Document Clustering: A Review

Sunita Bisht
Research Scholar
Deptt. of Computer Science Engineering
Bipin Tripathi Kumaon Institute of Technology
Dwarahat, Uttarakhand, India

Amit Paul
Asst. Professor
Deptt. of Computer Science Engineering
Bipin Tripathi Kumaon Institute of Technology
Dwarahat, Uttarakhand, India

ABSTRACT: As the internet is exploding with huge volume of text documents, the need of grouping similar documents together for versatile applications have hold the attention of researchers in this area. Document clustering can facilitate the tasks of document organization and web browsing, search engine results, corpus summarization, documents classification, information retrieval and filtering. However several attempts have been made to develop efficient document clustering algorithms but most of the clustering methods suffer from challenges in dealing with problems of high dimensionality, scalability, accuracy and meaningful cluster labels. This paper intends to provide a brief summary over methods studied and current state of documents clustering research, including basic traditional methods as well as advanced fuzzy based, GA, PSO, HS oriented techniques etc. Also document representation model and its challenges, dimensionality reduction mechanisms, issues in document clustering, and cluster quality evaluation criteria are discussed.

Keywords: document clustering, hierarchical clustering, partitioning clustering, frequent item set, vector space model.

1. INTRODUCTION

Clustering is the unsupervised classification process that group data objects into classes or clusters such that objects within the same cluster are highly similar to one another while highly dissimilar to objects in other classes. Thus clustering facilitates an efficient visualization of the documents in a collection by grouping similar and relevant documents together in one group or cluster. Clustering should not be confused with classification since unlike classification no labelled documents are provided in clustering. The topic of clustering has been widely studied in many disciplines and has become an active area of research in the field of data mining.

The very first challenge in any clustering approach is to decide which features of a document are to be considered discriminatory, for which we need a document model. However majority of clustering approaches have used vector space model to represent each document as vector. Section 2 provides a brief introduction of this model as well as challenges faced by vector space model.

Document clustering is being studied from many decades but still there are certain issues that must be solved to get fast and efficient clustering. Section 3 introduces these challenges in document clustering. There is an overwhelming amount of literature on document clustering and exploring this vast literature is very complicated. Hence this paper makes no attempt to present all existing algorithm, but to a certain extent present main classes of algorithm. Section 4 presents taxonomy of clustering approaches and major techniques in use. Clustering techniques are formalized with the goal of attaining high intra-cluster similarity and low inter-cluster similarity. Section 5 describes some commonly used metrics to assess the quality of clustering. Finally section 6 presents some concluding remarks.

2. DOCUMENT REPRESENTATION

Bulk of clustering approaches has employed Vector Space Model for document representation. In vector model each document is conceptually represented as a vector of keywords extracted from documents with associated weight representing the importance of key term in document as well as in whole document corpus. It was proposed as a model for Information Retrieval (IR) by Salton, Wong, Yang (1975) and was used first time in the System for the Mechanical Analysis and Retrieval of Text (SMART) information retrieval system (Salton, 1971b) (on which see Dubin (2004)) [1]. Under vector model, a collection of n documents with m unique words is represented as an m×n matrix, where each document is a vector of m dimension.

Several term-weighting schemes have been used such as binary term weighting, simple term frequency, tf-idf term weighting etc. Among them the most popularly used scheme is tf-idf term weighting scheme (Salton and McGill, 1986), which assigns a high weight to a term if it occurs frequently in document but rarely in whole document collection. As per the scheme, weight of a term composed of frequency of term in document multiplied by inverse frequency of term in whole corpus. Precisely the weight of a term j, in document i, with N, no of documents in collection is given as:

$$W_{ij} = tf_{ij} * idf_{ij} = tf_{ij} * \log\frac{N}{df_j}$$

This scheme is based on the assumption that word which occur frequently in document but rarely in entire collection are of highly discriminating power. Under all schemes, it is usual to normalize document vectors to unit length [2]. In order to perform clustering, similarity between documents must be measured. There exist a no of possible measures for computing the similarity between documents, but the most prominent measure is cosine similarity.

Given two document vectors $\underset{t_a}{\rightarrow}$ and $\underset{t_b}{\rightarrow}$, their cosine similarity is:

$$cosine\left(\underset{t_a}{\rightarrow},\underset{t_b}{\rightarrow}\right) = \frac{\left(\underset{t_a}{\rightarrow}.\underset{t_b}{\rightarrow}\right)}{\left(\left|\underset{t_a}{\rightarrow}\right| * \left|\underset{t_b}{\rightarrow}\right|\right)}$$

However the model itself is well established, but vector representation of document under traditional vector space model suffers from certain challenges.
First, in a collection of heterogeneous topics no of unique terms will be quite large, hence results in high dimensional document vector. This is also called as curse of dimensionality. Various document pre processing steps are performed to tackle this issue.
Second, the term-by-document matrix resulting from corpus under traditional vector model will be highly sparse. To address this issue various dimensionality reduction techniques have been used.
Third, in vector space terms represent the dimension and are considered linearly independent ie. their relation to each other is not taken into account. Since no word order information is encoded, traditional vector model is also known as *bag of words* model.
Fourth, in traditional vector space model the similarity calculation is based on word overlap measure. Thus two documents with similar topics but different vocabulary will not be considered relevant, only documents sharing vocabulary will be considered similar. Several approaches have been formalised to address these issues such as LSA, PCA, and SVD. Also a lexical database WORDNET has been used to overcome the problem of synonym and hypernym. A brief overview of these approaches is discussed in next section.

2.1 Pre Processing

Pre processing comprises of some steps that take a plain text document as input and returns a set of tokens as output. These steps typically consist of:

Filtering- It is the process of removing special characters and punctuation marks from input documents that are not significant enough to hold any discriminative power under vector model.

Tokenization- This step splits the sentences into tokens, usually words.

Stemming- It is the process of reducing words to their base word or stem. For example words "relating", "related", "relations" all are converted to their root word "relate". Porter's algorithm [3] is the de facto stemming algorithm.

Stopword Removal- Stopwords are mostly occurred insignificant word, which do not convey any meaning as a dimension in vector model. Thus it is desirable to remove these words from the list forming set of unique words. A most common strategy to remove stopwords is to compare each term with a list of known stopwords.

Pruning- It is the process of removing words that appear with very low frequency in the collection. Because, even if they had any discriminating power, result too small clusters to be useful. In general a pre specified threshold is used for this purpose.

2.2 Dimensionality Reduction

Although pre processing results in achieving a significant reduction in vector space, but for efficient clustering we need further diminution in dimensional space. There has been recent interest in producing an optimal low rank approximation of the term-by-document matrix and generating an optimal clustering of the dataset. With a given term-by-matrix **V**, the goal of any dimensionality reduction technique is to produce a rank **k** approximation of **V**, $\mathbf{V_k}$ with manageable error. A universal measure to assess the quality of this approximation is Frobenius Norm, which is defined as:

$$|V - V_k| = \sqrt{\sum_{v \in V} \sum_{v_k \in V_k} (v - v_k)^2}$$

Smaller the value of Frobenius norm, better the matrix $\mathbf{V_k}$ approximates to original matrix **V**. This section describes two matrix factorization techniques for the purpose of dimension reduction.

2.2.1 NMF (Non Negative Matrix Factorization)

In traditional vector space model each vector component is given a positive weight, if it is present in document or zero value otherwise, thus resultant term-by-document matrix always has positive entries. This characteristic of non negativity is preserved in NMF, which makes it different from other rank reduction techniques in data mining, eg. PCA. NMF is the process of finding a low rank approximation of term-by-document matrix V, by factorizing V into the product (WH) of two reduced matrices W & H.

$$V_{mn} = W_{mk} . H_{kn}$$

Each column of W is a basis vector ie. it contains an encoding of the concepts from V and each column of H contains an encoding of linear combination of basis vectors that approximate corresponding column of V. The dimensions of W and H are m×k and k×n respectively. Here k is reduced rank or the selected no of features.
However the appropriate value of k depends on application and is also influenced by nature of collection itself [4]. For document clustering k is no of features to be extracted or no of clusters required. In NMF, each k dimensional column vector in H corresponds to a document and actual clustering procedure is performed using these reduced document vectors.
Usually H is initialized to 0, and W contains some randomly generated values, with each $W_{ij}>0$. These initial estimates are improved by subsequent iteration of algorithm. Some of the NMF algorithms used are, multiplicative update algorithm by Lee and Seung [5], sparse encoding by Hoyer [6], gradient descent with constrained least squares by Pauca [7] and alternating least squares algorithm by Pattero [8].

2.2.2 SVD (Singular Value Decomposition)

SVD is a dimension reduction method that takes a high dimensional and highly variable set of data points as input and reduces it to a lower dimensional space rendering substructure of data more clearly. SVD is based on the theorems of linear algebra. The purpose of SVD is to decompose a rectangular matrix into the product of three matrices- an orthogonal matrix U, a diagonal matrix S, and a transpose matrix V, given as:

$$A_{mn} = U_{mm} S_{mn} V_{nn}^T$$

Where $U^T U = I$

$V^T V = I$

Column of U are orthonormal eigen vectors of AA^T

Column of V are orthonormal eigen vectors of $A^T A$

S is a diagonal matrix of singular values which are square roots of common eigen values from U and V in descending order [9].

SVD finds the features or concepts and also the relations between terms in the term-by-document matrix A and order them by weight. Only first k singular values in S and corresponding vectors from U and V are used for further processing. Here the selection of value of k is very crucial to the performance of clustering, because if k is very large, the hidden semantic structure cannot be exposed since documents and words are not projected near to each other. Also if k is very small, too many words will be superimposed on one another and again demolish the semantic relation.

After decomposition these three sub matrices represent the reduced semantic space, where now documents are represented as row vectors in V and words are represented as row vectors in U. Note that documents are represented as row vectors in V because we are considering V instead of V^T. Also the document similarity can be achieved by comparing rows in matrix VS and word similarity is measured by comparing row similarity in matrix US.

After SVD, the resultant vector contains component ranked from most significant to least significant. Also deletion of elements which do not exhibit meaningful implication in dimensional representation, word vectors can be represented more effectively. Now the word vectors are shorter and contain only elements that accounts for most important correlation among words in original matrix.

This mechanism offers us two benefits. First SVD algorithm collapse down the high dimensional vector space into an optimal approximation while try to preserve as much information as possible about relative distance between document vectors. Second it helps to expose the latent semantic relation between text units in original term-by-document matrix.

2.3 Using Wordnet

Most of the existing text clustering techniques only relates documents that uses identical terminology while ignore the conceptual similarity as a result of which two documents even depicting same topic but using different vocabulary cannot be considered similar. Semantic relation between terms are not taken into account i.e. Two terms with a close semantic relation and two terms with no semantic relation, both are treated in the same way. Use of Wordnet facilitates a clustering algorithm to take into consideration the semantic relation between terms. Wordnet is a large electronic database of English, developed by Miller et al. It is not just an alphabetical list of words with their meaning. In Wordnet nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms called synset, each expresses a different concept. These synsets are interlinked by means of some conceptual- semantic and lexical relations. Wordnet includes following semantic relations: synonymy, antonymy, hyponymy, meronymy, troponymy, and entailment. Methods available to compute the semantic similarity between terms using Wordnet can be categorized into four different classes- Path based, Information content based, Gloss based, Vector based.

In Path based methods length of the path between concepts is used to measure similarity. In Information content based methods information content of most specific shared parents is used to compute the relatedness of two concepts. In Gloss based methods, glosses of concepts are used to verify relatedness of concepts. In Vector based methods vectors derived from gloss are used to compute relatedness between terms [10].

3. CHALLENGES IN DOCUMENT CLUSTERING

However in literature a lot of approaches have been proposed for clustering text documents but still algorithms lacks in satisfying some specific features or challenges that need to be tackled to acquire high quality clustering solution.

1) - High Dimensionality-

In document vector each term can be regarded as a dimension and there are lots of terms in a document. Clustering algorithms can handle the dimensional space efficiently for small data set but it becomes complicated for large data collection. Selection of appropriate features for document representation can help, but it is challenging to extract the most valuable feature set.

2) - Scalability-

Many clustering techniques perform well on small data set but fail to achieve efficiency for large data collection. High scalable clustering algorithms are desired to resolve this issue.

3) - Knowledge Of Input Parameters-

Many clustering algorithms require user to provide certain information before clustering for example- no of clusters. Accuracy of clustering results may be sensitive to such input parameters, but identifying exact value of such parameters before execution is pretty difficult. Hence this may degrade clustering quality.

4) - Meaningful Cluster Labels-

Good cluster labelling can guide user in process of browsing the clustering by providing a brief but meaningful cluster description. Thus clustering methods should provide proper labels to clusters that must be understandable also to non experts.

5) - Accuracy-

It is the most desirable feature. The outcome of clustering procedure must be accurate despite of the complexities of algorithm. The resultant clusters should possess high clustering quality ie should reflect high intra-cluster similarity and low inter-cluster similarity.

4. CLUSTERING TECHNIQUES

Clustering has become an increasingly important topic in the field of data mining and information retrieval with the explosion of information available over internet. Organization of this vast amount of data into related groups or clusters can aid users in accessing increasingly large volume of data more effectively. Overview of a range of techniques is provided by Zamir et al (1997), Willet (1998), and Jain et al (1999).

According to [11], different clustering approaches available can be broadly classified into two classes- Hard (disjoint) and Soft (overlapping) clustering. Hard clustering algorithm does a hard assignment by assigning each document to exactly one cluster hence produces a set of disjoint clusters. Soft clustering algorithms perform a soft assignment means a document may appear in more than one cluster thus generate a set of overlapping clusters. Soft clustering algorithms are further classified as Hierarchical, Partitioning and Frequent-Itemset based clustering. Hierarchical clustering procedure organizes a collection of data objects into a tree of clusters known as dendrogram. These methods can be further categorized as Agglomerative hierarchical clustering and Divisive hierarchical clustering methods depending on whether the hierarchical decomposition is made in bottom up or top down fashion respectively. Partitioning clustering procedure creates a flat, non-hierarchical clustering solution. K-means and its variants are the most well known methods in this category. K-means algorithm iteratively refines initially chosen random set of k-initial centroids while minimizing the average distance of documents to their closed centroid. Frequent-Itemset based clustering procedures use frequent itemset generated by association rule mining for clustering the documents [11].

4.1 Hierarchical Clustering Techniques

Agglomerative hierarchical clustering is most popular approach in hierarchical clustering methods.
Algorithms of this family construct the hierarchy in bottom-up fashion by repetitively computing similarity between pairs of clusters and then merging the most similar pair. The steps of basic agglomerative hierarchical clustering approach are:

HIERARCHICAL AGGLOMERATIVE CLUSTERING ALGORITHM

1. Construct the similarity matrix containing the similarity between each pair of documents.
2. Initially consider each document as a cluster.
3. Find the most similar pair of cluster using similarity matrix. Merge these two clusters into one cluster and update the similarity matrix to reflect this change.
4. If anyone of the following condition satisfies then stop
   Otherwise go to step 3.
   - 4.1 All documents are in one cluster.
   - 4.2 Specified level of hierarchy is reached.

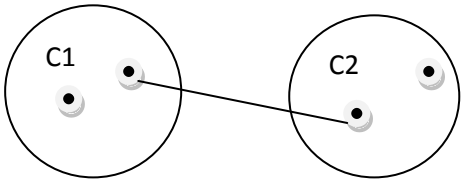
Different variations of agglomerative approach exist depending on similarity measure scheme employed. Three most commonly used methods to measure inter-cluster similarity are:

1) - Single linkage (Nearest neighbour) scheme

In this scheme proximity of two clusters is defined as the minimum of distance between any two data points across two clusters.

$$Sim(C_i, C_j) = min_{x \in C_i, x' \in C_j} dist(x, x')$$

This scheme considers only that area where two clusters come nearest to each other, other more distant parts of cluster and cluster's overall structure are not taken into consideration. But this scheme has two limitations- first is assessment of cluster quality is based on only two most similar pair of data points in two clusters. Second is it suffers from *chaining* effect. Graphically it can be represented as-



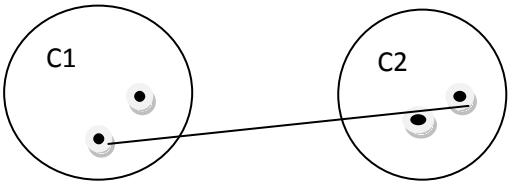
Single-linkage: Maximum similarity

2) - Complete linkage scheme-

In this scheme proximity of two clusters is defined as the maximum of distance between any two data points across two clusters.

$$Sim(C_i, C_j) = max_{x \in C_i, x' \in C_j} dist(x, x')$$

This scheme also suffers from some limitations- first is assessment of cluster quality is based on only two most dissimilar data points in two cluster. However this scheme removes the chaining effect of single linkage scheme, but has its own shortcoming. It plays too much attention to outliers ie. points that do not fit well into global structure of cluster. Graphically it can be represented as-



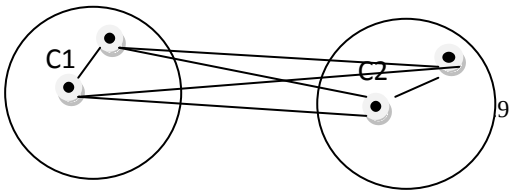
Complete-linkage: Minimum similarity

3) - Average linkage (UPGMA) scheme-

In this scheme proximity of two clusters is defined as the average of distance between all data points across two clusters.

$$Sim(C_i, C_j) = \frac{\sum x \in C_i, x' \in C_j dist(x, x')}{|C_i| . |C_j|}$$

The outcome of this scheme exist somewhere in between single-linkage and complete linkage. In order to judge the coherence of resulting merged cluster it considers all document-document similarity between two clusters hence evaluates cluster quality based on all similarity between documents, thus avoids pitfalls of single linkage and complete linkage [12]. Steinbach et al [13] showed that UPGMA scheme is most accurate one in agglomerative hierarchical clustering. Graphically it can be represented as-



Average-linkage: Average of all similarities

Many researchers have used this approach for clustering. In [14] Steve J et al proposed a *Phrase based hierarchical clustering approach* which clusters key phrases with which documents can later be associated rather than the most common way of clustering documents which are later labelled with key phrases

4.2 Partitioning Clustering Techniques

A partitioning clustering algorithm creates a single partition or one-level partitioning of data objects rather than generating a hierarchical clustering structure as done by hierarchical clustering technique. It is profitable to use partitioning method for large data set because constructing dendrogram for large document collection is computationally difficult. At the same time one major problem associated with partitioning method is the need to specify desired no of clusters at the start of execution. An incorrect estimation of this input parameter may result in poor clustering accuracy.
K-means and its variants are the most well known methods in this category.

K-MEANS CLUSTERING ALGORITHM

K-means (McQueen 1967) is the simplest and most commonly used flat clustering algorithm. The steps of basic k-means algorithm are-

ALGORITHM

1. Randomly select k- clusters as initial cluster centres.
2. Assign each document to their closest cluster.
3. Recomputed cluster centres based on new members of cluster.
4. Repeat step 3) until convergence criteria is met.

Typical convergence criteria could be fixed no of iterations; decrease in RSS falls below a threshold, assignment of documents to clusters doesn't changes between iterations.
A measure of how well the centre of cluster represents its members is RSS (Residual Sum of Squares) and the objective of K-means algorithm is to minimize RSS value. Cluster centres can be defined as mean or centroid of documents in cluster. K-means algorithm suffers from two major limitations- sensitivity to initial seed selection and convergence to local minimum if initial seeds are not properly chosen. In order to overcome these limitations certain variants of basic k-means have been proposed in recent years. Most of these variants use certain optimization technique for optimal clustering.
In [13], a simple and efficient variant of k-means, Bisecting K-means (BKM) is introduced. It is a divisive hierarchical clustering algorithm with linear time complexity. Initially the whole document set is considered one cluster. Then the algorithm recursively selects the largest cluster and uses basic k- mean algorithm to split it into two sub clusters until the desired no of clusters is reached. Bisecting k-means is proved to be superior to UPGMA and regular k-means. The better performance of bisecting K-means is because of production of relatively uniform size clusters. In [15] R. Kashef et al presented an enhanced version of BKM, Cooperative Bisecting K-means (CBKM) clustering algorithm, which concurrently combines the results of the BKM and KM at each level of the binary hierarchical tree using cooperative and merging matrices. Experimental results show that the CBKM attains better clustering quality than KM and BKM.
There have been many attempts to use GA (Genetic Algorithm) for clustering [16], [17], [18]. In [19] K Krishna et al presented a novel algorithm GKA (genetic k-means algorithm) for document clustering, which is the hybridization of computationally attractive and simple k-means with GA. In GKA author used the k-mean operator instead of crossover operator as used in conventional GA. Also they define a biased mutation operator specific to clustering. Thus GKA combines the simplicity of k-means algorithm with robustness of GA to produce globally optimal partition of given data into specified no of clusters. Author used the finite Markov chain theory to derive conditions on parameters of GKA for its convergence on globally optimal partitions.
In order to solve the problem of initial seed selection, Sung-Hyon et al in [20] proposed two algorithms GAIK (Genetic Algorithm Initializes K-means) and KIGA (K- means Initializes Genetic Algorithm). Via these two algorithms, author tries to show the feasibility of applying GA as an efficient initialization method for KM clustering technique. GAIK is a combination of K-means and GKA, where GKA is executed first to create initial values to K-means to start with, instead of choosing random values. This hybrid approach resulted in reducing the no of iterations required by K-means to converge to local minimum, whereas in KIGA, K-means is used first to initialize the GA clustering technique.
The Particle Swarm Optimization (PSO) algorithm which is a stochastic optimization technique can also be used to produce initial cluster centroids for k-means to overcome the problem of initial seed selection. In [21] author presented a novel document clustering algorithm based on PSO optimization technique with the objective to discover proper centroid of cluster for minimizing intra cluster distance and maximizing inter cluster distance. Contrary to localized search of k-means algorithm, PSO clustering algorithm performs globalized search. The PSO algorithm comprises of two stages: *global searching stage*, which guarantees that each particle searches widely enough to cover entire problem space and *local refining stage*, which ensures that each particle converges to optima. But it requires more no of iterations and computation than regular k- means. In [22] Xiaohui Cui et al proposed a hybrid Particle Swarm Optimization (PSO) +K-means document clustering algorithm which combines the ability of globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm and thus avoids the drawback of both algorithms. Algorithm consists of two modules: PSO module and K-Means module. At the initial stage PSO module is executed for a short period of time to discover the vicinity of optimal solution by global searching. Then K-Means module is applied for refining and generating final result. It uses the result from PSO module as initial seeds for k-means algorithm. In experimental evaluation hybrid PSO+K-means algorithm performed better than PSO and K-Means algorithm.
In [23] M. Mahdavi et al presented a novel document clustering approach Harmony K-Means clustering Algorithm (HKA) based on Harmony Search (HS) optimization method, which models the clustering problem as an optimization of an objective function. In the proposed approach HS algorithm is employed for global optimization and in order to improve the speed of convergence of HKA, KM algorithm is added to PAR process of HS for better fine tuning of algorithm. With the help

of finite Markov chain model it has been proved that HKA converges to local optimum. Also experimental results reveals that HKA algorithm provides better time complexity and partitioning accuracy.

4.3 Frequent Itemset Based Clustering Techniques

Most of the algorithms do not really address the special challenges of text clustering ie. High dimensionality, large databases, understandable description of clusters produced. This has motivated the development of new special document clustering approach which is not based on vector space model. This novel approach is based on the idea of frequent term set proposed by Agrawal et al. Frequent term sets are the set of terms co-occurring in more than a threshold percentage of all documents of a collection [24]. It provides a natural solution to reduce the high dimensional vector space. The key concept is to not consider the full high dimensional vector space but consider only low-dimensional frequent term sets to cluster the documents. A frequent term set do not represent a cluster but only describes a cluster, corresponding documents covered by the set forms the cluster.

In [25] Beil et al proposed two algorithms FTC and HFTC based on the notion of frequent terms sets. He made first attempt to use concept of frequent itemsets for document clustering. Algorithm FTC produces a *flat clustering* using a greedy approach with the goal of minimizing cluster overlapping. It starts with an empty set and continues to select one frequent term-set from remaining frequent term sets until the whole document collection is contained in the cover of chosen frequent term sets. In each step FTC makes a greedy choice to select the candidate frequent term set with cover having minimum overlap with already selected cluster candidates. Whereas algorithm HFTC which is based on FTC produces a hierarchical clustering by applying some condition on how to make selection of frequent term sets at each step.

In [24] Fung, Wang and Ester showed that HFTC is not scalable for large document collection and proposed hierarchical clustering scheme FIHC using frequent itemset. They used the notion of frequent itemset to construct hierarchical tree structure for clustering. Proposed algorithm carried out the construction of cluster in two phases.

First phase is the *Constructing Initial Cluster*, during which an initial cluster is formed for each *global frequent itemset* that includes all the documents that contain the itemset. Author used the algorithm proposed by Agrawal et al [26] for finding global frequent itemset. Second phase is *Making Clusters Disjoint.* In order to make each cluster disjoint, this phase identifies for each document the best initial cluster using following score measure-

$$Score(C_i \leftarrow doc_j) = \left[\sum_{x} n(x) * cluster\_support(x)\right] - \left[\sum_{x'} n(x') * global\_support(x')\right] \quad (3)$$

A cluster $C_i$is considered good for a documentdoc_j, if there are many frequent global items in doc_j that is also cluster frequent forC_i. This idea is implemented through this function in which first term reward cluster C_i if a global frequent item x in doc_j is cluster frequent inC_i and the second term penalizes clusterC_i, if a global frequent item x' in doc_j is not cluster frequent inC_i.

After producing the set of clusters, a hierarchical tree of clusters is constructed based on similarity between clusters. This tree structure utilizes the parent-child relationship between clusters based on the global items shared between clusters of two levels. In case if tree contains too many clusters two tree pruning methods are also proposed i.e. Child pruning and sibling merging to efficiently shorten and narrow a tree by merging most similar clusters together. Experimental evaluation showed that this algorithm outperformed HFTC in terms of accuracy, efficiency and scalability.

In [27] C C Ling et al proposed an efficient HAC algorithm based on fuzzy frequent item sets which uses fuzzy association rule mining to discover fuzzy frequent item sets to improve clustering quality of FIHC. In the proposed work clustering proceeds in three phases. In first phase document pre processing is performed and key terms are extracted. Second phase employs fuzzy association rule mining to determine a set of fuzzy frequent item set which contains key terms extracted in previous phase and used as labels of candidate clusters. In the third and final phase documents are clustered into hierarchical tree structure based on candidate clusters.

In [11] C C Ling et al proposed an effective Fuzzy based Multi label Document Clustering (FMDC) approach that combines fuzzy association rule mining with an existing ontology Wordnet. The proposed approach comprises of four modules. First is *document analysis module* during which key terms are extracted from set of documents and terms satisfying a minimum threshold criteria are selected for next phase. Second is *term onto construction module*. This module utilizes Wordnet to enrich the document representation with hpernym to find semantically related documents. Third phase is *candidate clusters extraction module*. This phase accepts structured document term vectors generated in previous phase as input, applies fuzzy data mining algorithm to generate fuzzy frequent item sets and output a candidate cluster. Fourth phase is *overlapping clusters generation module*, which assigns each document to multiple clusters. This phase produces a Document-Cluster matrix (DCM) to represent the degree of importance of a document to a candidate cluster. They also showed that clustering accuracy of proposed algorithm is better than FIHC, k-means, Bisecting K-means, and UPGMA.

C C Ling [28] in 2011 proposed a more efficient document clustering approach F^2IDC (Fuzzy Frequent Itemset based Document Clustering) which was the extension of their previous work. This approach is based on the fuzzy association rule mining in conjunction with Wordnet for clustering text documents. F^2IDC framework has three stages. First is *document pre processing* stage during which a set of key terms are extracted and selected from each document in the corpus. Second is *document representation and enrichment* stage in which basic vector representation of document is enriched using Wordnet by adding generality of terms via corresponding hypernym of Wordnet. Each key term is linked up to the top 5 levels of hypernyms. Third is *document clustering stage* during which first a membership function and a fuzzy data mining algorithm is defined to create a set fuzzy frequent itemstes, then documents are assigned to candidate clusters.

5. EVALUATION OF CLUSTER QUALITY

For clustering, two broad measures are used to gauge the cluster quality or goodness. One type is *internal quality measure*, which allows us to compare different sets of clusters without having reference to external knowledge. Other type is *external quality measure*, which allows us to evaluate the working of clustering by comparing the groups produced by clustering method with known classes. As there are many different quality measure, the performance and relative ranking of various clustering approaches can vary significantly depending on which measure is used. This section introduces four external quality measures.

5.1 Entropy

Entropy of clustering indicates how various semantic classes are distributed within each cluster. Given a particular cluster S_r, of size n_r , the entropy of cluster is calculated as:

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

Where q= no of classes in dataset

n_r^i= no of documents of i^{th} class that are assigned to r^{th} cluster .

Now the entropy of entire clustering is calculated as the sum of individual cluster entropies weighted according to cluster size.

$$\text{Entropy} = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r)$$

Smaller entropy indicates better clustering solution. Hence resultant clustering is considered good if one cluster contains words from one single class. In such case the entropy of clustering solution is zero.

5.2 Purity

One of the ways of measuring the quality of a clustering solution is cluster purity. Purity of clustering is defined as:

$$Pu(S_r) = \frac{1}{n_r}\max_i n_r^i$$

It gives the fraction of overall cluster size that the largest class of words assigned to that cluster represent. Now purity of entire clustering solution is given as weighted sum of individual cluster purity.

$$\text{Purity} = \sum_{r=1}^{k}\frac{n_r}{n}Pu(S_r)$$

In general, larger the purity value, better the clustering solution.

5.3 Random Index

During clustering we intend to assign two documents to the same cluster if and only if they are similar. A true positive (TP) decision assigns two similar documents to the same cluster; a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors we can commit. A (FP) decision assigns two dissimilar documents to the same cluster. A (FN) decision assigns two similar documents to different clusters. Random Index measures the accuracy of clustering result in terms of percentage of decision that is correct.

This notion can be made clearer with the help of following contingency table:

	Relevant	Non relevant
Retrieved	True positive(TP)	False positive(FP)
Not retrieved	False negative(FN)	True negative(TN)

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

5.4 F-Measure

Precision and *Recall* are the two basic and most frequent measures for assessing effectiveness of information retrieval. *Precision (P)* is the fraction of retrieved documents that are relevant and *Recall (R)* is the fraction of relevant documents that are retrieved. There is an inverse relationship between the two quantities. Recall is a non decreasing function of no of documents retrieved, while precision usually decreases as no of documents retrieved increases. But in general we need both, some amount of recall with only a tolerable percentage of false positive. A single measure that trades off precision versus recall is F-measure which is the weighted harmonic mean of precision and recall.

$$F = \frac{(\beta^2+1)PR}{\beta^2 P+R} \quad \text{where} \quad \beta^2 = \frac{1-\propto}{\propto}$$

When β = 1, it is called balanced F-measure since it equally weights precision and recall. It is given as:

$$F_{\beta=1} = \frac{2PR}{P+R}$$

However even weighting is not desired in most information retrieval, so values of β < 1 is used to emphasize precision, while values of β > 1 is used to emphasize recall [12].

6. CONCLUSION

Document clustering is a fundamental and crucial operation in various applications such as document organization, corpus summarization, information retrieval and filtering, automatic topic extraction. This study demonstrates the methods of well known categories of document clustering ie. Hierarchical, Partitioning and Frequent-itemsets based. We have tried to provide an exhaustive overview of various document clustering methods studied and researched since last few years, including basic traditional methods as well as advanced fuzzy based, GA, PSO, HS oriented techniques etc. Also we have explained Vector Space Model for document representation and challenges faced by it, dimensionality reduction mechanisms, cluster quality measures etc. The significance of document clustering approaches will continue to grow with rapid growth of information available online. Hence exploiting an effective and efficient method in text document clustering would be an essential direction for research in text clustering.

7. REFERENCES

[1] Johanna Geiß. July 2011. Latent semantic clustering for multi-documents summarization. UCAM-CL-TR-802 ISSN 1476-2986.

[2] Nicholas O. Andrews and Edward A. Fox. October 16, 2007. Recent Developments in Document Clustering. Technical Report TR-07-35, Computer Science, Virginia Tech.

[3] M. F. Porter. 1980. An algorithm for suffix stripping. Program, 14(3):130-137.

[4] Farial Shahnaz and Michael W. Berry. March 2006. Document Clustering Using NonNegative Matrix Factorization. Information Processing and Management: an International Journal, Volume 42 Issue 2, Pages 373-386.

[5] Lee, D & Seung. 2001. Algorithms for non-negative matrix factorization. In T. G. Dietterich and V. Tresp, editors, Advances in Neural Information Processing Systems, volume 13, Proceedings of the 2000 Conference: 556-562, The MIT Press.

[6] Hoyer, P. 2002. Non-Negative Sparse Coding. In Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, Martigny, Switzerland.

[7] Pauca, V, Shahnaz, F, Berry, MW & Plemmons R. April 22-24, 2004. Text Mining Using Non-Negative Matrix Factorizations. In Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, FL.

[8] Amy, L & Carl, M. 2006. ALS Algorithms Nonnegative Matrix Factorization Text Mining.

[9] Kirk Baker. March 29, 2005. Singular Value Decomposition Tutorial.

[10] Nguyen Chi Thanh and Koichi Yamada. September 2011. Document Representation and Clustering with Wordnet Based Similarity Rough Set Model. IJCSI Vol. 8, Issue 5, ISSN: 1694-0814.

[11] Chun-Ling Chen, Frank S.C. Tseng and Tyne Liang. September 2010. An integration of WordNet and fuzzy association rule mining for multi-label document clustering. Journal of Data & Knowledge Engineering, Volume 69, Issue 11, Pages 1208-1226.

[12] Chistopher D. Manning, Prabhakar Raghvan, Hinrich Schutze, April 1, 2009. Introduction to Information Retrieval. Cambridge University Press, online edition(c).

[13] Michael Steinbach, George Karypis, Vipin Kumar. 2000. A Comparison of Document Clustering Techniques. Proc. Of the 6th ACM SIGMOD int'l conf. on Knowledge Discovery and Data Mining (KDD).

[14] Steve Jones and Malika Mahoui. October 2000. Hierarchical Document Clustering Using Automatically Extracted Keyphrases. Computer Science Working Papers 00/13, University of Waikato, Department of Computer Science.

[15] R. Kashef and M.S.Kamel. Nov. 2009. Enhanced bisecting k-means clustering using intermediate cooperation. Journal of Pattern Recognition, vol. 42, issue 11, pp. 2557-2569.

[16] J. N. Bhuyan, V. V. Raghavan, and V. K. Elayavalli. 1991. Genetic algorithm for clustering with an ordered representation. In Proc. 4th Int. Conf. Genetic Algorithms. San Mateo, CA: Morgan Kaufman.

[17] D. R. Jones and M. A. Beltramo. 1991. Solving partitioning problems with genetic algorithms. In Proc. 4th Int. Conf. Genetic Algorithms. San Mateo, CA: Morgan Kaufman.

[18] G.P.Babu, Apr. 1994. Connectionist and evolutionary approaches for pattern clustering. Ph.D. dissertation, Dept. Comput. Sci. Automat., Indian Inst. Sci., Bangalore.

[19] K. Krishna and M. Narasimha Murty, June 1999. Genetic K-Means Algorithm. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 29, No. 3.

[20] Basher Al-Shboul, and Sung-Hyon Myaeng. 2009. Initializing K-Means using Genetic Algorithms. World Academy of Science, Engineering and Technology 54.

[21] Xiaohui Cui, Thomas E. Potok, Paul Palathingal. 2005. Document Clustering using Particle Swarm Optimization. IEEE Swarm Intelligence Symposium, (SIS).

[22] Xiaohui Cui and Thomas E. Potok, 2005. Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm. Journal of Computer Sciences (Special Issue): 27-33, ISSN 1549-3636.

[23] Mehrdad Mahdavi and Hassan Abolhassani, 2008. Harmony K-means algorithm for document clustering. In Data Mining and Knowledge Discovery (Springer) , 370-391.

[24] Benjamin C.M. Fung, Ke Wang, Martin Ester. 2003. Hierarchical Document Clustering Using Frequent Itemsets. In Proc. SIAM International Conference on Data Mining (SDM).

[25] Florian Beil Martin Ester Xiaowei Xu, 2002. Frequent Term-Based Text Clustering. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 436-442.

[26] R. Agrawal and R. Srikant, 12-15 1994. Fast algorithm for mining association rules. In J. B. Bo cca, M. Jarke, and C. Zaniolo, editors, Proc. 20th Int. Conf. Very Large Data Bases, VLDB , pages 487–499. Morgan Kaufmann.

[27] Chun-Ling Chen, Frank S.C. Tseng and Tyne Liang. 2010. Mining fuzzy frequent itemsets for hierarchical document clustering. International Journal of Information Processing and Management, 46, 193-211.

[28] Chun-Ling Chen, Frank S.C. Tseng and Tyne Liang. September 2011. An integration of fuzzy association rules and WordNet for document clustering. Journal of Knowledge and Information Systems - Special Issue on Data Warehousing and Knowledge Discovery from Sensors and Streams, Volume 28, Issue 3, Pages 687-708.