Implementation of Web Document Clustering Methods for Forensic Analysis

Karan Kadu[1] Santosh Nhavkar[2] Rajesh Shirke[3] Swapnesh Kothari[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]University of Pune, Zeal Collage of Engineering and Research, Narhe- Pune

Abstract— Web documents are diversified and complicated. Web documents involve complex associations and the linking to the other documents is also complicated. The interactions of terms in documents demonstrate imprecise and obscure meanings. There is need of systematic and effectual clustering methods to uncover untapped and rational meanings in context. Fuzzy clustering algorithm can be used to find the contextual meaning in the web documents. The main theme of clustering based techniques is to extract features from the web documents using conditional random field methods and build a fuzzy linguistic topological space based on the associations of features [1]. The associations of co-occurring features organize a hierarchy of connected semantic complexes called 'CONCEPTS,' wherein a fuzzy linguistic measure is applied on each complex to evaluate the relevance of a document belonging to a topic, and the difference between the other topics [1]. Web contents are able to be clustered into topics in the hierarchy depending on their fuzzy linguistic measures; web users can further explore the CONCEPTS of web contents accordingly [1]. Aside from web text domains, the algorithm can be applied to other applications such as forensic analysis, data mining, bioinformatics, social media, market analysis, banking sector and so forth.

Key words: Fuzzy Semantic, Stemming Algorithms, Clustering, Forensic Analysis

I. INTRODUCTION

Computer forensic analysis involves examination of thousands of files. Much of the data in these files consists of unstructured text, which makes it difficult for computer examiners to analyze it.

The Documents under analysis can be used to discover new and useful knowledge by applying algorithms for clustering documents. Also, we reported and discussed several practical results that can be very useful for researchers and practitioners of forensic computing.

II. RELEVANCE IN CURRENT SCENARIO:

Web clustering involves examination of hundreds of thousands of to reach to a conclusion. Much of the data in these files consists of unstructured text, which makes it difficult for computer examiners to analyze it. Automated methods of analysis are considered in this context.

We consider an approach that employs document clustering algorithms for forensic analysis of computers seized in police investigations.

III. METHODOLOGY

The magnitude of data in digital world is growing rapidly day by day, which directly impact the forensic analysis. So it is necessary to find a quick procedure to group the required documents. For clustering the documents, algorithms like k-mean, agglomerative clustering are used. Previous algorithms dealt with problems like handling outliers, data preparation etc. We propose a system that processes unstructured data to structured data and then extracts four features from each document like title sentences, numeric words, proper nouns and term weights. In the Proposed System we only consider extensions such as.pdf, .doc, .txt. Finally, a score matrix is created of all the documents by comparing them with each other to produce a score matrix that holds aggregate feature score. These scored values are grouped and they are used to represent the clustered documents with maximum accuracy

A. Preprocessing

Data preprocessing involves processing of raw data to prepare it for another processing procedure. Data preprocessing transforms the data into a more effective format so that it can be processed easily by the user - for example, in a neural network. Preprocessing involves number of tools and methods such as: sampling, which selects a representative subset from a large set of data; transformation, which transforms raw data to produce a single input; de-noising, which separates corrupt data from data; normalization, which arranges data for more coherent access; and feature extraction, which gives specified data significant in a particular context.

The following type of preprocessing is performed on raw data:

1) Removal of Special Symbol:

Special symbol like! @, #, $, %, *, ? etc. are removed in this process. These symbols not needed for any process so there is no need to keep them in the document. The document is freed of noise of special symbol.

2) Removal of Stop Words:

In computing environment, stop words are words that are removed before or after processing of natural language data (text). The most common words in a language are usually referred to as stop words. Natural language processing tools do not contain any single universal list of stop words. Also some tools even do not use such a list. To support phrase search some tools avoid removing stop words.

We can choose any group of words as the stop words for a given cause. In some search engines, stop words can be the most common, short functions words such as the, is, at, which, and on. In such cases, stop words may cause problems while searching for phrases that involve them. Other search engines may remove common words—including lexical words, such as "want"—from a query for improving the performance.

3) Removal of Stemming Word:

In computing environment, stop words are words that are removed before or after processing of natural language data (text). The most common words in a language are usually referred to as stop words. Natural language processing tools do not contain any single universal list of stop words. Also some tools even do not use such a list. To support phrase search some tools avoid removing stop words.

All rights reserved by www.ijsrd.com

B. Feature Extraction

Feature extraction includes an initial set of measured data and creates derived values (features) which are meant to be informative, non-redundant. The derived features facilitate the subsequent learning and generalization steps, in sometimes leads to improved human interpretations. Feature extraction is comparable to dimensionality reduction.

Sometimes large amount of input data needs to be processed by an algorithm and the data may be redundant. Then it is possible to transform this data into a reduced set of features (also called as "features vector"). This process is known as feature extraction. The extracted features hold related information from the input data. The desired task can be performed by using this reduced representation instead of the using the whole initial data.

The following types of feature extraction are performed on preprocessed data:

1) Title sentence
2) Numerical Data
3) Proper Noun
4) Top Word

C. Fuzzy Clustering

One of the essential techniques to discover contextual meaning or semantics from the returned heterogeneous web pages is Web document clustering. User's queries are imprecise and obscure. There are numerous document clustering methods which have been proposed; some uses probabilistic models equipped with distance and similarity measures, and while others use matrix factorization, and data mining techniques such as SOM.

A document is usually represented as a feature vector, which can be viewed as a point in the multi-dimensional space, to match users' queries. A set of key terms or phrases are selected to arrange the feature vectors based on the differences between documents to capture semantics to fit users' intents. Many methods, including k-means, hierarchical clustering, and nearest-neighbor clustering are used for this purpose. A method known as Suffix-tree clustering performs document clustering based on similarities between the documents. It is a phrase-based approach.

For documents with imprecise information, the use of fuzzy set theory is desirable. Fuzzy c-means and fuzzy hierarchical clustering algorithms can be employed for document clustering. These algorithms have some drawbacks as they require prior knowledge about 'number of clusters' and 'initial cluster centroids. The ant-based fuzzy clustering algorithms and fuzzy k- means clustering algorithms were proposed to address these drawbacks. These algorithms can deal with unknown number of clusters. The main limitations of these methods were the similarity measures and bag of words used to capture the semantics in the collection of documents. According to Vector Space Model, the similarity between two documents can be measured with vector distance, such as Euclidean distance, Manhattan distance, and so on. These methods do not take contextual meaning into consideration.

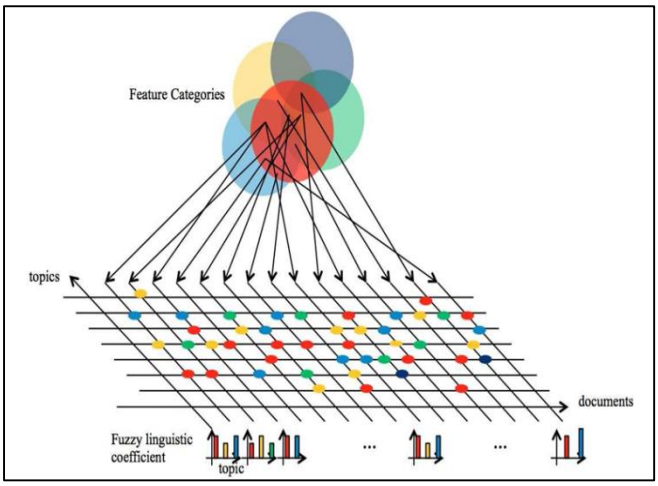

Fig. 1: Fuzzy Clustering

Fig. 1. This figure illustrates the structure of fuzzy linguistic topological space. Documents are composite of meaningful features that can be categorized into different topics with different possibilities. [1]

IV. FUTURE SCOPE

The basic ideas of computer forensic analysis is hundreds of thousands of files are usually examined to come to a conclusion. Much of the data in those files consists of unstructured text, whose analysis by computer examiners is difficult to be performed. In this context, automated methods of analysis are of great interest. In particular, algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis.
Besides the algorithm applicability in web text domains, it can be extended to other applications, such as forensic analysis, data mining, bioinformatics, content-based or collaborative information filtering, social media, trend analysis, market analysis, banking sector and so forth

V. CONCLUSION

This Paper show that by using fuzzy technique there will be web document clustering for forensic data which will be useful for police investigations. We presented an approach that applies document clustering methods to forensic analysis of computers seized in police investigations. Also, we reported and discussed several practical results that can be very useful for researchers and practitioners of forensic computing. More specifically, in our experiments the fuzzy algorithms show the better result than well-known algorithms like as Average Link and Complete Link. Despite their usually high computational costs, we have shown that they are particularly suitable for the studied application domain because the dendrograms that they provide offer summarized views of the documents being inspected, thus being helpful tools for forensic examiners that analyze textual documents from seized computers. As already observed in other application domains, dendrograms provide very informative descriptions and visualization capabilities of data clustering structures.

REFERENCES

[1] Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai, and Ajit Kumar, Discovering Latent Semantics in Web

All rights reserved by *www.ijsrd.com*

Documents using Fuzzy Clustering, DOI 10.1109/ TFUZZ.2015.2403878, IEEE Transactions on Fuzzy Systems

[2] LusFilipeda 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.

All rights reserved by *www.ijsrd.com*