ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 4, April 2018

Distance Measures For Text Cluster Mining

[1] Sruthy KG, [2] Dr.C.Nalini
[1] Mtech Student, Computer Science and Engineering, Bharath University
[2] Professor, Computer Science and Engineering, Bharath University

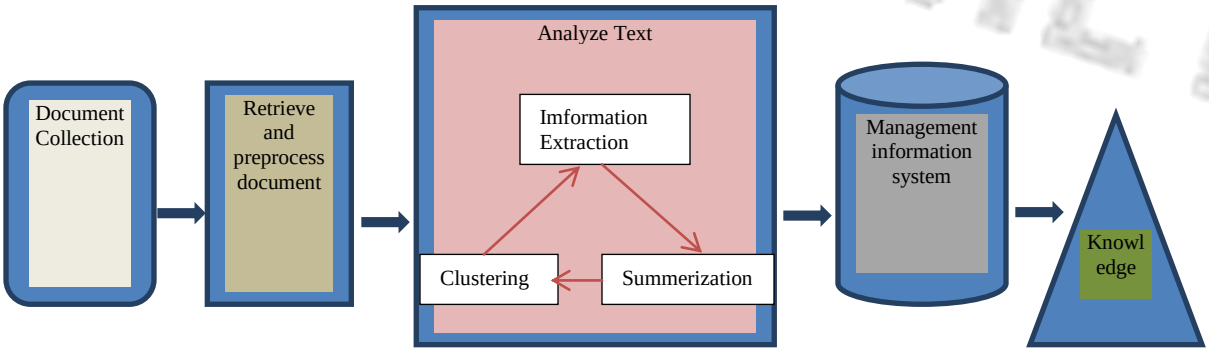
Abstract: ***The amount of text generated isincreasing day by day.This large volume of unstructured text cannot be merely processed and understand by computers. Therefore, efficient and effective techniques are required to find useful patterns.Text mining is the task of extracting meaningful information from text,.Clusteringis one of the data mining task, and is important for text mining. Text Clustering is mainly described as grouping of the similar documents a largecollection of unstructured documents. In this paper, we compare Euclidean distance and mahattan distance based onexecution time, number of iterations, number of clusters, sum of squared errors using clustering algorithm such as simple K means, and hierarchical clustering.***

Keywords: Text Mining, Clustering, Simple K means, Hierarchical clustering, Euclidean distance, Manhattan distance.

I. INTRODUCTION

With the wide use of internet, a huge amount of textual documents are present over internet. Text data is present in the form of enterprise information systems, digital documents and in personal files. As the size of text data is increasing, the handling and analysis of text data becomes very important. Text mining is being developed as technology to handle the large volumes of the text data. It is similar to data mining, except that data mining tools are designed to handle structured data, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. Different text mining functionalities are text clustering, text classification, text categorization.

The text miningtool would retrieve a particular document and preprocessit .After preprocessing it would go through a text analysis phase. In analysis phase, information is extracted and applying some techniques in the extracted information. The techniques may be clustering or classification and result is given to management information system. It yields to large amount of knowledge.



The objective of this research work is compare the Euclidean distance and manhattan distance using clustering algorithms such as simple k means and hierarchical clustering,then find the efficient algorithm. The remaining portion of the paper is discussed as follows. The proposed methodology is given in Section 3. Section 4 analyses the experimental results. Section 5 gives conclusion.

II. LITERATURE SURVEY:

Twinkle Svadas, JasminJha: Cluster mining done on text documents. Clustering of the text documents has becomean important technology over internet. Text Clustering is mainly described as grouping of the similar documents a large

collection of unstructured documents. Text document clustering is the most widely used method for generalizing large amount of information. This paper proposes a system to categorize the text documents and form the clusters.

AlexandreRibeiroAfonso,CláudioGottschalg Duque: This article reports the findings of an empirical study about Automated Text Clustering applied to scientific articles and newspaper texts in Brazilian Portuguese, the objective was to find the most effective computational method able to cluster the input of texts in their original groups. Considering the experiments carried out, the results of human text classification and automated clustering are distant; it was also observed that the clustering correctness results vary according to the number of input texts and their topics

Vinod S. Badgujar, Asha H. Pawar: This paper propose clustering algorithm that search into the documents with natural language contained and get the best words of their content to form a database knowledge that the first step to get the desired knowledge. They implemented the system using the K-means clustering algorithm. Moreover the future work uses the search engine to make searches classify the information introduced by the last user and searching in the exact cluster.

Anna Huang: In this paper, compare and analyze the effectivenessof these measures in partitional clustering for text document datasets. The experiments utilize the standard Kmeans
algorithm and report results on seven text document datasets and five distance/similarity measures that have been most commonly used in text clustering.

Mehdi Allahyari, SeyedaminPouriyeh: Text mining is the task of extracting meaningful information from text, which has gained significantattentions in recent years.In this paper, we describe several ofthe most fundamental text mining tasks and techniques including text pre-processing, classification and clustering. Additionally, we briefly explain text mining in biomedical and health care domains.

III. METHODOLOGY:

3.3 Dataset:

The eco hotel taken as a dataset for text mining. It consist one attribute and four hundred and one instances. This document comes under business area.

3.4 Text mining approaches:

Text Mining: It refers to the process of extracting high quality of information from text ,semi-structured text and unstructured text resources.
Information Retrieval (IR): Information Retrieval is the activity of finding information resources from a collection of unstructured data sets that satisfies the information need. So information retrieval focused on information access rather than analyzing information and finding hidden patterns.
Natural Language Processing (NLP): Natural Language Processing is sub-field of computer science, artificial intelligence and linguistics which aims at understanding of natural language using computers. Many text mining algorithms make use of NLP techniques, as a part of speech tagging (POG), syntactic parsing.
Information Extraction from text (IE): Information Extraction is the task of automatically extracting informationfrom unstructured or semi-structured documents .It usually serves as a starting point for other text mining algorithms.
Text Summarization: Text summarization is nothing butsummarize the text documents in order to get a large document or a collection of documents. There are two categories of summarization techniques,extractive summarization where a summary comprises information units extracted from the original text, and abstractive summarization where a summary may contain "synthesized" information that may not occur in the original document.
Unsupervised Learning Methods: Unsupervised learning methods are techniques trying to find hidden structure out of unlabelled data. Clustering and topic modelling are the two commonly used unsupervised learning algorithms used in the context of text data. Clustering is the task of segmenting a collection of documents into partitions where documents in the same group (cluster) are more similar to each other than those inother clusters.
Supervised Learning Methods: Supervised learning methods are machine learning techniques to learn a classifier from the training data in order to perform predictions on unseen data. The major supervised methods are nearestneighbor classifiers, decision trees, rule-based classifiers and probabilistic classifiers.

3.5 Text preprocessing:

Preprocessing is one of the key components in many text miningalgorithms. The preprocessing step usuallyconsists0020tokenization, filtering, lemmatizationand stemming.

Tokenization: Tokenization is theprocess of breaking a stream of textual content up into words, terms, symbols, or some other meaningful elements called tokens.
Filtering: Filtering is usually done on documents to removesome of the words. A common filtering is stop-words removal. Stopwords are the words frequently appear in the text without having much content information.
Lemmatization: Lemmatization is the task that considers themorphological analysis of the words, i.e. grouping together the variousinflected forms of a word so they can be analyzed as a singleitem.
Stemming: Stemming methods aim at obtaining stem (root) of derived words. Stemming algorithms are indeed language dependent.
String to wordvector: Converts string attributes into a set of attributes representingword occurrence information from the text contained in the strings.

3.6 Clustering:

Clustering is a process of partitioning a set of datainto a set of meaningful sub-classes, calledclusters.The main advantage of clustering is that, it is adaptable to changes and helps single out useful features that distinguish different groups. It is broadly used in market research, pattern recognition, data analysis and image processing. Clustering method can be classified as partitioning method, hierarchical method, density- based method, grid based method, model based method, constrained based method. This paper mainly focusing on partitioning method and hierarchical method.

Partitioning method: The algorithm used for partitioning method are simple K means. The k-means clustering, partitionsn documents in the context of text data into k clusters. Each partition will represent a cluster and k<=n. In k means, each group contains at least one object and each object must belong to exactly one group. The main disadvantage of k-means clustering is that it isindeed very sensitive to the initial choice of the number of k.

k-means clustering algorithm
Input :Document set D, similarity measure S, number k of cluster
Output: Set of k clusters
initialization
Select randomly k data points as starting centroids.
while not converged do
Assign documents to the centroids based on the closest similarity.
Calculate the the cluster centroids for all the clusters.
end
returnk clusters

Hierarchical method: This method creates a hierarchical decomposition of the given set of data objects. There are two approaches here, agglomerative and divisive. Agglomerative approach is also known as bottom up approach, each document is initiallyconsidered as an individual cluster. Then successively the most similar clusters are merged together until all documents are embraced in one cluster. There are three different merging methods for agglomerative algorithms: 1) Single Linkage Clustering: In this technique, the similarity between two groups of documents isthe highest similarity between any pair of documents from thesegroups. 2) Group-Average Linkage Clustering: In group-average clustering, the similarity between two cluster is the average similarity between pairs of documents in these groups. 3) Complete Linkage Clustering: In this method, the similarity between two clusters is the worst case similarity between any pair of documents in these groups.

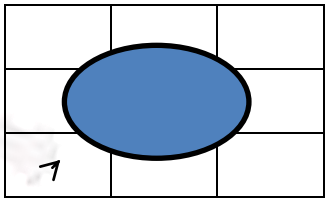
ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 4, April 2018

Similarity measures: The measure reflects the degree of closeness orseparation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters
embedded in the data. In general, similarity/distance measures map the distance orsimilarity between the symbolic description of two objects into a single numeric value, which depends on two factors— the properties of the two objects and the measure itself.This paper compare Euclidean distance and manhattan distance.

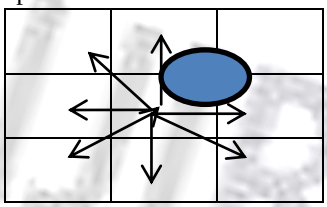
Euclidean Distance: Euclidean distance is a standard metric for geometrical problems.It is the ordinary distance between two points and canbe easily measured with a ruler in two- or three-dimensionalspace. It mainly used in clustering problems,including clustering text.It is the defaultdistance measure in K-means algorithm.Measuring distance between text documents, given two documentsd_a and d_b represented by their term vectorst_a andt_b respectively, the Euclidean distance of the two documentsis defined as

$$D_E(t_a, t_b) = (\sum_{t=1}^{m} |w\,t, a - wt, b|^2)^{1/2}$$

where the term set is T = {t1, . . . , tm}. As mentioned previously, we use the tfidf value as term weights, that iswt,a = tfidf(da, t).



Manhattan Distance: The distance between two points is the sum of the differences of their coordinates.



IV. EXPERIMENTAL RESULTS

This work is implemented in Weka tool. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression,clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.The experimental comparison of clustering algorithms are done based on the Euclidean distance and manhattan distance with execution time, number of iterations, number of clusters, sum of squared errors.

Table 1: Distance Measures for simple k means Algorithms

Algorithms Simple K Means	**Execution time(sec)**	**Numbe of Iterations**	**Sum of squared errors**
Euclidean distance	**0**	**2**	**252**
Manhattan distance	**0.02**	**3**	**291**

Table 2: Distance Measures for hierarchical clustering Algorithms

Algorithms Hierarchical clustering	Execution time(sec)	Numbe of Clusters
Euclidean distance	**0**	**0**
Manhattan distance	**0.01**	**0**

From the experimental result, In simple k means and hierarchical clustering Euclidean distance give the best performance, so it is considered as best distance measures.

V. RESULT AND DISCUSSION

After analysis, the Euclidean distance give the best performance and minimum execution time for both algorithms. So it is considered as best distance measure.

VI. CONCLUSION

In this research work eco hotel document used for text cluster mining. We compare Euclidean distance and manhattan distance based onexecution time, number of iterations, number of clusters, sum of squared errors. From the results, it can be concluded that Euclidean distance give best performance and minimum execution time than manhattan distance.In future better distance measures could also be used to improve the various performance factors and execution time.

REFERENCES

[1]Twinkle Svadas, JasminJha,"*Document Cluster Mining on Text Documents"*,IJCSMC, Vol. 4, Issue. 6, June 2015, pg.778 – 782
[2]AlexandreRibeiroAfonso,CláudioGottschalg Duque"*Automated text clustering of newspaper and scientific texts in brazilianportuguese: analysis and comparison of methods"*, JISTEM - Journal of Information Systems and Technology Management
[3]Vinod S. Badgujar, Asha H. Pawar, *"Search Engine Using Clustering and Text Mining"*, International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 08 | Nov-2015
[4]Anna Huang, "*Similarity Measures for Text Document Clustering*"*NZCSRSC 2008*, April 2008
[5]Mehdi Allahyari,SeyedaminPouriyeh*"A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques",*KDD Bigdas, August 2017,
[6]Vishal Gupta, Gurpreet S. Lehal*"A Survey of Text Mining Techniques andApplications*", journal of emerging technologies in web intelligence, vol. 1, no. 1, august 2009
[7]S. M. Kamruzzaman,FarhanaHaider,AhmedRyadhHasan, "*Text Classification Using Data Mining"*, ICTM 2005
[8]Aurangzeb Khan, BaharumBaharudin*, "A Review of Machine Learning Algorithms for Text-Documents Classification"*,journal of advances in information technology, vol. 1, no. 1, february 2010.
[9] Jincy B. Chrystal,Stephy Joseph,"*text mining and classification ofProduct reviews using structured Support vector machine"*