Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Design of a FUZZY logic based Categorical Text Clustering Algorithm for Information Retrieval

Jagatheesan S.M, Thiagarasu. V

Associate Professor of Computer Science

Gobi Arts & Science College, Gobi-63845. INDIA.

Abstract - Information is retrieved from wide range dataset by applying effective algorithms like k-modes, ROCK and STIRR algorithms. The usage of database in the emerging technology is extreme and on the other hand the major drawbacks with database are incorrect or incomplete or outdated data. In this paper, the clustering of text by using the above mentioned algorithms has been discussed. The method of categorizing the data is a major part in the information retrieval. While some of the algorithms available at present cannot handle categorical data the others are unable to handle the stability problems and also have efficiency issues. An importance is shown in increasing the scalability of data mining for the clustering of data by combining multiple attributes. The existing algorithms have facing the major problems in handling the larger and complex database which are difficult to cluster. Those database and structure are more complicated to understand. In order to improve the scalability of clustering in various aspects like memory requirement, execution time, and improving the quality, a novel FUZZY logic method has been introduced in categorical text clustering. Most of the Fuzzy logic approaches cannot satisfy the sentence clustering and an algorithm has been designed with pair wise similarities between the data objects. The centrality measures and result of the algorithm to clustering approach is capable in recognizing the semantically related sentences by implementing multiple mining tasks.

Keywords: Information retrieval, data mining, text clustering, fuzzy logic and pair wise similarity

I. INTRODUCTION

Clustering is an efficient method for analyzing unsupervised text. Generally content may be explored with large amount of sentences and various clustering techniques are discussed. Clustering is a method for grouping of object set those may be similar or related to them. The main thing in this clustering is similarity or correspondent meaning of particular term. Text processing is based on the sentence clustering which reduces the complexities in data mining. The structural summaries are analyzed and are helpful for characterizing among the information content of the data, and eventually in the database design. Existing algorithm for categorizing those data is used in an effective manner. The structure clustering is based on the information's which are obtained from different kinds of sources. This research helps in web searches and mining to be more similar than the traditional one. The information of similarity and non-similarity must be clear before doing the research on clustering. The initial stage in this process is data collections on which several data from various source can be obtained those are separated from individual set of attributes. Certainly a data should be in order before implemented on a clustering algorithm for these data characteristics and dimensionalities are seriously analyzed. Next thing is cluster tendency on which the data to be cluster or not. The problem in this research is to determine the attributes in a proper way that is how far or how similar the data are from one another. The minimized or maximized appropriates are to be noted during the final results.

The most common algorithms such as K-mode [1], ROCK (RObust Clustering using linKs) [2] and STIRR (Sieving Through Iterated Relational Reinforcement) [3] are discussed. The difficulties in clustering algorithm are language variability which has same meaning but it is phrased on two ways. By making the sentence smaller it would be exact matching of their terms. By doing this, one can expect the cluster which can be closely matched to the concepts described based on the query terms. But most of the documents have irrelevant details of topics or themes, and many sentences will be related to some degree. The calculation between the pair wise similarities or dissimilarities can be performed with data points that should be done from attribute data based on similarities such as cosine similarity. This can be also applicable for relational clustering algorithms.

II. LITERATURE REVIEW

A .K- Modes Algorithm

This is the extension of k-means algorithm, generally in k-means used for partition which are reasonably efficient in the sense of within-class variance. The procedure used in the k-means is easy to program and computation. It can be applicable for analyzing similarity nonlinear forecast, approximating multivariate, and grouping for testing nonparametric independence among several variables. The result has been obtained based on the asymptotic behavior [4]. In the K-mode algorithm various distance measures are implemented. Let us take the data objects as *a* and *b*, then their distance is expressed by

$$d(a,b) = \sum_{i=1}^{n} \delta(a_i, b_i)$$

in which

$$\delta(a_i, b_i) = \begin{cases} 0 \; if \; a_i = b_i \\ 1 \; if \; a_i \neq b_i \end{cases}$$

The above equation explains the difference on two data objects having the corresponding attributes. The values of the attributes in the dataset are (1; 2; 18; 3; 18) and the value of the mode is 18. Then the dimensionality is *n*, for every clustering *c*, 1 ≤*c* ≤ k, which can be defined vector as $Q^c = Cx_1^c, x_2^c \ldots \ldots x_n^c)$ the entries are based on individual attributes.

Then the expression of Q^c is;

$$E = \sum_{c=1}^{k} \sum_{a}^{k} \; d(a, Q)$$

It can be applicable for large number of inputs. A query redirection method has been proposed to improve the K-means clustering algorithm performance and accuracy in distributed environment [9]. 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 [10]. Different existing Text Mining Algorithms are briefly reviewed stating the merits / demerits of the algorithms [11].

B. ROCK Algorithm

The ROCK algorithm is a robust hierarchical clustering algorithm in which it links the distance based on the notion of links and the number of links between two tuples is the number of common neighbors they have in the dataset during the cluster [2]. These are discussed with the non-metric similarity measures which are based on the relevant situations. 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 [7]**.** Reference [5] generates better quality result when compared to the traditional methods. The ROCK algorithm is effective on vector represents a tuple in the data where the entries are identifying as categorical values. Then the equation is

$$E = \sum_{i=1}^{k} n_i . \sum_{a,b \epsilon c} link\left(\frac{a_q . a_{qr.}}{n^{1+2f(\emptyset)}}\right)$$

In order to merge the cluster obtained the equation is expressed as

$$g(c_i, c_j) = \frac{link\,[c_{i,c_j}]}{\left(n_{i+n_j}\right)^{1+2f(0)-n_{i-nj}}}$$

In the expression the $[c_i, c_j]$ is the number of cross links between the clusters.

$$\text{Link}[cc_i, c_j] = \sum_{xQ \epsilon ci \epsilon cj} \quad link\,(a,b)$$

This result shows the best pair of clusters to be merged. QROCK has also been developed to compute the clusters by determining the connected components of the graph [8]. This leads to a very efficient method of obtaining the clusters giving a drastic reduction of the computing time of the ROCK algorithm.

C. STIRR algorithm

STIRR is a famous algorithm for clustering in which is more effective on categorizing the data rather than the other algorithms. In this algorithm the clustering sets is purely done through their patterns of co-occurrence, without invoking the artificial linear order or numerical structure. STIRR algorithm was developed to cluster databases with mixed data by transforming numeric attributes into nominal ones through discretization [12]. There is no mathematical expression to assess the similarity of values in STIRR. They are considered to be similar if the values with which they show together in the database have a huge overlie, regardless of the fact that the objects themselves might never co-occur. As a result of this approach a good partitions of an undirected graph is attain to the eigenvectors of certain matrices derived from the graph [Figure 1].

Tuple	Attribute a	b	c
1.	A	W	1
2.	A	X	1
3.	B	W	2
4.	B	X	2
5.	C	Y	3
6.	C	Z	3

is represented as

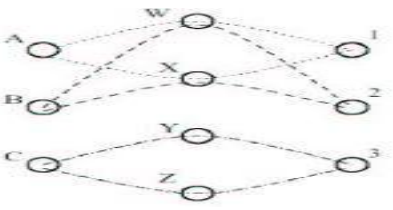

Figure 1. Clustering attributes and undirected graph [12]

III. COMPARITIVE ANALYSIS

To cluster the categorical data, these algorithms introduced various approaches in order to tackle the problems. Their performance gives various solutions in respect to the time, power and memory which also improves the quality of the clustering. The comparison results shows k-modes and its prototypes are scalable on discussing with the datasets. The hierarchical algorithm ROCK is based on the attribute values and its occurrence is examined with number of other attribute values with which it exists. The results are also scalable on comparing with other sampling techniques but the efficiency is less than the k-modes. The STIRR algorithm results have acquired either a positive or negative weight on two clusters. In resulting stage it requires a costly post-processing stage. The different partitions of their data set, which leads to the conclusion but those clustering, must be in meaningful. The best thing in STIRR algorithm is it converges quickly and identifies clusters in the presence of irrelevant values.

TABLE I
COMPARISION OF CLUSTERING METHODS

Clustering Methods				
Algorithms	**Input Parameters**	**Optimized For**	**Outlier Handling**	**Computational Complexity (number of in-memory operations)**
k-modes	Number of cluster	Data Sets with Well-separated Clusters	No	$\varphi(n)$
ROCK	Number of Cluster, Similarity Threshold	Small Data Sets with Noise	Yes	$\varphi(n^2 + nm_m m_a + n^2 \log n)$
STIRR	Initial Configuration, Combining Operator, Stopping Criteria	Large Data Sets with Well separated Clusters	No	$\varphi(n)$

The K-means algorithm requires more memory operation and also it is applicable for large inputs. For a large dataset, the STIRR needs one pass over the data set and a linear number of in-memory operations. These algorithms have still facing some disadvantage like text clustering, similarity analyses in clustering tuples and is not able to produce more than two clusters of attribute values. Another thing is ROCK is not suitable for large dataset and other drawbacks like no common quality measure. The comparison of three algorithm based on single dataset is too difficult. The proposed works are based on these issues which gives a scalable results and also improvement in the quality of the techniques.

IV. DESIGN OF FUZZY BASED ALGORITHM

Based on the above observation, FUZZY based 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 supports sentence clustering. The algorithm indentifies 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. FUZZY uses weighting schemes for information retrieval in order to assess the importance of whole attributes and individual values. The works is intended immediate retrieval of response based on the input query. Based on the query the clusters are done which related to the concepts based on the given queries by the user. The page rank algorithm was proposed [6]

$$xy(v_i) = (1-d) * \sum_{j \in n(vi)} \frac{1}{|out(vj)|} xy(vj)$$

Where v_i points in and v_j points out on the set of vertices. The similarity between the vj and v_i based on the similarity as store these are stored on the matrix form such as W= (w_{ij}) that refers to the affinity matrix. This can expressed on the below equations.

$$xy(vi) = (1-d) + d * \sum_{j=1}^{N} (w_{ji} \frac{xy(vj)}{\sum_{k=1}^{N} w_{jk}})$$

The overall Fuzzy based categorical text clustering algorithm is given in Figure2.

Step 1: Fragment each sentence into single words and those words are passed in to the word net
Step 2: Filter conjunctions and keywords
Step 3: Retrieval decision is made by comparing the terms of query with index terms
Step 4: Validate the duplication in the word net and determine the frequency of occurrence

Step 5: Form semantic clustering based on Fuzzy based pruning approach
Step 6: The similarity matrix can be formed using sim={simxy } where x and y is the similarity between the objects.
Step 7: The weight between the matrices is w^{c}_{ij} where c is the cluster
Step 8: Similarity matrix is calculated using $Sim_{x\&y}$ (w1,w2)=1/(IC(w1)+IC(w2)-2*IC(DCS(w1,w2))) where w are the words from the information content, IC is the information content and DCS is deepest common similarity.
Step 9: IC(w) is calculated by IC(w)=-log P(w) that is probability of the word w appear in the IC information content.

Figure 2. Fuzzy based categorical text clustering algorithm

V. CONCLUSION

The combined algorithms of k-modes, ROCK, STIRR and their limitation motivated the authors in developing the efficiency of text clustering using FUZZY logic which enhanced the performance of clustering than the existing methods. In this section, hard clustering is a challenging task on the dataset of famous quotations. This problem is commenced by using fuzzy logic to apply at relational clustering problems. This work proposes how categorical data fuzzy clustering can be implemented to classify categorical data and the basic idea of this approach is to extract a set of words and then transform them into categorical data vectors. This algorithm is applicable for general text mining settings based on query-directed text mining. This work is deeply worked on to improve the performance of better sentence similarity measures based on improved word sense disambiguation. This will be easy to investigate and applied on the real-time scenario.

REFERENCES

[1] Zhexue Huang. "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*", Data Mining and Knowledge Discovery*, **2**(3): pp.283-304, 1998.
[2] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim. " *ROCK: A Robust Clustering Algorithm for Categorical Attributes", Proceedings of the 15th International Conference on Data Engineering*, pp.512-521, Sydney, Australia, 23-26, 1999.
[3] David Gibson, Jon M. Kleinberg and Prabhakar Raghavan. "Clustering Categorical Data : An Approach Based on Dynamical Systems",*Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, pp. 311-322, New York, NY, USA, 24-27 August 1998.
[4] J.B MacQueen. "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings of the Fifth Berkeley Symposium. Mathematics, Statistics and Probability*, pp. 281-297, 1967.
[5] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim. " ROCK: A Robust Clustering Algorithm for Categorical Attributes ", *Stanford University Stanford*, CA 94305, shim@bell-labs.com, 2007
[6] S. Brin and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks and ISDN Systems*, vol. 30, pp. 107-117, 1998.
[7]. 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.
[8]. 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.
[9]. 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
[10]. R.Jensi and and 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.
[11]. 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.
[12]. D. Gibson, J. Kleinberg, and P. Raghavan. "Clustering Categorical Data: An Approach Based on Dynamical Systems", *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB'98)*, pp. 311–323, 1998.

