ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Development of Fuzzy based categorical Text Clustering Algorithm for Information Retrieval

S.M. Jagatheesan[1], V. Thiagarasu[2]

Associate Professor, Dept. of Computer Science, Gobi Arts & Science College, Gobichettipalayam - 638453, India[1]

Associate Professor, Dept. of Computer Science, Gobi Arts & Science College, Gobichettipalayam - 638453, India[2]

ABSTRACT: Similarities play a vital role in clustering text on the prediction, in order to produce an efficient result when compared to the existing algorithms like k-modes, ROCK and STIRR. Future selection is important for making a subset according to the dataset. In order to overcome the problems in the existing system, single cluster and multiple clustering methods are proposed in order to cluster the famous quotes with multiple semantic associations. But the problems on overlapping between the quotes are analyzed and the sentence similarities for information retrieval are measured. A FUZZY logic in finding the similarities to form a cluster, based on the relational prototypes has been proposed. A semantic clustering and FUZZY based pruning approach is practiced to bring more accuracy in mining process. FUZZY makes possible on using more complex prototypes that should be represent on the clustered text. The algorithm identifies the semantically related sentences and avoids duplication on the given data set. The information retrieval based on the keyword in which filtering is processed on the benchmark dataset. The result states the information retrieval based on the FUZZY algorithm maximizes the effectiveness.

Keywords: Benchmark dataset, Feature subset, filtering, FUZZY based clustering

I. INTRODUCTION

In text processing, the major part is sentence clustering and sentence clustering is nothing but grouping of sentences which are similar meanings into clusters. The task is performed by applying standard clustering algorithms to group sentences into clusters. There are various traditional methods are followed in practice that represents the sentences as vectors in term space and applies best clustering algorithm to achieve the result accuracy. The most common algorithms such as K-mode, ROCK (RObust Clustering using linKs) and STIRR (Sieving Through Iterated Relational Reinforcement) have been commonly used for text clustering. A query redirection method has been proposed to improve the K-means clustering algorithm performance and accuracy in distributed environment [1]. A brief survey on optimization approaches to text document clustering was carried out which limits to provide clustering on semantic to make the quality of text document clustering [2]. Different existing Text Mining Algorithms are briefly reviewed stating the merits / demerits of the algorithms [3]. ROCK clustering has been developed to decrease the query response time by searching the documents in the resulted clusters instead of searching the whole database [4]**.** QROCK has also been developed to compute the clusters by determining the connected components of the graph [5]. This leads to a very efficient method of obtaining the clusters giving a drastic reduction of the computing time of the ROCK algorithm. The difficulties in clustering algorithm are language variability which has same meaning but it is phrased on two ways. A novel sentence clustering scheme has been presented based on FUZZY logic on sentences over the term clusters. The major challenge in implementing the sentence clustering approach is language variability, where the same meaning can be phrased in various ways. The shorter the sentences are the less effective becomes exact matching of their terms. Various traditional methods in text processing are mainly focused on reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility [6], [7]. The embedded methods involves in feature selection which to be a part of training process that are usually meant for learning algorithms [8]. But the traditional machine learning algorithms like decision trees or artificial neural networks are discussed which are all depend on the embedded approaches [9].

In combining filter and wrapper methods, achieve effective result was achieved with a particular learning algorithm with similar time complexity of the filter methods [10], [11]. In respect to the previous filter feature selection

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

methods, the cluster analysis has been evaluated to be the more effective than traditional feature selection algorithms. Many researchers employed the distributional clustering of words to reduce the dimensionality of text data [12], [13]. In the cluster analysis approach, the graph-theoretic methods are examined which was used in many applications. The result is more effective when compared to the human performance [14]. The general graph-theoretic clustering: compute a neighborhood graph of instances, and then delete any edge in the graph that is much longer or shorter to its neighbors. The final result is compared to the forest and each tree in the forest represents a cluster. In the same manner, in this research, the graph-theoretic clustering method is applied for features selections. Text clustering at the document level is well established in the Information Retrieval (IR) literature, where documents are considered as a data points in these the high dimensional vector space in which each dimension are corresponds to a unique keyword [15], leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents. This type of data, refer to as "attribute data", is amenable to clustering by a large range of algorithms.

The challenges in sentence level clustering in which larger segment of text, such as documents that have highlighted important points. In information retrieval, the vector space model is adequately to capture much of the semantic content of the document-level text because these are based on the word co-occurrence [16]. References [17], [18] both used PageRank for ranking sentences for the purpose of extractive text summarization. In their approach, the important sentence is at central on which the ranking is based on the centrality. When considering the feature set selection algorithm those are effective in eliminating irrelevant features but that not succeeded in handling redundant features [19]-[24]. Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by [25].

II. PROPOSED ALGORITHM

The initial approach in this method is feature reduction. After giving the input that is the dataset it undergoes several preprocessing based on that a result will be generated. The result is passed to the word net that avoids the words which are all not meaningful and unwanted words. Then, the resultant data is processed based on the fuzzy logic to get the desired output.

A. *Preprocessing*

All the web data are Meta data or xml format on that tags are removed along with the unwanted symbols and other specifications. The each of the sentences is fragmented into single words and those words are passed for next process: stop words and stemming.

1) *Stop Words:* Filtering of conjunctions and words given for removal process. The stop words can make the search process to tedious rather than simple. Keywords are simpler for information retrieval in a large database. This method is designed in such way that it removes the words which are to be removed from the given dataset. Fig. 1 shows the keywords after filtering from the stop words from the given dataset.

2) *Stemming:* Stemming is the process of removing the prefixes and suffixes. The stemming algorithm generally used to find the root word from the sub word. The retrieval decision is made by comparing the terms of the query with the index terms. In mining, every word has a root word.

For example, "stemming", "stemmed" & "stemmer" are the words from the root word *stem*. It makes the retrieval process simple by making the searching keyword minimum and based on the root word and a cluster is formed. It also reduces the dictionary size, that is, the number of distinct terms which effective use the storage area and maximize the processing speed. Fig. 2 shows the keyword after completion of stemming process.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

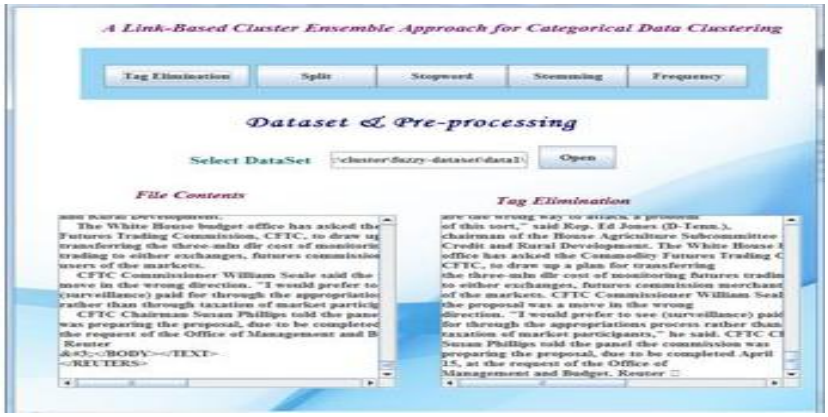

Fig. 1 Shows the stop word process

B. *Feature Reduction*

The data from the above phase is sent to word net. The word net is analyzed to check whether the extracted keyword is meaningful in order to form the cluster or not. At the same time, it validates weather any duplication is present or not. For this process, the frequency analysis term is used in order to define the frequency of a particular word (Fig. 3).

C. *Feature Clustering*

In feature clustering, separated keywords are analyzed from the above process and based on that result the feature occurrence is traced. The traced word from selection method which is cluster based on the occurrence of given dataset. The word pattern is compared with similarity search which generally based on the benchmark dataset. In data mining, directly finding a clustering process which minimizing the blocks such process is considered as a scary process. To make the process more efficient, the inner product of features from distinct one is explored. It requires minimum time for clustering rather than the traditional methods. For this, a variety of benchmark data set is given as an input and the feature clustering is formed after analysing the preprocess approaches.

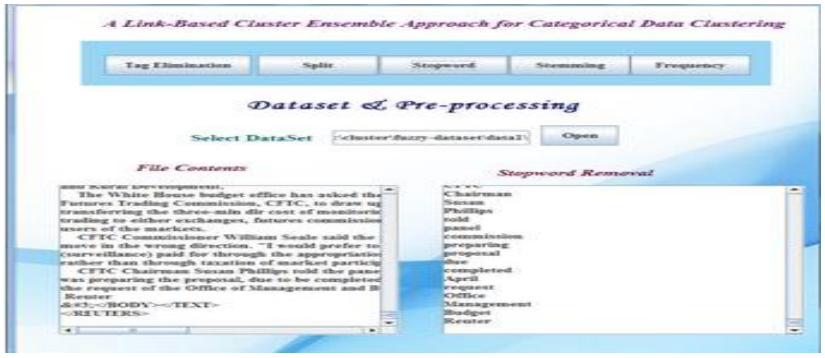

Fig. 2. Keywords after the stemming Process

The results across datasets are very reliable. A FUZZY algorithm has been proposed in which the result belongs to a single cluster. A semantic clustering and FUZZY based pruning approach is practiced to bring more accuracy in mining process. Generally, fuzzy clustering based on the prototypes or mixtures of Gaussians which does not support the sentence clustering. The proposed algorithm identifies the semantically related sentences and avoids duplication on the given data set. The information retrieval is based on the keyword in which filtering is processed on the benchmark dataset. FUZZY based approach uses weighting schemes from information retrieval, in order to assess the importance of whole attributes and individual values. The works is intended immediate retrieval of response based

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

on the input query. The clusters are identified which related to the concepts based on the given queries by the user. The mixing co-efficient of the values is expressed for page rank along with the similarities within the objects.

Fig. 3. Result of feature reduction

Then the equation can be expressed in a similarity matrix such as $sim = \{sim_{xy}\}$ where x and y is the similarity between the objects. Then, the weight between the matrixes is w_{ij}^{c} where c is the cluster.

Then,

$$sim_{x\&y}(w1, w2) = \frac{1}{IC(w1) + IC(w2) - 2 * IC(DCS(w1, w2))}$$

where the w is the words from the information content, IC is the information content and DCS is deepest common similarity. Then the IC(w) can be calculated by IC(w)=-log P(w) that is probability of the word w appear in the IC information content. Fig. 4 shows the feature clustering implementation.

Fig. 4. Feature clustering

III. DATA CLUSTERING

All the above processes are based on the data which are given as the input. The main objective of clustering is that it enables to understand large collections of data. Redundancy is calculated based on a prearranged model for the data. The structural clues in a data instance, which may contain errors, missing values, and duplicate records, are identified. The information may be of two types: structural information and un-structural information in which the structural information in order to decompose large software systems. At the same time, the un-structural information states such as file names or ownership information, have also demonstrated. The main thing is data clustering to assess

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

the importance of whole attributes and individual values. The data sets of interest are becoming larger, and their dimensionality prevents easy analysis and validation of the results.

A. *Measuring Keyword similarities*

Similarity is the term of information theory and similarity can be applied to multiple domains where various similarity measures had previously been proposed earlier. Generally, the actual similarity is not directly derived from the formula. All similarities are based on the set assumptions and a similarity is based on the commonalities between two terms.

Fig. 5. Result of hard, soft & medium weight

The maximum similarity is the identical match between two terms. A similarity measure can then be derived from those assumptions. A similarity is based on the comparisons of information in the commonality and the amount of information in the description of the two objects. The similarity paves a way to know how much more information is needed to determine what these two objects are. According to the process, the string similarity is art of retrieving from a word list the words that are derived from the same root as a given word. The methodology of similarity measure is position the words in the word list on descending order of their similarity to the given word. The similarity measure should be based on the words which was derived from the same root as the given word should appear early in the ranking. The similarities are categorized into three types such hard weight, soft weight & medium weight [Fig. 5].

IV. EXPERIMENTAL RESULTS

The algorithm based on the fuzzy logic is examined with the benchmark dataset. The similarity measures which shows the efficiency of the result accuracy, the number of selected features, the proportion of selected features and the equivalent runtime for each feature selection algorithm on each data set. The mean accuracy of each classification algorithm has been obtained under each feature selection algorithm and each data set.

The result classifies entities taking the class of the nearest associated vectors in the benchmark training set via distance metrics. This is the main feature of the proposed algorithm when compared to the existing algorithms. The obtained result shows the results of selected features, the time to obtain the feature subset and the classification accuracy (Fig. 6). It gradually increases the efficiency of clustering that compared to the traditional one. The cluster obtained is based on the mean and deviation based on those calculations.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Fig. 6. Cluster information

V. CONCLUSION & FUTURE WORK

The experimental results show that the clustering of sentence using FUZZY rule work in an efficient manner on considering with the feature extraction based on the processing time and overlapping clusters. A mean and deviance result gives the similarity measures on the basis of hard, soft and medium similarities. On the experimental result, the cluster information obtained gives the number of times the word existence on the given benchmark dataset.

This work further can be extended to produce the efficiency of FUZZY clustering on the basis of centrality measures that can be represented in graphical ways. Future research can also deal with hierarchical fuzzy relational clustering algorithm in an effective manner.

REFERENCES

1. Manpreet kaur and Usvir Kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 7, pp. 1454-1459, 2013.
2. Jensi, R. and d G.Wiselin Jiji, "A Survey On Optimization Approaches To Text Document Clustring", International Journal on Computational Sciences & Applications (IJCSA), Vol.3, No.6, pp. 31-44, 2013.
3. Sayantani Ghosh, Sudipta Roy and Samir K. Bandyopadhyay, " A tutorial review on Text Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, 2012.
4. Ashwina Tyagi and Sheetal Sharma, "Implementation of ROCK Clustering Algorithm For The Optimization Of Query Searching Time", International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 05, pp.809-815, 2012.
5. M. Dutta, A. Kakoti Mahanta and Arun K. Pujari, "QROCK : A Quick Version of the ROCK Algorithm for Clustering of Categorical Data", Proceedings of SDIS'01, National Workshop on Soft Data Mining and Intelligent Systems, Tezpur, India, 2001.
6. H. Liu, H. Motoda, and L. Yu, "Selective Sampling Approach to Active Feature Selection", Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74, 2004.
7. L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation", Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.
8. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", J. Machine Learning Research, vol 3, pp. 1157- 1182, 2003.
9. Generalization as Search, "Artificial Intelligence", 1998.
10. M. Dash and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.
11. S. Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection", Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

12. F. Pereira, N. Tishby, and L. Lee, "Distributional Clustering of English Words", Proc. 31st Ann. Meeting on Assoc. for Computational Linguistics, pp. 183-190, 1993.

13. L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification", Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.

14. J.W. Jaromczyk and G.T. Toussaint, "Relative Neighborhood Graphs and Their Relatives", Proc. IEEE, vol. 80, no. 9, pp. 1502-1517, 1992.

15. G. Salton, "Automatic Text Processing: The Transformation, Analysis,and Retrieval of Information by Computer", Addison-Wesley, 1989.

16. C.D. Manning, P. Raghavan, and H. Schu¨ tze, "Introduction to Information Retrieval", Cambridge Univ. Press, 2008.

17. R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts", Proc. Conference Empirical Methods in Natural Language (EMNLP), pp. 404-411, 2004.

18. G. Erkan and D.R. Radev, "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization", J. Artificial Intelligence Research, vol. 22, pp. 457-479, 2004.

19. G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", J. Machine Learning Research, vol. 3, pp. 1289-1305, 2003.

20. M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning", Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.

21. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF", Proc. European Conf. Machine Learning, pp. 171-182, 1994.

22. K. Kira and L.A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm", Proc. 10th Nat'l Conf. Artificial Intelligence, pp. 129-134, 1992.

23. M. Modrzejewski, "Feature Selection Using Rough Sets Theory", Proc. European Conf. Machine Learning, pp. 213-226, 1993.

24. M. Scherf and W. Brauer, "Feature Selection by Means of a Feature Weighting Approach", Technical Report FKI-221-97, Institut fur Informatik, Technische Universitat Munchen, 1997.

25. P. Pereira Langley, "Selection of Relevant Features in Machine Learning", Proc. AAAI Fall Symp. Relevance, pp. 1-5, 1994.

BIOGRAPHY

S.M. Jagatheesan received his Master Degree in Mathematics from Gobi Arts & Science College, Gobichettipalayam in 1984, M.Phil degree in Computer Science from Bharathiar University, Coimbatore in 1996 He is presently working as an Associate Professor in Computer Science, Gobi Arts & Science College since 1988. His current research centered on Data Mining and warehousing.

V. Thiagarasu received his Master Degree in Mathematics from Gobi Arts & Science College, Gobichettipalayam in 1985, M.Phil and Ph.D in Computer Science from Bharathiar University, Coimbatore in 1996 and 2010 respectively. He is presently working as an Associate Professor in Computer Science, Gobi Arts & Science College since 1989. He has also completed a UGC sponsored minor research project during 2004. His current research centered on the Networking, Multi-agent system, Concurrent Engineering and Project Scheduling.