Hierarchical Viewpoint Based Clustering

[1]S. Susmitha, [2]A. Isabella

[1,2]Dept. of CSE, A.S.L. Pauls College of Engineering and Technology, TamilNadu, India

Abstract

Clustering is an important technique in the data mining. The main goal of the clustering is to find the similarity between the data points and grouping them the data into a single groups or sub groups in clustering process. In this paper investigate k-means algorithm, to implement the document clustering with Multiviewpoint based similarity measure. Similarity between a pair of data points can be defined either explicit or implicit to find the optimal solution for clustering processes. To resolve this problem, proposed system which is developing a novel hierarchical algorithm for document clustering which produces superlative efficiency and performance which is mainly focuses 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. The hierarchical is like building a tree-based hierarchical taxonomy from a set of documents.

Keywords

Multiviewpoint, Hierarchical Algorithm, K- Means Algorithm, Agglomerative, Taxonomy

I. Introduction

Clustering is an important topic in data mining. The main aim of the clustering is to grouping the similar data points and form a single group or sub group in clustering process. Clustering based document are published in every year. In the exiting process, k-means algorithm to implement the similarity based clustering in different viewpoint. To find the optimal solution. The k-means algorithm is the one of the algorithm in the recent study[1]. 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.

The future is aggregated by investigations as of the over and compartable examine conclusion. 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.

Document clustering has become an increasingly important task in analyzing huge numbers of documents spread among various sites. The difficult aspect is to organize the documents in a way that results in better investigate without introducing much extra cost and complication. The Cluster theory is fundamental to the issue of improved efficiency. It states that related documents tend to be more similar to each other than to non-relevant documents and therefore tend to appear in the same clusters. If the cluster theory holds for a particular document collection, then important documents will be well separated from non-relevant ones. A important document may be ranked low in a best-match search because it may require some of the query terms. In a clustered collection, this related document may be clustered together with other appropriate items that do have the required query conditions and could therefore be retrieved through a clustered search. According to best-match IR systems, if a document does not control any of the query terms then its similarity to the query will be zero and this document will not be retrieved in reaction to the query. Document clustering offers an another file group to that of best-match recovery and it has the potential to address this issue, thereby increase the usefulness of an IR system.

II. Related Work

Research on multi-view learning in the semi-supervised setting has been launched by two manuscripts, Yarowsky [10] and Blum and Mitchell [3]. 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 illustrate a co-training algorithm which augments the training set of two classifiers with the n_p positive and n_n negative highest assurance examples from the unlabeled data in each of iteration for each view. The two classifiers work on special views and a new training example is absolutely based on the decision of one classifier.

Collins and Singer [4] 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 [6, 11].

Dasgupta et al. [5] offer PAC limits for the generality error of co-training in terms of the agreement rate of hypotheses in two

 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

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 [15]: 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) [5] 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 [7], where a dendrogram is constructing clusters by iteratively merging closest examples. Connected clustering algorithms that work in a multi-view location contain reinforcement clustering [9] and a multiview description of DBSCAN [8].

III. 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 method in which we create 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 procedure in which, the information that is logically similar is actually 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 kind the chief matters resultant from seek out, the matter repossessed to be clustered and worn to produce an design (e.g., explicitly) protest 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.

A. 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.

IV. Proposed Method

A. 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: Find 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 inter cluster similarity measure and $\|D_r\|^2$ is denotes the intra cluster 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, this will improve the clustering accuracy.

B. 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.

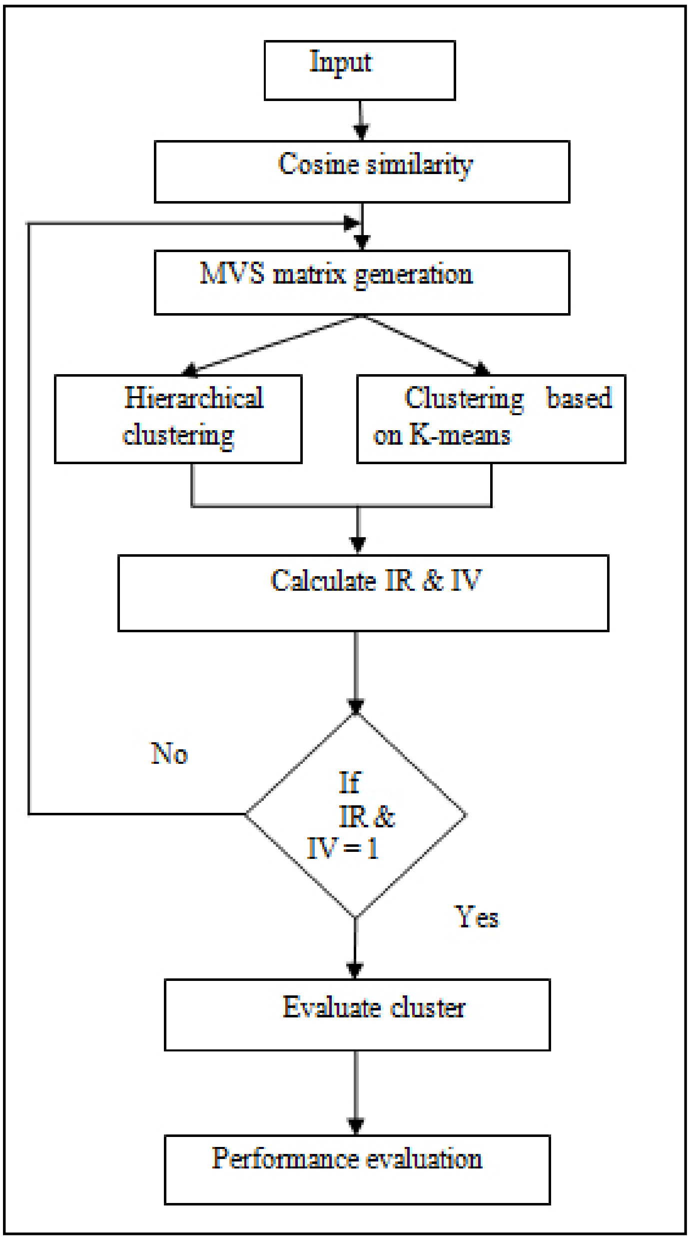

Fig. 1: System Flow Diagram

C. Design Layout

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

V. Methodology

The project has been discussed through 6 major modules. Each module is designed in such a way that it performs its assigned task.

A. Preprocessing

Data pre-processing is an often neglected but important step in this project. In this module we are doing the preprocessing of two data sets of breast and wine. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. In order to improve the quality of data the preprocessing step done at first. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. In this the parsing is the first step done when the document enters the process state. Parsing is defined as the separation or identification of Meta tags in a HTML document. Here, the raw HTML file is read and it is parsed through all the nodes in the tree structure.

B. Document Similarity Identification

The similarity between two documents is found by the cosine-similarity measure technique. The weights in the cosine-similarity are found from the TF-IDF measure between the phrases (meta-tags) of the two documents.

This is done by computing the term weights involved.

$$TF = C / T \tag{1}$$

$$IDF = D / DF \tag{2}$$

D a quotient of the total number of documents
DF a number of times each word is found in the entire corpus
C a quotient of no of times a word appears in each document
T a total number of words in the document

$$TFIDF = TF * IDF \tag{3}$$

The cosine similarity in (3) can be expressed in the following form without changing its meaning:

Sim (d_i, dj) =

$$\cos(d_i\text{-}0, dj\text{-}0) = (d_i\text{-}0)^t (d_j\text{-}0) \tag{4}$$

C. Clustering With Multi-Viewpoint Using K-Means

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. The similar documents are grouped together in a cluster, if their cosine similarity measure is less than a specified threshold.

It is used in the traditional k-means algorithm. The objective of k-means is to minimize the Euclidean distance between objects of a cluster and that cluster's centroid

$$Sim(d_i,d_j) = \cos(d_i,d_j) = d_i^t d_j \qquad (5)$$

Cosine measure is used in a variant of k-means called spherical k-means. While k-means aims to minimize Euclidean distance, spherical k-means intends to maximize the cosine similarity between documents in a cluster and that cluster's centroids

$$max \sum_{r=1}^{k} \sum_{d_i \in S_r} \frac{d^t C_r}{||C_r||} \qquad (6)$$

D. Clustering with Multi-Viewpoint Using Hierarchical Clustering

Hierarchical clustering techniques proceed by either a series of successive merges or a series of successive divisions. For both methods, the number of clusters is needed to select a clustering from the hierarchy. However the difference between the levels of the hierarchy may be an indication of the correct number of clusters. Hierarchical clustering generates a hierarchical tree of clusters. This algorithm is developing a novel hierarchal algorithm for document clustering. They used cluster overlapping phenomenon to design cluster merging criteria. The system computes the overlap rate in order to improve time efficiency and the veracity and the line passed through the two cluster's center instead of the ridge curve. Based on the hierarchical clustering method it used the Expectation-Maximization (EM) algorithm in the Gaussian mixture model to count the parameters and make the two sub-clusters combined when their overlap is the largest.

E. Ir and Iv Calculation

Having defined our similarity measure, we now formulate our clustering criterion functions. The first function, called IR, is the cluster size-weighted sum of average pairwise similarities of documents in the same cluster. First, let us express this sum in a general form by function F.

$$F = \sum_{r=1}^{k} n_r \left[\frac{1}{n_r^t} \sum_{d_i,d_j \in S_r} Sim(d_i, d_j) \right] \qquad (7)$$

The final form of our criterion function IR is

$$I_R = \sum_{r=1}^{k} \frac{1}{n_r^{1-\alpha}} \left[\frac{n+n_r}{n-n_r} \|D_r\|^2 - \left(\frac{n+n_r}{n-n_r} - 1 \right) D_r^t D \right] \qquad (8)$$

Again, we could eliminate n because it is a constant. Maximizing G is equivalent to maximizing IV below

$$I_V = \sum_{r=1}^{k} \left[\frac{n+\|D_r\|}{n-n_r} \|D_r\| - \left(\frac{n+\|D_r\|}{n-n_r} - 1 \right) \frac{D_r^t D}{\|D_r\|} \right] \qquad (9)$$

F. Performance Evaluation

Finally in this we are comparing the existing and proposed system based on some parameters to evaluate the performance.

VI. 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.

References

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

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

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

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

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

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

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

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

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

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

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

[12] Shengrui Wang, 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.

[13] U.Brefeld, T.Scheffer,"Co-EM support vector learning", In proc. of the Int. conf. on machine Learning, 2004.

 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

I am susmitha, completed my B.Sc Computer science in Lakshmi Narayanan Vishalakshi College of Arts & Science for women from Bharathiyar university Coimbatore, Tamilnadu in 2005, the M.C.A(Master of computer application) in School of distance education, Bharathiyar university Coimbatore, Tamilnadu in 2008. In between worked in Air Force Wives Welfare Association Institute as a Computer teacher (2006-2007) handled in computer languages. After that Worked in Nehru Institute of Engineering & Technology as a Computer Science Lecturer (2008-2010). At present M.E (Master of computer science and engineering) studying in A.S.L Pauls College of engineering and technology, solavampalayalam village & panchatyat (Elur), pollachi taluk, Coimbatore, Tamilnadu in (2010-2013).