ISSN:0975-9646

Heuristic Approach for Document Clustering in Forensic Analysis

Tushar Thopte[#1], Yogesh Indani[#2], Murlidhar Jangale[#3],Sharad Gaikwad[#4]

[#1234]*Fellow internship*

[#]*Innovatus Technologies, Pune, Maharashtra, India*

***Abstract*— The volume of data in digital world is growing exponentially, which has direct impact on forensic analysis. So there is a diverse need to find the quick method that can group the required documents. Numbers of algorithms like k-mean, agglomerative clustering are used for clustering purpose. Previously used algorithms deals with issues like handling outliers, data preparation etc. The Proposed System is pre-process unstructured document to structured data, then our idea extract four features of each document like title sentences, numeric words, proper nouns and term weights. This makes it much simpler than any other methods. The Proposed System neglecting unwanted extension's considering only extensions which are rich in text like .pdf, .doc, .txt. As the final step of clustering, system creates a score matrix of all the documents by comparing with one another to yield a score matrix which contains aggregate feature score. The grouping of these scored values represents the most accurate clustered documents**.

***Keywords*— Document clustering, Pre-Processing, Feature extraction, Cluster matrix.**

I. INTRODUCTION

As the time passes numbers of documents to be processed are increasing dramatically hence it is challenge for the forensic analysers to analyse this large set of documents. Proposed methodology groups the documents by using feature extraction method.

Clustering algorithms are widely used for large amount of unstructured data where there is no knowledge of data in advance [9] [10]. The main strategy behind the clustering methods is that the documents in the same cluster are more similar than other clusters. Mainly two types of clustering [6] algorithms i.e. hierarchical clustering and partition clustering are used for same purpose. Partition algorithms like k-means, k-medoid and hierarchical clustering algorithms like agglomerative clustering are the best methods for clustering, where hierarchical clustering having higher edge over partition clustering.

Proposed idea uses the characters of both the above mentioned types of clustering techniques, where system gets a directory containing no of documents as an input from which clusters are generated.

As an initial step of our system it needs to pre-process and then extract features by processing various methods. Stemming is one of the most widely used methods to pre-process any word, where it actually trims the word to its base form. Many famous algorithms are existing like port stemmer who trims almost all the words in English language with medium accuracy, so using these kinds of built in algorithms can greatly affect outcome clusters.

On the side of feature extraction, noun detection is the crucial part, where many NLP tools like wordnet greatly contribute for this process. The drawback of this is lies in its complexity of configuration with the system.

To introduce the whole process in simple narration, system performance is considered for a set of document as D= {d1, d2, d3}. The working pattern of our system describes as below with some simple scenario.

Documents	Content
D1	Philip Hughes dies at 25 by hitting the ball on his head in last match. Sean Abbott hits ball on his head.
D2	Sunil gavaskar answers the bowlers by his bat.He never wears the helmet.
D3	Now a day's almost all the digital devices makes use of java language because of its usability.

Table I: initial document content.

Documents	Content
D1	Philips Hughes die 25 hit boll head match. Abbott hit ball head
D2	Sunil gavaskar boll bat. Wear helmet
D3	Digital device java language

Table II: pre-processed data.

Documents	Title Sentence	Numerical Data	Proper Noun	Term weights
D1	Philips Hugh hit ball head match	25	Philips Hugh	Hit head Hugh
D2	Sunil gavaskar boll bat	0	Sunil gavaskar	Gavaskar,bat, boll
D3	Digital device java language	0	Java	Device, java, language

Table III: extracted features

	Document 1 (6,1,2,3)	Document 2 (4,0,2,3)	Document 3 (4,0,1,3)
Document 1(TS,P,N,TW) (6,1,2,3)	0	(1/6+0/1+0/2+0/3)/4=0.04	(0/6+0/1+0/2+0/3)/4=0
Document 2(4,0,2,3)	(1/4+0/0+0/3+0/3)/4=0.06	0	(0/1+0/1+0/3+0/3)/4=0
Document 3 (4,0,1,3)	(0/1+0/1+0/4+0/3)/4=0	(0/1+0/1+0/3+0/3)/4=0	0

Table IV: weighted feature score matrix.

Where, Title Sentence (TS), Numerical Data (P), Proper Noun (N), Term weights (TW).

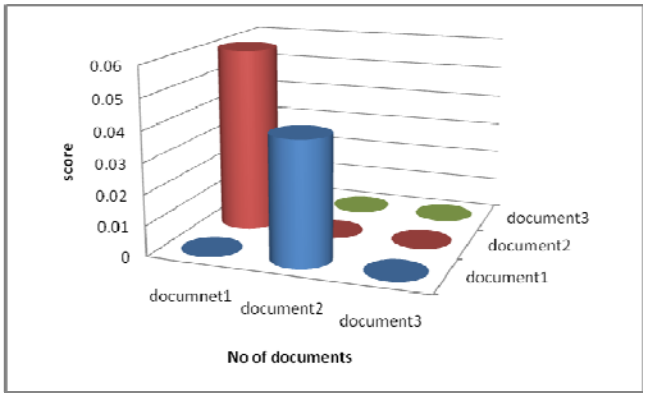

Figure 1: cluster formation with weighted score feature.

The above plot in figure 1 clearly indicates document 2 is having clearly correlation with document 1, so a cluster c1 will be created C_1= {D_1, D_2}. Similarly C_2 will be C_2= {D_1, D_2}.

The rest of the paper is organized as follows: Section II discusses some related work and section III presents the design of our approach. The details of the results and some discussions on this approach are presented in section IV as Results and Discussions. Section V concludes our system with some elaboration.

II. Related Work

Pre-processing is required for document clustering to remove unwanted data, [1] describes the need of pre-processing and how exactly the pre-processing is done. [2] illustrates the feature extraction using Fuzzy C - Means clustering which conclude that accuracy obtained using Fuzzy C-Means clustering for generic feature extraction is very close to the accuracy of classification obtained by using problem- specific feature extraction such as, ANN, SVM, BC, etc.

[3] Explains two Eigen vector based approaches. One is parametric and optimizes the ratio between-class variance to within-class variance of the transformed data. The Second approach is a nonparametric modification of the first one based on local calculation of the between-class covariance matrix.

The author in [4] presents a novel technique for texture extraction and classification. The proposed feature extraction technique uses 2D–DFT transformation. A combination of this technique and a Hamming Distance based neural network for classification of extracted features is investigated. The results obtained are very promising and showed that the proposed 2D-DFT based feature extractor has improved the classification rate significantly. The classification rate using the proposed technique based on 2D-DFT is approximately 26% higher than that of the algorithm without using 2D-DFT.

[5] Here feature extraction is done by using time series with wavelet and Fourier decomposition. Paper presents a new method for choosing the best coefficient instead of first coefficient as in earlier methods. [6] Represents a result of study which is conducted to find the difference of two clustering techniques i.e. hierarchical clustering and partition clustering.

[7] Talks about the techniques of agglomerative clustering algorithm, which comes under hierarchical clustering. Its merits and demerits. [8] Defines k means clustering which shows a better performance than partition clustering.

[9] Presents an approach to analyse the clusters. Different clustering techniques are well compared in [10] where it presents an advantage and disadvantage of one method with other methods. Here in this paper [11] NMF based feature extraction method is proposed which reduces the size of feature vectors compare to other methods.

III. Proposed Methodology

In this section, the paper discusses the heuristic approach for document clustering in forensic analysis with the below mentioned steps and this steps can be illustrated in figure 2 also.

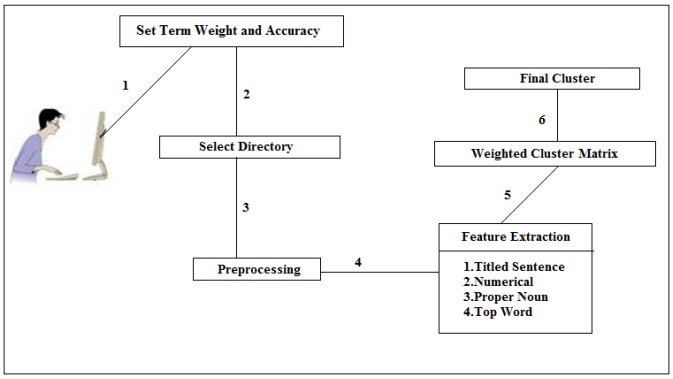

Figure 2: overview of our approach

Step 1:-Pre-processing.

This is the initial step which actually prepares the data for feature extraction to reduce the computation time by performing some steps as described below.

- ***Discarding special symbols.***
  Here all the special symbols from the documents are removed.
- ***Stop word removal.***
  In any document narration the conjunction words does not play much role in the meaning of the document, so by discarding these words (like: is, the, for, an) from the documents which greatly reduces the overhead of processing
- ***Stemming***
  Many of the elongated words in the English language generally fail to provide proper meaning in the given scenario and also they increases the computational time. So it is necessary to bring the words to their base form by replacing its extended characters with desired characters (Example: studied to study, where **ied** is replaced with **y**).

Step 2:-Feature extraction.

As mentioned in earlier segments the features of a document play a very great role in semantic clustering techniques. So our system makes use of four different features, which are extracted from the documents by using different techniques as mentioned below.

- ***Title sentence***
  In any document the title sentences are actually providing the most of the abstract of the narration. So the extraction of title sentence is contributing an important role in clustering. System considers the very first sentence of the document as the title sentence. Another use of title sentence is to assign a proper name to the outcome cluster.
- ***Numerical data.***
  The numbers in any narration greatly affects the quality of the document. So the system identifies the numbers and extract from a document to form a numerical vector.
- ***Proper noun.***
  Identifying and extracting a proper noun from a document always need a great support from a built in dictionary or an API. The problem lies to use a dictionary or an API is its integration complications. So our approach use a bag of all possible English words which are explicitly collected from a renowned sources and added in the database. System develops a procedure where it finds the word for its unavailability in the database to tag the word as a noun.
- ***Term weight.***
  The most repetitive words in document are obviously the important words. So system identifies the list of most repeated words and considers some top **n** elements (where n is user defined) as the important word for document to store in vector.

Step 3:- In this process of clustering all the feature vectors are merged to form a single merged feature vector. Now each index of this merged feature vector indicates the name of the respective document, which actually contains the four extracted feature values as mentioned in previous steps. Now each documents feature values are compared with one another and then they are aggregated to form a feature score matrix of all the input documents. This complete process can be analysed in the algorithm 1.

Step 4:- Here for every single document (d_i) the feature score values are compare with user defined accuracy with all other documents (d_1, d2....d_n).Then the feature score values higher than the user accuracy are selected as a cluster member for the respective document (d_i).So this formed cluster is labelled with title sentence of document d_i which actually indicate the cluster name. This can be visualized by the algorithm 1.

Algorithm 1

Input: Merged Feature vector F_v
User Accuracy as U_a
Output: Cluster Set C= {c1, c2, c3….c_n}
0: start
1: create matrix M of length F_v
2: **For** i=0 to F_v length (for each row)
3: **For** j=0 to F_v length (for each column)
4: F_{vr}= element of one row
5: F_{vc}=element of one column
6: Compare features and get score as S_c
7: Average Score as $A_{sc}=S_c/4$
8: add Average to matrix M
9: **End Inner For**
10: **End Outer For**
11: for every file in M's Rows if ($A_{sc}>=U_a$) then add into cluster C_i
12: **return** cluster set **C**
13: Stop

IV. Reults And Discussions

To show the effectiveness of proposed system some experiments are conducted on java based windows machine. To measure the performance of the system we set the bench mark for different number of document each of around 5 MB of size. System considers pdf, doc and txt extension documents for clustering and dynamically performs all the operation which is mentioned in above section and yield clustered documents with the following report

No of Documents	Time(s)
10	25
20	33
30	54
40	68
50	95

Table V : evaluation of time for different no of documents

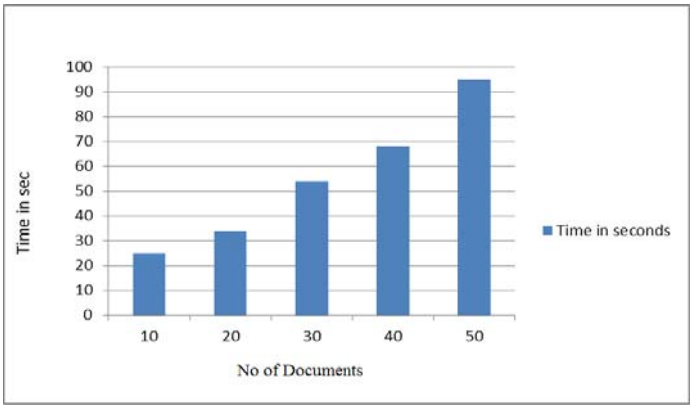

Figure 3:-performance measurement for different no of documents.

The above plot expresses result of clustering time which is not directly proportional to number of documents. So this indicates the system performs clustering based on semantic features not on the size of documents.

V. Conclusion and Future Scope

Many of document clustering technique are not based on semantic features of the content, rather than this they depend on structure of the document. So the idea proposed in this paper is perfectly tuned with semantic feature like Title sentence, Numerical data, and Proper noun and Term weight. The score matrix which is generated due to comparison of semantic feature of different documents is the master stroke of our idea. By using this score matrix documents are clustered for the user defined accuracy in all scenarios.

More feature like sentence similarity, thematic word, and verb score can be used to enhance the system of document clustering.

REFERENCES

[1] Navin Kumar Tyagi1, A.K. Solanki2& Sanjay Tyagi3," An algorithmic approach to data processing in web usage mining ",International journal of information technology and knowledge management

[2] Srinivasa K G * , Venugopal K R 1 and L M Patnaik 2, "Feature Extraction using Fuzzy C - Means Clustering for Data Mining Systems".

IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.3A, March 2006

Alexey Tsymbal 1,3 , Seppo Puuronen 1 , Mykola Pechenizkiy 2 , Matthias Baumgarten 3 , David Patterson 3, "Eigenvector-based Feature Extraction for Classification". FLAIRS-2002

[3] Yu Tao, Vallipuram Muthukkumarasamy, Brijesh Verma and Michael Blumenstein, "A Texture Feature Extraction Technique Using 2D–DFT and Hamming Distance". Compuatational intelligence and multimedia appliactions ,2003.

[4] Fabian M¨orchen ∗, "Time series feature extraction for data mining using DWT and DFT (November 5, 2003)".

[5] Ying Zhao and George KarypisC, "Comparison of Agglomerative and Partitional Document Clustering Algorithms ".

[6] K.Sasirekha,P.Baby,"Agglomerative Hierarchical Clustering Algorithm- A Review". Proceedings of the eleventh international conference on Information and knowledge management 2002

[7] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation". IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002

[8] B. S. Everitt, S. Landau, and M. Leese, Cluster Analysis. London, U.K.: Arnold, 2001.

[9] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[10] Osama Abu Samu computer science department yarmouk university,"comparison between data clustering algorithm", may 2 2007

[11] paresh chandra barman,Md. Sipon Miah, Bikash Chndra Singth,"Feature extraction clustering in text miningusing NMF basis probability", Ulab journal of science and engineering ,november 2,2011.