"Real World Document Clustering Using Modified Balanced Iterative Reducing and Clustering using Hierarchies"

Mrs. Sunita N. Chaudhari, Prof. Praveen Kumar Gautam

*1Research Scholar, Truba College of Engineering & Technology, Indore, M.P, India.

2Associate Professor, Truba College of Engineering & Technology, Indore, M.P, India.

*Department of Computer Science & Engineering

Abstract - *Clustering is "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are coherent internally, but clearly dissimilar to the objects belonging to other clusters. Document clustering is used in many fields such as data mining and information retrieval. to compare the clustering results of K-Mean approach ,agglomerative approach , partitioned approach for each of the criterion functions using real-world documents, and to establish the right clustering algorithm to produce high quality clustering of real-world document*

The goal of a document clustering method is to reduce intra-cluster distances between documents, while exploiting 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. The large variety of documents makes it almost unfeasible to create a general algorithm which can work best in case of all kinds of datasets.

Keywords: Document Clustering, K-Mean, B IRCH, Vector Space, Parsing

1. INTRODUCTION:

Document clustering makes gathering of indistinguishable documents in one cluster means they are similar to each other by any of the property which they exhibits and it get separated or indistinguishable with the documents in that of supplementary groups (clusters)..Clustering is the one of the technique used in wide variety of areas like in retrieving information, in pattern identification, to analysis images and videos, in bioinformatics, prediction of weather etc.

1.1 Document Parsing

As we have collection of raw and unstructured data {d1. . . dn}, we have to make it refined ,firstly by parsing it and then transformed it into a data model and which further used by algorithm. There are steps of procedures to each document:

Tokenization: In this procedure we convert the data of a document into cycle of term or in sequencing the term and also representing any word and phrase as in any other form to further use to characterize the document. It may used in many cases where we have to maintain the information according to relative ordering of terms, depends mainly on the selection suitable data model.

Stemming: In English language, we are using some process like Porter suffix stripping algorithm to strip out the common morphological and similar endings (e.g. "programming"! "Program"). In other languages also there is wide variety of open and usual techniques are available for stripping out the documents.

Stop-word removal:

To reduce noise from document clustering we have to eliminate some basic functional or stop words (e.g. "the", "if") which repeated in any document so many times. As they have no discriminating power so we have to treated as noise.

1.2 Vector Space Model

Each document d_j is denoted by a vector $x_j=\{f_1,\ldots f_m\}$ in a multidimensional term space, where number of total unique terms across all the documents denoted as m and frequency of occurrences of i^{th} term is represented by f_i. For convenience, the complete model is usually stored as a single term-document matrix .Before the documents can be clustered based on the similarities between vectors, we have to convert entire set of documents into a set of feature vectors{x1,...,xn}.

A = [x1 x2 . . . xn] belongs to $IR^{m\times n}$

1.3 Term Weighting

This function is typically composed of two components: term frequency (tf) and inverse document frequency (idf). Here term frequency is used to finding the term that occurs frequently means iteratively in any single document while idf is used to reduce the influence of terms occurring in several documents simultaneously, which is not useful in discriminating the underlying classes in the data.
The weighted frequency value for the i-th term in the document dj is defined as:

$$tfidf(i, j) = ltf(i, j) \cdot (\log(n/dfi))$$

Where fi is the number of occurrences of the i^{th} term in d_j and df_i is the total number of documents in the dataset, which contain that term.

1.4 Similarity Measures

For discovery of appropriate grouping of documents is finding by using suitable measure for increasing the strength association between pairs of documents. To avoid unnecessary computations during the subsequent process of clustering we needs a single pair wise similarity or dissimilarity said to be part of preprocessing phase of the clustering. During the execution the similarity measures applied between pairs of documents. While a wide range of techniques for assessing similarity have been proposed in different application fields, we consider here three metrics that are interesting from the perspective of researchers working with text datasets.

A. Euclidean distance

Computing the squared l2 norm is one of the most popular measures working with real –valued feature vectors call as Euclidean distance.
$ed(xi, xj) = ||xi − xj || = Xml=1((xi,l − xj,l)^2)^{1/2}$

B. Cosine similarity

As Euclidean distance not work well for high-dimensional data but it is used in many domains but Cosine similarity is used due to importance it places on absent value .So to measure the cosine angle between their corresponding vectors we used it as an alternative to compute the similarity between
$cos(xi, xj) = (xi, xj) / (||xi|| \cdot ||xj ||)$

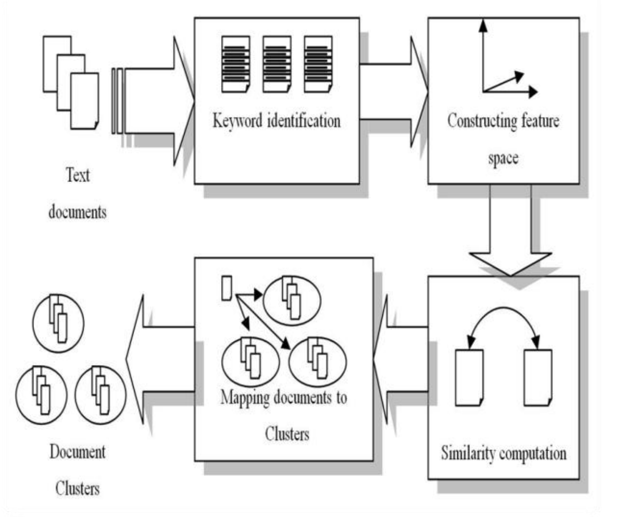

Figure1.1 Document Clustering Definition

2. LITERATURE SURVEY:

Charu C. Aggarwal ,Yuchen Zhao, Philip S. Yu(2014)" On Text Clustering with Side Information" In this paper, a method for text clustering with the use of side-information is presented. Many forms of text databases contain a large amount of side-information or meta information, which may be used in order to improve the clustering process.

Chandan Jadon, Ajay Khunteta(2013)" A New Approach of Document Clustering" In this the approach of document representation used we have a document -term matrix .The rows represent the documents and the columns represent the terms number(which are fix for each term).The terms are arranged in such a manner , the term number is first in the list whose weight is highest and they are arranged in decreasing order of weight(frequency in document).

Anusua Trivedi Piyush Rai Scott L. DuVall(2010)" Exploiting Tag and Word Correlations for Improved Webpage Clustering "In this paper, present a subspace based feature extraction approach which leverages tag information to complement the page-contents of a webpage to extract highly discriminative features, with the goal of improved clustering performance. In this approach, page-text and tags as two separate views of the data, and learn a shared subspace that maximizes the correlation between the two views.

Bibhu Prasad Mohanty & Pradeep Kumar Mallick(2012)"A New Approach of Text Clustering Using

Parallelization "In this paper proposes to parallelize a method which uses a tree based summarization technique to store cluster summaries in a tree stored in the memory at all times of processing. The results show that method shows good accuracy along with a good speed up in calculating clusters measures used in this clustering. Various approaches could be used for clustering

M.Deepa, P. Revathy(2012)" Validation of Document Clustering based on Purity and Entropy measures" By this paper we says that the classical clustering algorithms assign each data to exactly one cluster, but fuzzy c-means allow data belong to different clusters.

3. EXISTING SYSTEM:

1) **K-Means Algorithm :** The K-Means algorithm aims to partition a set of items, according to their attribute/features, into *k* clusters, where *k* said to be predefined or user-defined constant.

2) **Basic K-Means Algorithm**

1. select k number of clusters to be determined
2. Choose k objects arbitrarily as the initial cluster center
3. reiterate

3.1. Allot each object o their closest cluster
3.2. Compute new clusters i.e. calculate mean points. Until
4.1. (Centroids do not vary position any more) OR
4.2. No object changes its cluster. (We can classify stopping criteria as well.)

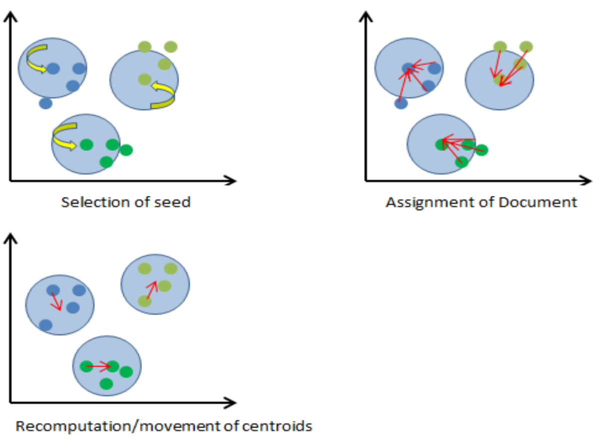

Figure 1.2 Kmean Clustering

BIRCH (balanced iterative reducing and clustering using hierarchies):

BIRCH (is an unsupervised data mining algorithm used to achieve hierarchical clustering over particularly huge data-sets.

A CF tree is a height-balanced tree by means of two parameters: branching factor B and threshold T. Each non leaf node comprise at most B entries of the type [CF_i, $child_i$], where i=1, 2,..B" $child_i$" is a pointer to its i-th child node, plus CF_i is the CF of the sub cluster symbolize by this child. So a nonleaf node stand for a cluster made up of the entire sub clusters represented via its entries. A leaf node contain at most L entries, each of the type [CF_i], where i=1,2,...,L. In addition, all leaf node has two indicators, "prev" and "next" which are used to sequence all leaf nodes mutually for proficient scans. A leaf node also represents a cluster made up of all the sub clusters represented via its entries. But all entries in a leaf node must assure a threshold necessity, by value to a threshold value T: the diameter (or radius) has to be a lesser amount of than T.

We at the moment present the algorithm for inserting an entry in a CF tree. Given entry "Ent", it carries on as below:

- Recognizing the suitable leaf: Initiating from the root, it recursively go down the CF tree by choosing the closest child node according to a selected distance metric: D0, D1, D2, D3, D4 as defined
- Altering the leaf: When it reaches a leaf node, it discovers the closest leaf entry, say L_l, and tests whether L_l can "absorb" "Ent" without violating the threshold clause.
- A Merging Refinement: Splits are produced due to the page size, which is free of the clustering belongings of the data. In the existence of skewed data input order, this can influence the clustering quality, and also shrink space utilization.
- A simple additional merging step frequently helps to upgrade these problems: Assume that there is a leaf split, and the spreading of this split stops at some nonleaf node N_J, i.e., N_J can contain the additional entry resultant from the split. We now scan node N_J to locate the two closest entries.

4. PROPOSED SYSTEM

The Modify BIRCH algorithm is very scalable with respect to the number of records in a datasets. The complexity of phase1 algorithm is clearly linear with respect to the dataset size. Further, it alleviates the drawbacks of linkage metric-based algorithm, which cannot undo the splitting or merging of nodes.

Pseudo Code of Modify BIRCH algorithm

BIRCH (data D, B, T)
Phase 1
1. N= {{}} # initial CF tree with an empty leaf node.

2. for each point p in D:
3. M=leaf node in N/2 that closest to p.
4. Add p to m
5. Compute diameter M of m
6. If M> T:
7. Split (m) # may require splitting of ancestors of m
phase 2
8. Apply another clustering algorithm to cluster the leaves of N/2

We apply both the algorithms partitioning i.e. K-means and Hierarchical i.e. BIRCH for the different sizes of records. Both the algorithms K-means and BIRCH need number of words which were already converted in number form by vector space represented as an input. In the K-means clustering algorithm set of primary centroids are essential. In hierarchical method we require set of data as input which they converted in tree form for applying clustering algorithm. In BIRCH it does not require initial centroids as it scan whole data at once.

5. RESULT:

We apply both the algorithms partitioning i.e. K-means and Hierarchical i.e. BIRCH for the different sizes of records. Both the algorithms K-means and BIRCH need number of words which were already converted in number form by vector space represented as an input. In the K-means clustering algorithm set of primary centroids are essential. In hierarchical method we require set of data as input which they converted in tree form for applying clustering algorithm. In BIRCH it does not require initial centroids as it scan whole data at once.

The K-means clustering algorithm is executed several times for the different data values of initial centroids. In each experiment the time was computed and taken the average time of all experiments. The experiments results show that the BIRCH algorithm is producing better results in less amounts of computational time compared to the k-means algorithm.

As the size of documents rises the performance of hierarchical algorithm goes increasing and time for execution get reduces. Hierarchical algorithm also increases quality of cluster and decreases its time of execution as compared to K-mean algorithm. Hierarchical algorithm is more time efficient as compared to k-mean algorithm

Table 5.1Finding the time of k-means and Birch algorithm

Document Name	Document Size	K-means Algorithm (time in millisecond)	BIRCH Algorithm (time in millisecond)
Document 1	591 bytes	124	18
Document 2	423 bytes	90	12
Document 3	501 bytes	106	15

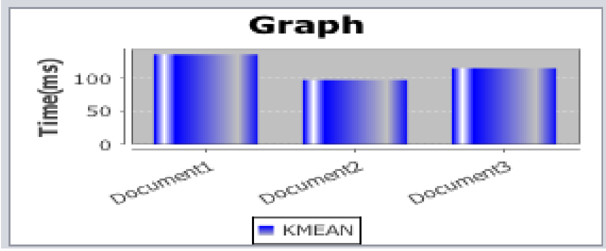

Figure 5.1 Snapshot of the performances of the k-means algorithm in terms of the time.

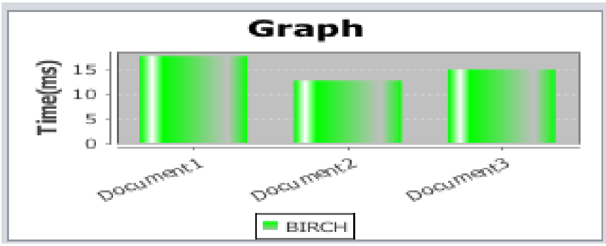

Figure 5.2 Snapshot of the performances of the BIRCH algorithm in terms of the time

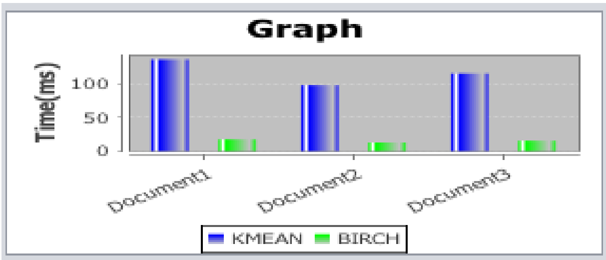

Figure 5.3 Snapshot of the evaluation of the K-means algorithm and BIRCH in terms of the time

6. CONCLUSION

One of the partition clustering algorithms is K-mean clustering algorithm, which depends on initial clusters. In basic K-mean clustering, initial clusters are based on algorithm arbitrarily selected centroids. The k-means algorithm is generally used for clustering huge and small sets of records. But the standard algorithm do not at all times guarantee good quality outcomes as the precision of the concluding clusters depend on the selection of initial centroids. Moreover, the computational complexity of the standard algorithm is high to the need to reassign the data points a number of times, through each iteration of the loop. Our aim was to improve the time polices of the algorithm.

REFERENCES:

[1] Wang, J. and X. Su, "An improved K-means clustering algorithm," in 3rd International Conference on Communication Software and Networks (ICCSN), Xi'an, 2011.

[2] Na, S. and L. Xumin, "Research on K-means Clustering Algorithm An Improved K-means Clustering Algorithm," in Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), Jinggangshan,2010

[3] Shi Na , Liu Xumin , Guan yong "An Improved K-means algorithm "IITSI 2010

[4] Shuhua Ren, Alin Fan "K-means Clustering Algorithm Based on Coefficient of Variation" (ISP) IEEE 2011

[5] K. A. Abdul Nazeer, M. P. Sebastian "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm" IEEE (WCE) 2009.

[6] Mahmud, M.S.; Rahman; Akhtar "Improvement of K-means algorithm with better initial centroid based on weighted average" IEEE (ICECE) 2012.

[7]Jinmie Feng ; Zhimao Lu " A K-mean clustering algorithm based on maximum triangle rule" IEEE (ICMA)2013

[8] M. P. S Bhatia, Deepika Khurana Analysis of Initial Centers for k-Means Clustering Algorithm International Journal of Computer Applications (0975 –8887) Volume 71– No.5, May 2013

[9] Dr. M.P.S Bhatia1 and Deepika Khurana Experimental study of Data clustering using k-Means and modified algorithms International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013

[10]. Kohei Arai and Ali Ridho Barakbah Hierarchical K-means: an algorithm for centroids Initialization for K-means Rep. Fac. Sci. Engrg. Reports of the Faculty of Science and Engineering, Saga Univ. Saga University, Vol. 36, No.1,p.p25-312007.

[11]. Napoleon, D. and P.G. Lakshmi An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points , 2010 in Trendz in Information Sciences and Computing (TISC), Chennai

[12] Malay K. Pakhira "A Modified *k*-means Algorithm to Avoid Empty Clusters " In International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009.

[13] P. Maji and S. K. Pal,"Rough–fuzzy C-medoids algorithm and selection of biobasis for amino acid sequence analysis,"IEEE Trans. Knowl. Data Eng., vol. 19, no. 6, pp. 859–872, Jun. 2007.

[14] Gurjit Singh and Navjot Kaur, "Hybrid Clustering Algorithm with Modifications Enhanced K-means and hierarchical algorithm", in International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 5, May 2013.

[15] Ke Sun, Jie Liu, Xueying Wang, K mean cluster algorithm with refined initial center

point, in Journal of Shenyang Normal University(Natural Science), 27(4), 448-451,2009

[16] Zhaoxia Tang, K-means clustering algorithm based on improved genetic algorithm, in: Journal of Chengdu University (Natural Science Edition), 30(2), 162-164,2011.

[17] Chunfei Zhang∗, Zhiyi Fang∗ "An Improved K-means Clustering Algorithm" in Journal of Information & Computational Science 10: 1 193–199,2013