Survey on Clustering for Data mining Text Clustering for Digital Analysis

P.Seethamani[1], S.Christina Magneta[2] and Dr.C.Sundar*[3]

[1]PG Scholar, Department of Computer Science and Engineering, Christian College of Engineering and Technology, Dindigul, Tamilnadu - 624619, India.
Seethamani564@gmail.com

[2],[3]Department of Computer Science and Engineering, Christian College of Engineering and Technology, Dindigul, Tamilnadu - 624619, India.
christinarvs@gmail.com

**Corresponding Author E-mail:* sundarc007@yahoo.com

Abstract— **The computer analysis, hundreds of thousands of files are regularly examined. A large amount of the data in those files are contains of unstructured text passage, whose analysis by computer examiners is complicated to be performed. In this framework, computerized methods of analysis are of great attention. In exacting, algorithms for clustering documents can smooth the progress of the discovery of new and useful knowledge from the documents less than analysis. We illustrate the proposed approach by carrying out widespread testing with clustering algorithms K-means++, and CSPA applied to Text-document for document analysis. In adding together, two qualified validity indexes were used to automatically estimation of the large number of clusters. Correlated studies in the narrative are significantly more restricted than our study.**

Keywords— **High dimensional data, Subspace clustering, Density based clustering, Hubness based clustering, Popular Nearest neighbors.**

I. Introduction

In generally, the process of analysing data from different levels and methods it into useful information is called as data mining. It is also known as data or knowledge discovery. It allows users to analyse the data from many different dimensions or angles, categorize it, and summarize the relationships identified. Officially, data mining is the method of finding the association or correlations or patterns among by a computer are termed as data, which includes operational data, non-operational data and Meta data. Data mining functionalities are characterization and discrimination, classification and prediction, cluster analysis, outerlier analysis, trend and evolution analysis.

Clustering is a partition of data into groups of analogous objects. Each group, called cluster, consists of objects that are similar amongst them and dissimilar compared to objects of other groups. Clustering is the process of grouping objects into classes of same objects. The objective of clustering is to determine the intrinsic grouping in a set of unlabeled data. There are four groups in clustering algorithms: *partitional, hierarchical, density based, and subspace clustering.* Clustering is the crucial task to cluster the high dimensional data. The data items are collected authorizing to logical relationships or end user likings, which can be defined by the cluster.

Theoretically, a cluster is collection of items which are *related* between them and *unrelated* to the items belonging to other clusters. The role of clustering is to find the structure in a collection of unlabelled data. A set of unlabeled data can be considered the most important goal of clustering. The observation having more dimensions usually leads to the so-called curse of dimensionality this shows many standards machine-learning an algorithm becomes impaired. This is due to two common effects: *the empty space phenomenon and concentration of distances.*

In this paper we discuss about how to cluster the high dimensional data effectively and accurately. High dimensional data sets show a tendency of sparse, because the number of points required to represent any distribution grows exponentially with the number of dimensions. The problem of high dimensional data is omnipresent and abundant. This shows a bad density approximate for high-dimensional data and to create strain for density-based approaches and concluding counterintuitive property of high-dimensional data descript all distances between data points become harder to discriminate as dimensionality increases, this can leads to problems with distance-based algorithm.

The proposed technique to analyse the documents use of clustering methods. In the analysis process use two different clustering algorithms k-means++ and CSPA algorithms. From this we can analyse document-analysis effectively and accurately. *K-means++* clustering algorithm aims to partition *n* observations into *k* clusters ,which each observation belongs to the clusters with the nearest mean. The *CSPA algorithm* essentially finds a consensus clustering from a cluster ensemble formed by a set of different data partitions. This represents pair-wise similarities between objects. The similarity between two objects is simply the fraction of the clustering solutions in which those two objects lie in the same cluster.

II. LITERATURE SURVEY

In the literature survey we study about different type of 10 papers and their techniques related to clustering high dimensional data. From that analyze different clustering techniques for clustering data sets.

N. Tomasev and M. Radovanovi´c at etl [1] proposed role of Hubness in Clustering High- Dimensional Data.Many domains naturally consist of High dimensional data and that can arise every day. that is great challenge for traditional data mining techniques. In this paper they use novel point of view on the difficulty of clustering high dimensional data.Instead of that attempt to avoiding the curse of dimensionality by observe a lower particularly by show that Hubness. From this the bent of high-dimensional data to include points that frequently occurs in k-nearest-neighbor list of other points that can be exploited in clustering. The algorithms are used to deliver the effective results that are HPC, HPKM and GHPC. The enforcing K-nearest-neighbor consistency such as K-means explored. The typical use of k-nearest-neighbor list is to construct a k-NN graph and reduce the problem to graph clustering.

The Hubness has been investigated in other related fields: Classification, image feature representation, data reduction, collaborative filtering, text retrieval, and music retrieval. Proposing numerous Hubness-based clustering algorithms and testing on high dimensional data. (K-nearest neighbor, Centroid based algorithm (GHPC-Global Hubness-Proportional clustering).The proposed GHPKM method had proven to be more robust then K-means++ baseline on both synthetic and real-world data and the presence of high levels of artificially introduced noise. The GHPKM and GHPC excel are expected to improvement by providing intercluster distance. GHPC offer greatest improvement in noise detection and clustering high dimensional data.

V.Suganthi and S.Tamilarasi [2] they propose a studing about clustering high dimensional data using hubness phenomenon.The non-trivial process of extracting information in the large database, this technique is the data mining in real world data set. The data ordnance has a high dimensional data, which made a problem of computationally infeasible. This problem clustering plays a major job in management low dimensional data and high dimensional data. The low dimensional data made a task extremely simple and easy to cluster but high dimensional data is a important statement to cluster and it has to determine using Hubness phenomenon.

This paper is used to determine and manage hyperspheric cluster. In this paper the system to use hub based clustering to improve the quality of cluster in their effectiveness and accuracy and to avoid only hyper-spherical cluster.The various clustering algorithms are used for predicting the high dimensional data. The high dimensional data can affect by the curse of dimensionality. The algorithms are a). Subspace clustering algorithm, b). Density based clustering algorithm, c). Hubness based clustering algorithm, D). Popular Nearest neighbors in High Dimensional Data [2].

A. *Subspace clustering algorithm* to detecting and comparing the cluster with three different methods such as cell based, density based, and clustering oriented. The overall quality can be judge by the measure called accuracy. The accuracy specified by correctly predicted objects divided by all objects[2].

B. *Density based clustering algorithm* use SUBCLU algorithm for identifying the cluster in high dimensional data. The gene expression is taken as data set for generate cluster results.

C. *Hubness Based clustering algorithm* using the hubness information k-Nearest Neighbor is used for managing high dimensional data. Hub is a data point that frequently occurred in a k-nearest neighbor list and rarely occurred points or May outliers are called as antihubs. Hubness has three algorithms such as hw- kNN, h-FNN, NHBNN.

D. *Popular Nearest Neighbors in High Dimensional Data* has been performed a theoretical and experimental analysis of hubness and its triggers, techniques for clustering, Classification and information retrieval. It has observed the following methods for reduce the dimensionality: (a). Principal – component analysis, (b). Independent component analysis, (c). Stochastic neighbour embedding, (d). Isomap and (e). Diffusion map.

Singh Vijendra [3], proposed the efficient clustering for high dimensional data that use of subspace based and density based clustering.Existing System.Finding a cluster in high dimensional data and clustering problem with high dimensional data and approaches to solve this problem use the following clustering methods such *as subspace based clustering and density based clustering* approaches. *Subspace based clustering algorithms* focus the search for applicable dimensions allowing them to discover clusters that exist in multiple probably overlap subspace and identifies subspace clusters. It can be classified into two categories are partition based approaches and grid based approaches.

The partition based algorithms partition the set of objects into mutually exclusive groups. Each group along with the subset of dimensions. The grid based algorithm considers the cluster problem as the data matrix as a high dimensional grid and the clustering process as a search for dense regions in the grid. *Density based clustering* supported capably to decrease the runtime of clustering. It groups the neighboring objects into cluster based on local density conditions rather than proximity between objects. Density based methods have noise tolerance and can discover no convex clusters. It is similar to hierarchical and partitioning methods, density based techniques encounter the difficulties in high dimensional spaces and reduces any clustering tendency[3].

C. C. Aggarwal and P. S. Yu [4] finding generalized projected clusters in high dimensional spaces.This paper used to finding arbitrarily oriented projected clusters in high dimensional spaces. It may be view as a way of demanding to redefine clustering for high dimensional applications by search for unseen subspaces with clusters which are formed by inter-attribute correlations. Estimated clustering which are capable to assemble clusters in randomly associated subspace of lower dimensionality. Using absolute cluster feature vectors categorize to make the algorithm scalable for extremely large database. The successively time and space necessities of the algorithm are regulating, and are expected to tradeoff with enhanced accuracy. The idea of eliminating the most sparse subspace for each cluster, and projected the points into those subspaces, which the greatest similarity. The sparsity of high dimensional data prevents the detection of natural cluster for full-dimensional problems. Use this technique for effective high dimensional data visualization.

Michael Steinbach and Levent Ertoz at etl [5] told about the challenges in clustering high dimensional data.Cluster analysis divides data into groups for the purpose of summarization or improved understanding. The main aim of cluster analysis used to group connected documents for browsing, to discover genes and proteins they have similar functionality. The goal is the objects in a group should be similar to one another and diverse from objects in other groups. This paper provided a brief introduction to cluster analysis with an emphasis on the dare of clustering high dimensional data. The principal challenge in extending cluster analysis to high dimensional data to overcome the “curse of dimensionality “. The several working definition cluster analysis is discussed they are well-separated cluster, center-based cluster, contiguous cluster, density-based cluster, and similarity based cluster.

Cluster analysis has been used in a wide range of fields: *psychology and other social sciences, pattern recognition, biology, information retrieval, statistics, information retrieval, machine learning, and data mining.* In this several clustering algorithms are discussed they are hierarchical and partitional clustering (K-means, MIN, MAX, and Group average. Grid based clustering for high dimensional data (CLIQUE, MAF, DENCLUE, and OptiGrid), Noise modeling in wavelet space, and new introduced technique is Concept based clustering. Resolved issues in cluster analysis: Scalability to large data sets, independence of the order of input, effective means of evaluating the validity of clusters that are produced, easy interpretability of results, an ability to function in an incremental manner, and robustness in the presence of different underlying data and cluster characteristics[5].

Dragomir R. Radev and Hongyan Jing at etl [6] proposed the centroid-based summarization of multiple documents using utility-based evaluation, sentence extraction and user studies.Multi document summarizer called MEAD, which generate summaries using cluster centroids formed by topic finding and tracking system. In this express two innovative techniques based on sentence utility and subsumption.It summarizes cluster of news articles without human intervention grouped by a topic recognition system. MEAD uses in sequence from the centroids of the clusters to pick sentence that are most expected to be appropriate to the cluster topic. It used a new utility-based procedure, CBSU for the assessment of MEAD and of summarizers in general. From this we can determine how the multi-document summaries are built and evaluated. The *Topic detection and multi-document summarization* is the process of identifies all articles on an promising event is called *tropic Detection and Tracking.* A new technique for multi-document summarization, called Centroid-based summarization which uses as participation the centroids of the cluster produced by the CIDR to identify which sentence are central to the topic of cluster, rather than the individual articles.

The main objectives of this paper are the improvement of a centroid-based multi-document summarizer, the utilize of cluster-based sentence utility (CBSU) and cross-sentence informational subsumption (CSIS) for assessment of particular and multi-document summaries, two user studies that support our findings and an estimation of MEAD. Then finally suggest a utility-based evaluation scheme. It can be used to calculate both multi-documents and single-documents summaries. Informational content of sentence use the following techniques are *Cluster-based sentence utility(CBSU), Cross-sentence informational subsumption(CSIS), Equivalence classes of sentences, Comparison with MMR*. MEAD a centroid-based multi-document summarizer use some cluster algorithms are centroid-based algorithm and redundancy-based algorithms. The techniques for evaluating summaries are single-document summaries and utility-based evaluation of both single and multidocument summaries[6].

David Arthur and Sergei Vassilvitskii [7] tell as about the k-means++ and the advantages of careful seeding.The k-means method seeks to minimize the average squared distance points in the same cluster. But it offers no exactness guarantees, its cleanness and speed .By augmenting k-means with a easy, randomized seeding method, from that obtain an algorithm that is O(log k)-competitive with the most favorable clustering. That technique improves both speed and the accuracy of k-means. The k-means to be depend almost linearly on n in practice. The k-means algorithm is a simple and fast algorithm; it offers approximation guarantees at all. In which they choose another algorithm for improvement of accuracy. The k-means++ augments the k-means algorithm by choosing the initial cluster centers according to the D2 metric, and not uniformly at random from the data. The new seeding method yields a much better performing algorithm, and consistently finds a better clustering with a lower potential than k-means. The main advantage is algorithms are test random, that report minimum and average potential, as well as the mean time to complete.

Nenad Tomaˇsev and Miloˇs Radovanovi´c at etl [8] proposed the concept of hubness-based fuzzy measures for high dimensional and k-nearest neighbor classification.The data of concern today in data-mining application was complex and was usually represented by many different features. High dimensional data can be processed by the design of machine-learning algorithms. It observes the arising high dimensional properties could be exploited n improving the overall algorithm design. The high dimensional data is related to nearest-neighbor learning methods, is known as hubness and refers to the emergence of very influential nodes in k-nearest neighbor graphs. In this propose several fuzzy measures for k-nearest neighbor classification, all based on hubness, which express fuzziness of elements appearing in k-neighborhoods of other points. The related works on the *hubness based fuzzy measures are hubness-weighted k-NN, Fuzzy nearest neighbor algorithm and proposed hubness-based fuzzy measures.* The experimental evaluations based on the following data sets are UCI data sets, imageNet data, and scalability and probability landscapes. The main advantage of the fuzzy approach lies in the mentioned interpretability of the results. The main advantage of the fuzzy approach lies in the mentioned interoperability of the results, and takes advantage of high intrinsic dimensionality of the data.

Maurizio Filippone and FrancescoCamastra at etl [9] proposed a survey of kernel and spectral methods based on clustering.In this paper spectral and kernel method for clustering have been reviewed attention to fuzzy kernel method for clustering and to the connection between kernel and spectral approaches. This fully presents a survey of kernel and spectral clustering method. The two approaches clever to construct nonlinear separating hypersurfaces linking clusters. The kernel clustering methods are the kernel versions of many classical clustering algorithms in these examples are K-means, SOM, and Neural Gas. The spectral clustering arise in spectral graph theory and graph cut problem. The fuzzy kernel clustering methods were offered as extensions of kernel k-means clustering algorithms. Unsupervised data analysis uses the clustering algorithms, which provides useful tool to explore data structures.

The clustering method addressed in many contexts such as data mining, document retrieval, image segmentation and pattern classification. The clustering techniques can be divided into two categories: *hierarchical and partitioning.* Hierarchical clustering techniques were able to finding the structures which can be further divided in substructures and so on recursively. The hierarchical structure of groups has been known as dendrogram. Partitioning clustering methods to obtain a single partition of data without any sub-partition and or based on the optimization of an appropriate objective function. Spectral clustering methods arise from the spectral graph theory. Both approaches lies in their ability to construct an adjacency structure between data avoiding to deal with a prefixed shape of clusters[9].

Arthur Flexer and Dominik Schnitzer at etl [10] proposed a mirex meta-analysis of hubness technic in audio music similarity.MIREX “Audio Music likeness and reclamation” task for a Meta analysis of the hub phenomenon. Hub songs come out related to an undesirably high number of other songs due to a difficulty of measuring distances in high dimensional spaces. That hub songs display fewer perceptual match to the songs they are close to, according to an audio likeness function, than non-hub songs. Proper modeling of music similarity is at the heart of every application allowing automatic organization and processing of music data base. Hub songs are defined as songs which are according to the audio similarity function, similar to very many other songs and keep appearing unwontedly in recommendation lists preventing other songs from being recommended at all. *All results evaluation measures for all algorithms use the mutual proximity.* The following algorithms are hubness across algorithms, hubness and perceptual quality, and reducing hubness. Application of the freshly introduced technique of “mutual proximity” is able to decisively improve the separation of similarity audios.

Sindhupriya. R, and Ignatius Selvarani [11] proposed the technic to find high dimensional data using k-means based clustering.In this paper using Centroids-based approach that use hubs to inexact local data centers improvement. In that test GHPC (Global Hubness-Proportional Clustering) had a best performance in various test settings, on synthetic and real-world data, as well as in the existence of high levels of artificially initiate noise. The proposed algorithm represented to use hubness for improving high dimensional data clustering. User can use high dimensional data with the maximum number of data in the multiple columns and rows used for high dimensional data clustering into the real data dimensional using k-means algorithm. The k-means algorithm using the clustering with real data depends upon density based approach to find nearest neighbor node. Clustering of high dimensional data to transfer analyze process to find out how much amount of noise occur in data. User get the search data entity value how much relation between the maximum distances of data noise occur, it should become data losses. Analyze process noise data neglecting into high dimensional data process. To finalize clustering of real data into without losses to get this user real data.

III. PROPOSED SYSTEM

Clustering algorithms are typically used for exploratory data analysis, where there is little or no prior knowledge about the data. From a more technical viewpoint, our datasets consist of unlabeled objects, the classes or categories of documents that can be found are a priori unknown. In this context, the use of clustering algorithms, which are capable of finding latent patterns from text documents found in computers. The proposed system can implement the document analysis use of clustering algorithms. From we can identify cluster the documents and the document contain which type of information, which can focus what type of concepts.

In this system we use two clustering algorithms. They are K-means++ and CSPA. Use of those algorithms can achieve efficiency and accuracy in clustering document analysis. *K-means++* clustering algorithm aims to partition n observations into k clusters ,which each observation belongs to the clusters with the nearest mean. The *CSPA algorithm* essentially finds a consensus clustering from a cluster ensemble formed by a set of different data partitions. This represents pair-wise similarities between objects. The similarity between two objects is simply the fraction of the clustering solutions in which those two objects lie in the same cluster.

IV. Conclusion

From the literature survey, we discussed about the existing clustering method, which are mainly used to understand about clustering and clustered data. In recent trends, Text-document data as real world dataset for clustering becomes difficult because it is a very high dimensional data. To eradicate this problem, the proposed system use a hub based clustering algorithm to automatically determine the number of cluster from the Text-document data. The proposed approach by delivery out extensive experimentation with clustering algorithms K-means++, and CSPA applied to datasets obtained from computers held in real-world data set. The computer documents are clustered and the document can be identified from that the document contains information about which.

References

[1]. N. Tomasev, M. Radovanovi´c, D. Mladeni´c, and M. Ivanovi´c, “The role of hubness in clustering high-dimensional data,” in *Proc. 15th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Part I,* 2011, pp. 183–195.

[2]. V.Suganthi and S.Tamilarasi,” A Study on Clustering High Dimensional Data Using Hubness Phenomenon” *in ISOR Journal of Computer Engineering(IOSR-JCE),(May-Apr2014),PP 22-30.*

[3]. Singh Vijendra,“Efficient Clustering for High Dimensional Data: Subspace Based Clustering and Density Based Clustering.” *In Information Technology Journal 10(6),2011.*

[4]. C. C. Aggarwal and P. S. Yu, “Finding generalized projected clusters in high dimensional spaces,” in *Proc. 26th ACM SIGMOD Int. Conf. on Management of Data*, 2000.

[5]. Michael Steinbach,Levent Ertoz,and Vipin Kumar,”The Challenges of Clustering High Dimensional Data”.

[6]. Dragomir R. Radev, Hongyan Jing, Malgorzata Budzikowska,”Centroid-based summarization of multiple documents : sentence extraction , utility-based evalution,and user studies. “

[7]. David Arthur and Sergei Vassilvitskii, ”K-means++:The advantages of Careful Seeding”.

[8]. Nenad Toma˘sev · Milo˘s Radovanovi´c · Dunja Mladeni´c · Mirjana Ivanovi´c, ”Hubness-based fuzzy measures for high-dimensional k-nearest neighbour classification” *in Int.J. Mach. Learn. & Cyber 2011.*

[9]. Maurizio Filippone,FrancescoCamastra,Francesco Masulli, and Stefano Rovetta,“A survey of Kernel and spectral methods for clustering.” *In 2007.*

[10]. Arthur Flexer, Dominik Schnitzer, and Jan Schluter, “A MIREX Meta-analysis of Hubness in Audio Music Similarity“,*in 13th International Society for Music Information Retrieval Conference(ISMIR 2012).*

[11]. Sindhupriya. R, and Ignatius Selvarani. X,” K-Means Based Clustering In High Dimensional Data “ , *in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 2, February 2014.*