ISSN 2395-1621

Text Mining-Gathering Vital Information from Textual Data

[#1]Sagar Rajput, [#2]Gaurav Pawar,[#3]Vinay Lokhande

[#4]Prof Aarti Dandavate
1sagar.rajput91@gmail.com
2gauravpawar045@gmail.com
3vinaylokhande16@gmail.com

[#123]Department of Computer Engineering
[#4]Prof. Department of Computer Engineering
Dhole Patil College of Engineering ,Wagholi, Pune

ABSTRACT

Text mining is a procedure of extracting useful information from heaps of textual data. Since a lot of information gets generated in modern world activities like social media,E-mails, texting etc;In modern day world it is however possible to store such a huge amounts of data in low cost disks but extracting useful data is a prime challenge. Using text mining we are able to extract relevant and useful information from textual data as per user's interest. An innovative clustering method named Correlation preserving indexing (CPI) is mentioned in this paper as a solution for finding intrinsic geometrical structure of the document space which is often embedded in the similarities between the documents. It can effectively deal with data with very large size. Effectiveness of this algorithm can be found by number of experiments on various data sets using the existing text mining methods and our proposed method, however with a disadvantage of having computational complexity.K-means algorithm is also used to find low dimensional representation of the documents to decrease computational complexity. Both these algorithms for dimensional representation of the documents are compared on the basis of their accuracy. Using both the above stated algorithms we are able to extract relevant information from huge amounts of textual data thereby saving user's time in manually analyzing huge volume of available data and then extracting meaning information.

ARTICLE INFO

Article History

Received :1st April 2016

Received in revised form :

3rd April 2016

Accepted : 4th April 2016

Published online :

6th April 2016

Keywords— Clustering, k-means, Correlation preserving indexing.

I. INTRODUCTION

Text Mining aims to automatically group related documents (or) text files into clusters it is one of the most important tasks in machine learning and artificial intelligence and has received much attention in recent years. Based on various distance measures, a number of methods have been proposed to handle document clustering. A typical and widely used distance measure is the Euclidean distance. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to find a low dimensional representation of the documents to reduce computation complexity. Low computation cost is achieved in spectral clustering methods, in which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters.

II. LITERATURE SURVEY

Text mining is a burgeoning field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analysing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling even if success is only partial. The phrase "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information.Text mining aims to automatically group related documents into clusters it is one of the most important tasks in machine learning and artificial intelligence and has received much attention in recent years. Based on various distance measures, a number of methods have been proposed to handle text mining. A typical and widely used distance measure is the Euclidean distance. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to find a low dimensional representation of the documents to reduce computation complexity. Low computation cost is achieved in spectral clustering methods, in which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. Latent semantic indexing (LSI) is one of the effective spectral clustering methods, aimed at finding the best subspace approximation to the original document space by minimizing the global reconstruction error (Euclidean distance). However, because of the high dimensionality of the document space, a certain representation of documents usually resides a nonlinear manifold embedded in the similarities between the data points. Unfortunately, the Euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents. Thus, it is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them. An effective document clustering method must be able to find a low dimensional representation of the documents that can best preserve the similarities between the data points. Locality preserving indexing (LPI) method is a different spectral clustering method based on graph partitioning theory. The LPI method applies a weighted function to each pair wise distance attempting to focus on capturing the similarity structure, rather than the dissimilarity structure, of the documents. However, it does not overcome the essential limitation of Euclidean distance. Furthermore, the selection of the weighted functions is often a difficult task. In recent years, some studies suggest that correlation as a similarity measure can capture the intrinsic structure embedded in high-dimensional data, especially when the input data is sparse. In probability theory and statistics, correlation indicates the strength and direction of a linear relationship between two random variables which reveals the nature of data represented by the classical geometric concept of an "angle". It is a scale invariant association measure usually used to calculate the similarity between two vectors. In many cases, correlation can effectively represent the distributional structure of the input data which conventional Euclidean distance cannot explain.

III. PROPOSED SYSTEM:

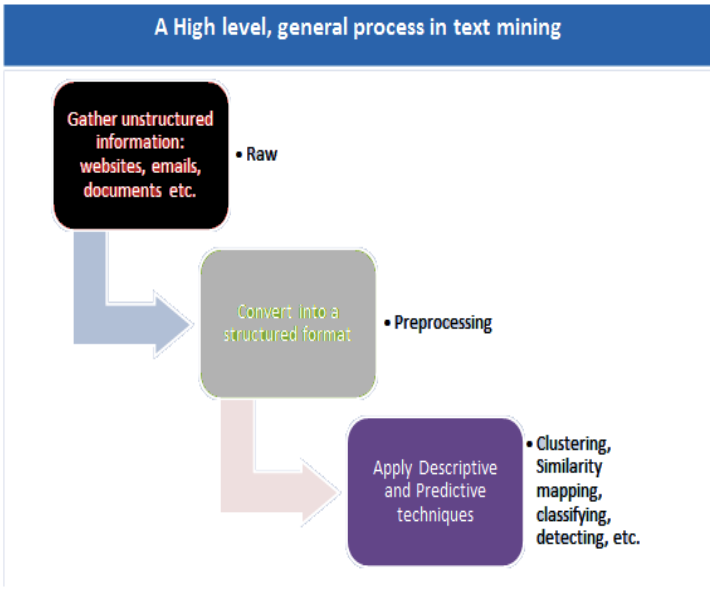

Figure 1

Functional requirements:

The functional requirements for text mining are:
1. Pre-processing
2. Text mining based on Correlation Preserving Indexing
3. K-means on Document sets
4. Classification of Documents into clusters

1. Pre-processing:

A new text mining method based on correlation preserving indexing (CPI), which explicitly considers the manifold structure embedded in the similarities between the documents. It aims to find an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches. This is different from LSI and LPI, which are based on a dissimilarity measure (Euclidean distance), and are focused on detecting the intrinsic structure between widely separated documents rather than on detecting the intrinsic structure between nearby documents. The similarity-measure-based CPI method focuses on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely separated documents.

Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents, CPI can effectively detect the intrinsic semantic structure of the high-dimensional document space. At this point, it is similar to Latent Dirichlet Allocation (LDA) which attempts to capture significant intra-document statistical structure (intrinsic

semantic structure embedded in the similarities between the documents) via the mixture distribution model.

2. Text mining based on Correlation Preserving Indexing:

In high-dimensional document space, the semantic structure is usually implicit. It is desirable to find a low dimensional semantic subspace in which the semantic structure can become clear. Hence, discovering the intrinsic structure of the document space is often a primary concern of text mining. Since the manifold structure is often embedded in the similarities between the documents, correlation as a similarity measure is suitable for capturing the manifold structure embedded in the high-dimensional document space

3. K-means on Document sets:

The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to find a low dimensional representation of the documents to reduce computation complexity.

4. Classification of Documents into clusters:

Online text mining aims to group documents into clusters, which belongs to unsupervised learning. However, it can following side information:

1. If two documents are close be transformed into semi-supervised learning by using the to each other in the original document space, then they tend to be grouped into the same cluster.
2. If two documents are far away from each other in the original document space, they tend to be grouped into different clusters.

Modules:

1. Text mining.
2. Text mining results.
3. Text mining based on CPI.
4. Correlation measure.

Module description:

1. Text mining: Input text file is provided.

2. Text mining results: Produces result of text Mining.

3. Correlation measure: It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches.

4. Text mining based on CPI: It gives document info and cluster info.

IV. ALGORITHM

1. Correlation preserving index
2. k-means

1. Correlation preserving index(CPI):

We present a new text mining method based on correlation preserving indexing. It simultaneously maximizes the correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. Consequently, a low-dimensional semantic subspace is derived where the documents corresponding to the same semantics are close to each other. Extensive experiments on NG20, Reuters and OHSUMED corpora show that the proposed CPI method outperforms other classical clustering methods. Furthermore, the CPI method has good generalization capability and thus it can effectively deal with data with very large size.

2. k-means:

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

V. CONCLUSION

There are many examples of text-based documents (all in electronic format) like e-mails, web pages etc, there is not enough time or patience to read. So in this paper we extract the most required information from text document according to user interest. A new spectral clustering method called correlation preserving indexing (CPI) is used which is performed in the correlation similarity measure space. Here documents reside in local patches and outside these patches. In this method the correlation between the documents is maximized for local patches, and the correlation between the documents is minimized outside these patches. The proposed CPI method can find the intrinsic structures embedded in high-dimensional document space. Also k-means algorithm is used for clustering as specified. Hence we will be able to find important information which can be used for fraud detection for investigation of claims, banking, customer care, news gathering etc.

REFERENCES

[1]. Manu Konchady, Text mining Application Programming, 1st Edition.

[2]. Jiawei Han and Micheline Kamber, Data Mining concepts and techniques, 2nd Edition, Elsevier publisher.

[3]. Text Mining –Finding nuggets in mountains of textual data [pdf file]

[Online]: http://www.mpi-inf.mpg.de/departments/d5/teaching/ss00/proseminar-papers/paperbag-sammelsurium/kdd99-p398-dorre.pdf

[4]. Clustering Techniques

[Online]:http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf

[5] K-Means algorithm

[Online]: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[6]. K-Means

[Online]: http://www.cs.uvm.edu/~xwu/kdd/Slides/Kmeans-ICDM06.pdf

