Scientific Journal of Impact Factor (SJIF): 3.134

E-ISSN (O): 2348-4470
P-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 12,December -2015

Fuzzy Clustering For Forensic Science Documents

Manisha Bhore[1], Prajakta Satpute[2], Ashwini Shinde[3] ,Raveena Rao[4]

[1,2,3,4]*Computer Engg., AISSMSCOE*

Abstract — *In this paper we have come up with a system that will work on clustering for forensic science documents. They are widely used for the investigation process in forensic crime branch, police investigations. Many systems are existing in market but they deal with non semantic clustering of documents and also require a lot of time and space complexity. So to overcome this disadvantage a proposed system puts forward the idea of fuzzy clustering for forensic science documents by using web crawler, feature extraction and fuzzy clustering, fuzzy logic and weighted score matrix. This system will eventually help in reducing a great amount of time and effort in the investigation process.*

Keywords- *Pre-processing, Feature Extraction, Fuzzy Clustering, Fuzzy logic, Web crawler*

I. INTRODUCTION

Web document clustering is the process of partitioning the web documents into subclasses and each subclasses are known as clusters The clustering process uses various methods to cluster a document such as partitioning methods ,hierarchical method , density based method and grid based method .Clustering technique has widely shown its application in many fields like marketing ,biology, libraries, insurance ,city planning , world wide web ,etc .In this system the various methodologies used to cluster the documents are crawling , preprocessing, feature extraction , master matrix generation and fuzzy logic generation. Various tools are available in market which tackle only with non semantic clustering. The content has to be inserted by the user himself, for this the user has to study the topic thoroughly and also invest a lot of time and efforts. These problems are taken care by this system. This system is efficient because it generates clustering of documents automatically when the web pages are given as input. This system extracts the text which is important and related to the topic and also displays it in the document.

II. PROPOSED SYSTEM

In this system when the folder is given as input it goes through various stages. The text which is given as input can be of various format like pdf or document files.

1.1. Preprocessing.

The aim of this stage is to preprocess the input files. This is done by following methods.

1.1.1. Special symbol removal.

In this process the special symbols like space, exclamatory mark, commas, semicolon etc. are removed.

1.1.2. Stop word removal.

Stop word are those words which when removed will not change the desired meaning of this sentence. Hence stop word removal will increase the processing speed.

1.1.3. Stemming.

In the stemming process the words are brought to its base or root form. By using this process the reduction of overhead are possible and the accuracy is increased.

1.2. Feature Extraction.

The aim of this stage is to extract the important features from documents. This is done by following methods.

1.2.1. Numerical data.

In this process the numerical score of sentence is calculated. This score is obtained through calculating the number of numbers present or occurring in the sentence. Based on the extracted score it is decided whether to include the sentence or not.

1.2.2. Proper noun.

In this stage scores of proper noun occurring in the sentence is calculated

1.2.3. Top word.
Top word is the process in which the number of times a particular word occurring in sentence is calculated. If the weight of the term is high then it is considered to be important.

1.2.4. Title sentence.
In this stage the title sentence is extracted.

1.3. Master Matrix Generation.
In this stage matrix is generated according to their scores.

1.4. Fuzzy Logic.
In this stage clustering is done into five types. Firstly very low that is 0, then based on low, then on medium, then high and last very high that is 1.Generated scores of the sentences are checked according to the clustering score is termed as Very low if it has a score 0 and is termed as Very high if it has a score 1.Therefore if the sentence is having score 0 that means the sentence is having less importance and if the sentence is having score 1 then it means the sentence is having more importance.

III. BLOCK DIAGRAM

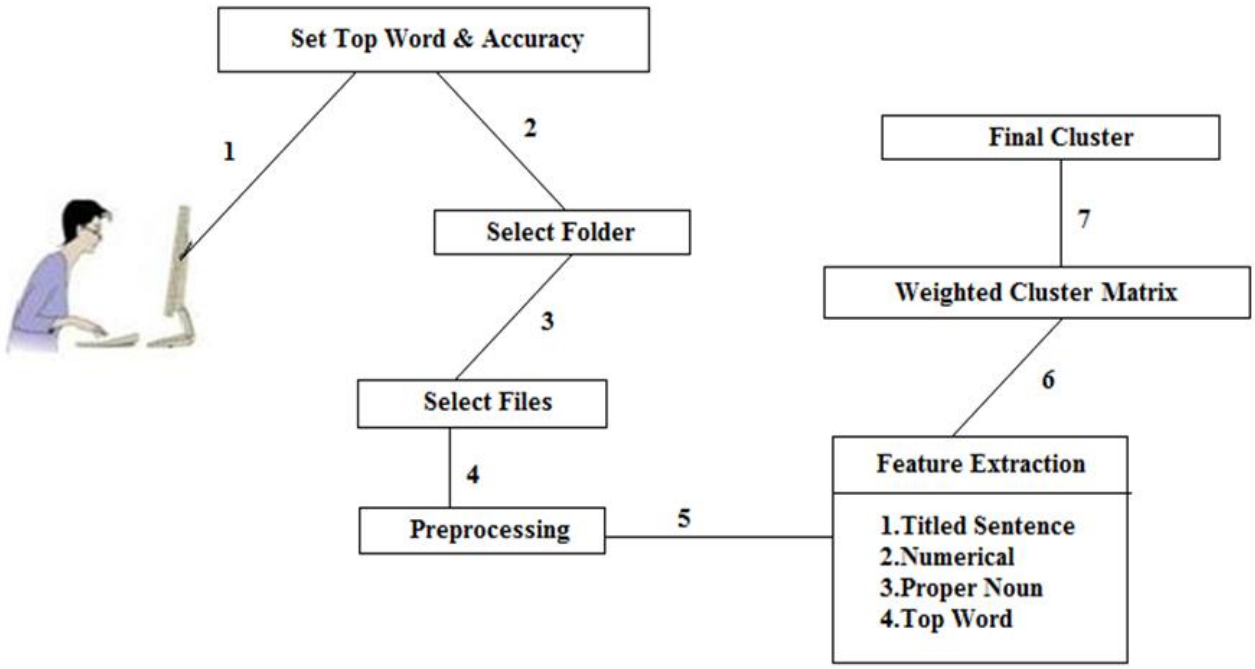

Fig 1: Block Diagram

IV. ALGORITHMS

1.1. Algorithm for Pre processing

Step 0: Start
Step 1: Get contents of Query
Step 2: split in Words
Step 3: Remove Special Symbols
Step 4: Identify Stopwords
Step 5: Remove Stopwords
Step 6: Identify Stemming Substring
Step 7: Replace Substring to desire String

Step 8: Concatenate Strings
Step 9: Preprocessed String
Step 10: Stop

1.2. Algorithm to find stop words

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its frequency
Step 6: Add frequency in List Called L
Step 7: end of for
Step 8: return L
Step 9: stop

1.3. Algorithm to find noun

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for i=0 to N (Where N is length of V)
Step 5: for i word of N check for its occurrence in Dictionary
Step 6: if present then return true
Step 7: else return false
Step 8: stop

V.CONCLUSION

To overcome some major disorders in document clustering this paper tries to propose a method using best idea which includes feature extraction based on fuzzy logic and weighted score matrix. This system will save the users effort and it will also save time in investigation process used in police investigation. The result produced by this system is quickly obtained and efficient by the above methods discussed.

VI.FUTURESCOPE

In future this system can be improved by modifying it to work efficiently on graphical, audio, video and image files. This system can be enhanced to work as an independent application programming interface. This system can also be enhanced and raised up as a cloud service provider.

REFERENCES

[1] Lus Filipe da Cruz Nassif and Eduardo Raul Hruschka, “Docu-ment clustering for forensic analysis: An approach for improving computer inspection”, IEEE transactions on information foren-sics and security, vol. 8, no. 1, January 2013
[2] L. Liu, J. Kang, J. Yu, and Z. Wang, “A comparative study on unsupervised feature selection methods for text clustering”, IEEE Int. Conf. Natural Language Processing and Knowl-edge Engineering, 2005, pp. 597601.
[3] A. L. N. Fred and A. K. Jain,” Combining multiple clusterings using evidence accumulation”, IEEE Trans. Pattern Anal. Mach.Intell., vol. 27, no. 6, pp. 835850, Jun. 2005.
[4] B. K. L. Fei, J. H. P. Elo_, H. S. Venter, and M. S. Oliver,“Exploring forensic data with self-organizing maps”, in Proc. IFIP Int. Conf. Digital Forensics, 2005, pp. 113123.B.E. Seminar Data mining

