ISSN No. 0976-5697

Volume 4, No. 2, Jan-Feb 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Text Document Featured Clustering and Classification using Fuzzy Logic

Ms. S. Nalini*
Assistant Professor
Department of Computer Science & Engineering
Velammal Institute of Technology,Chennai
nalini.tamilmani@gmail.com

Mrs. S. Selvakanmani
Assistant Professor
Department of Computer Science & Engineering
Velammal Institute of Technology,Chennai
sskanmani6@yahoo.com

Mrs. A. V. Kalpana
Assistant Professor
Department of Computer Science & Engineering
Velammal Institute of Technology,Chennai
kalpanavijay21@gmail.com

Abstract: Recent years huge amount of information are available in the World Wide web. 'Text Document Featured Clustering and Classification Using Fuzzy Logic' aims at obtaining effective clustering of text documents which can simply browsing of large volume of data set using semantic relationship between the words. Text clustering is mainly used for clustering a set of documents based on the user typed key term. 'A Fuzzy similarity-based algorithm' automatically classify a group of documents into set of groups with frequent concept. The set of documents and key terms which are entered by the user are pre-processed. The proposed system uses Fuzzy Featured Clustering(FFC) algorithm to identify the semantic relationship of words to create concepts. Text is classified based on features into number of clusters. The frequency of words are weighted by weighted matrix. The relationship between words like synonyms, hypernyms, hyponyms can also be identified. The proposed system is found more accurate, scalable and effective when compared to existing text clustering algorithms.

Keywords: Data Mining, Fuzzy Featured Clustering, Text Clustering, Classification, Fuzzy Logic

I. INTRODUCTION

Text mining is the process of unstructured or textual information. It extracts meaningful indices from the text. Text mining sometimes referred to as text data mining or text analytics. It refers to the process of deriving high-quality information from text. The purpose of Text Mining is to develop information contained in textual documents in various ways. It includes the discovery of patterns and trends in data, associations among entities, predictive rules. It can be implemented in the field of pattern recognition, neural networks, natural language processing, information retrieval and machine learning.

Text mining includes text categorization[1][3], text clustering, document summarization, etc. Text categorization means separation or ordering of object(or things) into classes from a predefined set automatically. It is otherwise known as text classification[6]**.** It is classified into two types. Apriori classification means classes are created without looking at the data(non-empirically). Posteriori classification means classes are created empirically(by looking at the data). Decision tree and Naive Bayes are important technique of Text classification.

Text Clustering or cluster analysis is the task of assigning a set of objects into groups (called clusters). It is used to organize objects into meaningful groups. K-means and Agglomerative methods are used in clustering. Text clustering and classification are the important techniques that partition object into meaningful disjoint subgroups.

II. PROPOSED SYSTEM

The proposed system has a frequent concept to cluster the text documents. Fuzzy Featured Clustering (FFC) algorithm is used to cluster the same similarity words into one group. This system reduces the dimensality of vectors.

a. The relationship between words are analyzed and clustered into synonyms, hypernyms, and hyponyms. They are clustered into meaningful groups.
b. The next process is the high dimensionality of text documents are reduced[2][4]. A clustering algorithm works with frequent concepts rather than frequent items used in traditional text mining techniques.
c. When compared with other text clustering algorithms, the Frequent Concepts based Document Classification algorithms is more efficient and accurate.

III. METHODOLOGY

This section presents the system architecture of Text document featured clustering, as shown in Figure 1.

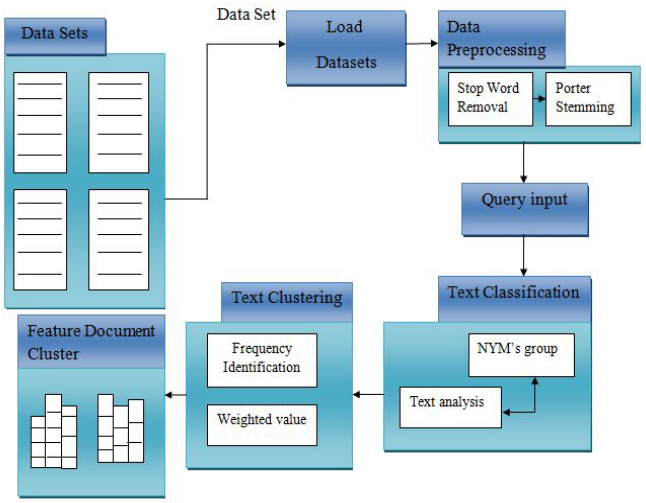

Figure 1. System Architecture

This proposed system is comprised namely, Load documents, Data Pre-processing, Text classification, Text Clustering.

A. Load Datasets:

An object (data record) typically has dozens of attributes and the domain for each attribute can be large data records. The dimensionality of the feature vector is usually huge. For example, 20 Newsgroups and Reuters 21578 top-10, which are two real-world data sets, both have more than 15,000 features.

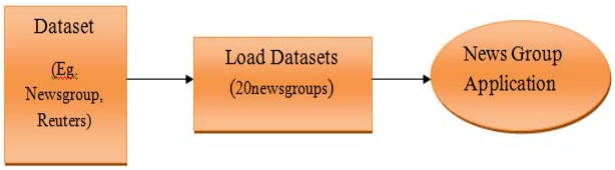

Figure.2 Load documents

B. Data Pre-Processing:

a. Stop Word Removal:

Stop word removal is a process of removing words. It can be implemented in the natural language data. It is controlled by human input and not automated. These are some of the most common, short function words, such as *the, is, at, which* and *on*.

b. Porter Stemming:

Stemmers utilize a lookup table which hold the relations between root forms and inflected forms. Lookup table is queried to find a matching inflection. The associated root is returned, if a matching inflection is found.

Example: A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

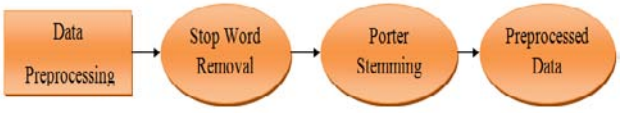

Figure.3 Data Pre-processing

C. Text Classification-Fcdc:

a. Nym's Group:

Words ending in nym's are often used to describe different classes of words, and the relationships between words. They can be classified as Hypernym, Hyponym and Synonym.

(a). A **Hypernym** is a word that has a more general meaning than another.

Example:1.In the relationship between chair and furniture, **furniture** is a hypernym.

2. In the relationship between horse and animal, **animal** is a hypernym.

(b). A **Hyponym** is a word that has a more specific meaning than another.

Example:1.In the relationship between chair and furniture, **chair** is a hyponym.

2. In the relationship between horse and animal, **horse** is a hyponym. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

b. Text Analysis:

The Artificial-Intelligence literature contains many definitions of ontology (WordNet).Word Net is a large lexical database **(**A lexical database is an organized description of the lexemes of a language**)** of English. Word Net is a semantic lexicon for the English language, which puts words in semantic relations to each other, mainly by using the concepts of hypernym and hyponym. The featured results obtained from Word Net produces sentence-based, document-based, corpus-based analysis. The combined approach concept analysis has higher quality than those produced by a single-term analysis similarity.

D. Text Clustering-Ffc:

Fuzzy Feature Clustering algorithm[5] is an incremental approach to reduce the number of features for the text classification task and form efficient clustering

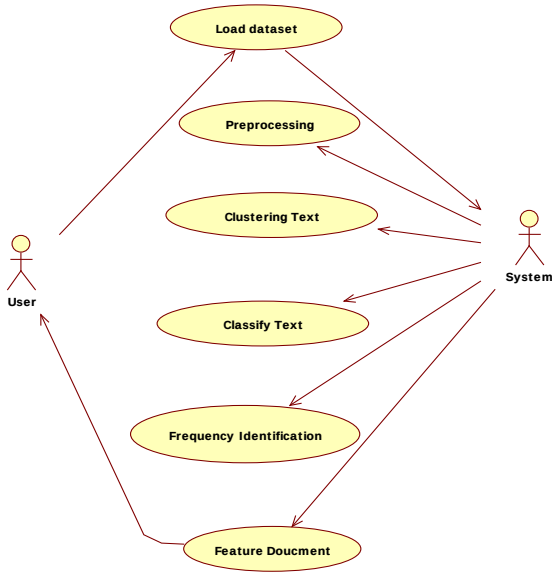

Figure. 4. Use case Diagram for Text Clustering

a. Frequency Identification:

Proposed system combines the definitions of term frequent and inverse document frequency to produce a composite weight for user term in each document. Traditional dimensionality reduction methods for text clustering, such as document frequency, mutual information, information gain provide required data for clustering the documents

b. Weighted Value:

Weighted Value is obtained by the membership function of the FFC algorithm and is represented in a weighted matrix.

c. Membership Function:

Let k be the number of currently existing clusters. The clusters are $G_{1,}G_{2,\dots,}G_{K,}$respectively. Each cluster G_J has mean m_j=<$m_{j1,}m_{j2,\dots}m_{jp}$> and deviation $\sigma_j = \sigma_{j1,}\sigma_{j2,} \dots, \sigma_{jp}$>.let S_J be the size of cluster G_J.initially, we have k=0, so, no clusters exist at the beginning. For each word pattern x_I =< $x_{i1,}x_{i2,} \dots, x_{ip}$ >, 1≤ i ≤m, we calculate according to the similarity of x_i to each existing cluster, i.e,

$$\mu G_J(x_i) = \prod_{q=1}^{p} exp\left[-\left(\frac{x_{iq}-m_{jq}}{\sigma_{jq}}\right)^2\right] \text{ For } 1 \le j \le k,$$

We say that x_i passes the similarity test on cluster G_j if $\mu G_j(x_i) \ge p$.

Formula

$$t_{ij}=\begin{cases}1, & if j = \arg max_{1\le\alpha\le k}(\mu G_\alpha(X_i)),\\ 0, & \text{Otherwise}\end{cases}$$

$$t_{ij} = \mu G_j(X_i)$$

$$t_{ij} = (\gamma) \times t_{ij}^{H} + (1-\gamma) t_{ij}^{S}$$

Where t_{ij} is used for the calculation of hard, soft and mixed weighting methods in the given order respectively.

The extracted feature[7] corresponding to a cluster is a weighted combination of the words contained in the cluster. There are three ways of weighting hard, soft, and mixed. In hard weighting method each word of the original feature belongs to exactly one word cluster. In soft weighting method number of clusters is small and contains more details of document cluster.

Mixed weighting method is a combination of hard and soft weighting where the user can decide the type of clustering needed. Feature clustering is an efficient approach for feature reduction, which groups all features into some clusters, where features in a cluster are similar to each other.

IV. EXPERIMENTAL RESULTS

Final result is mainly concentrated on obtaining the clusters of the document set and the relationship weighted graph. For hypernym, hyponym and synonym 3 different cluster types and graphs are obtained individually.

These results are based on the semantic relationship of user typed key terms and the Wordnet results hypernym, hyponym and synonym of the given term.

The following is a result obtained from 50 document sets which are loaded into the database. The user typed key terms used here are 'good reality people'.

For these 3 input terms, system produced the results hypernym, hyponym and synonym of each terms, clusters are formed based on the result and relationship graph.

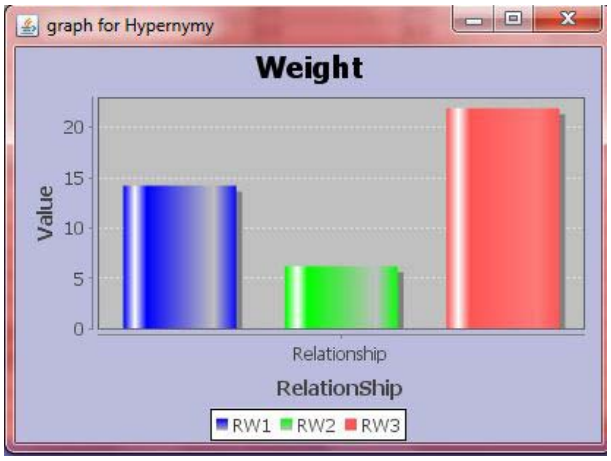

Figure 5. Graph for Hypernymy

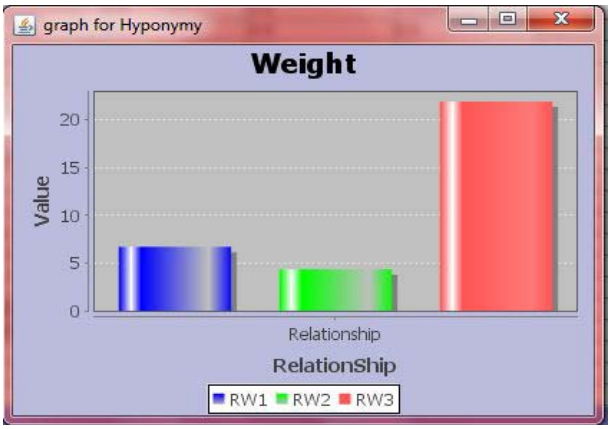

Figure 6. Graph for Hyponymy

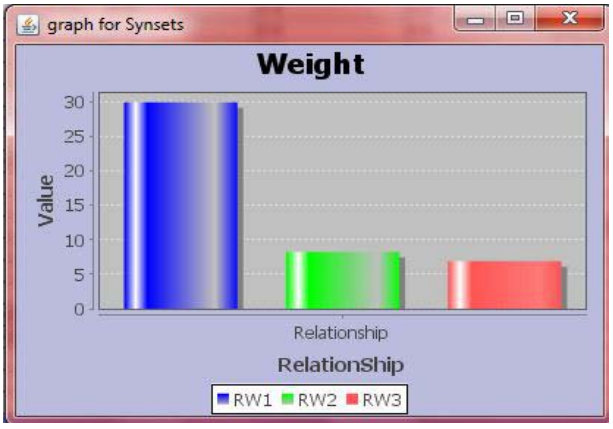

Figure 7. Graph for Synsets

The above graphs 5,6 and 7 are prove that three different types of results can be obtained. The nature of graph depends on the values of feature cluster and value of weight.

V. CONCLUSION AND FUTURE ENHANCEMENTS

A Fuzzy Featured Flustering (FFC) algorithm which is a feature evaluation approach of text clustering is used. The

derived membership function of this method matches closely and describes properly the real distribution of the training data (word sets).Also a Frequent Concept Based Document Classification (FCDC) approach which reflects the use of various conventional classification based on the meanings of the text words is used.

The semantic relationship between various word sets have been derived which has contributed for making an efficient feature cluster.

Cluster thus formed had semantic relationship between the terms and found to simplify the text document search. Similarity based clustering is one of the techniques that can also be applied in machine learning research. This problem can be applied to other problems such as image segmentation, data sampling, fuzzy modeling, web mining etc.

It is found that when a document set is transformed to a collection of word patterns, the relevance among word patterns can be measured and the word patterns can be grouped by any similarity based approach.

This concept has its own ways of enhancement in many of today's applications and research. In many scientific and commercial applications, this concept can be blended with natural language processing (NLP) and other translates entities.

VI. REFERENCES

[1] H. Al-Mubaid and S.A. Umair, "A New Text Categorization Technique Using Distributional Clustering and Learning Logic", IEEE Trans. Knowledge and Data Eng., vol. 18, no. 9, pp. 1156-1165, Sept. 2006.

[2] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi,and Z. Chen, "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing", IEEE Trans. Knowledge and Data Eng., vol. 18, no. 3, pp. 320-333, Mar. 2006.

[3] E.F. Combarro, E. Montanes, I. Diaz, J. Ranilla, and R. Mones,"Introducing a Family of Linear Measures for Feature Selection in Text Categorization", IEEE Trans. Knowledge and Data Eng., vol. 17, no. 9, pp. 1223-1232, Sept. 2005.

[4] H. Kim, P. Howland, and H. Park, "Dimension Reduction in Text Classification with Support Vector Machines", J. Machine Learning Research, vol. 6, pp. 37-53, 2005.

[5] Jung-Yi Jiang, Ren-JiaLiou, and Shie-Jue Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification", IEEE Transactions On Knowledge And Data Engineering, VOL. 23, NO. 3, MARCH 2011.

[6] J.S. Wang and C.S.G. Lee, "Self-Adaptive Neurofuzzy Inference Systems for Classification Applications", IEEE Trans. Fuzzy Systems, vol. 10, no. 6, pp. 790-802, Dec. 2002.

[7] K. Daphne and M. Sahami, "Toward Optimal Feature Selection," Proc. 13th Int'l Conf. Machine Learning, pp. 284-292, 1996.