Analysis of Different Clustering Techniques in Data and Text Mining

Ms.S.Prabha
Associate Professor, Department of Information Technology
K.S.Rangasamy College of Technology
Namakkal (Dt), Tamil Nadu, India.
prabha.dw@gmail.com

Dr.K.Duraiswamy
Dean
K.S.Rangasamy College of Technology
Namakkal (Dt), Tamil Nadu, India.

Ms.M.Sharmila
PG Scholar, Department of Information Technology
K.S.Rangasamy College of Technology
Namakkal (Dt), Tamil Nadu, India.
sharmi28.it@gmail.com

***Abstract--* In recent days clustering becomes important in pattern detection, unsupervised learning process, data concept construction, information retrieval, text mining, web analysis, marketing and medical diagnostic. The purpose of this paper is an attempt to reconnoiter some of the important clustering techniques in the data mining literature and to compare some aspects of clustering algorithms which contains performance, order of input, accuracy, scalability, shapes discovered, dimensionality and dealing with noisy data. The algorithms are Partitional approach, hierarchical approach, seeded approach, ontology approach, concept based approach.**

Keywords- Clustering; Semi Supervised; Ontology; Semantic; Constraint; Similarity measures;

I. INTRODUCTION

The application side work mostly focuses on document clustering or classification which has become a significant technique for document organization, extracting topics and interests and speedy information retrieval. Cluster based retrieval gives faster search, better navigation, improved recall and analysis. The issues of clustering lies in the representatives for clustering whether you use vector space or normalization and it needs a notion of similarity or distance. The documents are related based on either semantic similarity or statistical similarity. The basic intension is to produce good clustering with high quality in which intra- cluster similarity is high and inter- cluster is low. The measured quality of clustering depends on the representatives and similarity measures.

II. DIFFERENT TYPES OF CLUSTERING TECHNIQUES

A. Partitional Clustering

The Partitional clustering algorithms [1] partition the given n data sets or data tuples into k partitions (k≤ n) where each partition represents a sub-set or a cluster. The data objects which are partitioned should follow the below mentioned criteria,

- At least one data object should reside in each cluster or sub-set
- Data objects should belong to only one cluster group

The second criteria may be relaxed in some soft clustering algorithms. There are many Partitional algorithms are available. The following are widely used methods, one is iterative or reallocation and another one is single pass method. Reallocation methods are used to improve the results of partitioning. In this method data objects are being shifted from one cluster to another. But single pass methods [2] are used in the initial stage of iterative methods. When compared to single pass methods iterative methods are widely used. In the Partitional algorithms distance between the data object and the centroid should be minimum in order to obtain better results. K-means, k-medoid and some of their variations are mostly used.

1) K-Means

The main aim of k-means algorithm [1] [3] is to partition the input data objects into k clusters. The number of clusters (k) is given as input to this algorithm. The initial step in this algorithm is to select k objects randomly. Then iteratively refining the data objects till some threshold level or stop criteria is reached. K-means works as follows, Step1: Chose k objects randomly from set 'D' which contains n objects.

Step2: Based on the mean value of the objects in the sub-set, reassign the objects to the sub-set to which the object is most similar.

Step 3: Recalculate the mean values for all sub-sets or clusters.

Step 4: Repeat the above step until no change in the object assignment.

2) *K-Medoid*

Because of outliers, the quality of k-means algorithm [1] [14] is distorted. Instead of calculating mean values, the real objects in the data set are taken for the formation of clusters. All other objects are assigned to the most similar data objects which is taken as representative. Reassign the representative and data objects till the quality of the result is increased. Another variation ok k-means is bisecting k-means. In this method data objects are partitioned into two (bi) clusters. In those two partitions, any one of the partition is chosen and again it is bisected. But, if the largest cluster is selected for bisection, effective and balanced clustering will result. This bisection continues until 'k' clusters results. The main advantage of Partitional algorithms [2] is they are simple. They have many disadvantages like not scalable to large database, noise or outliers will affect the Partitional algorithms easily and not suitable for finding complex shapes of clusters.

*B. **Hierarchical Clustering***

Hierarchical clustering algorithms [14] attempt to form a tree of clusters by grouping data objects. When a set of N items to be clustered, and an N*N distance matrix, the following are the steps for hierarchical clustering [25],

Step1: Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

Step2: Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

Step3: Compute distances (similarities) between the new cluster and each of the old clusters.

Step4: Repeat steps 2 and 3 until all items are clustered into K number of clusters

Hierarchical clustering is categorized into two,

- Agglomerative approach
- Divisive approach

1) *Agglomerative Approach*

It is a bottom-up method [14] in which it starts with a singleton. The data objects are combined till all the data objects become a single cluster or until some termination conditions are reached. Initially, each object in the set is taken as a single cluster. The objects are merged based on the inter cluster similarity.

2) *Divisive Approach*

This is a top-down method [3] in which it starts with a big cluster at the initial stage. Then the data objects are divided into smaller clusters according to the similarity between the objects. The procedure of dividing continues till each data object belongs to a single cluster. Both the methods stops when k number of clusters is achieved and the user can define the number of clusters to be the condition which stops the merge or split operation. The main advantage of Hierarchical clustering [14] is flexibility to any level of granularity, ability to handle multiple similarity or distance, versatile and is applicable to different attribute types. The main disadvantage is that it is relatively unstable and unreliable.

*C. **Frequent Itemset-based Clustering***

Frequent item set-based clustering [4] [5] identifies frequent sets of keywords that often occur jointly in the document set. These sets are used as soft clusters because they share one or more keywords which are supposed to be related. The main advantage of this method is that the clusters are labelled by keywords shared by the documents in which it occurs. The limitation [6] is dependency of overlapping between keywords assigned to documents to form relevant clusters and also works on limited number of keywords assigned per document.

*D. **Seeded Clustering***

Seeded approach [8] [10] is a type of semi-supervised learning method. If both labeled and unlabeled data is used, it is termed as semi-supervised learning. It is also known as transductive learning. The label or prior information which is used in transductive methods will yield good clusters. These approaches are used only in the starting stages of the clustering algorithms. Hence, according to the algorithm, the nature of the seeded method may vary. That is the initialization step of the algorithms may change with respect to the label provided by the user. Due to the use of these methods in initial stages, the quality of the clustering results will be improved. Many of the algorithms use seeded methods to enhance their output. Two among them are as follows,

1) *Seeded Approach in K-Means*

In K-means algorithm, seeded approach [9] is used to initialize the cluster at the first step or in first iteration. Seeded method is used by Basu et al. to initialize the K-Means algorithm instead of starting K-Means with K random means. If seeded method is used in K-means, the resulting cluster quality is high.

2) *Seeded Approach in Affinity Propagation*

In affinity propagation algorithm, seeded clustering [9] [10] is combined with affinity propagation (AP) to increase the convergence rate of AP algorithm and also enhance the clustering outcome. The number of iterations which are used in original AP algorithm [22] is reduced due to the use of seeded clustering. In this method, asymmetric similarity measurement is used to capture the structural information of texts. Semi supervised clustering algorithm aimed to address the complexity problem in text clustering which results from the high dimension and sparse matrix computations. The main features of this algorithm are tri-set computation, similarity computation, seeds construction and messages transmission.

We exploit the knowledge from a large number of unlabeled object and a few labeled objects using this method. We use the labeled objects to construct efficient initial “seeds” for our [22] affinity propagation clustering algorithm. This method finds the representative features quickly in labeled objects. The seeds are made up of these features and their values in different clusters. They should be more representative than normal objects. The seeds will be chosen as exemplars and it helps to get exact cluster numbers. Seed affinity propagation reduces the computing complexity and improves accuracy.

E. ***Ontology based Clustering***

Ontology [12] [13] is a collection of concepts and their interrelationships which can collectively provide an abstract view of an application domain [11]. In other words it is a collection of terms, attributes and their relationships. The relationship may be ‘is-a’ or ‘has-a’ type relationship. Recently this method is used in most of the document clustering techniques. High quality ontologies are crucial for many applications, and their construction, integration, and evolution greatly depends on the availability of a well-defined semantics and powerful reasoning tools. There are many strategies available for compiling ontology into the text representations, focusing on concepts, disambiguation, and hyponyms.

The following are the important terms which are used in the ontology approach

- *Holonym*: It defines the relationship [7] between a term denoting a whole and the term denoting its part or member of a whole. It defines the relationship between the term as a whole and its part.

  Example: ‘Tree’ is a Holonym of ‘root’.

- *Meronym*: A term that denotes a part of the whole that is denoted by another term. A Meronym is a term which specifies a part or a member of something.

  Example: Word ‘shoulder’ is Meronym of word ‘body’.

- *Hyponym*: A hyponym is a term which denotes a more specific term or denotes a sub ordinate grouping word or phrase.

  Example: Lion is hyponym of animal.

- *Hypernym*: A term [7] whose meaning includes the meaning of other words or a super set of something.

  Example: animal is a Hypernym of dog.

F. ***Concept based Clustering***

The machine system [15] should interpret and understand the semantic or meaning of the concepts. In order to do this the semantic relationship between the concepts should be identified and defined. Concept based approaches [17] [20] are used to discover the topic of the document. In documents, one term is much more important when compared to other word or terms. Important and non-important terms are differentiated by analyzing the semantics of the sentences in the document in the concept based approaches.

G. ***Constraint based Clustering***

Constraint based Clustering [16] has either a set of must-link constraints, cannot-link constraints or both, with a clustering method. Both the constraints define a relationship between two data instances. A must-link constraint is used to specify that the two instances in the must-link relation should be associated with the same cluster. A cannot-link constraint [19] is used to specify that the two instances in the cannot-link relation should *not* be associated with the same cluster. The clustering method [18] uses these sets of constraints to find clusters in a data set that satisfy the specified must-link and cannot-link constraints. In some constrained clustering method if no such clustering exists for the specified constraints, it aborts. Others it will try to minimize the constraint violation should it be not able to find a clustering which satisfies the constraints.

III. PERFORMANCE EVALUATION

This section represents the estimation of the quality of above stated clustering techniques using few validity indices and datasets taken from UCI Machine Learning Repository.

A. *Precision*

Precision is the fraction of the documents retrieved that are relevant to the user's information need. Precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is calculated as given below,

$$Precision\ P(i,j) = \frac{N_{ij}}{N_j} \tag{1}$$

where N_{ij} is the total number of objects of class i in cluster j and N_j is the number of objects in cluster j .

B. *Recall*

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved. In binary classification, recall is often called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision. The recall measure can be calculated as follows,

$$Recall\ R(i,j) = \frac{N_{ij}}{N_i} \tag{2}$$

where N_{ij} is the total number of objects of class i in cluster j and N_i is the number of objects in class i .

C. *Fall-out*

The proportion of non-relevant documents that are retrieved, out of all non-relevant documents are available. In binary classification, fall-out is closely related to specificity and is equal to (1-specificity). It can be looked at as the probability that a non-relevant document is retrieved by the query. It is trivial to achieve fall-out of 0% by returning zero documents in response to any query. It can be measured as given below,

$$Fallout = \frac{n-\sum_{i=1}^{n} d_i}{L-R} \tag{3}$$

where n denotes the number of documents in the resulted output, L denotes the size of the dataset which includes the collection of documents, d_i represents the relevance level of the particular document in the output according to the given query, and R denotes the number of retrieved documents.

D. *Accuracy*

Accuracy is the term in which it measures the degree of proximity of a quantity to the quantity’s actual true label values. In other words it is defined as the number of exactly determined data objects of cluster results in contrast to the known true labels divided by the total number of instances in the dataset.

$$Accuracy = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{M_i}}{\mathrm{D}} \tag{4}$$

where D represents the total number of data objects in the dataset, and $\mathrm{M_i}$ illustrates the majority of the data objects points to exact true labels.

E. *Compactess*

Compactness is the measurement of average distance between every pair of data objects belonging to the same cluster. Specifically the members of every cluster should be as close as possible. Hence it is stated that the lower value of the compactness measure tends to be the better cluster configuration.

$$Compactness = \frac{1}{D} \sum_{k=1}^{k} n_k \left(\frac{\sum_{Xi,Xj \in Ck} d(Xi,Xj)}{n_k\ (n_k - 1)/2} \right) \tag{5}$$

where K denotes the number of clusters, n_k is the number of data objects in the cluster k, $d(Xi, Xj)$ is the distance between the data points $Xi\ and\ Xj,$ and D is the total number of instances in the dataset.

F. *F-Measure*

F-Measure is the term which denotes the harmonic combination of the precision and the recall values used in the information extraction. This F-Measure can be mainly used to examine the quality of the clustering solutions of some algorithms based on document clustering. The corresponding F-Measure of cluster j and class i is defined as,

$$F(i,j) = \frac{2*P(i,j)*R(i,j)}{P(i,j)+R(i,j)} \tag{6}$$

The global F-Measure can be calculated as given below,

$$F = \sum_i \frac{N_i}{N} \max_j \left(F(i,j)\right) \tag{7}$$

G. *Entropy*

Entropy is the measure of the quality, uniformity or the purity of the cluster. The smaller entropy measure illustrates the better performance. It can be measured as follows,

$$E_j = -\sum_i p_{ij} \log\,(p_{ij}) \tag{8}$$

where p_{ij} denotes the probability of an object belonging to a class i in cluster j. The global entropy for the clusters can be calculated as follows,

$$E = \sum_{j=1}^{m} \left(\frac{N_j}{N} * E_j\right) \tag{9}$$

where N is the total number of objects in the dataset, N_j is the number of objects in cluster j and m is the number of clusters.

IV. EXPERIMENTAL RESULTS

Based on the above stated validity measures the following Table I and Table II compares the performance and the accuracy levels of different clustering techniques over the several examined UCI [21] repositorydatasets.

TABLE I. COMPARATIVE ANALYSIS OF CLUSTERING TECHNIQUES USING VALIDITY MEASURES

Dataset	Clustering Techniques	Precision	Recall	Fall-out	Accuracy	Compactness
Iris	K-Means	0.98	0.89	0.85	0.92	0.71
	K-Medoids	0.91	0.71	0.67	0.89	0.78
	Agglomerative Approach	0.85	0.89	0.59	0.79	0.62
	Divisive Approach	0.67	0.86	0.85	0.86	0.61
Wine	K-Means	0.95	0.84	0.44	0.60	0.93
	K-Medoids	0.89	0.81	0.57	0.78	0.89
	Agglomerative Approach	0.67	0.87	0.48	0.67	0.81
	Divisive Approach	0.78	0.56	0.60	0.59	0.76
Soybean	K-Means	0.90	0.74	0.73	0.87	0.74
	K-Medoids	0.82	0.58	0.67	0.61	0.61
	Agglomerative Approach	0.87	0.61	0.77	0.65	0.44
	Divisive Approach	0.67	0.63	0.84	0.78	0.65
Breast Cancer	K-Means	0.78	0.85	0.88	0.80	0.94
	K-Medoids	0.64	0.67	0.74	0.86	0.86
	Agglomerative Approach	0.68	0.69	0.45	0.89	0.85
	Divisive Approach	0.69	0.77	0.64	0.82	0.81

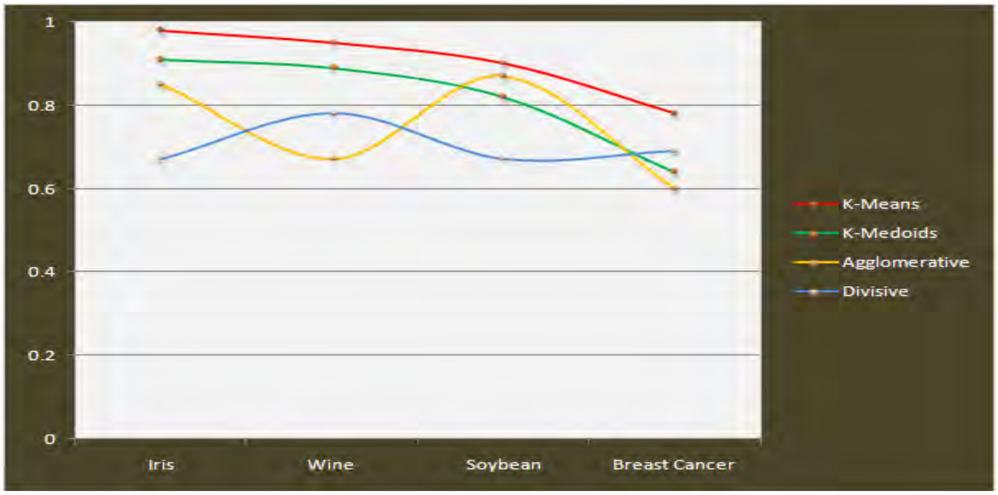

Figure 1. Precision Comparison

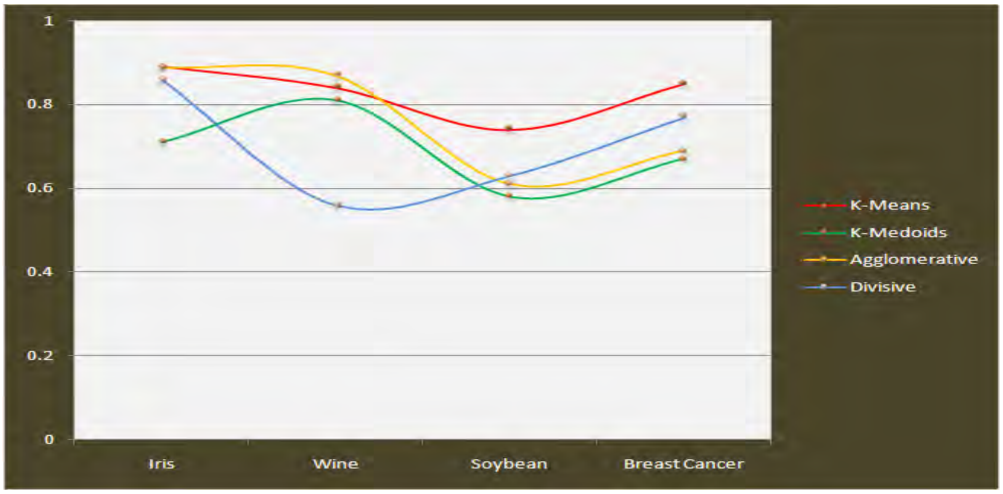

Figure 2. Recall Comparison

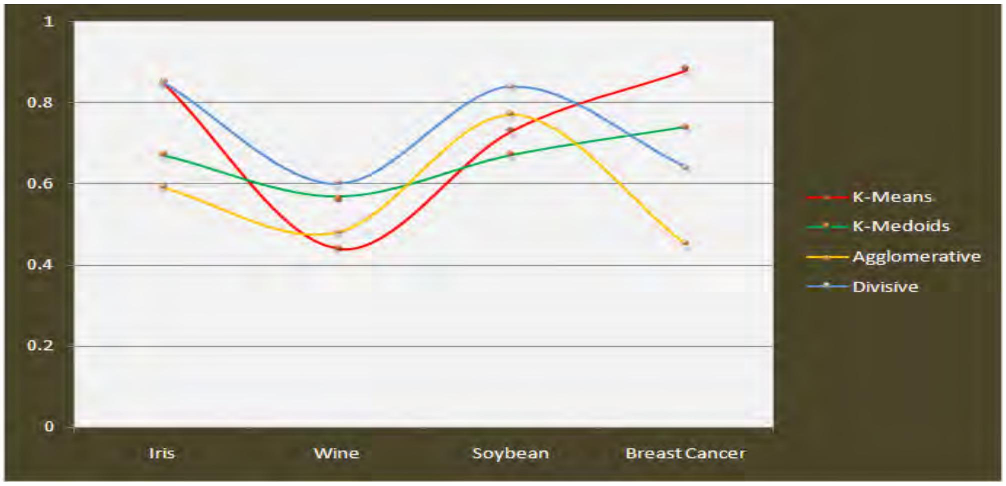

Figure 3. Fall-out Comparison

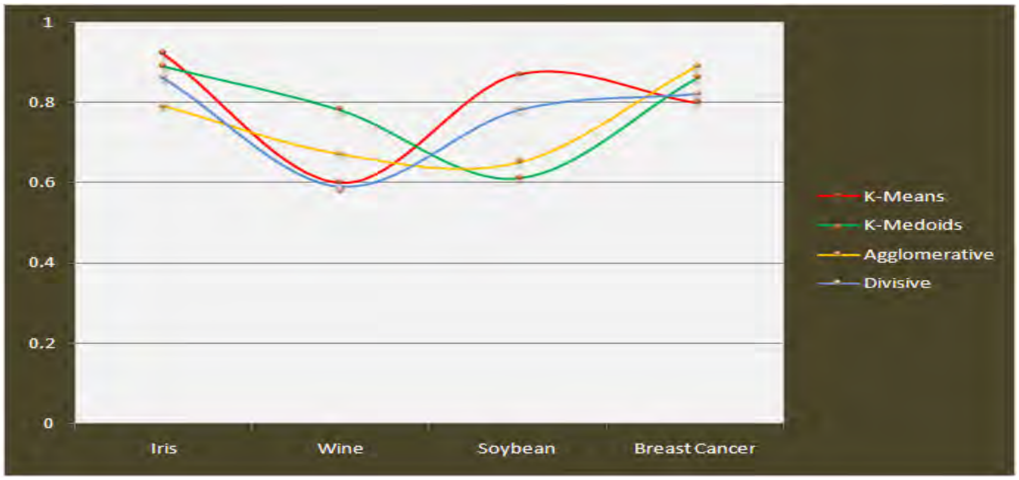

Figure 4. Accuracy Comparison

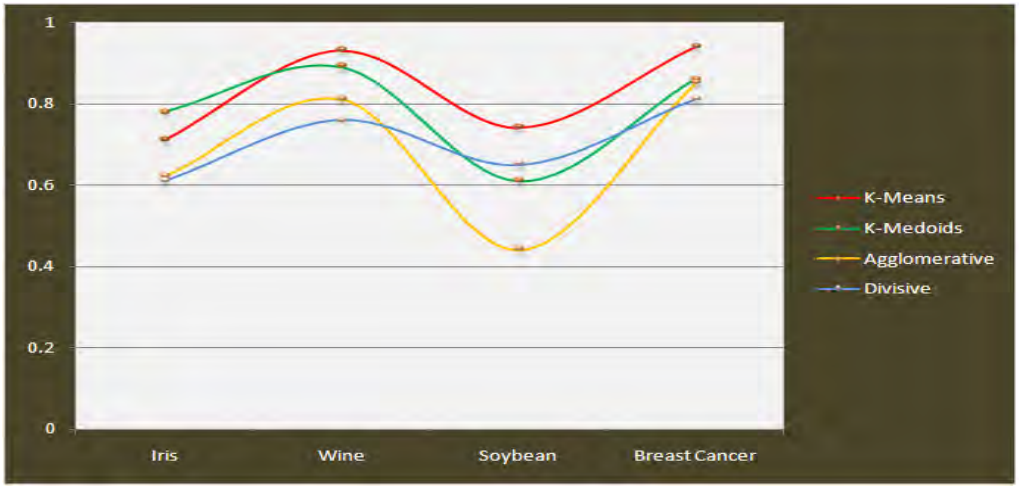

Figure 5. Compactness Comparison

TABLE II COMPARATIVE ANALYSIS OF CLUSTERING TECHNIQUES USING TEXT DATASET

Dataset	Clustering Techniques	Precision	Recall	F-Measure	Entropy
Reuter_50_50	Frequent Item-set Clustering	0.64	0.57	0.62	0.49
	Seeded Clustering	0.91	0.67	0.77	0.51
	Ontology based Clustering	0.68	0.87	0.73	0.62
	Concept based Clustering	0.83	0.88	0.85	0.44
	Constraint based Clustering	0.77	0.89	0.82	0.48
20Newsgroup	Frequent Item-set Clustering	0.59	0.71	0.56	0.38
	Seeded Clustering	0.85	0.96	0.90	0.49
	Ontology based Clustering	0.79	0.88	0.83	0.60
	Concept based Clustering	0.57	0.67	0.61	0.45
	Constraint based Clustering	0.78	0.88	0.82	0.57

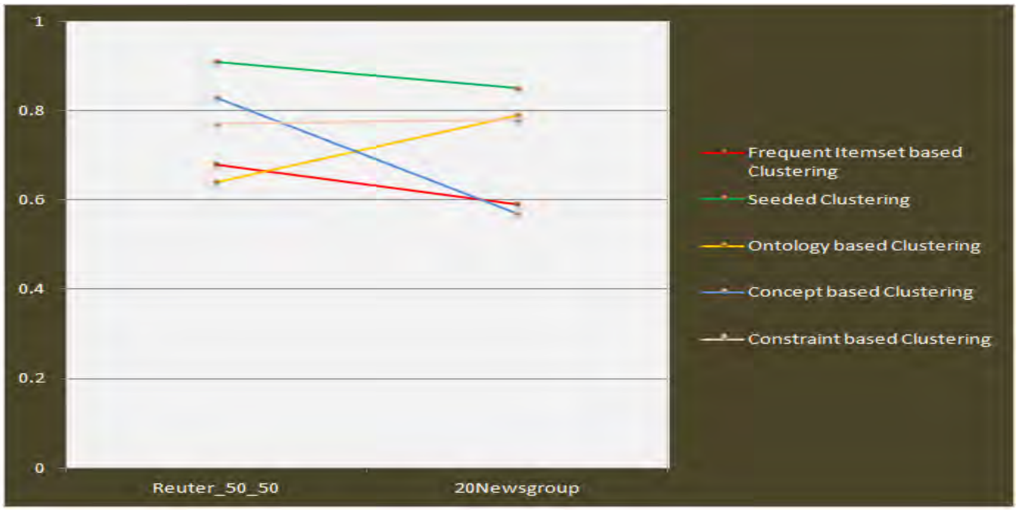

Figure 6. Precision Comparison on Text dataset

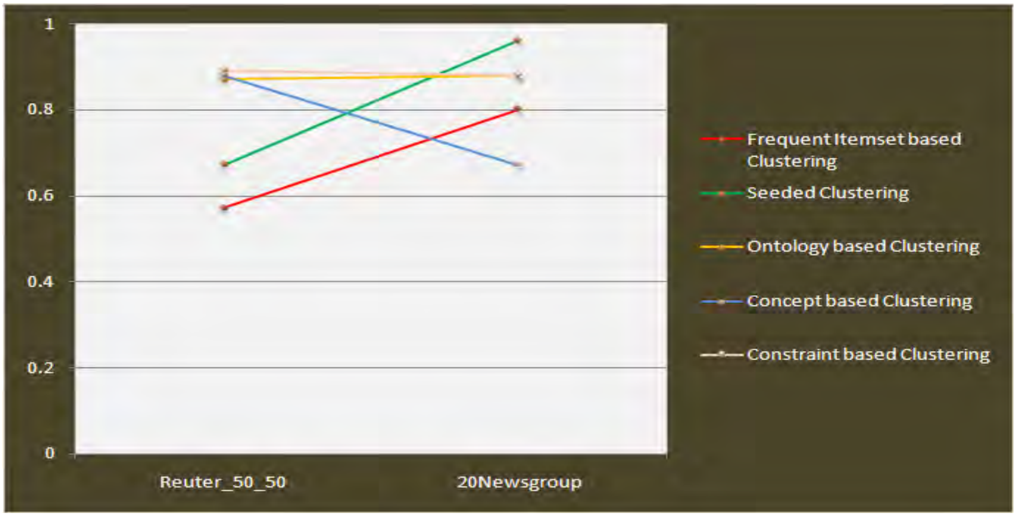

Figure 7. Recall Comparison on Text dataset

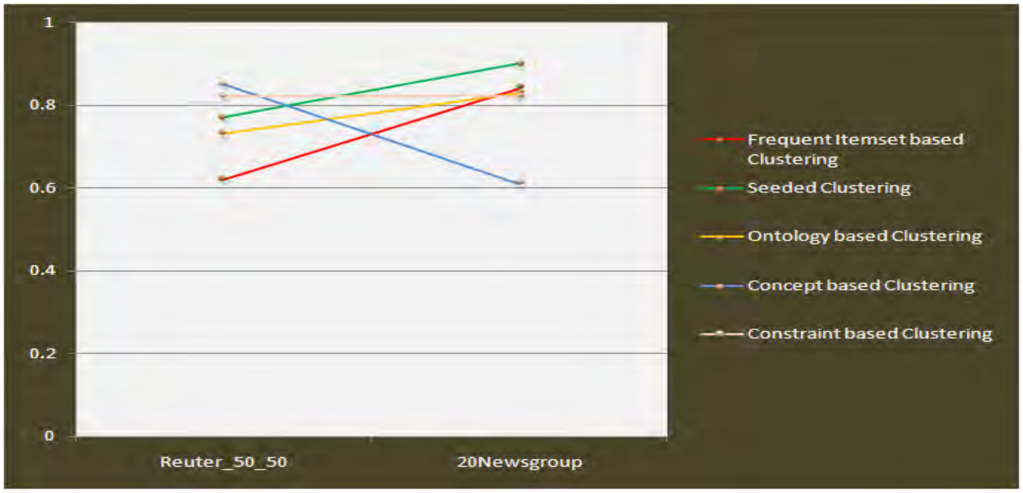

Figure 8. F-Measure Comparison on Text dataset

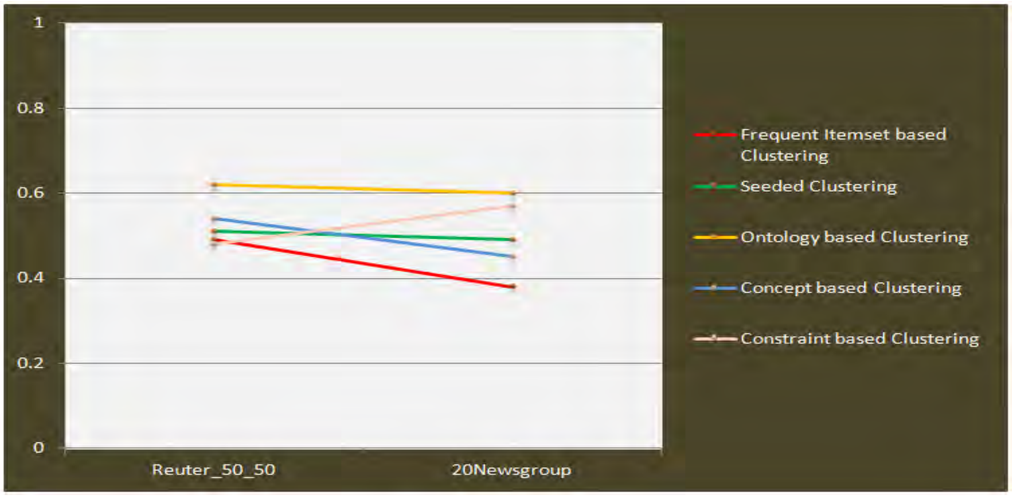

Figure 9. Entropy Comparison on Text dataset

V. CONCLUSION

Thus the process of Clustering tends to be the most supporting task for efficient and speedy information retrieval, and also in extracting accurate results form large dimensional data. Due to the enormous means of information embedded in huge data warehouses maintained in several domains, the usage of clustering technique has been a mandatory task for grouping relevant information in a single cluster and irrelevant data in other groups. This requirement paves the way for evolving various methods of clustering technique. Furthermore, the community of data mining puts a lot of efforts on developing fast and time consuming clustering algorithms for grouping both numerical and text data in very large datasets. Hence in this survey some of the clustering methods and its working process along with the features are highly elucidated. Moreover in order to examine the quality of each clustering methods, experiments are performed on few UCI repository datasets. The empirical result shows that some of the techniques need to improve the accuracy levels. This analysis really makes better understanding for the readers as well as it helps the clustering researchers to invent more clustering algorithms and also improves the quality of existing methods in future.

REFERENCES

[1] Li Xinwu, “Research on Text Clustering Algorithm Based on Improved K_means”, 2009 ETP International Conference on Future Computer and Communication

[2] M. Steinbach, G. Karypis, and V. Kumar, “A Comparison of Document Clustering Techniques,” Proc. Knowledge Discovery and Data Mining (KDD) Workshop Text Mining, Aug. 2000.

[3] Yuqin Li , Xueqiang Lv , Yufang Liu , Shuicai Shi , “Research on Text Clustering Based on Concept Weight”, 2010 Fourth International Conference on Genetic and Evolutionary Computing.

[4] Rekha Baghel, Dr. Renu Dhir, “A Frequent Concepts Based Document Clustering Algorithm” , International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010.

[5] Florian Beil, Martin Ester, Xiaowei Xu, “Frequent Term-Based Text Clustering”, 2002 ACM 1-58113-567-X/02/0007.

[6] Peter D. Turney , Patrick Pantel, “From Frequency to Meaning: Vector Space Models of Semantics”,Journal of Arti_cial Intelligence Research 37 (2010) 141-188 Submitted 10/09; published 02/10.

[7] Prajna Bodapati, Shashi Mogalla, “Document Clustering Technique based on Noun Hypernyms”, IJECT Vol. 2, SP-1, Dec . 2011.

[8] Jie Ji, Tony Y. T. Chan, Qiangfu Zhao, “Fast Document Clustering Based on Weighted Comparative Advantage”, Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009

[9] M. Junker, M. Sintek, and M. Rinck, “Learning for Text Categorization and Information Extraction with ILP,” Proc. First Workshop Learning Language in Logic, 1999.

[10] Taewon Lee, Bongki Moon, Sukho Lee “Bulk insertion for R-trees by seeded clustering”, Data and Knowledge Engineering Elsevier September 2005 86-106.

[11] S.C. Punitha, K. Mugunthadevi, M. Punithavalli, “Impact of Ontology based Approach on Document Clustering”, International Journal of Computer Applications (0975 – 8887) Volume 22– No.2, May 2011.

[12] Thangamani.M, Dr. Thangaraj.P ,”Ontology Based Fuzzy Document Clustering Scheme”, Modern Applied Science, Vol. 4, No. 7; July 2010 148 ISSN.

[13] J. Jayabharathy, S. Kanmani and A. Ayeshaa Parveen’ “Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature”, 978-1-61284-486-2/11/$26.00 ©2011 IEEE.

[14] Han J, Kambr M. “Data Mining: Concepts and Techniques”. Hand Book. Beijing: Higher Education Press, 2001.

[15] Nicholas O. Andrews, Edward A. Fox, “Recent Developments in Document Clustering”, October 16, 2007.

[16] H. Jin, M.-L. Wong, and K.S. Leung, “Scalable Model-Based Clustering for Large Databases Based on Data Summarization,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11,pp. 1710-1719, Nov. 2005.

[17] S. Shehata, F. Karray, and M. Kamel, “An Efficient Concept-Based Mining Model for Enhancing Text Clustering,” IEEE Transactions on data and knowledge engineering,Vol. 22, No. 10, october 2010.

[18] Tatiane M. Nogueira, Heloisa A. Camargo and Solange O. Rezende, “Fuzzy Rules for Document Classification to Improve Information Retrieval”, International Journal of Computer Information Systems and Industrial Management Applications ISSN 2150-7988 Volume 3 (2011) pp. 210-217.

[19] Jiabin Deng, JuanLi Hu, Hehua Chi, Juebo Wu, “An Improved Fuzzy Clustering Method for Text Mining”, 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing.

[20] Amir Hamzah, Adhi Susanto, F.Soesianto, Jazi Eko Istyanto, “Concept based Text Document Clustering” Proceedings of International Conference on Electrical Engineering and Informatics, Indonesia June 17-19 2007.

[21] A. Asuncion and D.J. Newman, “UCI Machine Learning Repository,” School of Information and Computer Science, Univ. of California, http://www.ics.uci.edu/~mlearn/MLRepository. html, 2007.

[22] Renchu Gaun, Xiaohu Shi, Mauriozio Marchese, Chen Yang, and Yanchum Liang, “Text Clustering with Seeds Affinity Propagation” IEEE Transaction on Knowledge and Data Engineering Vol 23 No. 4 2011.