IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-1(1) E-ISSN: 2347-2693

A Clustering Framework for Large Document Datasets

Krishna Kumar Mohbey*, G.S. Thakur
Maulana Azad National Institute of Technology Bhopal, India

www.ijcaonline.org

Received: 02/July /2013 Revised: 13/July/2013 Accepted: 14/Aug/2013 Published: 30/Sept/203

Abstract Document set is the collection of different types of document. Each document contains special type of information, which is beneficial for the peoples. We have the need of document clustering by their similarity. Document may contain data related to the blogs, website access pattern, any transaction or simply text. By the clustering of similar documents one can find the future trends of the people and it is also useful for the business point of view. In this paper, we have proposed a clustering approach for large size document sets. This proposed approach immediately assign document into appropriate cluster. Experiments are conducted with the twenty newsgroup dataset using java and MATLAB software. Comparisons are also performed with the existing methods. Experimental results show the effectiveness of the proposed approach for large document sets.

Keywords— *Large document set, Similarity measurement, Term extraction, Dendrogram*

I. INTRODUCTION

Document clustering has been widely used in different fields of social science, computer science, medical, technology etc. documents are the most available form of the data for various purposes, it is generally used to find out the habits and interest of the particular users. Documents are the collections of different kinds of data and it is collect from the various sources like social networking sites, blogs, online data accessing or online transactions. Document clustering problems appear in many different fields like data mining, pattern recognition, statistical data analysis, bioinformatics etc. fig 1show the general process of document clustering.

Today mostly people store the large amount of documents for different purpose and uses for future analysis and management [3]. Peoples have the requirement to categorize these data into different sets according to their need, is called clusters. This categorization process is mostly done by the people on the bases of similarities or dissimilarities based on the some rules or standards. The process of data classification may be supervised or unsupervised it depends on whether they assign New inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively [11], [12], [4].

In supervised document classification, the mapping from a set of input data vectors to a finite set of discrete class labels is modeled in terms of some mathematical function f=y(A,B), where B is a vector of adjustable parameters. The values of these parameters are determined (optimized) by an inductive learning algorithm (also termed inducer), whose aim is to minimize an empirical risk functional (related to an inductive principle) on a finite data set of input–output examples, where is the finite cardinality of the available representative data set [11].

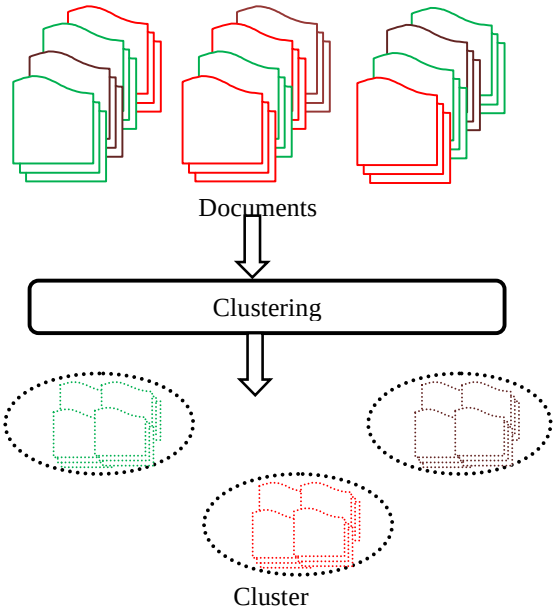

Fig.1 document clustering process

Clustering or exploratory is an unsupervised classification data analysis process in which no labeled data are available [7], [9]. The goal of clustering is to separate a finite unlabeled document set into a finite and discrete set of objects rather than provide an accurate characterization of unobserved samples generated from the same probability distribution [13], [12].Backer and Jain [10] defined that “in clustering process an document or group of object is divided into a number of more or less similar subgroups On the bases of similarity measurement.

There are lots of Clustering algorithms Available today which are used to partition documents into a certain number

Corresponding Author: *Krishna Kumar Mohbey*

of clusters/groups/ Categories. Most researchers describe a cluster by considering the object similarity and the external separation [8], [9], i.e., patterns in the same cluster should be similar to each other, while patterns in different clusters or groups will be dissimilar. Both the similarity and the dissimilarity should be examinable in a clear and meaningful way. Clusters are useful to search relevant document from the available groups and it provides the searching efficiency because it is applied on the smaller collection instead of whole document collections.

There are different clustering techniques [1] each have their own working procedure, performance and issues. Some Clustering techniques include on [3], [4], [6], [7], [9] references. Before applying clustering method it is important to use data mining preprocessing activities such as stemming and removing stop words [16] and cleaning of the data.

The rest of the paper is organized as follows. Section II describes the various clustering algorithms and related works for the document classification. Section III described a brief background of the proposed clustering architecture. Experimental results and conclusions are discussed in section IV and section V respectively.

II. Related Works

In this section, recent works related to the document clustering are discussed. The main objective of the document clustering is to classify a set of documents into a cluster and assign meaningful classes. Document classification can be dividing as follows [14]. Different clustering methods are shown in fig 2

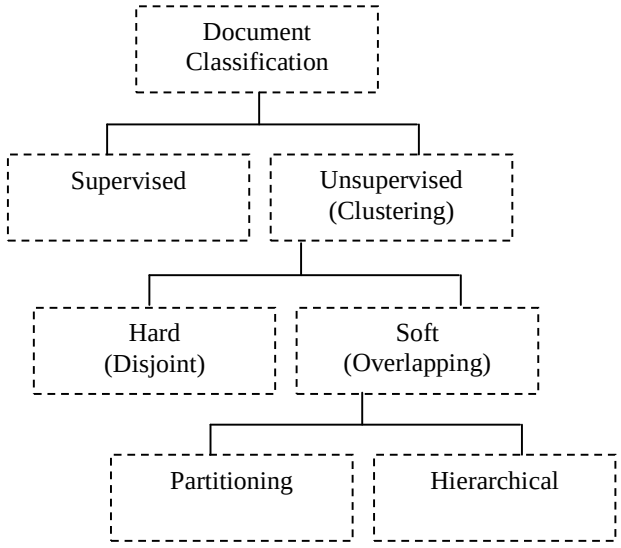

Fig.2 types of document classification

In supervised document classification predefine classes set are available but in unsupervised document clustering there are no predefine classes. Similarity measure is applied for document clustering, in which similar documents are together into a cluster.

In hard clustering algorithms each documents is exactly assign into a cluster while; in soft clustering it is possible to be a document into multiple clusters. Partitioning clustering algorithms like K-mean [14]; partition the set of documents into a number of clusters by moving documents from one to other cluster. Hierarchical clustering algorithms like Single link, complete link and average link prepare hierarchical tree of clusters [15].

Dash et. All[17] proposed a fast hierarchical clustering method based on partially overlapping partitions. Nanni[19] exploits triangle inequality property of metric space to speed up hierarchical clustering methods. Vijaya et all [20] proposed a hybrid clustering technique to speed up protein sequence classification. Few other clustering methods had been proposed to combine partition clustering method with hierarchical clustering methods [2]. The above mentioned work suffers from lack of efficiency and accuracy. The high complexity and low accuracy are still issues and challenges in the clustering. One of the main issue is related to multiple time document set scanning. Document scanning takes more time compare to document clustering.

III. Proposed Clustering Architecture

Fig. 3 shows the proposed architecture for the document clustering; which consist of Document preprocessing, document term extraction and similarity measure and cluster finding phases. The following definition is used in document clustering architecture:

Definition 3.1 (Document). A document represents to text which is the combination of term and frequency; it is denoted as

$$D = \{(t_1, f_1), (t_2, f_2), (t_3, f_3)..(t_n, f_n)\}$$

Definition 3.2 (Document Set). A document set is a collection or different documents; it is denoted as

$$D = \{D_1, D_2, D_3, ..., D_n\}$$

Definition 3.3 (Term Set). The Term set of a document set D is the set of terms appeared in D; it is denoted as

$$TD = \{t1, t2, t3, ..., tn\}$$

Definition 3.4 (Term Matrix). A Term matrix is a matrix or two-dimensional array contains the term and frequency of the documents.

Definition 3.5 (Euclidean Distance Matrix). An n × n matrix $D = [d_{ij}]$ is called a Euclidean distance matrix if there exists a set of n vectors, say $\{x_1 \ldots x_n\}$, in a finite dimensional inner product space such that $d_{ij} = |x_i - xj|^2$ [18].

Definition 3.6 (Dendogram). It is used to show the same information as the graph, however distance or threshold are in vertical and points are at the horizontal axis [5]. The height at which two clusters are merged is the dendogram reflects the distance of the two clusters.

Text Pre-processing: Text pre-processing means transform documents into a suitable representation for the clustering task. The text documents have different stop words, punctuation marks, special character and digits and other characters. This module is removed HTML Tages, Stop words from the text Documents. After removing stop words, word stemming is performed .Word stemming is the process of suffix removal to general word stems. A stem is a natural group of words with similar meaning.In the text-pre-processing step we performed the following task:
Removal of HTML tags and special character
Removal stop words
Word stemming

The following algorithms are used for removing stop words and stemming.

Algorithm1: This algorithm remove stop words & special characters

Input:A document set D and List of Stop words L
$D=\{d_1,d_2,d_3,\ldots,d_k\}$;where $1<=k<=i$
t_{ij} is the j^{th} term in i^{th} document
Output: All valid stem text term in D

1. for (all d_i in D) do
2. for (1 to j) do
3. Extract t_{ij} from d_i
4. If(t_{ij} in list L)
5. Remove t_{ij} from d_i
6. End for
7. End for

Term Extraction: term extraction is the greatest challenge of document classification, so it became major issue in classification. This step performs two functions- indexing and term selection. In indexing method we assign the value to the terms in the documents. After indexing, term selection method is applied. It is the process of removing indiscriminate terms from the documents to improve the document classification accuracy and reduce the computational complexity. The algorithm 2 is proposed for indexing and term extraction.

Algorithm 2: Algorithm for Term Extraction

Input : A document Set D and y minimum threshold value, N is the counter
$D=\{d_1,d_2,d_3,\ldots,d_k\}$;where $1<=k<=i$
t_{ij} is the j^{th} term in i^{th} document
Output: Documents set D with Terms

1. for (all d_i in D) do
2. for (1 to j) do
3. Count total occurrence of t_{ij} in document d_i
4. Assign the total occurrence of t_{ij} in N
5. If (N<y)
6. Remove t_{ji} from the document d_i
7. End for
8. End for

This proposed document clustering architecture works in the following steps-

i. In document preprocessing it necessary to clean up the document by applying algorithms for stop word removing and stemming. For these tasks we maintain a predefined stop word list and stemming algorithms.

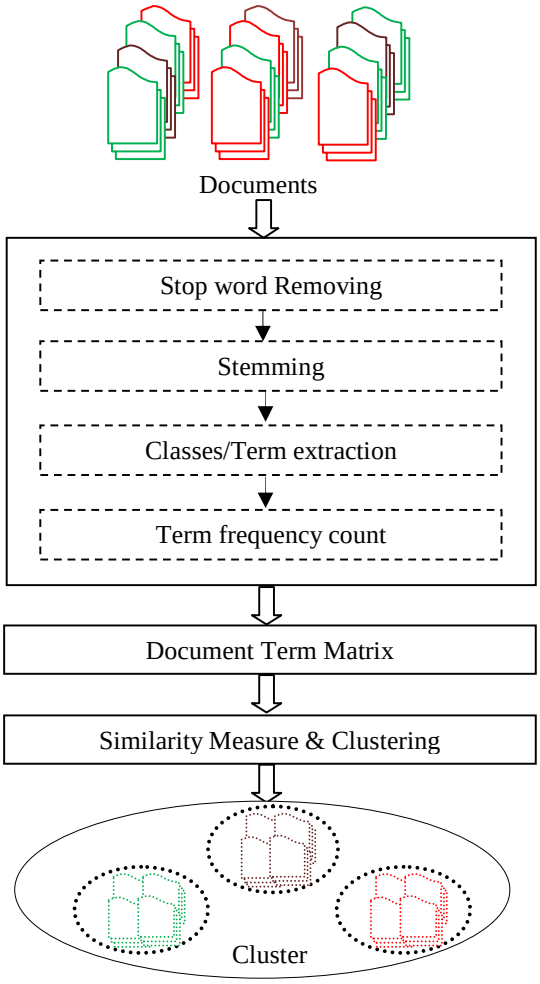

Fig.3 document clustering framework

ii. When documents are completely preprocessed then can be transferred to the term extraction phase; in this phase terms or features are extracted from the document sets. And frequency of the terms is counted by the document sets. These terms and frequencies of the documents can be represented as table 1.

iii. Similarity measurement is performed on the term and document sets which is obtained into document term matrix.

iv. For all documents similarity calculation and cluster preparation process are applied. This process continues until all documents are combining into similar clusters.

TABLE I
DOCUMENT TERM MATRIX

Doc	T1	T2	T3	T4
D1	5	7	10	5
D2	4	5	6	8
D3	7	8	9	15
D4	4	5	8	5
D5	5	8	9	10

IV. Experimental Results

In this section, we evaluated the proposed framework experimentally. We have use twenty newsgroups dataset for document classification. This data set consists of 20000 documents taken from Usenet newsgroup. Stop word removing and stemming algorithms are performed using Java language and executed on Intel Pentium 4 CPU (3.0 GHz) with 1 GB RAM PC and Document similarity measure and clustering algorithms are performed using MATLAB 2009b version on same PC configuration. Fig 4 and Fig 5 shows the output of cluster preparation and dendogram.

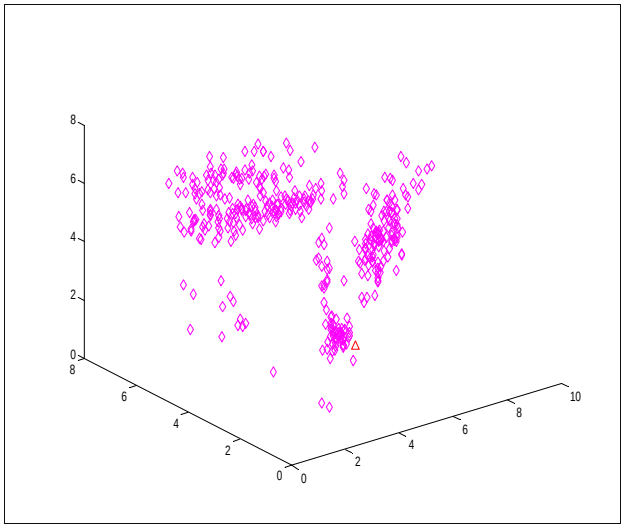

Fig.4 clusters obtained by document clustering

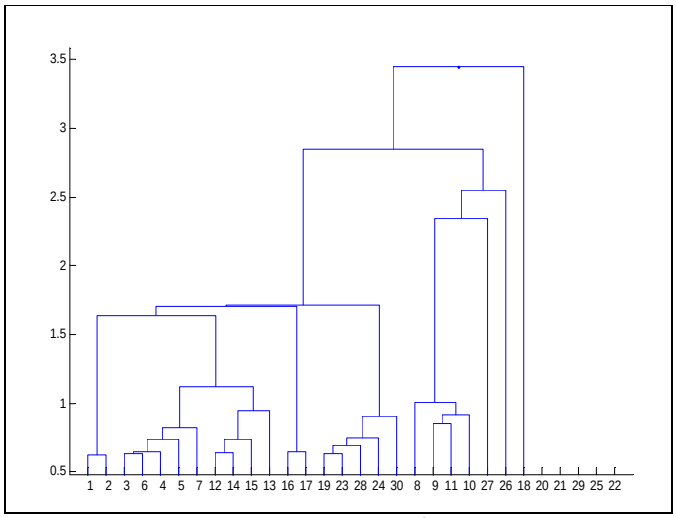

Fig.5 dendogram of the clusters

Experiments are also conducted with large datasets on Intel (3.6 GHz) with 8 GB RAM PC. Fig 6 shows the execution time of the different document set size.

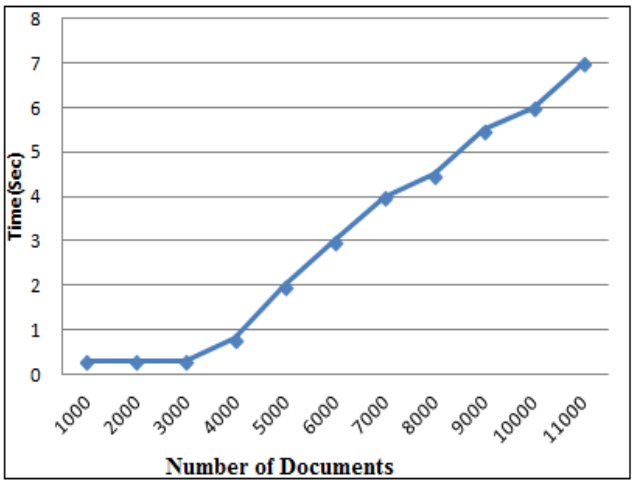

Fig.6. execution time with different size documents

V. Conclusion

In this paper, we have proposed a clustering framework for large document sets. The process of large document clustering begins with document preprocessing and cleaning and terminates with preparing clusters according to similarity between higher terms. We have used a predefined threshold for extracting terms from the databases. This proposed framework is successful in terms of noise reduction and it assigns documents into most similar cluster in lesser timing. This proposed approach is also compared with the classical Kmean clustering method, in this comparison we found that this proposed framework is efficient and have higher performance.

REFERENCES

[1] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, Survey of Clustering Algorithms, IEEE Transactions on Neural Networks Vol. 16, No. 3, May 2005.

[2] Bidyut kr. Patra,Sukumar Nandi,P.Viswanath, A distance based clustering method for arbitrary shaped clusters in large datasets,Pattern Recognition 44(2011) 2862-2870.

[3] M. Anderberg, Cluster Analysis for Applications. New York: Academic,1973.

[4] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed. NewYork: Wiley, 2001.

[5] Jin Chen, Alan M. MacEachren, and Donna J. Peuquet, "Constructing Overview + Detail Dendrogram-Matrix Views ", IEEE Transactions on Visualization and Computer Graphics, Vol .15, No.6 ,Nov 2009.

[6] B. Duran and P. Odell, Cluster Analysis: A Survey. New York:Springer-Verlag, 1974.

[7] B. Everitt, S. Landau, and M. Leese, Cluster Analysis. London: Arnold, 2001.

[8] P. Hansen and B. Jaumard, "Cluster analysis and Math- ematical programming," Math. Program., vol. 79, pp. 191–215, 1997.

[9] A. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[10] E. Backer and A. Jain, "A clustering performance measure based on fuzzy set decomposition," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-3, no. 1, pp. 66–75, Jan. 1981.

[11] C. Bishop, Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995.

[12] V. Cherkassky and F. Mulier, Learning From Data: Concepts, Theory, and Methods. New York: Wiley, 1998.

[13] A. Baraldi and E. Alpaydin, "Constructive feedforward ART clustering networks—Part I and II," IEEE Trans. Neural Netw., vol. 13, no. 3, pp. 645–677, May 2002.

[15] M. Steinbach, G.Karypis, V.Kumar, A Comparison of document clustering techniques, Proc. Of the 6th ACM SIGKDD int'l conf. on Knowledge Discovery and Data Mining(KDD), 2000.

[16] P.Willet, Recent trends in hierarchical document clustering: a critical review, Information processing & Management 24(5) (1988), pp 577-597.

[17] Ghanshyam Thakur, Rekha Thakur and R.C. Jain, "Association Rule Generation from Textual Document" International Journal of Soft Computing, 2: 2007 pp. 346-348.

[18] M. Dash, H.Liu, P. Scheuermann, K.L. Tan, fast hierarchical clustering and its validation, Data & Knowledge Engineering 44(1) (2003) pp. 109-138.

[19] R. Balaji And R.B. Bapat, Block Distance Matrices, Electronic Journal of Linear Algebra ISSN 1081-3810 A publication of the International Linear Algebra Society Volume 16, pp. 435-443, December 2007.

[20] M.Nanni, speeding-up hierarchical agglomerative clustering in presence of expensive metrics, in proc. Of Ninth Pacific-Asia conference on knowledge discovery and Data mining (PAKDD)2005, pp. 378-387.

[21] P.A.Vijaya, M.N.Murty, D.K. Subramanian, Efficient bottom up hybrid hierarchical clustering techniques for protein sequence classification, pattern Recognition 39 (12) (2006), pp.2344-2355.