IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A FRAMEWORK FOR MINING OF TEXT DATA WITH THE APPLICATION OF SIDE INFORMATION

Ms. Neha Tiwari*, Prof. Garima Singh
* Department of CSE W.C.O.E.M. Nagpur.
Department of CSE W.C.O.E.M. Nagpur.

ABSTRACT

Now a days data is not purely available in text form.It also contains a lot of side information ,which may be of different kinds,such as web logs which tells about the user access behavior,links in the document and some non-textual attributes hidden in the text document.This type of document contains the information about the clustering process.But sometime the information is noisy,so it becomes risky to involve side information into the clustering process because it may add noise to the process or may improve the quality of clustering process. So to increase the advantages of using side information we need a principle way to perform clustering process.In this paper,we design an algorithm which increases the efficiency of using side information and to decrease

KEYWORDS: Text Mining,Stemming,Stopping,Clustering,Real dataset.

INTRODUCTION

Data Mining can be defined as the type of database analysis that attempts to extract useful patterns or relationships in a group of data. A major goal of data mining is to extract previously unknown useful relationships among different data.Many Web applications,digital collections and social networks faces the problem of text clustering. As most of the data are available in the form of online collections in information retrieval communities and in database, so this is primarily designed for pure text clustering for real data set. In many applications side information is available with the documents because text documents typically occurs in variety of applications which contains other kinds of database attributes useful for clustering process.But it becomes risky when side information is noisy. Our goal is to determine a clustering in which side-information and text attributes provide similar features about the clusters and ignore which creates conflicts hints about the clusters. Document Clustering of document is a process which includes unsupervised document organization, fast information retrieval and automatic topic extraction.For Example, in web search huge numbers of pages are returned when user enters a query making it difficult for user to browse or extract needed information where as clustering produces results automatically grouped into list of meaningful categories. noise.we have presented an experimental result on real datasets to gain the advantages. Clustering is a division of data into groups of similar objects. Each group, called cluster,consists of objects that are similar between themselves and dissimilar to objects of other groups. Document Clustering is a technique used in unsupervised document organization for identifying clusters or forming group of documents such that the documents in the same cluster are more similar to one another than they are to the documents in other cluster.This technique can be used in information retrieval to automatically categorize large collection of retrieval results by grouping similar type of documents together that helps users browsing of retrieval results.

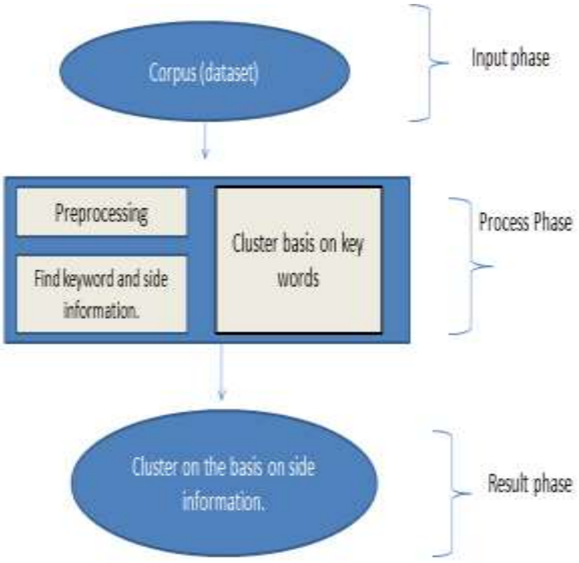

Corpus (dataset)
Input phase
Preprocessing
Find keyword and side information.
Cluster basis on key words
Process Phase
Cluster on the basis on side information.
Result phase

Figure: System Flow

In the figure above we have divided our implementation work into three phases,

1. Input Phase
2. Process Phase
3. Result Phase

The Input phase consist of Corpus(dataset) which is provided to the Process phase where preprocessing takes place.After preprocessing,we used to find the keywords in the dataset and side information is extracted.On the basis of these keywords clusters are formed. This completes the Process phase. At last Clusters are formed on the basis of side information which is done in Result phase.We can use different types of dataset for extracting side information which are as follows:

- The Cora: The Cora data set contains scientific publications in the computer science domain.
- DBLP: A subset extracted from DBLP that contains four data mining related research areas, which are database,data mining, information retrieval and machine learning.
- IMDB data sets: The Internet Movie DataBase (IMDB) is an online collection of movie information.

RELATED WORK

Literature review in the area of Mining indicates that there are several ways of mining text data so that efficient Clusters should be formed and better results should be achieved.

Charu C. Aggarwal, Yuchen Zhao and Philip S. Yu 2014 [1] designed an algorithm which combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach. They then show how to extend the approach to the classification problem. They presented an experimental results on a number of real data sets in order to illustrate the advantages of using such an approach. They presented methods for mining text data with the use of side-information. Many forms of text databases contain a large amount of side-information or meta-information, which might be used in order to improve the clustering process. In order to design the clustering method, They combined an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information. This general approach is used in order to design both clustering and classification algorithms. They presented results on real data sets illustrating the effectiveness of their approach. The results showed that the use of side-information can greatly enhance the quality of text clustering and classification,while maintaining a high level of efficiency.

Charu C. Aggarwal 2006 [2] proposes a method for clustering massive-domain data streams with the use of sketches.They prove probabilistic results which show that a sketch-based clustering method can provide similar results to an infinite space clustering algorithm with high probability. They presented a method for massive-domain clustering of data streams. Such scenarios can arise in situations where the number of possible data values is very large in the different dimensions is very large or the underlying hardware is very space-constrained. They proposed a CSketch algorithm which maintains a high processing rate. Thus, the CSketch algorithm is a robust algorithm, which closely minimizes a clustering algorithm based on exact statistics maintenance while maintaining its computational efficiency across a wide range of parameters.

Charu C. Aggarwal,Yuchen Zhao,Philip S.Yu 2012 [4] explains a first approach of using other kinds of attributes in conjunction with text clustering. They showed the advantages of using such an approach over pure text based clustering.They prove that this approach is especially useful, when the auxiliary information is highly informative, and provides effective guidance in creating more coherent clusters. They design an algorithm which combines probabilistic models with classical partitioning algorithms in order to create an effective clustering approach. They presented experimental results on a number of real data sets in order to illustrate the advantages of using such an approach.

R. Angelova and S. Siersdorfer 2006 [5] focus towards the problem of automatically structuring linked document collections by using clustering. In contrast to traditional clustering, they studied the clustering problem in the light of available link structure information for the data set (e.g., hyperlinks among web documents). Their approach was based on iterative relaxation of cluster assignments, and which could be built on top of any clustering algorithm. That technique results in higher cluster purity, better overall accuracy,and made self-organization more robust.The conducted experiments on three sets of data obtained from the database of scientific papers DBLP, the movie database IMDB, and the online encyclopedia Wikipedia. They proposes two beneficial extensions. They aim to ignore the unnecessary and most probably noisy information behind all irrelevant links in a neighborhood by assigning to each edge e a weight we equal to the cosine similarity between the feature vectors of the documents connected by the edge. They also explore further the hypothesis that neighboring documents should receive similar cluster assignments.

D.Cutting, D. Karger, J. Pedersen, and J. Tukey 1992 [3] explains the Scatter/Gather method which demonstrates that document clustering can be effective information access tool in its own right. They presented a document browsing technique that employs document clustering as its primary operation,they also presented fast clustering algorithms that support this interactive browsing paradigm.

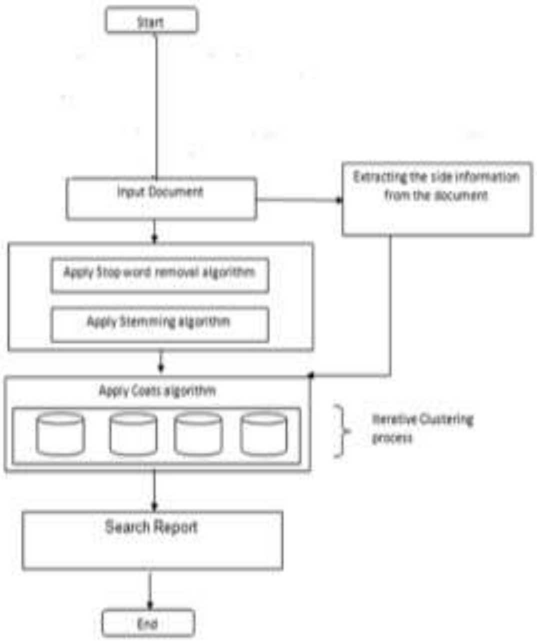
IMPLEMENTATION

In this paper, we are enhancing an approach for clustering text data with side information,here we are using various techniques like Stemming, Stopping and proceeding with clustering process by forming clusters which are formed by using K-means algorithm and finally mining text data by using a modified COATES algorithm, which is modified by using the JACARD function which calculates the minimum distances and helps in forming clusters. It also helps in reducing the calculation of probabilistic approaches and calculates the frequency or probability of the string on its own. It is primarily designed for the problem of pure text clustering, in the absence of other kinds of attributes .It also helps in reducing noise from data.So,that side information can be effectively use for mining the data. We are using a real dataset in this paper. We are also providing an analysis result in this paper.

TECHNOLOGY USED

Stop Word Removal Technique:

Sometimes a very common word, which would appear to be of little significance in helping to select documents matching user's need, is completely excluded from the vocabulary. These words are called "stop words" and the technique is called "stop word removal".The general strategy for determining a "stop list" is to sort the terms by collection


Start
Input Document
Extracting the side information from the document
Apply Stop word removal algorithm
Apply Stemming algorithm
Apply Coats algorithm
Iterative Clustering process
Search Report
End

frequency and then to make the most frequently used terms, as a stop list, the members of which are discarded during indexing.Some of the examples of stop-word are: a, an, the, and, are, as, at, be, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with etc.

Stemming Process
Keyword stemming is the procedure of using a popular keyword and changing it so as to create more hits from search engines and also adding prefix, suffix in the keyword. Keyword stemming is a method using different keyword variation in your article and pages. Keyword stemming is a useful tool for web pages and search engine optimization. The process of keyword stemming involves taking a basic but popular keyword pertaining to a particular website and adding a prefix, suffix, to make the keyword into a new word.

Clustering Algorithm
K-means is one of the simplest unsupervised learning algorithms to group similar data objects. It was developed by J.MacQueen (1967) and then by J.A.Hartigan and M.A.Wong around 1975 K-means forms clusters for n objects based on the attributes into k partitions where k<n. The algorithm starts by partitioning the input points into k initial sets, either at random or using heuristic data. It then calculates the mean point or centroid of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for new clusters, and the algorithm is repeated by alternate application of these two steps until convergence, which is obtained when the points no longer switch clusters. The centroids should be placed in a cunning way as different centroid location provides different results. The main goal using K-means algorithm is to minimize the objective function, shown below

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2.$$

Where $||xi(j) – cj||^2$ is a distance measure between a datapoint xi $^{(j)}$ and cluster center c_j, showing the distance between n data points to their respective cluster centroids. This above equation clearly specifies that clusters are formed by minimizing the distance between the centroid and the data point. The algorithm begins with assigning k centroids choosen randomly in a plane. All the points in the data set are assigned to a centroid that is nearest to it forming clusters.

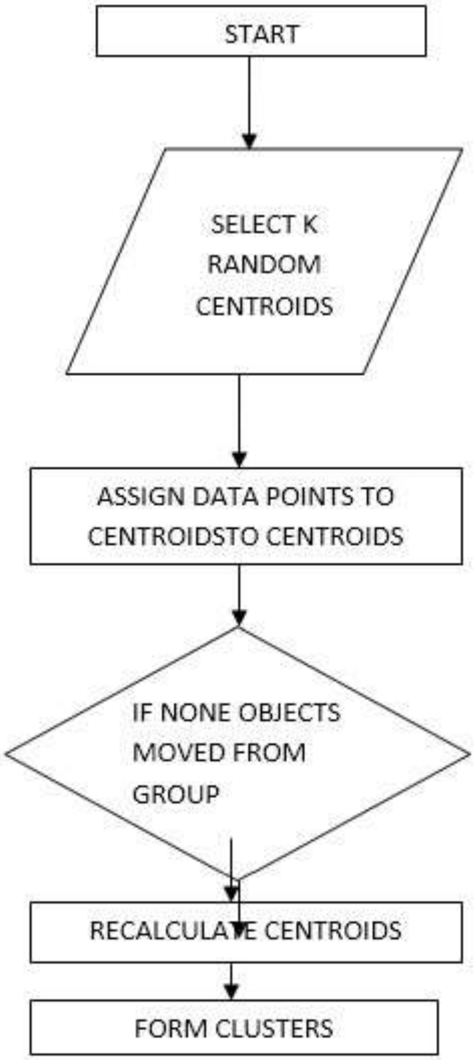

START
SELECT K RANDOM CENTROIDS
ASSIGN DATA POINTS TO CENTROIDSTO CENTROIDS
IF NONE OBJECTS MOVED FROM GROUP
RECALCULATE CENTROIDS
FORM CLUSTERS

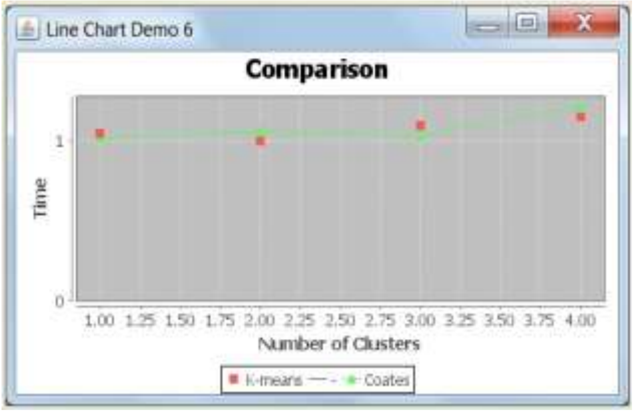
Figure: K-Means Algorithm

COATES ALGORITHM

Here we are using a COATES algorithm for text clustering with side-information.Which corresponds to the fact that it is a COntent and Auxiliary attribute based Text cluStering algorithm. We assume that an input to the algorithm is the number of clusters k. As in the case of all text-clustering algorithms, it is assumed that stop-words have been removed, and stemming has been performed in order to improve the discriminatory power of the attributes. The algorithm requires two phases:

- Initialization:
  We use a lightweight initialization phase in which a standard text clustering approach is used without any side-information.
- Main Phase:
  The main phase of the algorithm is executed after the first phase. This phase starts off with these initial groups, and iteratively reconstructs these clusters with the use of both the text content and the auxiliary information. This phase performs alternating iterations which use the text content and auxiliary attribute information in order to improve the quality of the clustering.
- Analysis phase :

  In analysis phase,we have obtain an efficient result by using COATES algorithm.As shown in graph below there is a comparison between COATES algorithm and K-means algorithm.By applying Jacard function with COATES algorithm information retrieval becomes easy and text data are easily mined with the help of side information and better results are obtained with efficient quality.



REFERENCES

[1] Charu C. Aggarwal, Yuchen Zhao and Philip S. Yu ,"On the Use of Side Information for Mining Text Data", IEEE transactions on knowledge and data engineering, vol. 26, no. 6, June 2014

[2] C. C. Aggarwal and P. S. Yu, "A framework for clustering massive text and categorical data streams," in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.

[3] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, "Scatter/Gather:A cluster-based approach to browsing large document collections," in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.

[4] C.C. Aggarwal and P.S.Yu,"On text clustering with side information," in Proc. IEEE ICDE Conf., Washington, DC, USA, 2012.

[5] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in Proc. CIKM Conf., New York, NY, USA, 2006, pp. 778–779.

[6] "Effects of similarity metrics on document clustering"2009 UNLV Theses/Dissertations/Professional Papers/Capstones