International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 8 5073 - 5075

Survey of Document Clustering Approach for Real World Objects (Documents)

Sandeep Kumar
Computer Science & Engineering Department
United College of Engineering & Research, UPTU
Allahabad, India
sandeepkr.2050@gmail.com

Associate Prof. Sanjay Pandey
Computer Science & Engineering Department
United College of Engineering & Research,
UPTU,Allahabad, India
sanjaypandey@united.ac.in

Abstract:- Since the amount of text data stored in computer repositories is growing every day, we need more than ever a reliable way to assemble or classify text documents. Clustering can provide a means of introducing some form of organization to the data, which can also serve to highlight significant patterns and trends. Document clustering is used in many fields such as data mining and information retrieval. This thesis presents the results of an experimental study of some common document clustering techniques. In particular, we compare the two main approaches of document clustering, agglomerative hierarchical clustering BIRCH and Partitional clustering algorithm K-means. As a result of comparing both algorithms we attempt to establish appropriate clustering technique to generate qualitative clustering of real world document.

Keywords:- *Document Clustering, Vector Space Model, Matrix Representation, K-means, Partition Clustering, Hierarchal clustering*

*****

I. Introduction

In this century we are surrounded by tremendous quantity of information and data that means we are at information age. And in this age we have not known that in what way we organizes our data as it is appears unbounded means there is not any limit on it. As we are going towards digitization, data and useful information finding in big samples of data becomes difficult by man power. So we need some techniques to get rid out of sampling data from chaos of data. And data retrieving from huge data problem linked with number of fields like in retrieving useful news along with their archives, in whether forecasting, in sampling and retrieving scientific data etc.

Data mining is being positioned keen on apply and studied for databases, as well as relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web.

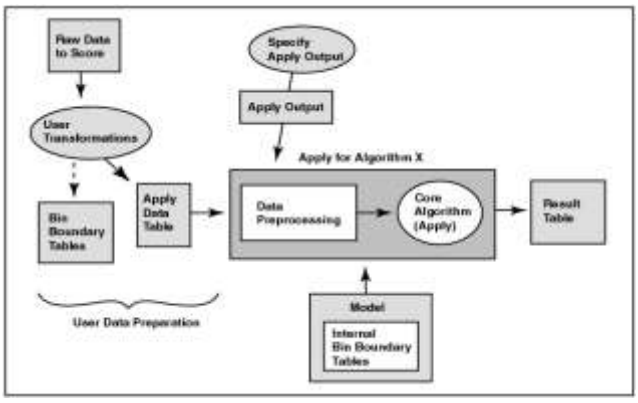

Figure 1.1 Data Mining Definition

Data clustering algorithms can be broadly classified into following categories:

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods
- Frequent pattern-based clustering
- Constraint-based clustering

With partitional clustering the algorithm creates a set of data non-overlapping subsets (clusters) such that each data object is in exactly one subset. These approaches require selecting a value for the desired number of clusters to be generated. A few popular heuristic clustering methods are k-means and a variant of k-means-bisecting k-means, k-medoids, PAM (Kaufman and Rousseeuw, 1987), CLARA (Kaufmann and Rousseeuw, 1990), CLARANS (Ng and Han, 1994) etc. With hierarchical clustering the algorithm creates a nested set of clusters that are organized as a tree. Such hierarchical algorithms can be agglomerative or divisive approaches.

Agglomerative algorithms, also called the bottom-up algorithms, initially treat each object as a separate cluster and successively merge the couple of clusters that are close to one another to create new clusters until all of the clusters are merged into one. Divisive algorithms, also called the top-down algorithms, proceed with all of the objects in the same cluster and in each successive iteration a cluster is split up using a flat clustering algorithm recursively until each object is in its own singleton cluster. The popular hierarchical methods are BIRCH, ROCK, Chameleon and UPGMA. An experimental study of hierarchical and partitional clustering algorithms was done by and proved that bisecting kmeans technique works better than the standard kmeans approach and the hierarchical approaches.

Density-based clustering methods group the data objects with arbitrary shapes. Clustering is done according to a density (number of objects), (i.e.) density-based connectivity. The popular density-based methods are DBSCAN and its extension, OPTICS and DENCLUE [3]. Grid-based clustering methods use multi resolution grid structure to cluster the data objects. The benefit of this method is its speed in processing time. Some examples include STING, Wave Cluster. Model-based methods use a model for each cluster and determine the fit of the data to the given model. It is also used to automatically determine the number of clusters. Expectation-Maximization, COBWEB and SOM (Self-Organizing Map) are typical examples of model-based methods.

Frequent pattern-based clustering uses patterns which are extracted from subsets of dimensions, to group the data objects. Constraint-based clustering methods perform clustering based on the user-specified or application-specific constraints. It imposes user's constraints on clustering such as user's requirement or explains properties of the required clustering results.[7]Among all these methods, this paper is aimed to explore three methods –K-means, Clarans which is partitioning based clustering method and Hierarchal based clustering methods .We compare them by using some criterion function.

FIG 1.2 THE STAGES OF DOCUMENT CLUSTERING

II. Literature Review

Venu Satuluri, Srinivasan Parthasarathy (2011) et. al. Symmetrizations for Clustering Directed Graphs" this paper investigates various ways of symmetrizing a directed graph into an undirected graph so that previous work on clustering undirected graphs may subsequently be leveraged. Direct application of these similarity measures to modern large-scale power- law networks is problematic because of the presence of hub nodes, which become connected to the vast majority of the network in the transformed undirected graph. Analyze this problem and propose a Degree-discounted similarity measure which is much more suitable for large-scale network sand show extensive empirical validation.

This paper also remove draw backs of is that there exist meaningful clusters which do not necessarily have a low directed normalized cut. The prime examples here are groups of vertices which do not point to one another, but all of which point a common set of vertices (which may belong to a different cluster).

Chandan Jadon, Ajay Khunteta(2013) et. al. 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). So here clustering approach as well as document representing approach gets compare with k-means clustering algorithm to focus on getting less time complexity. In this paper there is given that how data is firstly preprocessed to make for further processing and then it get represented by either vector space model or matrix representation.

Charu C. Aggarwal ,Yuchen Zhao, Philip S. Yu(2014) et. al." 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. In order to design the clustering method, combined an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information. Present results on real data sets illustrating the effectiveness of approach. The results show that the use of side-information can greatly enhance the quality of text clustering, while maintaining a high level of efficiency.

Bader Aljaber, Nicola Stokes ,James Bailey ,Jian Pei(2008)et. al. Document Clustering of Scientific Texts Using Citation Contexts" This paper investigate the power of these citation-specific word features, and compare them with the original document's textual representation in a document clustering task on two collections of labeled scientific journal papers from two distinct domains: High Energy Physics and Genomics .And compare these text-based clustering techniques with a link-based clustering algorithm.

III. Methodology

BIRCH Clustering Algorithm
Phase 1

The core task of Phase 1 is to examine all data and construct an initial in memory CF tree via the identified capacity of page of memory and reprocess the space again and again on disk.

Phase 2

Phase 2 is elective. We have monitored that the existing universal or semi-global clustering technique implement in Phase 3 have diverse input extent ranges within which they execute well in terms of both speed and excellence.

Phase 3

We acquire a set, of clusters that, confines the key distribution pattern in the data, though small and restrict inexactness might be present because of the uncommon misplacement difficulty mentioned in and the fact that Phase 3 is implemented on a coarse review of the data.

 ISSN: 2321-8169
5073 - 5075

PHASE 4

In this phase the centroids of clusters generated by phase 3 are used for further processing means we have allocated the data points to their position by following the constraints.

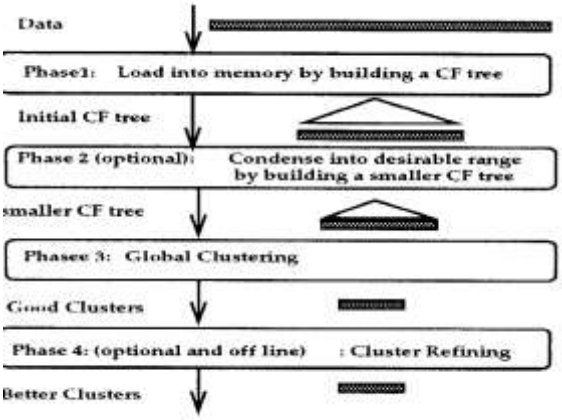

Fig 1.3 BIRCH Overview

The 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.

IV. Conclusion

In this paper, we will like to evaluate the performance of K-means, partition approach and Agglomerative hierarchal approach by comparing them. Along with comparing clustering approaches we find that which is appropriate clustering algorithm to produce high feature quality clustering of real world documents. And also establishing a new algorithm which we try to make more efficient than existing clustering approaches which we already discusses in this paper.

Refrences

[1] Shi Zhong,"A k-means algorithm to improve the Efficiency Using Normal Distribution Data Points",(IJCSE) International Journal on Computer Science and Engineering, 2010.

[2] Xufei Wang, Jiliang Tang and Huan Liu, "Document Clustering via Matrix Multiplication" 2011 11th IEEE InternationalConference On Data Mining.

[3] Book: Information Retrieval, Algorithms and heuristics by David A.Grossman and Ophir Frieder.Published by Springer International.

[4] Anil K. Jain, "Pattern Recognition Letters", Journal Elsevier, Pattern Recognition Letters 31 (2010) 651–666.

[5] Atika Mustafa, Ali Akbar, and Ahmer Sultan"Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009.

[6] L. Wanner, "Introduction to Clustering Techniques",International Union of Local Authorities, July, 2004.

[7] T. Velmurugan,and T. Santhanam, "A Survey of Partition based Clustering Algorithms in DataMining: An Experimental Approach" An experimental approach.

[8] Porter, M.F.: An algorithm for suffix stripping.Program, Vol. 14, No. 3, 1980

[9] Na Wang; Pengyuan Wang; Baowei Zhang; , "An improved TF-IDF weights function based on information theory," Computer and Communication Technologies in Agriculture Engineering (CCTAE), 2010 International Conference On , vol.3, no., pp.439- 441, 12-13 June 2010.

[10] Lee, D.L.; Huei Chuang; Seamons, K.; , "Document ranking and the vector-space model,", IEEE , vol.14, no.2, pp.67-75, Mar/Apr 1997.

[11] Shobha S. Raskar, D. M. Thakore "Text Mining and Clustering Analysis", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.6, June 2011.

[12] Mrs .S.C.Punitha and Dr.M.Punithavalli, "A Comparative Study to Find A Suitable Method for Text Document Clustering", International Journal of Computer Science & Technology (IJCSIT) Vol 3,No 6 December 2011.