A Novel & More Efficient Clustering Technique for Document Clustering Methodology

[1]Patil Pravin Ishwar, [2]Prof. Gajendra Singh

ABSTRACT: **Document clustering is becoming more and more important with the abundance of text documents available through World Wide Web and corporate document management systems. But there are still some major drawbacks in the existing text clustering techniques that greatly affect their practical applicability. This paper presents a new method for document clustering. The improved clustering method uses a unique method of centroid selection. In this way, it improves the accuracy of the document clustering.**

1. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software [1,2] is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. **Generally, any of four types of relationships are sought:**

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

2. Clustering:

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering.

Clustering is the [3] most common form of unsupervised learning and this is the major difference between clustering and classification. No super-vision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. Clustering is sometimes erroneously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so called training set, i.e. a set of data correctly labeled by hand, and then replicates the learnt behavior on unlabeled data.

Document clustering is also applicable in producing the hierarchical grouping of document [4,5]. In order to search and retrieve then information efficiently in Document Management Systems (DMS), the metadata set should be created for the documents with adequate details. But just one metadata set is not enough for the whole document management systems. This is because various document types need different attributes to be distinguished appropriately. So clustering of documents is an automatic grouping of text documents into clusters such that documents within a cluster have high resemblance in comparison to one another, but are different from documents in other clusters. Hierarchical document clustering [6] categorizes clusters into a tree or a hierarchy that benefits browsing.

Information Retrieval (IR) [7] is the field of computer science that focuses on the processing of documents in such a way that the document can be quickly retrieved based on keywords specified in a user's query. IR technology is the foundation of web-based search engines and plays a key role in biomedical research, as it is the basis of software that aids literature search.

Document clustering is becoming more and more important with the abundance of text documents available through World Wide Web and corporate document management systems. But there are still some major drawbacks in the existing text clustering techniques that greatly affect their practical applicability. The drawbacks in the existing clustering approaches are listed below:

- Text clustering that yields a clear cut output has got to be the most favorable. However, documents can be regarded differently by people with different needs vis-à-vis the clustering of texts. For example, a businessman looks at business documents not in the same way as a technologist sees them [8,10,11]. So clustering tasks depend on intrinsic parameters that make way for a diversity of views.
- Text clustering is a clustering task in a high-dimensional space, where each word is seen as an important attribute for a text. Empirical and mathematical analysis have revealed that clustering in high-dimensional spaces is very complex, as every data point is likely to have the same distance from all the other data points [9,12,13].
- Text clustering is often useless, unless it is integrated with reason for particular texts are grouped into a particular cluster. It means that one output preferred from clustering in practical settings is the explanation why a particular cluster result was created rather than the result itself. One usual technique for producing explanations is the learning of rules based on the cluster results. But this technique suffers from a high number of features chosen for computing clusters.

3. Comparison of various systems:

K-means is the most important flat clustering algorithm. The objective function of Kmeans is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C

- K-means has problems when clusters are of differing Sizes, Densities, Non-globular shapes
- Problems with outliers
- Empty clusters

The K Nearest Neighbour (K-NN) suffers from the following drawbacks:

- Because all the work is done at run-time, k-NN can have poor run-time performance if the training set is large.
- k-NN is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification. This can be ameliorated by careful feature selection or feature weighting
- On very difficult classification tasks, k-NN may be outperformed by more exotic techniques such as Support Vector Machines or Neural Networks.

4. Proposed System:

We have proposed an improved clustering method. The steps are as follows:

Output: D = {d1, d2, d3,..., di,..., dn } //set of documents di = { x1, x2, x3,..., xi,..., xm } k // Number of desired clusters.

Input: A set of k clusters.

Working Procedure

1: Calculate distance for each document or data point from the origin

2: Arrange the distance (obtained in step 1) in ascending order.

3: Split the sorted list in K equal size sub sets. Also the middle point of each sub set is taken as the centroid of that set.

4: repeat this step for all data points. Now the distance between each data point & all the centroids is calculated. Then the data set is assigned to the closest cluster.

5: in this step, the centroids of all the clusters are recalculated.

6: Now for all data points. Now the distance between each data point & all the centroids is calculated. If this distance is less than or equal to the present nearest distance then the data point stays in the same cluster. Else it is shifted to the nearest new cluster.

Conclusion:

This paper proposes a novel technique for the document clustering. The accuracy of the proposed method is better as compared to the present method. As clustering plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the characteristics of existing methods. These existing approaches form the basis for the various innovations in the field of clustering. From the existing clustering techniques, it is clearly observed that the clustering techniques based on GA, fuzzy and ontology provide significant results and performance. Hence, this research concentrates mainly on the semantic clustering based on GA, NDRGA, ACO and fuzzy ontology clustering for better performance.

References:

[1] Guo-Yan Huang, Da-Peng Liang, Chang-Zhen Hu and Jia-Dong Ren, "An algorithm for clustering heterogeneous data streams with uncertainty", 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 4, pp. 2059-2064, 2010.

[2] Li Taoying, Chne Yan, Qu Lili and Mu Xiangwei, "Incremental clustering for categorical data using clustering ensemble", 29th Chinese Control Conference (CCC), pp. 2519-2524, 2010.

[3] Likas, A., Vlassis, N. and Verbeek, J.J. "The Global k-means Clustering algorithm", Pattern Recognition , Vol. 36, No. 2, pp. 451-461, 2003.

[4] Lijuan Jiao and Liping Feng, "Text Classification Based on Ant Colony Optimization", Third International Conference on Information and Computing (ICIC), Vol. 3, pp.229 - 232, 2010.

[5] Macskassy, S.A., Banerjee, A. Davison, B.D. and Hirsh, H. "Human Performance On Clustering Web Pages: A Preliminary Study", In Proc. of KDD-1998, New York, USA, pp. 264-268, Menlo Park, CA, USA, 1998.

[6] Malay K. Pakhira, "A Modified k-means Algorithm to Avoid Empty", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 220-226, 2009.

[7] Meila, M. and Heckerman, D. "An experimental comparison of model-based clustering methods", Machine Learning, kluwer Academic publishers, Vol. 42, pp. 9-29, 2001.

[8] Miha Grcar, Marko Grobelnik and Dunja Mladenic, "Using Text Mining and Link Analysis for Software Mining", Lecture Notes in Computer Science, Vol. 4944, pp. 1-12, 2008.

[9] Murtagh, F. "A Survey of Recent Advances in Hierarchical Clustering Algorithms Which Use Cluster Centers", Comput. J, Vol. 26, pp. 354-359, 1984

[10] Pallav Roxy and Durga Toshniwal, "Clustering Unstructured Text Documents Using Fading Function", International Journal of Information and Mathematical Sciences, Vol. 5, No. 3, pp. 149-156, 2009

[11] Shehroz S. Khan and Amir Ahmad, "Cluster Center Initialization Algorithm for K-means Clustering", Pattern Recognition Letters, Vol. 25, No. 11, pp. 1293-1302, 2004.

[12] Shin-Jye Lee and Xiao-Jun Zeng, "A three-part input-output clustering-based approach to fuzzy system identification", 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 55-60, 2010.

[13] Ward Jr, J.H. "Hierarchical grouping to optimize an objective function", J. Am. Stat. Association, Vol. 58, pp. 236-244, 1963.