International Journal of scientific research and management (IJSRM)
||Volume||3||Issue||12||Pages|| 3806-3811||2015||
Website: www.ijsrm.in ISSN (e): 2321-3418

Improved Document Clustering Technique Using K-Mean Method

Purvi Khare
Scholar, M.tech (Software Engineering)
Computer Science and Engineering, Shri Vaishnav Institute of Technology (R.G.P.V), Indore, Madhya Pradesh, 453111, India
Purvikhare09@gmail.com

Abstract: Clustering of document is important for the purpose of document organization, summarization, topic extraction and information retrieval in an efficient way. Initially, clustering is applied for enhancing the information retrieval techniques. Of late, clustering techniques have been applied in the areas which involve browsing the gathered data or in categorizing the outcome provided by the search engines for the reply to the query raised by the users. In this paper, we are providing a methodology for more accurate document clustering. The experimental results have shown that the clustering accuracy cum purity of the proposed technique is better than that of the present clustering technique.

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 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

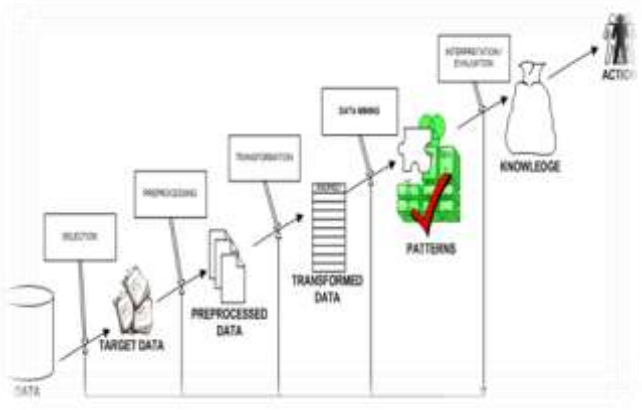

Figure 1: Structure of Data Mining

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 (Macskassy *et al.* 1998). 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 (Beyer *et al.* 1999).
- 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.

2. Related Work:

Document clustering is the process of categorizing text document into a systematic cluster or group, such that the documents in the same cluster are similar whereas the documents in the other clusters are dissimilar. It is one of the vital processes in text mining. Liping (2010) emphasized that the expansion of internet and computational processes has paved the way for various clustering techniques. Text mining especially has gained a lot of importance and it demands various tasks such as production of granular taxonomies, document summarization etc., for developing a higher quality information from text.

Likas *et al.* (2003) proposed the global K-means clustering technique that creates initial centers by recursively dividing data space into disjointed subspaces using the K-dimensional tree approach. The cutting hyper plane used in this approach is the plane that is perpendicular to the maximum variance axis derived by Principal Component Analysis (PCA). Partitioning was carried out as far as each of the leaf nodes possess less than a predefined number of data instances or the predefined number of buckets has been generated. The initial center for K-means is the centroids of data that are present in the final buckets. Shehroz Khan and Amir Ahmad (2004) stipulated iterative clustering techniques to calculate initial cluster centers for K-means. This process is feasible for clustering techniques for continuous data.

Agrawal *et al.* (2005) ascribed data mining applications and their various requirements on clustering techniques. The main requirements considered are their ability to identify clusters embedded in subspaces. The subspaces contain high dimensional data and scalability. They also consist of the comprehensible ability of results by end-users and distribution of unpredictable data transfer.

The main limitation of K-means approach is that it generates empty clusters based on initial center vectors. However, this drawback does not cause any significant problem for static execution of K-means algorithm and the problem can be overcome by executing K-means algorithm for a number of times. However, in a few applications, the cluster issue poses problems of erratic behavior of the system and affects the overall performance. Malay Pakhira et al. (2009) mooted a modified version of the K-means algorithm that effectively eradicates this empty cluster problem. In fact, in the experiments done in this regard, this algorithm showed better performance than that of traditional methods.

Uncertainty heterogeneous data streams (Charu Aggarwal *et .al* 2003) are seen in most of the applications. But the clustering quality of the existing approaches for clustering heterogeneous data streams with uncertainty is not satisfactory. Guo-Yan Huang *et al.* (2010) posited an approach for clustering heterogeneous data streams with uncertainty. A frequency histogram using H-UCF helps to trace characteristic categorical statistic. Initially, creating 'n' clusters by a K-prototype algorithm, the new approach proves to be more useful than UMicro in regard to clustering quality.

Alam *et al.* (2010) designed a novel clustering algorithm by blending partitional and hierarchical clustering called HPSO. It utilized the swarm intelligence of ants in a decentralized environment. This algorithm proved to be very effective as it performed clustering in a hierarchical manner.

Shin-Jye Lee *et al.* (2010) suggested clustering-based method to identify the fuzzy system. To initiate the task, it tried to present a modular approach, based on hybrid clustering technique. Next, finding the number and location of clusters seemed the primary concerns for evolving such a model. So, taking input, output, generalization and specialization, a HCA has been designed. This three-part input-output clustering algorithm adopts several clustering characteristics simultaneously to identify the problem.

Only a few researchers have focused attention on partitioning categorical data in an incremental mode. Designing an incremental clustering for categorical data is a vital issue. Li Taoying *et al.* (2010) lent support to an incremental clustering for categorical data using clustering ensemble. They initially reduced redundant attributes if required, and then made use of true values of

different attributes to form clustering memberships.

Crescenzi *et al.* (2004) cited an approach that automatically extracts data from large data-intensive web sites. The "data grabber" investigates a large web site and infers a scheme for it, describing it as a directed graph with nodes. It describes classes of structurally similar pages and arcs representing links between these pages. After locating the classes of interest, a library of wrappers can be created, one per class with the help of an external wrapper generator and in this way suitable data can be extracted.

Miha Grcar *et al.* (2008) mulled over a technique about the lack of software mining technique, which is a process of extracting knowledge out of source code. They presented a software mining mission with an integration of text mining and link study technique. This technique is concerned with the inter links between instances. Retrieval and knowledge based approaches are the two main tasks used in constructing a tool for software component. An ontology-learning framework named LATINO was developed by Grcar *et al.* (2006). LATINO, an open source purpose data mining platform, offers text mining, link analysis, machine learning, etc.

Similarity-based approach and model-based approaches (Meila and Heckerman 2001) are the two major categories of clustering approaches and these have been described by Pallav Roxy and Durga Toshniwal (2009). The former, capable of maximizing average similarities within clusters and minimizing the same among clusters, is a pairwise similarity clustering approach. The latter tries to generate techniques from the documents, each approach representing one document group in particular.

3. Proposed System

The outline of the proposed system is as follows:

1. Pre-Processing Module
2. Calculating the number of clusters
3. Clustering techniques

3.1 Preprocessing Module

Stopwords (prepositions, pronouns, articles, and irrelevant document metadata) have been removed. A typical method to remove stopwords is to compare each term with acompilation of known stopwords

Input: A document Data Base D and List of Stop words L

D= {d1,d2,d3,….,dk} ;where 1<=k<=i

tij is the jth term in ith document

Output: All valid stem text term in D

```
for (all di in D) do
for (1 to j) do
Extract tij from di
If(tij in list L)
Remove tij from di
End for
End for
```

Word stemming is the process of suffix removal to general word stems. A stem is a natural group of words with similar meaning. The process of reducing words to their base form, or stem. For example, thewords "connected," "connection", "connections" are all reduced to the stem "connect." Porter's algorithm is the de facto standard stemming algorithm [3].

The similarity between two documents I & j is computed as follows:

For i: = 0 to N (total documents)

For j: =0 to N (total documents)

Simvalue= ((doc[i]*doc[j]))/Math.sqrt(doc[i]*doc[j])

Add _simvalue to the list Build_matrix;

Next

Next

Where N is total number of documents

Doc[i] for i=1, 2.....n are document

3.2 Calculating the number of Clusters

In this step, only the number of clusters are feed as input. It is represented by K.

3.3 Clustering Techniques

K means & improved method are used.

Steps of K Means method:

This algorithm consists of four steps:

1. Initialization-In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. Classification- The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. Centroid Recalculation Clusters generated previously, the centriod is again repeatly calculated means recalculation of the centriod.

4. Convergence Condition Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied

Steps of improved method:

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

4 Results

We have used animal data set & several other data sets. We compared the performance of the k means & the proposed clustering technique. The sequence of execution is as follows:

- Removing Stop Words:
- Stemming
- Computing Term Frequency
- Distance Calculation
- Purity Checking

5 Clustering Accuracy

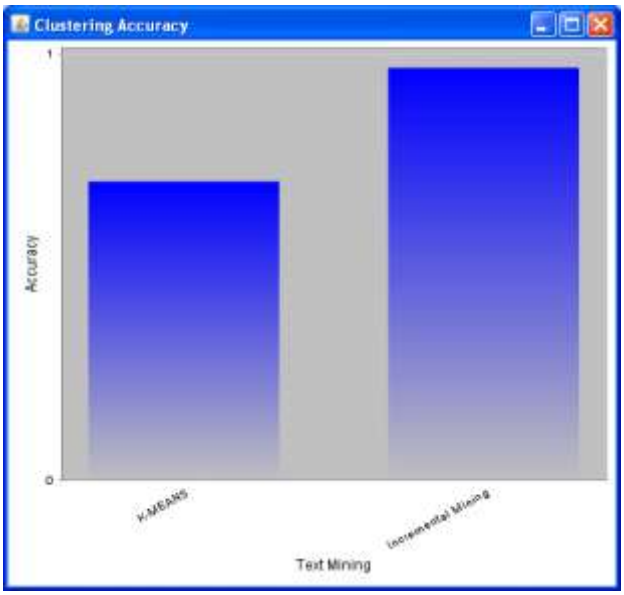

Figure 2: Clustering Accuracy

6 Conclusion

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.

In this paper, we have presented the implementation of the proposed model to perform the document clustering. The proposed incremental clustering technique is proposed with thet of the present K means clustering technique. It is observed that the purity of the proposed method is better.

REFERENCES

1. Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios, Raghavan and Prabhakar, "Automatic subspace clustering of high dimensional data", Data Mining and Knowledge Discovery (Springer Netherlands) Vol. 11, pp. 5-33, DOI: 10.1007/s10618-005-1396-1, 2005.

2. Alam, S., Dobbie, G., Riddle, P. and Naeem, M.A. "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 2, pp. 64-68, 2010.

3. Baeza-Yates, R.A. "Introduction to Data Structures and Algorithms Related to Information Retrieval", In Information Retrieval: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice- Hall, Inc., Upper Saddle River, New Jersey, pp. 13-27, 1992.

4. Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu, "A Framework for Clustering Evolving Data Streams", Proceedings of the 29th international conference on Very Large Data Bases (VLDB), pp. 81-92, 2003.

5. Crescenzi valter, Giansalvatore Mecca, Paolo Merialdo and Paolo Missier, "An Automatic Data Grabber for Large Web Sites", VLDB , pp. 1321-1324, 2004

6. Grcar, M., Mladenic, D., Grobelnik, M., Fortuna, B. and Brank, J. "Ontology Learning Implementation", Project report IST-2004-026460 TAO, WP 2, D2.2, 2006.

7. 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.

8. 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.

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

10. 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.

11. 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.

12. 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.

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

14. 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.

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

16. 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

17. 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.

18. 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.

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