International Journal of Computer Trends and Technology- volume4Issue3- 2013

An NMF and Hierarchical Based Clustering Approach to support Multiviewpoint-Based Similarity Measure

K.S.Jeen Marseline[1], A.Premalatha[2]

[1]Asst. Prof and Head
Department of Information & Computer Technologies
Sri Krishna Arts and Science College
Coimbatore, Tamil Nadu

[2]Assistant Professor. Department of Computer Science
LNV College of Arts And Science
Coimbatore, Tamil Nadu

Abstract: In data mining, clustering technique is an interesting and important technique. The main goal of the clustering is finding the similarity between the data points or similarity between the data within intrinsic data structure and grouping them the data into single groups (or) subgroups in clustering process. The existing Systems is mainly used for finding the next frequent item set using greedy method, greedy algorithm can reduce the overlapping between the documents in the itemset. The documents will contain both the item set and some remaining item sets. The result of the clustering process is based on the order for choosing the item sets in the greedy approach; it doesn't follow a sequential order when selecting clusters. This problem will lead to gain less optimal solution for clustering method. To resolve this problem, proposed system which is developing a novel hierarchal algorithm for document clustering which produces superlative efficiency and performance which is mainly focusing on making use of cluster overlapping phenomenon to design cluster merging criteria. Hierarchical Agglomerative clustering establishes through the positions as individual clusters and, by the side of every step, combines the mainly similar or neighboring pair of clusters. This needs a definition of cluster similarity or distance. With this we are proposing the multiview point clustering approach with the NMF clustering method. The experimental results will be displayed based on the clustering result of three algorithms.

Key Words: Clustering, Multi-view point, Hierarchical clustering, Hierarchical Agglomerative clustering, Cosine similarity, Non-Negative Matrix Factorization.

1. INTRODUCTION

The future is aggravated by investigations as of the over and comparable examine conclusions. It materializes to the environment of match appraise the stage an extremely significant position in the achievement or failure of a clustering method. Our first purpose is to obtain a novel method for measuring connection among data objects in light and high-dimensional field, mainly text documents. As of the proposed similarity measure, then devise new clustering criterion functions and initiate their relevant clustering algorithms, which are quick and scalable like k-means, other than be also competent of as long as high-quality and reliable performance. It expands two criterion functions for document clustering and their optimization algorithms. We augment the work by proposing a novel method to work out the go beyond charge with the intention of developing the time competence and "the accuracy" concentrated with Hierarchical Clustering Algorithms. Researches in together intra and inter of data and document clustering data demonstrate that this approach can get better the effectiveness of clustering and accumulate computing time.

In other words, there could be an important disparity among instinctively distinct clusters and the true clusters equivalent to the apparatus in the assortment. In document clustering no labeled documents are provided not like in document classification. Even if ordinary clustering techniques such as k-means be able to be applied to document clustering, they typically do not gratify the unusual necessities for clustering documents: high dimensionality, high quantity of data, relieve in support of browsing, and significant cluster labels. As well, several existing document clustering algorithms need the user to identify the number of clusters as an input constraint and are not strong adequate to hold different types of document locates in a real-world situation. Intended for instance, in various document sets the cluster amount varies as of few to thousands of documents. This discrepancy extremely decreases the clustering accuracy for several of the state-of-the art algorithms.

We also propose a novel document partitioning method based on the non-negative factorization of the term-document matrix of the given document corpus. In the latent semantic space derived by the non-negative matrix Factorization (NMF), each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value.

2. RELATED WORK

Research on multi-view learning in the semi-supervised setting has been launched by two manuscripts, Yarowsky [15] and Blum and Mitchell [6]. Yarowsky illustrates an algorithm for word sense disambiguation. It utilizes a classifier supported on the limited background of a declaration (vision one) and a second classifier using the sanity of further happenings of that declaration in the same document (view two), where both classifiers iteratively bootstrap each other. Blum and Mitchell introduce the term co-training as a general term for bootstrapping procedures in which two hypotheses are trained on distinct views. They describe a co-training algorithm which augments the training set of two classifiers with the n_p positive and n_n negative highest confidence examples from the unlabeled data in each of iteration for each view. The two classifiers work on different views and a new training example is completely based on the decision of one classifier.

Collins and Singer [8] propose an alteration of the co-training algorithm which clearly optimizes an intention function that dealings the measure of concurrence among the rules in dissimilar visions. They as well explain an addition to the AdaBoost algorithm that increases this objective purpose. Blum and Mitchell necessitate a qualified self-government supposition of the visions and provide an instinctive clarification on why their algorithm facility, in conditions of maximizing concurrence on unlabeled data. They as well status the Yarowsky algorithm cascade under the co-training background. The co-EM algorithm is a multi-view description of the Expectation Maximization algorithm for semi-supervised learning [16, 11, 7].

Dasgupta et al. [9] offer PAC limits for the generality error of co-training in terms of the agreement rate of hypotheses in two independent views. This also justifies the Collins and Singer method of directly optimizing the conformity rate of classifiers above the different visions. Clustering algorithms can be separated into two categories [7]: generative (or model-based) approaches and discriminative (or similarity-based) approaches. Model-based approaches endeavor to discover generative models as of the documents, through all models on behalf of one cluster. Frequently generative clustering approaches are depended on the Expectation Maximization (EM) [10] algorithm. The EM algorithm is an iterative statistical technique for maximum likelihood evaluation in locations with incomplete data.
Given a representation of data invention, and data with several missing values, EM will nearby maximize the likelihood of the model parameters and provide approximates for the missing values. Similarity-based clustering approaches optimize an objective function that engage the pair wise document similarities, seeking at maximizing the average similarities contained by clusters and minimize the average similarities between clusters. Mainly of the similarity based clustering algorithms pursue the hierarchical agglomerative approach [12], where a dendrogram is constructing clusters by iteratively merging closest examples. Connected clustering algorithms that work in a multi-view location contain reinforcement clustering [14] and a multiview description of DBSCAN [13].

3. PREVIOUS WORK

For a long time the concept of clustering has been around. It has more than a few applications, mainly in the situation of information retrieval and in organizing web possessions. The focal point of clustering is to situate information and in the current framework, to place mainly significant electronic assets. The research in clustering ultimately goes ahead to automatic indexing to index as well as to recover electronic proceedings. Clustering is a technique in which we construct cluster of objects that are someway similar in individuality. The crucial intend of the clustering is to supply a grouping of similar records. Clustering is frequently confused with classification, but there is some distinction between the two. In classification the objects are consigned to predefined classes, while in clustering the classes are produced. The tenure “class” is in truth often employed as synonym to the word “cluster”. In database management, data clustering is a technique in which, the information that is logically similar is physically stored together. So as to enhance the competence of search and the recovery in database management, the number of disk contacts is to be minimized. In clustering, as the objects of comparable properties are located in one class of substance, a single admittance to the disk can recover the whole class.

If the clustering obtains locate in some abstract algorithmic break, we may cluster an inhabitants into subsets with comparable distinctive, and then decrease the difficulty break by performing on only a delegate from each separation. Clustering is ultimately a procedure of dropping a mountain of information to convenient loads. For cognitive and computational simplification, these heaps might consist of "similar" objects. There are two advances to document clustering, mainly in information reclamation; they are known as expression and item clustering. Term clustering is a technique, which collections disused provisions and this assemblage diminish, blare and enlarge occurrence of obligation. If there are smaller amount bunches than there were innovative provisions, then the measurement is also concentrated. However semantic possessions suffer. There are many different algorithms accessible for phrase clustering.

These are factions, particular relation, and pin-ups and associated mechanism. Factions necessitate all objects in a cluster to be within the entrance of all other substance. In solitary linkage clustering the muscular constriction that each phrase in a class is comparable to every added phrase is comfortable. The regulation to engender particular association clusters is that any idiom that is analogous to several extra terms in the cluster can be additional to the cluster. The luminary practice selects a phrase and then spaces in the class all stipulations that is associated to that idiom (i.e. in consequence a luminary with the preferred phrase as the heart provisions not yet in curriculum are preferred as original starting points pending all stipulations are dispersed to a class. There are many dissimilar modules

that can be fashioned with the star procedure. Item clustering; on the other hand over lend a hand the user in make excursion germane material**.** It is used in two traditions: First is directly found supplementary things that capacity not have been institute by the query and to hand round as a starting point for hallucination of the hammer sleeve. Each item crowd together has a frequent semantic source containing comparable provisions and thus analogous perceptions. Second is to support the consumer in indulgent the chief matters resultant from seek out, the matter repossessed to be clustered and worn to produce an illustration (e.g., explicitly) demonstration of the clusters and their topics. This allows a user to navigate between topics, potentially showing topics the user had not considered. The subjects are not definite by the inquiry except via the transcript of the substance reclaimed.

While items in the catalog comprise been clustered, it is probable to regain all of the objects in a cluster, even if the exploration statement did not categorize them. When the abuser retrieves a powerfully applicable point, the consumer can appear at added items like it devoid of issuing a different investigate. When pertinent items are worn to produce a fresh uncertainty (i.e., important feedback), the recovered hits are comparable to what capacity be fashioned by a clustering algorithm. However, phrase clustering and article clustering in sagacity realize the equivalent intention flush although they are the opposite of every one added. The purpose of both is to conclude supplementary significant objects by a co-occurrence progression. For all of the expressions surrounded by the equivalent cluster, here will be momentous extend beyond of the position of things they are found in. Item clustering is supported upon the matching terms being found in the further items in the cluster. Thus location of items so as to reasoned a period clustering has a brawny possibility of being in the same item cluster based ahead the terms. For illustration, if a phrase cluster has 10 terms in it (assuming they are closely related), then at hand will be a set of items where every one item surrounds foremost detachments of the terms. From the entry perspective, the position of items so as to have the commonality of terms has a strapping opportunity to be positioned in the equivalent entry cluster.

3.1 Concept of Similarity Measurement

The perception of similarity is essentially vital in roughly each methodical pasture. Fuzzy set premise has also urbanized it's possess events of similarity, which discover claim in areas such as management, medication and meteorology. An imperative problem in molecular biology is to determine the succession similarity of couples of proteins. An appraisal or still a catalog of all the exploits of similarity is unfeasible. As an alternative, apparent resemblance is alert on. The amount to which populace distinguish two things as alike basically involves their cogent consideration and performance. Consultation between politicians or corporate executives may be viewed as a process of data collection and assessment of the similarity of hypothesized and real motivators. The appreciation of a fine fragrance can be understood in the same way. Similarity is a core element in achieving an understanding of variables that motivate behavior and mediate affect. In a lot of researches populace are inquired to construct straight or not direct decisions concerning the similarity of pairs of substance.

Cosine similarity assess of similarity among two vectors of *n* dimensions by discovery the cosine of the perspective among them, frequently used to evaluate documents in text mining. Given two vectors of attributes, *A* and *B*, the cosine similarity, θ, is signified using a dot product and magnitude as $similarity = \cos(\theta) = \frac{A.B}{\|A\|\|B\|}$. For content matching, the attribute vectors A and B are typically the tf vectors of the documents. The cosine similarity can be seen as a process of normalizing document length at some point in comparison.

4. PROPOSED METHOD

4.1 Hierarchical clustering overview

A hierarchical clustering algorithm generates a hierarchical corrosion of the given locate of data objects. Depending on the decay approach, hierarchical algorithms are confidential as agglomerative (merging) or divisive (splitting). The agglomerative approach creates through each data position in a disconnect cluster or through a definite large number of clusters. Each step of this move toward combines the two clusters that are the most similar. Thus after each step, the entire number of clusters reduces. This is frequent waiting the preferred number of clusters is attained or only one cluster relics. Through difference, the divisive advance creates by way of all data objects in the same cluster. In each step, one cluster is split into smaller clusters, until a termination condition holds. Agglomerative algorithms are more extensively utilized in observe. The most important work is to develop a novel hierarchal algorithm for document clustering which offers utmost competence and recital. It is chiefly listening carefully in revising and assembly exploit of cluster be related ping occurrence to intend cluster amalgamation criterion. Proposing a novel way to work out the overlie rate to facilitate get better occasion competence and ―the reality" is mostly determined. In the simplest folder, an optimization crisis consists of maximizing or minimizing a genuine purpose by systematically choosing contribution ideals from within an acceptable position and computing the value of the function. The generalization of optimization theory and techniques to other formulations comprises a large area of applied mathematics. Further normally, optimization comprises decision of choosing finest available values of various objective functions specified a defined field, counting a selection of unlike categories of objective functions and diverse types of domains.

Such a formulation is described as an optimization crisis or a statistical training problem in which a term not straight connected to mainframe programming, except at rest in exercise for case in linear programming a lot of genuine world and notional troubles might be formed in this common framework. Tribulations prepared method as force

minimization, communication of the assessment of the purpose as representing the energy of the scheme mortal modeled. Usually, A is some separation of the Euclidean function, frequently precise by a set of limitations, equalities or inequalities that the members of A have to convince. The field A is called the look for space or the option position, whereas the fundamentals are called contestant (maximization), or, in convinced fields, energy function, or energy function. A possible explanation that diminish (or maximizes, if that is the ambition the objective occupation is described an optimal solution. By caucus, the normal form of an optimization problem is stated in conditions of minimization.

Generally, unless both the purpose utility and the possible area are rounded in a minimization problem, there can be numerous local minima, where a local minimum x* is definite as a position for which there subsists some δ > 0 so that for all x such that ||X – X*|| <= ∂; the expression f(x*) <= f(x) holds; that is to say, on some region around x* all of the function values are greater than or equal to the value at that point. Local maxima are defined similarly. A great number of algorithms planned for solving non-convex problems – counting the preponderance of commercially obtainable solvers – are not accomplished of creation a difference between local optimal solutions and precise optimal solutions, and will pleasure the previous as authentic solutions to the innovative problem. The division of applied arithmetic is worried with the growth of deterministic algorithms that are competent of assurance convergence in limited time to the actual optimal solution of a non-convex problem is called global optimization.
The subsequent are the steps in an agglomerative hierarchical clustering algorithm for assemblage N objects.
Step 1: start with N clusters, apiece enclosing single point
Step 2: work out the distance between each one pair of clusters. These distances are typically accumulated in a symmetric distance matrix
Step 3: combine the two clusters through the minimum distance
Step 4: modernize the distance matrix
Step 5: do again steps 3 and 4 awaiting a particular cluster remains

In this the clustering is done to void the iteration process which we can see in the above steps. To achieve this goal we are setting two criterion functions called I_R and I_V. This is done with similarity measurement. The functions are given below: First is for I_R calculation let we express the sum in a general form by function OF: $OF = \sum_{r=1}^{k} n_r \left[\frac{1}{n_r^2}\sum_{d_i,d_j \in S_r} Sim(d_i, d_j)\right]$

After this calculation the objective function transformed into some suitable form such that it could facilitate the optimization procedure to be performed in a simple, fast and effective way according to the above equation. Then at last the final form of our criterion function I_R is as follows:

$$I_R = \sum_{r=1}^{k} \frac{1}{n_r^{1-\alpha}} \left[\frac{n+n_r}{n-n_r}\|D_r\|^2 - \left(\frac{n+n_r}{n-n_r} - 1\right) D_r^t D\right]$$

$D_r^t D$ Is represents the intercluster similarity measure and $\|D_r\|^2$ is denotes the intracluster similarity measure. After this I_V is defined as follows:

$$I_V = \sum_{r=1}^{k} \left[\frac{n+\|D_r\|}{n-n_r}\|D_r\| - \left(\frac{n+\|D_r\|}{n-n_r} - 1\right) D_r^t D / \|D_r\|\right]$$

This above equation is the objective function O after the two criterion function calculation is done. Here the clustering process done also considering web browsing time, which is stored in the web log as shown in table 1. This will improve the clustering accuracy.

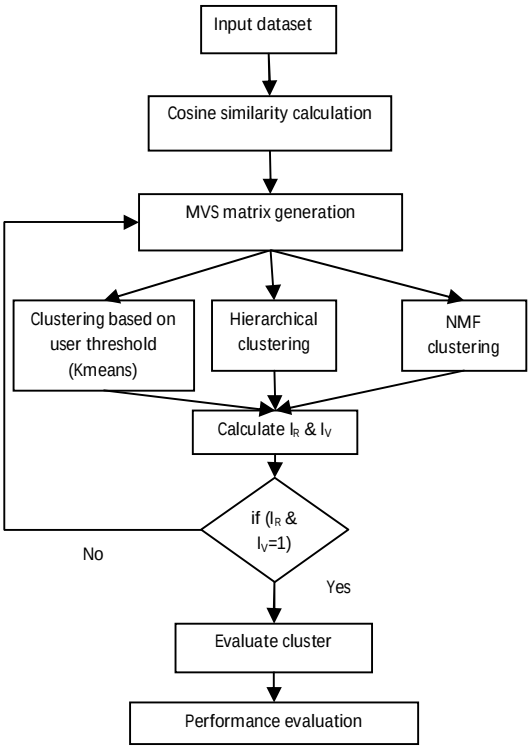

Fig 2: System Flow Diagram

4.2.2 ALGORITHM STEPS

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this:
STEP 1: 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.

STEP 2: Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less with the help of tf - itf.
STEP 3: Compute distances (similarities) between the new cluster and each of the old clusters.
STEP 4: Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

4.2.3 DESIGN LAYOUT

procedure INTIALIZATION
Select k seeds s1,….sk randomly
Cluster[di] ← p = $argmax_r\{s_r{}^t d_i\}$,∀ i=1,….n
$D_r \leftarrow \sum_{didi \in sr,nr} \leftarrow |\ sr\ |$, ∀ r=1,….k
end procedure
procedure REFINEMENT
repeat
{ [1: n]} ← random permutation of {1,…..,n}
for j←1:n **do** i ←v[j]
p ←$cluster$[d_i]
$\Delta I_p \leftarrow I\ (n_p - 1, D_p - d_i) - I\ (\ , D_p)$
$q \leftarrow argmax_{r,\neq p}\ \{\ I\ n_r+1, D_r+d_i - I\ n_r, D_r\ \}$
$\Delta I_q \leftarrow I\ n_q+1, +d_i - I(\ n_q, D_q)$
if $\Delta I_p + \Delta I_q > 0$ **then**
Move d_i **to cluster** q:$cluster$ [d_i] ←q
Update D_p , n_p , D_q , n_q

5. NON-NEGATIVE MATRIX FACTORIZATION CLUSTERING

Assume that a document corpus is comprised of *k* clusters each of which corresponds to a coherent topic. Each document in the corpus either completely belongs to a particular topic, or is more or less related to several topics. To accurately cluster the given document corpus, it is ideal to project the document corpus into a *k*-dimensional semantic space in which each axis corresponds to a particular topic. In such a semantic space, each document can be represented as a linear combination of the *k* topics. Because it is more natural to consider each document as an additive rather than subtractive mixture of the underlying topics, the linear combination coefficients should all take non-negative values. Furthermore, it is also quite common that the topics comprising a document corpus are not completely independent of each other, and there are some overlaps among them. In such a case, the axes of the semantic space that capture each of the topics are not necessarily orthogonal. Based on the above discussions, we propose to use non-negative matrix factorization (NMF) to find the latent semantic structure for the document corpus, and identify document clusters in the derived latent semantic space. NMF does not require the derived latent semantic space to be orthogonal, and it guarantees that each document takes only non-negative values in all the latent semantic directions.

These two characteristics make the NMF superior to the LSI and spectral clustering methods because of the following reasons. First, when overlap exists among clusters, NMF can still find a latent semantic direction for each cluster, while the orthogonal requirement by the SVD or the eigenvector computation makes the derived latent semantic directions less likely to correspond to each of the clusters. Second, with NMF, a document is an additive combination of the base latent semantics, which makes more sense in the text domain. Third, as the direct benefit of the above two NMF characteristics, the cluster membership of each document can be easily identified from NMF, while the latent semantic space derived by the LSI or the spectral clustering does not provide a direct indication of the data partitions, and consequently, traditional data clustering methods such as K-means have to be applied in this eigenvector space to find the final set of document clusters. The following subsections provide the detailed descriptions of the proposed document clustering method.
Document clustering can loosely be defined as “clustering of documents”. Clustering is a process of recognizing the similarity and/or dissimilarity between the given objects and thus, dividing them into meaningful subgroups sharing common characteristics. Good clusters are those in which the members inside the cluster have quite a deal of similar characteristics. Since clustering falls under unsupervised learning, predicting the documents to fall into certain class or group isn't carried out. Methods under document clustering can be categorized into two groups as follows:
This approach divides the documents into disjoint clusters. The various methods in this category are : k-means clustering, probabilistic clustering using the Naive Bayes or Gaussian model, latent semantic indexing (LSI), spectral clustering, non-negative matrix factorization (NMF).

Non-negative matrix factorization is a special type of matrix factorization where the constraint of non-negativity is on the lower ranked matrices. It decomposes a matrix V_{mn} into the product of two lower rank matrices W_{mk} and H_{kn}, such that V_{mn} is approximately equal to W_{mk} times H_{kn}. Where, k << min(m,n) and optimum value of k depends on the application and is also influenced by the nature of the collection itself. In the application of document clustering, k is the number of features to be extracted or it may be called the number of clusters required. V contains column as document vectors and rows as term vectors, the components of document vectors represent the relationship between the documents and the terms. W contains columns as feature vectors or the basis vectors which may not always be orthogonal (for example, when the features are not independent and have some have overlaps). H contains columns with weights associated with each basis vectors in W. Non-negative Matrix Factorization, a technique which makes use of an algorithm based on decomposition by parts of an extensive data matrix into a small number of relevant metagenes. NMF’s ability to identify expression patterns and make class discoveries have been shown to able to have greater robustness over popular clustering techniques such as HCL and SOM.

MeV’s NMF uses a multiplicative update algorithm, introduced by Lee and Seung in 2001, to factor a non-negative data matrix into two factor matrices referred to as W and H. Associated with each factorization is a user-specified rank. This represents the columns in W, the rows in H, and the number of clusters to which the samples are to be assigned. Starting with randomly seeded matrices and

using an iterative approach with a specified cost measurement we can reach a locally optimal solution for these factor matrices. H and W can then be evaluated as metagenes and metagenes expression patterns, respectively. Using a "winner-take-all" approach, samples can be assigned to clusters based on their highest metagenes expression. Multiple iterations of this process allow us to see the robustness of the cluster memberships. Additionally, running multiple ranks consecutively can allow for the comparison between differing numbers of classes using cophenetic correlation. NMF is most frequently used to make class discoveries through identification of molecular patterns. The module can also be used to cluster genes, generating metasamples rather than metagenes.

NMF is a matrix factorization algorithm that finds the positive factorization of a given positive matrix. Assume that the given document corpus consists of k document clusters. Here the goal is to factorize X into the non-negative $m \times k$ matrix U and the non-negative $k \times n$ matrix VT that minimize the following objective function:

$$J = \frac{1}{2}\|X - UV^T\|$$

Where $\|.\|$ denotes the squared sum of all the elements in the matrix. The objective function J can be re-written as:

$$J = \frac{1}{2}tr((X - UV^T)(X - UV^T)^T)$$

Non-negative matrix factorization (NMF) has previously been shown to be a useful decomposition for multivariate data. Two different multiplicative algorithms for NMF are analyzed. They differ only slightly in the multiplicative factor used in the update rules. One algorithm can be shown to minimize the conventional least squares error while the other minimizes the generalized Kullback-Leibler divergence. The monotonic convergence of both algorithms can be proven using an auxiliary function analogous to that used for proving convergence of the Expectation-Maximization algorithm. The algorithms can also be interpreted as diagonally rescaled gradient descent, where the rescaling factor is optimally chosen to ensure convergence.

6. RESULTS AND DISCUSSION

6.1 Recall Rate comparison

We analyze and compare the performance offered by existing algorithm and proposed algorithm. Here if the number of datasets increased the recall rate also increased linearly. The recall rate of the proposed is high. Based on the comparison and the results from the experiment show the proposed approach works better than the other existing systems.

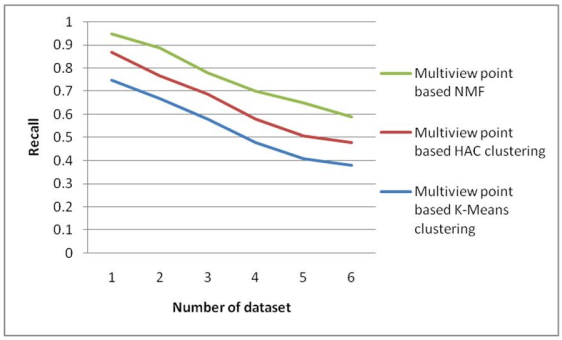

Fig 3: Recall rate comparison

6.2. Precision Rate

We analyze and compare the performance offered by existing algorithm and proposed algorithm. Here if the number of datasets increased the recall rate also increased linearly. The recall rate of the proposed algorithm is high. Based on the comparison and the results from the experiment show the proposed approach works better than the other existing systems.

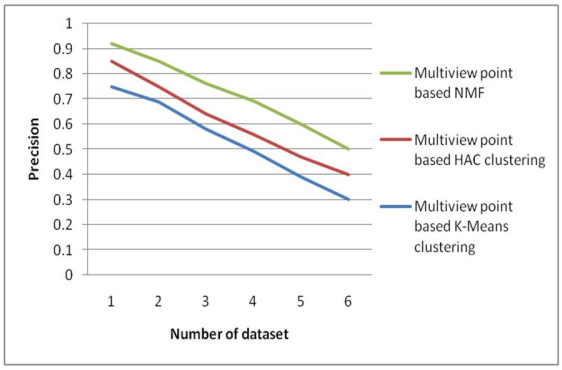

Fig 4: Precision rate comparison

7. CONCLUSION

The experience in general data sets and a document set indicates that the new method can decrease the time cost, reduce the space complexity and improve the accuracy of clustering. In this paper, selecting different dimensional space and frequency levels leads to different accuracy rate in the clustering results. We also developed an incremental insertion component for updating the comments-based hierarchy so that resources can be efficiently placed in the hierarchy as comments arise and without the need to re-generate the (potentially) expensive hierarchy.

7.1 FUTURE WORK

There are a number of future research directions to extend and improve this work. One direction is that this work might continue on is to improve on the accuracy of similarity calculation between documents by employing different similarity calculation strategies using genetic

algorithm which produces optimal result. Although the current scheme proved more accurate than traditional methods, there are still rooms for improvement.

REFERENCE

[1] Johnson,S.C., "Hierarchical Clustering Schemes" Psychometrika, 2:241-254. 1967

[2] Cole, A. J. & Wishart, D. An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. The Computer Journal 13(2):156- 163. (1970).

[3] D'andrade,R., "U-Statistic Hierarchical Clustering" Psychometrika, 4:58-67. 1978

[4] Jeff A. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. ICSI TR-97-021, U.C. Berkeley,1998.

[5] P. Berkhin. Survey of clustering data mining techniques.Unpublished manuscript, available from accrue.com, 2002.

[6] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the Conference on Computational Learning Theory, pages 92–100, 1998.

[7] U. Brefeld and T. Scheffer. Co-EM support vector learning. In Proc. of the Int. Conf. on Machine Learning, 2004.

[8] M. Collins and Y. Singer. Unsupervised models for named entity classification. In EMNLP, 1999.

[9] S. Dasgupta, M. Littman, and D. McAllester. PAC generalization bounds for co-training. In Proceedings of Neural Information Processing Systems (NIPS), 2001.

[10] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1977.

[11] R. Ghani. Combining labeled and unlabeled data for multiclass text categorization. In Proceedings of the International Conference on Machine Learning, 2002.

[12] A. Griffiths, L. Robinson, and P. Willett. Hierarchical agglomerative clustering methods for automatic document classification. Journal of Doc., 40(3):175–205, 1984.

[13] K. Kailing, H. Kriegel, A. Pryakhin, and M. Schubert. Clustering multi-represented objects with noise. In Proc. of the Pacific-Asia Conf. on Knowl. Disc. and Data Mining, 2004.

[14] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W. Ma.Recom: Reinforcement clustering of multi-type interrelated data objects. In Proceedings of the ACM SIGIR Conference on Information Retrieval, 2003.

[15] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proc. of the 33rd Annual Meeting of the Association for Comp. Linguistics, 1995.

[16] K. Nigam and R. Ghani. Analyzing the effectiveness andn applicability of co-training. In Proceedings of Information and Knowledge Management, 2000.

[17] I. Dhillon, D. Modha,"Concept decompositions for large sparse text data using clustering", Mach. Learn., Vol. 42, No. 1-2, pp. 143–175, 2001.

[18] W. Xu, X. Liu, Y. Gong,"Document clustering based on nonnegative matrix factorization", in SIGIR, 2003, pp. 267–273.

[19] Shengrui Wang and Haojun Sun. Measuring overlap- Rate for Cluster Merging in a Hierarchical Approach to Color Image Segmentation. International Journal of Fuzzy Systems,Vol.6,No.3,September 2004.

[20] A. Banerjee, I. Dhillon, J. Ghosh, S. Sra,"Clustering on the unit hypersphere using von Mises-Fisher distributions", J. Mach. Learn. Res., Vol. 6, pp. 1345–1382, Sep 2005.

[21] S. Zhong,"Efficient online spherical K-means clustering", in IEEE IJCNN, 2005, pp. 3180–3185.

[22] E. Pekalska, A. Harol, R. P. W. Duin, B. Spillmann, H. Bunke,"Non-Euclidean or non-metric measures can be informative", in Structural, Syntactic, and Statistical Pattern Recognition, ser. LNCS, Vol. 4109, 2006, pp. 871–880.

[23] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg,―Top 10 algorithms in data mining", Knowl. Inf. Syst., Vol. 14, No. 1, pp. 1–37, 2007.

[24] I. Guyon, U. von Luxburg, R. C. Williamson, "Clustering: Science or Art?", NIPS'09 Workshop on Clustering Theory,2009.

[25] M. Pelillo,"What is a cluster? Perspectives from game theory", in Proc. of the NIPS Workshop on Clustering Theory,2009.

[26] D. Lee, J. Lee,"Dynamic dissimilarity measure for support based clustering", IEEE Trans. on Knowl. and Data Eng., Vol. 22, No. 6, pp. 900–905, 2010.

[27] A. Banerjee, S. Merugu, I. Dhillon, J. Ghosh,"Clustering with Bregman divergences", J. Mach. Learn. Res., Vol. 6, pp. 1705–1749, Oct 2005. Volume 2, Issue 6, June 2012

[28] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401:788–791, 1999

[29] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems, volume 13, pages 556–562, 2001.

[30] X. Liu and Y. Gong. Document clustering with cluster refinement and model selection capabilities. In Proceedings of ACM SIGIR 2002, Tampere, Finland, Aug. 2002

[1]K.S. Jeen Marseline working as Assistant professor and head department of information and computer technology in Sri Krishna arts and science college, Coimbatore. She has 16 years of experience in teaching field. She is currently doing her research work in the field of image processing. She has published and presented papers in conferences and journals.

[2] A.Premalatha doing her M.phil in Sri Krishna arts and science college, Coimbatore. She is currently working as Assistant professor in LNV College of Arts And Science Coimbatore. Her research area is Data mining.