International Journal of Database Theory and Application
Vol.7, No.6 (2014), pp.233-240
http://dx.doi.org/10.14257/ijdta.2014.7.6.21

Performance Analysis of Clustering using Partitioning and Hierarchical Clustering Techniques

S. C. Punitha[1], P. Ranjith Jeba Thangaiah[2] and M. Punithavalli[3]

[1]*Research Scholar, Department of Computer Science and Engineering, Karunya University, Coimbatore, India*
[2] *Head, Department of Computer Applications, Karunya University Coimbatore, India*
[3] *Dean, Sri Ramakrishna College of Arts & Science, Coimbatore, India*

Abstract

Text clustering is the method of combining text or documents which are similar and dissimilar to one another. In several text tasks, this text mining is used such as extraction of information and concept/entity, summarization of documents, modeling of relation with entity, categorization/classification and clustering. This text mining categorizes only digital documents or text and it is a method of data mining. It is the method of combining text document into category and applied in various applications such as retrieval of information, web or corporate information systems. Clustering is also called unsupervised learning because like other document classification, no labeled documents are providing in clustering; hence, clustering is also known as unsupervised learning. A new method called Hierarchical Agglomerative Clustering (HAC) which manages clusters as tree like structure that make possible for browsing. In this HAC method, the nodes in the tree can be viewed as parent-child relationship i.e. topic-subtopic relationship in a hierarchy. HAC method starts with each example in its own cluster and iteratively combines them to form larger and larger clusters. The main focus of this work is to present a performance analysis of various techniques available for document clustering.

Keywords: *Document clustering, Text Mining, Hierarchical Agglomerative Clustering (HAC), K-Mean, Expectation Maximization (EM), Text Clustering with Feature Selection (TCFS)*

1. Introduction

Text mining is utilized in a variety of text related missions such as information extraction, document conclusion, entity relational modeling concept/entity extraction, clustering and categorization/classification. Huge number of various areas in text mining and information retrieval, this document clustering is used. Agglomerative can be categorized as greedy, in the algorithmic sense. A series of irreversible algorithm is used to construct the desired data structure. The remainder of this paper is organized as follows. Section 2 summarizes the concepts and literature survey. Section 3 discusses the proposed method, and section 4 provides the experiments with high accuracy. Finally, Section 5 presents the conclusions of the work.

ISSN: 2005-4270 IJDTA

2. Literature survey

Cutting *et al.* suggested that the large document collections can be browsed by clustering method [4]. Jain and Anil (2010) has proposed that clustering has a long and rich history in a variety of scientific fields. Dittenbach *et al.* generates clusters in a layered manner starting from the top most layers. Michael Steinbach et al compared the two main approaches to document clustering, agglomerative hierarchical clustering and K-means. Hierarchical clustering is often portrayed as the better quality clustering approach. Shehata *et al.* (2010) introduced a new concept-based mining model that analyzes terms on the sentence, document, and corpus levels is introduced. Vipin Kumar and Sonajharia Minz (2014) introduce the concepts of feature relevance, general procedures, evaluation criteria, and the characteristics of feature selection.

3. Methodology

Hierarchical clustering algorithms have been widely studied in the clustering for records of various kinds like text data, multidimensional numerical data and categorical data. For searching process, the association of different hierarchical clustering algorithms and the method of agglomerative hierarchical clustering are mainly useful to support a variety of searching methods because it naturally creates a tree-like hierarchy. In particular, the effectiveness of this technique is improving the search efficiency over a sequential scans method. The main idea of this algorithm is to merge documents based on their similarity into clusters. Single Linkage Clustering, Average Linkage Clustering and Complete Linkage Clustering are the different agglomerative clustering techniques.

Algorithmic steps for Agglomerative Hierarchical clustering

Let X = $\{x_1, x_2, x_3, ..., x_n\}$ be the set of data points.

1) Start with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

2) Determine the smallest distance pair of clusters in the current clustering, state that pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is in over all pairs of groups in the current clustering.

3) Increasing the sequence number: m = m +1.Combine clusters (r) and (s) into a one cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].

4) Updating the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance among the new cluster, denoted (r,s) and old or previous cluster(k) is described in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).

5) If all the data points are in one cluster then stop, else repeat from step 2).

3.1 K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. The objective function

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 ,$$

where $\left\|x_i^{(j)} - c_j\right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3.2 EM Algorithm

EM algorithm is an iterative method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values.

The EM algorithm has several appealing properties, some of which are:

1. It is numerically stable with each EM iteration increasing the likelihood.
2. Under fairly general conditions, it has reliable global convergence.
3. It is easily implemented, analytically and computationally. In particular, it is generally easy to program and requires small storage space. By watching the monotone increase in likelihood (if evaluated easily) over iterations, it is easy to monitor convergence and programming errors (McLachlan and Krishnan, 1997, Sect. 1.7).
4. The cost per iteration is generally low, which can offset the larger number of iterations need for the EM algorithm compared to other competing procedures.
5. It can be used to provide estimates of missing data.

3.3 TCFS Method

The methodology used by TCFS method is similar to that of HSTC method, but it differs in three ways. The first is in the preprocessing phase, next is the clustering progression and the last is utilize the clustering algorithm. Both utilize the equal resemblance measure, cosine distance. In preprocessing stage, apart from stop word eradication and stemming, a weight estimation function that evaluates the term weight and semantic weight be integrated. Term weight is predictable using TF/IDF values that make use of information about word and number of times (n) it emerges in the document. Using the term weight value a term cube is constructed. A term cube is a 3-D model representing the document, term and n relationship. The semantic weight is considered by concept extraction, concept or semantic weight computation and production of semantic cube. The concept extraction module is designed to identify concept in each document. This procedure is completed with the assist of the ontology set. The terms are matched with concepts, synonyms, meronyms and hypernyms in the ontology. The concept weight is anticipated with the notion and its element count. The semantic cube is constructed with concepts, semantic weight and document. In group processing which cluster the documents, two methods, that is, term clusters and semantic clustering method be utilized. Term clustering collects a document based on the word heaviness, though semantic clustering clusters documents based on the semantic weight.

4. Experimental Result

This section reports experimental results when applying the selected four algorithms to cluster text documents using MATLAB. This research applies HAC algorithms on 20 newsgroup datasets. First this work trims the dendrogram of 20 newsgroup dataset into several sub-dendrogram. The three performance metrics, namely, precision, recall and F-measure were used. The overall accuracy obtained by the selected algorithms for different number of clusters.

$$Precision = \frac{correctly\ recommented\ items}{total\ recommended\ items}$$

$$Recall = \frac{correctly\ recommended\ items}{total\ useful\ recommendations}$$

$$F - Measure = \ 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Table 1. Precision and Recall for Proposed HAC Approach

TECHNIQUES	PRECISION	RECALL
K-Means	51	84
EM	74	79
TCFS	86	64
HAC	92	50

Table 2. F-Measure for Proposed HAC Approach

TECHNIQUES	F-MEASURE
K-Means	0.49
EM	0.61
TCFS	0.72
HAC	0.98

In Table 1 show the precision and recall for document clustering using various techniques like K-means, EM algorithm, TCFS and proposed HAC. The precision and recall values are predict the best in this proposed HAC method. In Table 2 denotes the F-measure function for various clustering method. In HAC method is only the best measuring value when compared to others.

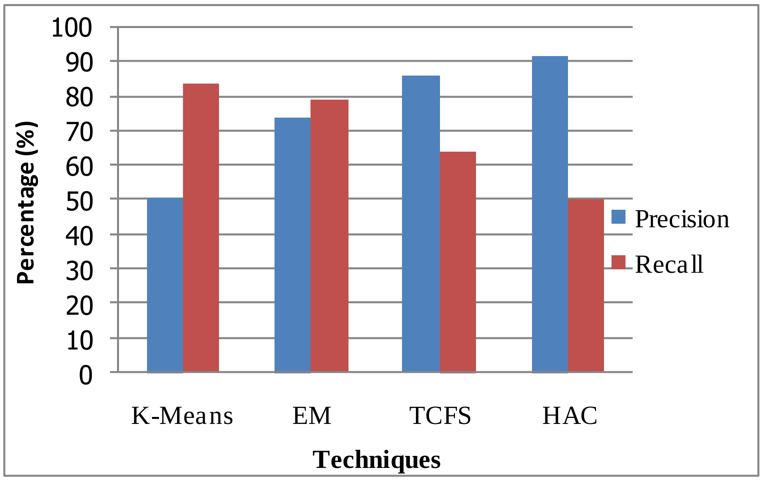

Figure 1. Precision and Recall for Proposed HAC

In Figure 1 shows the precision and recall value for proposed HAC technique. The precision value is higher than the other methods when using proposed HAC. The recall value is very low in proposed HAC when compared to others.

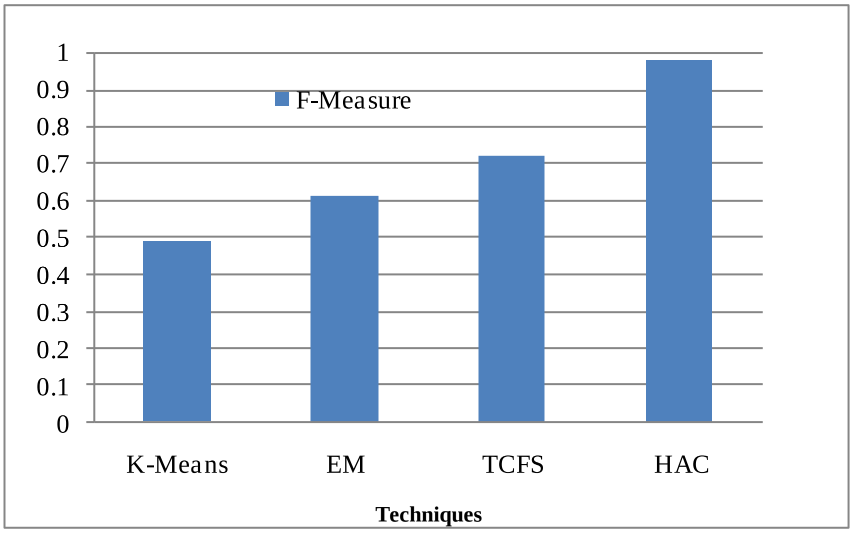

Figure 2. F-Measure for Proposed HAC

In Figure 2 represents the F-measure value for proposed HAC and where the HAC method is only the best when compared to other approaches.

5. Conclusion

Text document clustering is the task of grouping documents which have similar properties based on semantic and statistical content, is a major important component in many information organization and management tasks. An effective feature selection may used for good clustering and a correct choice of the algorithm for the clustering performance. Among the various classes of algorithms, the agglomerative hierarchical clustering approaches are the much popular in a huge amount of applications. This work discussed various algorithmic aspects, including well-definedness (*e.g.*, inversions) and computational properties. This work has also touched on a number of application domains, again in areas that reach back over some decades or many decades and more recent application domains.

References

[1] S. Shady, F. Karray and M. S. Kamel, “An efficient concept-based mining model for enhancing text clustering”, IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10 **(2010)**, pp. 1360-1371.

[2] C. C. Aggarwal and C. X. Zhai, “A survey of text clustering algorithms”, In Mining Text Data, Springer US, (**2012)**, pp. 77-128.

[3] W. Zhang, T. Yoshida, X. Tang and Q. Wang, “Text clustering using frequent itemsets”, Knowledge-Based Systems, vol. 23, no. 5, **(2010)**, pp. 379-388.

[4] S. M. Krishna and S. D. Bhavani, “An efficient approach for text clustering based on frequent itemsets”, European Journal of Scientific Research, vol. 42, no. 3, **(2010)**, pp. 399-410.

[5] A. K. Jain, “Data clustering: 50 years beyond K-means”, Pattern Recognition Letters, vol. 31, no. 8, **(2010)**, pp. 651-666.

[6] R. Gil-García and A. Pons-Porrata, “Dynamic hierarchical algorithms for document clustering”, Pattern Recognition Letters, vol. 31, no. 6, **(2010)**, pp. 469-477.

[7] L. Jing, M. K. Ng and J. Z. Huang, “Knowledge-based vector space model for text clustering”, Knowledge and information systems, vol. 25, no. 1, **(2010)**, pp. 35-55.

[8] S. Singh, A. Subramanya, F. Pereira and A. McCallum, “Large-scale cross-document coreference using distributed inference and hierarchical models”, In Proceedings of the 49th Annual Meeting of the Association

for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, (**2011)**, pp. 793-803.

[9] H. Gao, J. Jiang, L. She and Y. Fu, "A New Agglomerative Hierarchical Clustering Algorithm Implementation based on the Map Reduce Framework", JDCTA, vol. 4, no. 3, **(2010)**, pp. 95-100.

[10] C. -L. Chen, F. S. C. Tseng and T. Liang, "Mining fuzzy frequent itemsets for hierarchical document clustering", Information processing & management, vol. 46, no. 2, **(2010)**, pp. 193-211.

[11] B. Aljaber, N. Stokes, J. Bailey and J. Pei, "Document clustering of scientific texts using citation contexts", Information Retrieval, vol. 13, no. 2, **(2010)**, pp. 101-131.

[12] J. F. Cui and H. S. Chae, "Applying agglomerative hierarchical clustering algorithms to component identification for legacy systems", Information and Software Technology, vol. 53, no. 6, **(2011)**, pp. 601-614.

[13] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 1, **(2012)**, pp. 86-97.

[14] G. V. R. Kiran, R. Shankar and V. Pudi, "Frequent itemset based hierarchical document clustering using Wikipedia as external knowledge", InKnowledge-Based and Intelligent Information and Engineering Systems, Springer Berlin Heidelberg, (**2010)**, pp. 11-20.

[15] H. H. Malik, J. R. Kender, D. Fradkin and F. Moerchen, "Hierarchical document clustering using local patterns", Data Mining and Knowledge Discovery, vol. 21, no. 1, **(2010)**, pp. 153-185.