Web Document Clustering using Fuzzy Clustering : An Approach for Improving Document Inspection

Karan Kadu[1], Santosh Nhavkar[2], Rajesh Shirke[3], Swapnesh Kothari[4], Prof. Kavita Kimmatkar[5]
[1, 2,3 ,4, 5] Department of Computer Engineering
[1, 2, 3, 4, 5] SPPU, Zeal college of Engineering and Research, Narhe- Pune

Abstract*- Document Clustering is becoming vital for obtaining good results using unsupervised learning methods*

The approaches such as extractions and clustering are being increasingly used to improve the Document Clustering techniques. These approaches help reduce the problems in designing a general purpose document clustering.

The traditional fuzzy clustering methods are not suitable for sentence clustering because it is difficult to depict most of the sentence similarity measures in a common metric space. An enhanced Fuzzy clustering algorithm can be applied to the sentences of data sets to group the related sentences and documents.

This paper discusses two major sequential stages in Web Document Clustering "Extraction Features and Fuzzy Clustering Algorithm" as well as the major challenges and the key issues in designing extraction features and clustering algorithms.

These methods aid performance enhancement and help speed up the solving of crimes by the law enforcement officers and detectives.

In addition to web text domains, these algorithms can be incorporated for applications such as forensic analysis, data mining, bio-informatics, content-based or collaborative information filtering, social media, trend analysis, market analysis, banking sector and so forth.

Keywords*:* –Clustering, Fuzzy logic, Feature Extraction, Pre-processing, Stemming Algorithms

I. INTRODUCTION

Web documents are heterogeneous and complex. There exist complicated associations within one web document and linking to the others. The high interactions of terms within the documents show imprecise and abstruse meanings. A systematic and efficacious clustering method to discover latent and coherent meanings in context is a must. A way to discover the contextual meaning in the web documents includes using a fuzzy linguistic topological space along with a fuzzy clustering algorithm

Document Clustering is one of the most commonly used methods for detecting topics/events or types of crime documents. Document clustering composed of three main processes. The first process is pre-processing of documents which discard irrelevant words and symbols from the document. The second process is to extract the most important information called 'Feature' from the document. The last process of document clustering includes applying the Fuzzy clustering algorithm to the groups of documents consisting of topics/events or types documents based on the similarities among the documents.

This study looks forwards to restrict Document Clustering to two stages (Extraction of Document and Fuzzy cluster algorithm). Document Clustering is a popular area in the field of research and has been studied for ages, so it requires more improvements.

II. RELEVANCE IN CURRENT SCENARIO:

In Web clustering, thousands of files are usually scrutinized to reach a conclusion. The data in these files usually consists of unstructured text. It is difficult for computer examiners to perform the analysis of such documents. The Automated methods of analysis are of huge interest in such context. The algorithms for clustering documents can aid in the discovery of new and useful knowledge from the documents under analysis.

We present an approach for clustering algorithms for analysis of computer documents. The proposed approach has been previously illustrated by researchers by carrying out comprehensive analysis using well-known clustering algorithms. These include algorithms such as K-means, K - medoids, CSPA that is applied to data sets obtained from computers seized for performing investigations. The analysis has been performed with different compositions of parameters.

These methods have a lot of limitations associated with them. Our model depicts that Feature Extraction and

Fuzzy Logic algorithms for web document clustering give the best results. If suitably initialized, both the algorithms (Feature Extraction and Fuzzy Logic) can also yield very good results. Also, we present and discuss several practical results which can be helpful for researchers and practitioners of forensic computing.

III. METHODOLOGY

The first step includes creating an interactive web crawler. The crawler searches the different web pages and collects the data from them and then data are saved in text format. Now, the folder in which the web data is stored is given as the input to the system. The system performs pre-processing on the data for extracting the features like Term Weight, Numerical data, Title sentence, Nouns and then applies the fuzzy logic to get the feature scores classification pattern. This is given to the weighted matrix method which creates semantic clusters for the web page documents.

Here in this chapter, we are giving complete focus on the design of the system. Each and every stage of the proposed system is well narrated here. Along with the explanation, the complete system is well presented using the system architecture.

The following four steps show the working of the complete system:

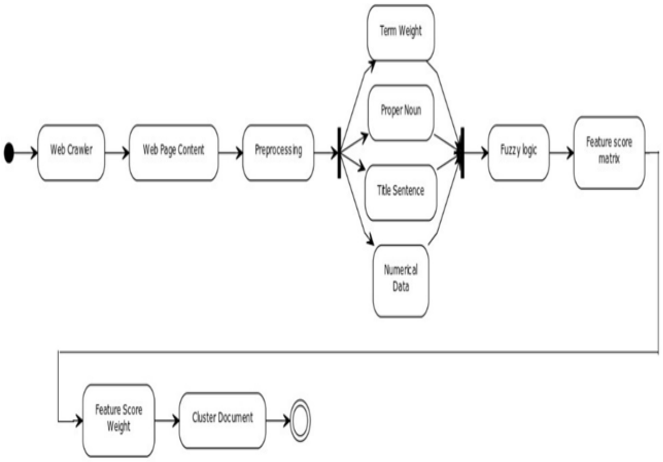


Figure1. System Architecture

1. Pre-processing.

Data Pre-processing improves the data quality. Pre-processing is vital step in data mining applications as it reduces the size of the data required for processing. This reduced size minimizes the cost and space complexity of the system. Generally, pre-processing is composed of three steps.

Special symbol removal:

Here, all the special symbols from the content are removed, e.g. !, @, #, $, % etc. These special symbols should be removed as they do not contribute to result generation.

Stop words:

Stop words are supporting words in the content used to bring the semantics in the sentence. Discarding the words doesn't change the meaning of the sentence too much. Hence, they are removed here by maintaining a repository for comparison. This repository has 500+ stop words.

Stemming:

The word is derived using stem. Generally, the words are derived from making the proper use of tenses. These stems unnecessarily increase the system cost. Hence, they are removed over here. There is no stemming algorithm which gives 100% accuracy.

2. Feature Extraction.

Feature extraction is an essential step in data mining. It extracts the required data i.e. features from the huge set of data. Here in our proposed work we extract four features:

Title sentence:

Title sentences represent the first sentence of the file content. The reason behind this extraction is to give a proper name to the cluster. Each cluster is named by the title sentences.

Numeric data:

Numeric data play an important role in file content as most of the important data are represented using numerical values. Therefore, the numerical values are extracted from the file content.

Proper nouns:

Proper nouns are the words which represent person or place. This extraction is performed using a dictionary. The API provides all the necessary functionalities to use this dictionary.

Top words:

Top words are important words in the sentence. Here, in this feature the frequency of the each word is found.

Consider, the word which repeats several times as it will have more weight in the file content.

3. Master matrix creation.

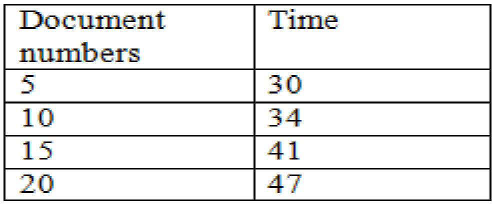
In this step, we take all the extracted features as input. Then, we create a matrix using these features. A Particular feature of each file is compared with the feature of the other file. In this way, all four features are compared with four features of another file. This comparison leads to feature score of each file with another file.

4. Fuzzy logic.

The input from the matrix is the generated score. We calculate the smallest and biggest scores. We calculate exactly five ranges starting from smallest the value and ending with the largest value. Now assign the score to score calculated in master matrix step and check the score in these five ranges. Once, the scores have been calculated a threshold of say ‘2' is set. The file having threshold more than two is added to the cluster and discards the file which fails to satisfy the condition.

IV. RESULTS AND DISCUSSIONS

To show the effectiveness of the system, we conduct an experiment is on Java 1.6 based machine using Net beans as an IDE. The Windows machine should preferably have 2GB RAM and 500GB HDD. After performing the experiment by providing files from different categories such as text, Pdf, and doc the following is observed.



Document numbers	Time
5	30
10	34
15	41
20	47

Table 1: Time required against numbers of documents

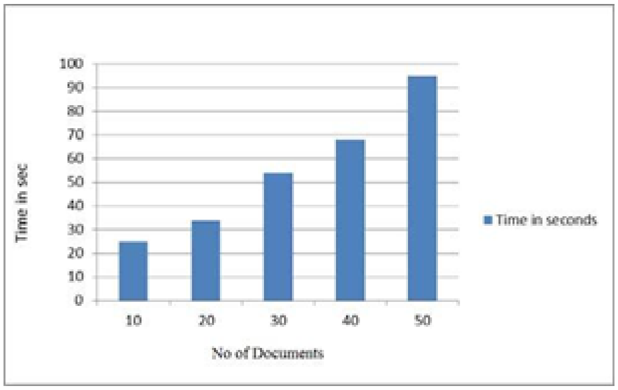


Figure 2: Performance measurement

The graph above represents the clustering time. From the graph, we can conclude that as the numbers of documents increase the required time is also increased.

V. CONCLUSION

The basic need of applying clustering techniques on web document arises mainly because it is difficult to cluster a huge number of the web pages semantically. The application developed examines thousands of files. Most of the data in these files consist of unstructured text, whose analysis by human examiners is difficult to be performed. We present an approach that applies web document clustering algorithms to web documents analysis. Also, the approach produces several practical results that can be very useful for researchers and practitioners of forensic computing.

In the future, additionally to web text domains, this application can be extended to applications such as forensic analysis, data mining, bio-informatics, content-based or collaborative information filtering, social media, trend analysis, market analysis, banking sector and so forth.

REFERENCES

[1] Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai and AjitKumar, Discovering Latent Semantics in Web Documents using Fuzzy Clustering, IEEE Transactions on Fuzzy Systems, DOI 10.1109/TFUZZ.2015.2403878.

[2] Luis Filipe da, Cruz Nassif and Eduardo Raul Hruschka, Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection, JANUARY 2013

[3] N. L. Beebe and J. G. Clark, Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, Digital Investigation, Elsevier, vol. 4, no. 1, pp. 4954,2007.Year of publication: 2007

[4] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and Intell. Security Inf. Syst., vol. 63, pp. 2936, 2009.

[5] K. Kishida, “High-speed rough clustering for very large document collections,”J. Amer. Soc. Inf. Sci., vol. 61, pp. 1092–1104, 2010, doi:10.1002/asi.2131.

[6] Aggarwal, C. C. Charu, and C. X. Zhai, Eds., “Chapter 4: A Survey of Text Clustering Algorithms,” in Mining Text Data, NewYork: Springer, 2012

[7] L. F. Nassif and E. R. Hruschka, “Document clustering for forensic computing: An approach for improving computer inspection,” in Proc.Tenth Int. Conf. Machine Learning and Applications (ICMLA), 2011,vol. 1, pp. 265–268, IEEE Press.

[8] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, “Mining write prints from anonymous e-mails for forensic investigation,” Digital Investigation, Elsevier, vol. 7, no. 1–2, pp. 56–64, 2010.

[9] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, “Relative clustering validity criteria: A comparative overview,” Statist. Anal.Data Mining, vol. 3, pp. 209–235, 2010.

[10] K. Stoffel, P. Cotofrei, and D. Han, “Fuzzy methods for forensic data analysis,” in Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition,2010, pp. 23–28

[11] R. Xu and D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009

[12] R. Hadjidj, M. Debbabi, H. Louis, F. Iqbal, A. Szporer, and D. Benredjem, “Towards an integrated e-mail forensic analysis framework,” Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137,2009