Volume 5, Issue 4, April 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Data Mining Objective Function with Feature Selection Algorithm using Document Clustering

P. Nithya
M.E/CSE
Gnanamani College of Technology
Namakkal, Tamilnadu, India

R. Umamaheswari
Asso.Professor/CSE
Gnanamani College of Technology
Namakkal, Tamilnadu, India

Dr. N. Shanthi
Professor and Dean/CSE
Nandha Engineering College
Erode, Tamilnadu, India

Abstract: ***Conventional feature selection classifiers work with known and precise data values. In recent data collection methods, appreciable amount of attributes are uncertain. The uncertain attributes, in almost all applications, have more influences on the data set on information classification and feature selection constructs. Uncertainty needs to be handled properly reasons for uncertainty are due to measurement errors, quantization errors, data staleness and multiple repeated measurements. Uncertainty of a document item is represented in terms of multiple values. Usually uncertain document are abstracted by statistical derivatives Complete information of the data item improves the accuracy of feature selection algorithm. In this proposal work is made to improve the pruning of feature selection algorithm by clustering with distance boundaries and partitioning of uncertain probability distribution values. Clustering techniques increase the speed of feature selection construction and minimize the pruning time to greater extent. Distance boundary clustering technique, works based on the criteria of lower and upper bounds distances of the uncertain attributes values.***

Keywords*: Cluster Interface,Document Clustering, Feature Selection, Objective Function Uncertainty.*

I. INTRODUCTION

Data mining is a process of analyzing data in order to fetch about trends or patterns from the data. Many techniques in data mining such as text mining and web mining also exist. Clustering is one of the important data mining or text mining algorithm that is used to group similar objects together. In other words, it is used to organize given objects into some significant sub groups that make further analysis on data easier. Clustering requires definition of a distance-measure which assigns a numeric value to the amount of difference between two documents and which the clustering algorithm uses for making different groups of a given dataset. No single distance measure works the best in all scenarios and selection depends on which distance measure best catches the spirit of important individual characters for the given document set. All clustering methods have to assume some cluster relationship among the data objects that they are applied on similarity between a pair of objects can be defined either explicitly or implicitly. A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity among data. Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and fixed in the clustering measure function. Intext document clustering , used to reorganizing the large amount of documents into a smaller number of manageable clusters. While several clustering methods and the associated similarity measures have been proposed in the past, the partition clustering algorithms are reported performing well on document clustering.

Text document clustering has proved to be an interesting research problem, to attracting much work in the domain of information retrieval. It aims at grouping similar documents in one class and separates this group as much as possible from the ones which contain information on entirely different topics. In the website has a huge application of text clustering. From clustering of results for users on search engines to grouping of comments to suggest products on online stores, the technique has massive direct applicability. An important problem that closely arises in the task is to specify what aspects of documents decide their nearness. Measures such as Cosine similarity, Jaccard coefficient have been suggested and widely used for this purpose and have proved to be efficient. In this experiment results of empirical study on five similarity measures namely Euclidean Measure, Cosine Similarity, Jaccard Coefficient and Pearson Correlation Coefficient using k-means clustering algorithm. To have consistent conclusions, we experiment on seven datasets with different characteristics. These includenewspaper articles, research articles, web pages etc. They come with different categories which is important later in evaluations when try to measure quality of obtained clusters based on purityand entropy values.

II. RELATED WORKS

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 of the two documents. This is done by computing the term weights involved.

TF = C / T

IDF = D / DF

A Similarity measure is a method which computes the degree of similaritybetween a pair of documents. The computation of similarity between two documents is briefly described here.

Euclidean distance between two documents are calculated,

$D_E(d_i, d_j)=\sqrt{\sum_{i=1}^{n}(d_i - d_j)^2}$

D is a set, which contains m text documents

D={d1,d2,.....di...dm} ; i=1,2,.....m.

There are n words among m text documents.

$di=\{w_{i1},w_{i2},....w_{ij},....w_{in}\}$

i=1,2,.....m;j=1,2,....n

This measure is defined as the root of square difference between the respective coordinates of d_i, d_j. Also, one can use weighted distance, and many other similarity/distance measures.

Cosine similarity measure the cosine of the angle between two documentsd_1, d_2 as follows

$S_{cos}(d_1,d_2)= \frac{d_1.d_2}{\sqrt{(d_1.d_1)(d_2.d_2)}}$

Pair wise-adaptive similarity dynamically selects a number of features out of d_1 and d2 is defined as

$S_{pair}(d_1, d_2)= \frac{d_{1,k}.d_{2,k}}{\sqrt{(d_1.d_1)(d_2.d_2)}}$

Where $d_{i,k}$ is a subset of d_i, i=1,2containing the values of the feature which are the union of the K largest features which are the union of the K largest features appearing in d_1 andd_2, respectively.

The Extended Jaccard coefficient is an extended version of the jaccard coefficient for data processing,

$S_{EJ}(d_1, d_2)= \frac{d_1.d_2}{d_1.d_1+d_2.d_2-d_1.d_2}$

Dice similarity also used to measure the similarity between the two documents. It is represented as follows,

$S_{Dic}(d_1, d_2)= \frac{2d_1.d_2}{d_1.d_1+d_2.d_2}$

Existing systems greedily picks the next frequent item set which represent the next cluster to minimize the overlapping between the documents that contain both the item set and some remaining item sets. In other words, the clustering result depends on the order of picking up the item sets, which in turns depends on the greedy heuristic. This method does not follow a sequential order of selecting clusters. Instead, we assign documents to the best cluster.

The proposed measure takes in to report the following three cases: a) The feature appears in both documents, b) the feature appears in only one document, and c) the feature appears in none of the documents.

$F(d_1,d_2)=\frac{\sum_{j=1}^{m} N*d_{1j},d_{2j}}{\sum_{j=1}^{m} N\cup d_{1j},d_{2j}}$

$$N_*(d_{1j},d_{2j})=\begin{cases} 0.5\left(1+exp\left\{-\frac{(d_{1j},d_{2j})}{\sigma_j}\right\}\right) \\ \quad if\ d_{1j},d_{2j}>0 \\ 0\ ,if\ d_{1j}=0 \text{ and } d_{2j}=0 \\ \quad -\lambda, other\ wise \end{cases}$$

$$N\cup(d_{1j},d_{2j})=\begin{cases} 0\ ,if\ d_{1j}=0 \text{ and } d_{2j}=0 \\ \quad 1, \text{otherwise} \end{cases}$$

Then our proposed similarity measure ,

$S_{SMTP}(d_1,d_2)=\frac{F(d_1,d_2)+\lambda}{1+\lambda}$

III. SYSTEM APPROACHES

The steps involved are as below:

1. Document preprocessingis which includes removing stop words and stemming process. This step is done to reduce high dimensionality of the data set and improve the computational time.

1.1) Filtering:Filtering removes special characters and punctuation from documents, which are not thought to hold any discriminative power under the vector model.

1.2) *Tokenization*:This step splits sentences into individual tokens, typically words.

1.3) Stop word removal:A stop word is defined as a term which is not thought to convey any meaning of dimension in the vector space. Stop words are the most common words (e.g., "and", "or", "in") in a language, but they do not convey any significant information so they are stripped from the document set.

1.4) Stemming:This is known as the process of reducing words to their base form, or stem. For example, the words "connected", "connection", "connections" are all reduced to the stem "connect". Porter's algorithm is the de facto standard stemming algorithm. A smaller number of distinct terms results in a saving of memory space and processing time.In preprocessing process is to calculate and generate the phrases, nodes and edges.

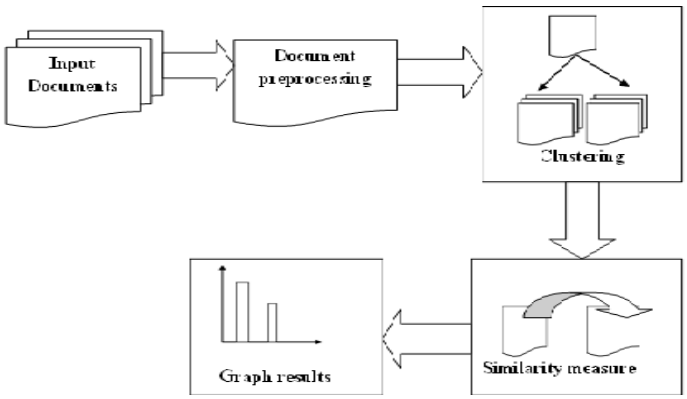

Figure 1.1: System Architecture

2. Clustering

Hierarchical based clustering technique is used to generate the collection form of words in the document. Whatever decides documents of different classes is the frequency with which the words are used. Furthermore, each document has only a subset of all words from the broad vocabulary. Also, because of the probabilistic nature of how words are distributed, any two documents may share many of the same words. In many cases, the agglomerative hierarchal clustering will often put documents of the same class in the same cluster, even at the earliest stages of the clustering process.

3. Similarity Measure

The Similarity measure used to find similarity between documents. Though, the characterization of a pair of documents being similar or dissimilar is not always perfect and normally varies with the actual problem setting. This contains the document similarity study and thereby finding the Overlapping Rate (OLP Rate).

4. Graph Results:

Clustering is a division of data into groups of similar objects. . Representing the data by fewer clusters necessarily loses certain good details, but accomplishes simplification between the similar documents are grouped together in a cluster, if their cosine similarity measure is less than a specified threshold. The graph results show the document item and the cluster formation.

IV. EXPERIMENTAL RESULTS

Accurate clustering needs a precise definition of the nearness between a pair of objects, in terms of either the pair wise similarity or distance. Before clustering, a similarity/distance measure must be determined. The measure redirects the degree of nearness or separation of the target items and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. Similarity between a pair of objects can be defined either explicitly or implicitly. A similarity measure is a method which computes the degree of similarity between a pair of text objects.

A reasonable measure of the distancebetween points, clusters are sometimes not well separated and so a new measure ofsimilarity between points was introduced that reflects the neighborhood of a point. While some techniques use one variable at a time, most approaches use all features at once to calculate the distance or similarity.In this scheme work, focus is given to derive a novel method for measuring similarity between data objects in sparse and high-dimensional field, particularly text documents. From the proposed similarity measure, then express new clustering measure functions and introduce their respective clustering algorithms, which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance.

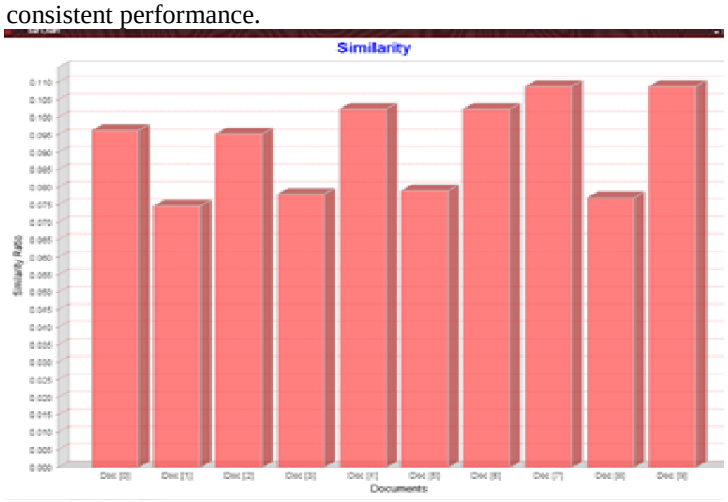

Figure 1.2 Similarity Graph

Fig1.2 shows the similarity graph between the documents and similarity rate.The main work is to develop a clusteringalgorithm for document clustering whichprovides maximum efficiency and performance.

V. CONCLUSION&FUTUREWORK

Certain a documents set, the ultimate consequence would be to have a given set of measures to choose a proper clustering method to apply. Taking a clustering algorithm, conversely, can be a challenging task. This report has a proposal of a novel hierarchical clustering algorithm based on the overlap rate for cluster integration. Experiment 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. Especially, in the document clustering, the newly proposed method calculating result show great advantages. The hierarchical based clustering algorithm provides a regularway of characteristic clusters and fulfillingthe basicconstraint of clustering as high within-cluster match and between-cluster dissimilarity.This process is to increase the clustering accuracy and less classification time. To extract the features reasonably will be considered in the future work. There are a number of future research directions to extend and improve this work.

REFERENCES

[1] C. Silva, U. Lotric, B. Ribeiro, and A. Dobnikar, "Distributed text classification with an ensemble kernel-based learning approach,"*IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 3,pp. 287–297, May 2010.

[2] D. Lee and J. Lee "Dynamic Dissimilarity measure for Support Based Clustering," IEEE tran .knowledge and Data Eng., vol. 2 no. 6, pp. 900-905, June.

[3] I.Ceema , M.Kavitha , G.Renukadevi, G.sripriya, S. RajeshKumar, "Clustering Web Documents using Hierarchical Method for Efficient Cluster Formation," *IJSETR, vol.1, no.5,pp.127-131, Nov 2012.*

[4] I.S. Dhillon, J. Kogan, and C. Nicholas, "Feature selection and document clustering," in *A Comprehensive Survey of Text Mining*, M. W. Berry, Ed. Heidelberg, Germany: Springer, 2003.pp. 83–93.

[5] T. Kanungo*et al.*, "An efficient k-means clustering algorithm:Analysis and implementation," *IEEE Trans. Pattern Anal. Mach.Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[6] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, "Document clustering in correlation similarity measure space," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1002–1013, Jun. 2012.

[7] Y. Zhao and G. Karypis, "Comparison of agglomerative and partitional document clustering algorithms," in *Proc. Workshop Clustering High Dimensional Data Its Appl. 2nd SIAM ICDM*, 2002,

[8] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee," A Similarity Measure for Text Classification and Clustering", IEEE Trans.Know.DataEng, *vol. 26, no. 7,pp .1575- 1590, july 2014.*

[9] P.Nithya , R.Umamaheswari and N.Shanthi, "An Enhanced Similarity Computation for Document Clustering Approaches ,"*IJ0SER vol 2 ,no.10 ,oct 2014.*

