 | Available online at: www.ijcseonline.org

Accepted: 11/Oct/2019, Published: 31/Oct/2019

***Abstract*—** Text summarization automatically creates a shorter version of one or more text documents. It is an effective way of finding relevant information from large set of documents. Text summarization techniques are categorized as Extractive summarization and Abstractive summarization. Extractive summarization methods evaluate text summarization by selecting sentences present in documents according to predefined set of rules. Abstractive summaries attempt to improve the coherence among sentences by eliminating redundancies and clarifying the content of sentences. It should also extract the information is such a way that the content would be in the interest of the user. In this paper we used tokenization for preprocessing of statements then calculate TF/IDF for feature extraction, K-means clustering to generate clusters containing high frequency statements and then NEWSUM algorithm for weighing of statements that are used for relevant content summarization. We also present experimental results on a number of real data sets in order to illustrate the advantages of using proposed approach

***Keywords*—** Text Mining, Text Summarization, Clustering, extractive summary, information extraction

I. INTRODUCTION

Information filtering is an effective approach to eliminate repetitive and unwanted information from a large set of documents which represent user's interest. Traditional rule based frameworks were developed using a term-based approach. The main advantage of the term-based approach is that it is faster. But term-based document representation suffers from the limitations of polysemy and synonymy. To curb these limitations of term-based approaches, pattern based techniques have been used to utilize patterns to represent user's interest and have achieved some refinements in efficacy [2] [3], since patterns carry more semantic meaning than terms. Some data mining techniques have also been developed to improve the quality of patterns like maximal patterns, closed patterns and master patterns for removing the redundant and noisy patterns[4][5].

The merging of data from multiple documents is called multi-document merger. Data is found in unstructured or structured form and many times we have to generate summary from multiple files in less time, so, multi-document merger technique is useful. Multi-document summarization generates information reports that are both concise and comprehensive. With different opinions being put together, every topic is described from multiple perspectives within a single document. While the main focus of brief summary is to make information search more comprehensible and minimize the time by extracting the most relevant source documents, an accurate multi-document summary should contain the required information, hence there will be no need to access original files when refinement is required.

Extractive summarizer selects the most relevant sentences within the document as well as maintaining a reduced redundancy within the summary. Now a day, most of the researchers focuses their research in automatic text summarization are extractive summarizations. Some of the basic extractive processes are as follows:

a. Coverage: extraction plays a major role in text summarization process. Firstly it finds out all the necessary information that covers the different topics in input documents. It is applicable on text paragraphs. Numerous methodologies have been proposed to recognize the most important information from the set documents.

b. Coherency: optimal ordering of retrieved sentences to formulate the coherent context flow is the complex issue. In single document summarization, one probable ordering sentence is given by the input text document itself. Still, this process is a nontrivial task.

c. Redundancy elimination: due to the length limitation needed for an effective summary, and the existence of the extracted sentences which contains the same information, most of the approaches use similarity to remove duplicate information from the documents.

This paper introduces an automatic text summarization approach to overcome the difficulties in the existing summarization approaches. Here, TF-IDF approach is utilized to identify the necessary keywords from the text. TF-IDF is used to estimate the distinguishing keyword features in a text and retrieves the keyword from the input based on this information. The features are generally independent and distributed. Scoring is estimated for the retrieved sentence to compute the word frequency. The combination of this scoring concept helps to improve the summarization accuracy. The proposed summarization method achieves better coverage and coherence using the TF-IDF, hence it automatically eliminates the redundancy in the input documents.

The remainder of this paper is organized as follows. Section 2 summarizes the related works in the multi document text summarization. Section 3 shows the proposed text summarization approach Section 4 describes the performance analysis. And finally, the paper ends with the conclusion and future work at Section 5.

II. LITERATURE SURVEY

Paper [1] proposes an improved extractive text summarization method for documents by enhancing the conventional lexical chain method to produce better relevant information. Author has investigated the approaches to extract sentences from the document(s) based on the distribution of lexical chains then built a transition probability distribution generator (TPDG) for n-gram keywords which learns the characteristics of the assigned keywords from the training data set. A new method of automatic keyword extraction also featured in the system based on the Markov chains process. Among the extracted n-gram keywords, only unigrams are selected to construct the lexical chain. Effectiveness and time consumption are the main issues in this paper.

Paper [2] proposed a framework for addressing the c ross-language document summarization task by extraction and ranking of multiple summaries in the target language

- Top- K ensemble ranking algorithm is used to rank sentences
- TF-IDF is used to word count and word level feature extraction

Framework extracts multiple candidate summaries by proposing several schemes for improving the upper-bound quality of the summaries. Then, proposed a new ensemble ranking method for ranking the candidate summaries by making use of bilingual features. Extensive experiments have been conducted on a benchmark dataset System is designed for multiple language document summarizations but the accuracy of summarization is not up to the mark

Paper [3] demonstrates how to process large data sets in parallel to address the volume problems associated with big data and generate summary using sentence ranking

- TF-IDF is used for document feature extraction
- MapReduce and Hadoop is used to process big data

Limitation of this framework is that, it is designed only for big data framework

Paper [4] proposes two stage structures

- Key sentence extraction using Levenshtein Distance formula
- Recurrent neural network for summarization of documents

In extraction phase system conceives a hybrid sentence similarity measure by combining sentence vector and Levenshtein distance and integrates into graph model to extract key sentences. In the second phase it constructs GRU as basic block, and put the representation of entire document based on LDA as a feature to support summarization. Only limitation of this paper is that there is possibility of occurrences of negative value in the decomposed GRU matrix.

III. PROPOSED SYSTEM

Proposed system involves the different module to generate the summary for given multiple documents. Previous system has some drawback such that it can take only text file as input. If we give other files such as PDF or word file as input then it cannot accept that file and shows the message only text files are allowed. To overcome these problems we proposed a new system that takes the input as text, PDF and word files. The system involves the following basic three phases.

1. Data Pre-processing Phase-

In this phase system first check the type of input files. If the input is text file it can directly retrieve the data and eliminate the stop words. Other than text file if the input is PDF or word file it first converts that file into text and after that retrieve the data and remove the stop words.

2. Feature extraction Phase-

In this phase system can extract relevant feature i.e. weighted feature. In this phase frequency of word is calculated using TFIDF.

3. Similarity Based Approach Phase-

In last phase system can generate summary from multiple documents by merging the multiple generated summary into one. For this purpose NEWSUM algorithm is used. Below figure shows the detailed design for proposed system.

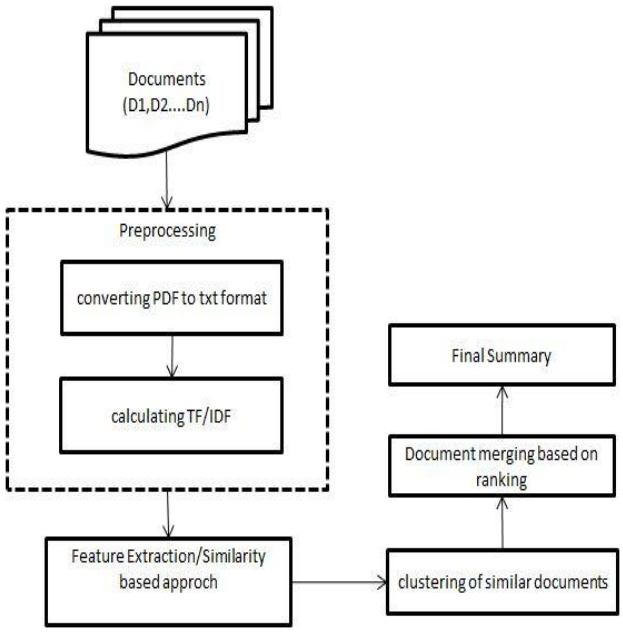


Fig.3.1. Detailed design for proposed system

The above diagram contains the three modules. These are as follows.

Module 1: Processing of Input multiple documents

a) Tokenization:

It breaks the text into separate lexical words that are separated by white space, comma, dash, dot etc. [3]

b) Stop Word Removal:

Stop words are words which are filtered out before or after processing of natural language data (text).Stop word removal is helpful is keyword searching. [2]

c) Stemming Suffixes:

Here enlisted suffixes are removed for topic detection

Example

Plays and playing → Play where "s" and "ing" are suffixes added to topic play need to be removed for accuracy purpose.

Module 2: Similarity based documents extraction from multiple documents

Cosine Similarity Approach:

Cosine Similarity measures the similarity between two sentences or documents in terms of the value within the range of [-1, 1] whichever you want to measure. That is the Cosine Similarity. Cosine Similarity extracted TF and IDF by using following formulae:

TFIDF:

TF (term, document) = Frequency of term / No of Terms

$$tfi = \left(\frac{ni}{\sum knk}\right) \quad \text{.....(i)}$$

IDF (inverse document frequency) calculates whether the word is rare or common in all documents. IDF (term, document) is obtained by dividing total number of Documents by the number of documents containing that term and taking log of that.

IDF (term, document) = log (Total No of Document / No of Document containing term)

$$idfi = \log\left(\frac{D}{\{d:t\epsilon d\}}\right) \quad \text{.....(ii)}$$

TF-IDF is the multiplication of the value of TF and IDF for a particular word. The value of TF-IDF increases with the number of occurrences within a document and with rarity of the term across the corps

$$tfidf = tf * idf \quad \text{..... (iii)}$$

Example:

Consider a document containing 100 words wherein the word "sachin" appears 3 times. The term frequency (tf) for "sachin" is then TF=(3/100)=0.03. Now, assume we have 100 documents and the word "sachin" appears in 10 of these. Then, the inverse document frequency (idf) is calculated as IDF = log(100 / 10) = 1.

Thus, the Tf-idf weight is the product of these quantities TF-IDF = 0.03 * 1 = 0.03.

Given a document containing terms with given frequencies: A(3), B(2), C(1) and total number of terms in document are 15. Assume collection contains 10,000 documents and document frequencies of these terms are:

A(50), B(1300), C(250).

Then, using above equation (i), (ii) and (iii) we calculate the tf-idf for terms A, B and C.

A: $tf = \frac{3}{15} = 0.2$; $idf = \log\left(\frac{10000}{50}\right) = 7.6$;

$tfidf = 0.2 * 7.6 = 1.52$

B: $tf = \frac{2}{15} = 0.133$; $idf = \log\left(\frac{10000}{1300}\right) = 2.9$;

$tfidf = 0.133 * 2.9 = 0.38$

C: $tf = \frac{1}{15} = 0.066$; $idf = \log\left(\frac{10000}{250}\right) = 5.3$;

$tfidf = 0.066 * 5.3 = 0.35$

NEWSUM Algorithm:

NEWSUM algorithm is a type of clustering algorithm and it uses sentence extraction to compose a summarization based upon the sentences that receives the highest score. It uses second order merge function and fβ-optimal merge function. By using these two algorithms, relevancy of multiple documents can be found out by taking hybrid form of both algorithms. First NEWSUM Algorithm forms different clusters and give a highest score sentences by using merge functions then Cosine Similarity extracts the term frequency and finally generates the output.

Algorithm: Clustering(List<String> Documents)
It clusters the document.
Input: Documents D = (D_1, D_2, D_3, D_4, …)
Output: Clusters C_1,C_2,C_3

For each documents D Read()
If (document extension = = {.doc,.pdf}) Then
ConvertToText()
Else SplitTerms()
Remove stopwords
Calculate TF-IDF for each term
Calculate similarity between each documents
Calculate threshold value from similarity
Classify documents in Clusters as per threshold

Module 3: Summary Generation
After checking similarity based approach and relevancy of documents, relevant sentences are extracted and merge the relevant sentences into one by using cosine similarity approach. Thus after merging the data it generates a final summary.

Algorithm: Summarization(List<String> L)
Input: Clusters List L=(C_1,C_2,C_3)
Output: Summary List S = [S']

While size of cluster C!=0
For each cluster C, do
Select a sentence S' with highest score from L[i]
S.Add(S')
End for
Return S

System State Transition
Our system has a finite state of events that executes one after the other and each states output is given as input to the next state. Here we are representing state transition for following reasons.

i) To identify the life cycle of an object.
ii)To identify the object that you will create during the development of classes in the program.
iii)To identity the actions or events.
iv)To identify the possible states for an object.

Let S be a system which is defined in the following manner
• S = {I, O, S} Where,
• I is Input,
• O is Output,
• S is the Transition State

Input:
• I ={ D, Ft, } Where,
• D : Document,
• Ft : Processed document,
• Ft ∈ D,
• Ft = { Ft1 , Ft2 ……. Ftn } where n ⩾1

Transition States:
Here, S is the state for system and S1, S2, … , Sf are the sub states for system S.
• S={ S1, S2…….Sf)
• Sd : Drop state dropping the sentences which are not extracted.

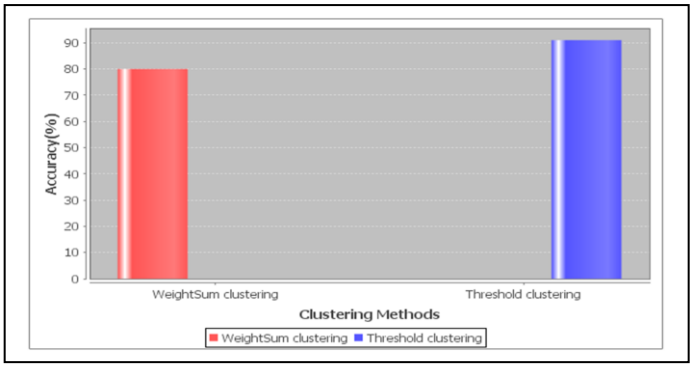
Output:
The output is generated when the system goes through out all the states. Irrelevant documents are dropped in drop state Sd. Thus output is:
• Generated summary draft from given documents.

IV. RESULT ANALYSIS

For result analysis we have created our own dataset and used one standard dataset OpinosisDataset1.2. System processes PDF, word and text documents. Here we use text documents for result analysis. Below graph1 shows the comparative analysis between two clustering methods, first is WeightSum method which generate the cluster based on the weight of documents, and other one is threshold clustering. Threshold clustering is proposed clustering algorithm. Here result and analysis is performed on OpinosisDataset1.2. [26]

Table 1 Accuracy of clustering methods

Clustering methods	Accuracy (%)
WeightSum	80
Threshold	91



Graph 1 Accuracy of clustering methods

Accuracy: It is the degree to which the result of a measurement, calculation, or specification conforms to the correct value (true). The proposed system gives maximum 91% Accuracy. To calculate the accuracy in percentage we can multiply by 100 to the result.

$$\text{Accuracy} = \frac{TP + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where,
TP (True Positive) – Correctly Identified.
FP (False Positive) – Incorrectly Identified.
TN (True Negative) – Correctly Rejected.
FN (False Negative) – Incorrectly Rejected.

Below graph 2 shows the accuracy of a proposed system on variable number of documents. The size of all documents is between 2 kb to 10kb. Here for calculating result we took first dataset that contains total 6 documents in that 1 file is irrelevant and 5 files are relevant and system merged total 4 documents.
TP – 4, FP – 0, TN – 1, FN – 1

$$\text{Accuracy} = \frac{4+1}{4+1+0+1} = 0.8333$$

After that for calculating result we took dataset that contains total 12 documents in that 3 files are irrelevant and 9 files are relevant and system merged total 7 documents.
TP – 7, FP – 0, TN – 3, FN – 2

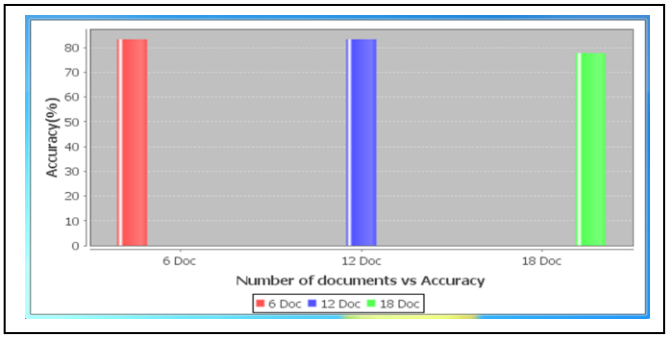
$$\text{Accuracy} = \frac{7+3}{7+3+0+2} = 0.8333$$

Again for calculating result we took dataset that contains total 18 documents in that 5 files are irrelevant and 13 are relevant and system merged total 9 documents except one that is irrelevant so,
TP – 9, FP – 0, TN – 5, FN – 4

$$\text{Accuracy} = \frac{9+5}{9+5+0+4} = 0.7777$$

Table 2 Accuracy

Number of documents	Accuracy (%)
6	83.33
12	83.33
18	77.77



Graph 2 Accuracy

When we calculate the accuracy of proposed framework for variable number of document, it shows that the accuracy is more compared to the previous module.

Values obtained on applying different summarization tools for analysing average Precision results are shown below in graph 3. In clustering we are generating two clusters on the basis of average cosine similarity between documents and in that first cluster contain documents with higher similarity, and other cluster contains the documents with lowest similarity. The documents with higher cosine similarity are selected for summarization purpose.

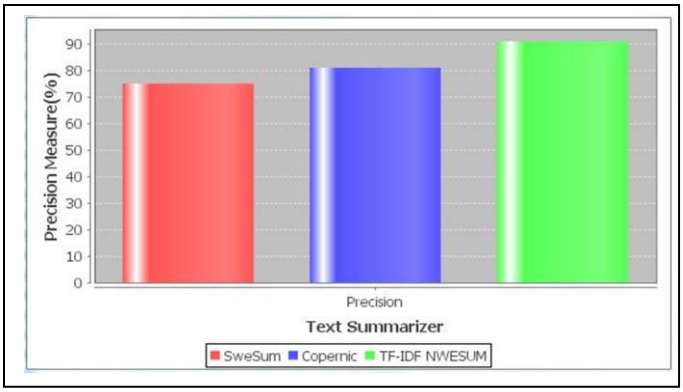
Here we calculate the precision value of proposed system and compare that value with standard tools. These standard tools are SweSum and Copernic. Below graph shows the precision measure of proposed system and two standard tools. The graph shows that the technique which uses SewSum and Copernic has less precision measure as compare to proposed system.
For calculating precision results we took first dataset that contains total 6 documents in that 1 file is irrelevant and 5 files are relevant.

Precision: It is degree to which repeated measurements under unchanged condition show the same results.

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Table 3 Precision scores Different Summarization techniques

Text Summarizer	Precision Measure (%)
SweSum	75
Copernic	81
TF-IDF NWESUM	91



Graph 3 Precision scores summarization techniques

Finally, from the above result analysis we can say that our proposed system has more accuracy compared to the previous module. Threshold clustering method gives the 91%

precision and 85% accuracy while generating the clusters. When we calculate the precision measure of proposed framework, it is more compared to the standard text summarization tools.

V. CONCLUSION AND FUTURE SCOPE

In this paper we proposed a framework for content based document summarization. Previous system has drawback like that it can take only text files as input. To overcome this problem we proposed a new framework. Not only text inputs but also various documents like PDF and Microsoft word document are taken in consideration. Here we use clustering algorithm. In order to design the clustering method, we combined an iterative partitioning technique with the help of threshold value. Threshold approach is used in order to design both clustering and classification algorithms. We present results on real data sets illustrating the effectiveness of our approach. The result shows that the use of proposed framework can greatly enhance the quality of text summarization, while maintaining a high level of efficiency. In future work, we will test the robustness of our proposed framework in other target languages, e.g. Chinese, Hindi and other most spoken languages. We will also try to use deep learning techniques for learning latent features to improve summary ranking. Further, we will explore different summarization methods to produce more diversified candidate summaries for ranking, and we believe the cross-language document summarization performance will be improved.

REFERENCES

[1] HtetMyet Lynn 1 , Chang Choi 2 , Pankoo Kim "*An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms*", Springer-Verlag Berlin Heidelberg 2017

[2] Xiaojun Wan 1 , FuliLuo 2 , Xue Sun Songfang Huang3 , Jin-ge Yao *"Cross-language document summarization via extraction and ranking of multiple summaries"* Springer- Verlag London 2018

[3] Andrew Mackey and Israel Cuevas *"AUTOMATIC TEXT SUMMARIZATION WITHIN BIG DATA FRAMEWORKS"*, ACM 2018

[4] Yong Zhang, Jinzhi Liao, Jiyuyang Tang *"Extractive Document Summarization based on hierarchical GRU",* International Conference on Robots & Intelligent System IEEE 2018

[5] Lili Wan *"Extractive Algorithm of English Text Summarization for English Teaching"* IEEE 2018

[6] Anurag Shandilya, Kripabandhu Ghosh, Saptarshi Ghosh *"Fairness of Extractive Text Summarization",* ACM 2018

[7] P.Krishnaveni, Dr. S. R. Balasundaram *"Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence"*, Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication

[8] Bagalkotkar, A., Kandelwal, A., Pandey, S., &Kamath, S. S. (2013, August). *"A Novel Technique for Efficient Text Document Summarization as a Service"*, In Advances in Computing and Communications (ICACC), 2013 Third International Conference on (pp. 50-53). IEEE.

[9] Ferreira, Rafael, Luciano de Souza Cabral, Rafael DueireLins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J. Simske, and Luciano Favaro. *"Assessing sentence scoring techniques for extractive text summarization."* Expert systems with applications 40, no. 14 (2013): 5755-5764.

[10] Gupta, V. K., &Siddiqui, T. J. (2012, December*). "Multi-document summarization using sentence clustering"*, In Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on (pp. 1-5). IEEE.

[11] Min-Yuh Day Department of Information Management Tamkang University New Taipei City, Taiwan myday@mail.tku.edu.tw Chao-Yu Chen Department of Information Management Tamkang University New Taipei City, Taiwan susan.cy.chen@gmail.tw *"Artificial Intelligence for Automatic Text Summarization"*,2018 IEEE International Conference on Information Reuse and Integration for Data Science

[12] Xiaoping SunandHaiZhuge*, Senior Member, IEEE Laboratory of Cyber-Physical-Social Intelligence, Guangzhou University, China Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, University of Chinese Academy of Sciences, Chinese Academy of Sciences, China System Analytics Research Institute, Aston University, UK *"Summarization of Scientific Paper through Reinforcement Ranking on Semantic Link Network"* ,IEEE 2018

[13] Ahmad T. Al-Taani (PhD, MSc, BSc) Professor of Computer Science (Artificial Intelligence) Faculty of Information Technology and Computer Sciences Yarmouk University, Jordan. ahmadta@yu.edu.jo *"Automatic Text Summarization Approaches"* ,IEEE 2017

[14] AlokRanjan Pal Dept. of Computer Science and Engineering College of Engineering and Management, KolaghatKolaghat, India chhaandasik@gmail.com DigantaSaha Dept. of Computer Science and Engineering Jadavpur University Kolkata, India neruda0101@yahoo.com *"An Approach to Automatic Text Summarization using WordNet",* IEEE 2014

[15] Prakhar Sethi1, Sameer Sonawane2, Saumitra Khanwalker3, R. B. Keskar4 Department of Computer Science Engineering, Visvesvaraya National Institute of Technology, India 1 prakhar.sethi2@gmail.com, 2 sameer9311@gmail.com, 3 theapogee2011@gmail.com, 4 rbkeskar@cse.vnit.ac.in"*Automatic Text Summarization of News Articles"*, IEEE 2017

[16] Yue Hu and Xiaojun Wan *"PPSGen: Learning-Based Presentation Slides Generation for Academic Papers"* , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 4, APRIL 2015

[17] Daan Van Britsom, AntoonBronselaer, and Guy De *Tre "Using Data Merging Techniques for Generating Multidocument Summarizations"* , IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 23, NO. 3, JUNE 2015

[18] NingZhong, Yuefeng Li, and Sheng-Tang Wu *"Effective Pattern Discovery for Text Mining*", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[19] Mohsen Pourvali and Mohammad SanieeAbadeh Department of Electrical & Computer Qazvin Branch Islamic Azad University Qazvin, Iran Department of Electrical and Computer Engineering at TarbiatModares University Tehran, Iran *"Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet and Wikipedia Knowledge Base"* , IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

[20] Daan Van Britsom, AntoonBronselaer, Guy De Tre´ Department of Telecommunications and Information Processing, Ghent University Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium *"Using data merging techniques for generating multi-document*

summarizations" ,IEEE TRANSACTIONS ON FUZZY SYSTEMS 2018

[21] Yang Gao,YueXu, Yuefengli, *"Pattern-based Topics for Document Modeling in Information Filtering"* in IEEE Transaction on Knoweledge and Data Engineering, vol.27,No.6,June 2015.

[22] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, *"Mining frequent patterns with counting inference,"* ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 66–75, 2000.

[23] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, *"Discriminative frequent pattern analysis for effective classification,"* in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716–725.

[24] R. J. BayardoJr, *"Efficiently mining long patterns from databases,"* in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 85–93.

[25] J. Han, H. Cheng, D. Xin, and X. Yan, *"Frequent pattern mining: Current status and future directions,"* Data Min. Knowledge. Discovery., vol. 15, no. 1, pp. 55–86, 2007.

[26] http://kavita-ganesan.com/opiniosis-opinion-dataset/

Authors Profile

Ms. Sajjan R.S. received her M.Tech in Computer Science and Engineering. She has a working experience of 15 years and is currently the H.O.D. of the Computer Science and Engineering Department. She is currently pursuing Ph.D. Her research interest is in Cloud Computing.

Ms. Shinde Meera Ganpat received her Bachelor of Engineering in Computer Science & Engineering from B.M.I.T. , Solapur. She is currently working toward the M.E degree in Computer Science & Engineering from Solapur University, Solapur. Her research interests lies in area of programming & datamining.

