ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)
Vol. 2, Issue 4 (Oct. - Dec. 2015)
International Journal of Advanced Research in Education & Technology (IJARET)

Mining Text Data with Auxiliary Attributes in Contents

[I]Greeshma RG, [II]Smitha ES
[I]P.G. Scholar, [II]Associate Professor
[I]Dept. of CSE, LBSITW, Poojappura,Thiruvananthapuram
[II]Dept. of IT, LBSITW, Poojappura, Thiruvananthapuram

Abstract
In the digital world there is tremendous growth of digital information. This information also contains side information with it. The side information is the auxiliary information which may also be useful. They are elements within a document which are not part of main body text. This side-information can be like document provenance information, links existing inside the document etc.. Such attributes will contain remarkable amount of information for clustering purposes. In this work links are treated as side attributes or auxiliary attributes. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Usually it is difficult to estimate the importance of this side-information when they are noisy. In these scenarios, there is a huge amount of risk involved in incorporating this side-information into the mining process, since they can add noise to the process rather than improving the quality of the mining process. Gini index is used as the feature selection method to filter the informative side-information from text documents. A standard way to perform the mining process, so that to make best use of the advantages based on this side information is needed. Here an algorithm is proposed to create an effective clustering approach, based on the combination of traditional partitioning algorithms with probabilistic models. A probabilistic model on the side information uses the partitioning approach.

Keyword
Classification, Clustering, Data mining, Side information, Text Mining.

I. Introduction

Data mining is the exercise of automatically searching large stores of data to discover patterns and trends with simple analysis. Many researchers have used techniques such as classification, outlier detection, clustering, regression analysis etc. The clustering is used some special application. Clustering is mechanisms of combining set of physical or abstract objects into classes of similar objects. There are different orders or groups which is called cluster, subsist of objects that are correlated within themselves and unrelated to objects of other order or groups. Text mining is the discovery of new, previously unknown information by automatically extracting information from different written resources. Text mining is a variation on data mining that practically find out interesting patterns from large databases. The use of digital information is increasing day by day. Since increasing the amount of information it needs to extract relevant information from this huge amount of data for text mining. This proves to a reason in creating scalable and efficient mining algorithms. The clustering of data in the pure form is done till now. But to manage such large quantity of data we require indexing the data according to the users need. Large number of web documents contains side information. This side information can be sometimes called meta-data. These meta-data are exactly matching to the various different kinds of attributes such as the origin or other information related to the origin of the document. Data such as location, possession or even temporal information may prove to be informative for mining purposes in other cases. [9] Documents may be linked with user-tags in many network and user-sharing applications. This may also be quite informative for doing effective text mining. The process of deriving high quality information from text is known as text data mining. The side-information can sometimes provide useful information for improving the quality of clustering process, but when the side-information is noisy it can be a risky approach. A method is used to discover the coherence of clustering characteristics of side information with the text content and at the same time reject those aspects in which incompatible clues are provided.A probabilistic model on the side information uses the partitioning information for the purpose of estimating the coherence of different clusters with side attributes. In this work the links in the documents can be treated as side attributes.

II. Literature Survey

Text clustering becomes a problem in many application domains due to the increasing amount of unstructured data. A general survey of text clustering algorithm can be found in [1]. In [2] major fundamental clustering methods are discussed. These methods can be classified as partitioning methods, hierarchical methods, density based methods and grid based methods [3]. K-means and K-medoids methods come under partitioning methods. They are distance based methods. In Distance-based Clustering Algorithms there is a use of similarity function which measures the closeness between the text objects takes place. The most well-known similarity function which is used commonly is the cosine similarity function. Hierarchical methods can be classified as agglomerative and divisive methods. In these methods clustering is a hierarchical decomposition.The general concept of agglomerative clustering is to successively merge documents into clusters based on their similarity with one another DBSCAN, DENCLUE are some of density based methods. With density based methods we can find arbitrarily shaped clusters. CLIQUE is one of the grid based methods. These methods use a multi-resolution grid data structure.

In [4] the cosine similarity to find the similarity of documents is explained. The documents can be represented as vectors. To compute the similarity between two vectors the following steps are used.

- Consider two vectors (say A and B).
- Take the union of those vectors.
- Find the dot product of vectors A and B.
- Calculate the magnitude of vector A and B.
- Multiple the magnitudes of A and B.
- Divide the dot product of A and B by the product of the magnitudes of A and B.

ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

$$\text{Similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

In [5] C.C.Aggarwal explained about the different clustering algorithms. Data classification is a two-step process consisting of learning step and classification step. In learning step a classification algorithm builds the classifier by learning from a training set. In classification step a model is used to predict class labels for given data. Some methods which are commonly used for text classification are as follows. Decision trees, Rule based classifier, SVM classifiers, Bayesian classifier etc. Decision tree is a hierarchical decomposition of training set in which a condition on attribute value is used in order to divide the data space hierarchically. In rule based classifier a set of rules are used to model the data space. SVM classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. Neural network classifiers are related to SVM classifiers, both are in the category of discriminative classifiers. Bayesian classifiers build a probabilistic classifier based on modelling the underlying word features in different classes.

In [6] a scatter-gather technique is discussed. It is a cluster based approach to browse large document collections. It uses document clustering as its primitive operation. Here initially system scatters the collection into clusters and present short summaries to the user. Based on the summaries one or more groups are selected. The selected groups are gathered to form sub collection. This is an iterative process. Here partitioning methods are defined to partition the collection into clusters. Buckshot and fractionation algorithms are used to find initial clusters. Buckshot is a fast clustering algorithm needed for re-clustering. Fractionation is a clustering algorithm with great accuracy.In [7] a co-clustering approach for documents and words is explained. Here documents and words are clustered simultaneously. The document collection can be represented as a word by document matrix. This word by document can then be represented as a bipartite graph. The dual clustering problem is done in terms of finding minimum cut in bipartite graph. A spectral algorithm is used to solve the partitioning problem.High dimensionality of feature space is a challenge for clustering algorithms [8].

Feature extraction and feature selection techniques are used to reduce feature space dimensionality. In feature extraction it extracts a set of new features from original features through some functional mapping. In feature selection it chooses a subset from the original feature set according to some criteria. Document frequency, information gain, term strength are some of the feature selection methods. Unsupervised feature selection methods are much worse than supervised feature selection. In order to utilize the efficient supervised method an iterative feature selection method that iteratively performs clustering and feature selection is proposed in this paper.

III. Side Information

Side information's are elements within a document which are not part of main body text. Such side-information may be of different kinds, such as document provenance information, the links in the document, user-access behavior from web logs, or other non-textual attributes which are embedded into the text document. Such attributes may contain a tremendous amount of information for clustering purposes. Here links in the documents are treated as side information.The relative importance of this side-information may be di cult to estimate, especially when some of the information is noisy. The relative importance of side information is difficult to cluster when information is complex. Sometimes side information may be useful in improving the quality of the clustering process. It can be a risky approach when the side information is noisy. It will either improve the quality of clustering or it may worsen the quality of mining process. So an approach is needed to determine a clustering in which text attributes and side information provide similar hints about the nature of underlying clusters and at the same time ignore those aspects in which conflicting hints are provided. In the proposed method, an algorithm which combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach. A probabilistic model on the side information uses the partitioning information for the purpose of estimating the connection of different clusters with side attributes.

IV. System Analysis

Let us denote S as corpus of text documents. The total number of documents is N, and they are denoted by T1 . . . TN. It is assumed that the set of distinct words in the entire corpus S is denoted by W. Associated with each document Ti, a set of side attributes can be represented by Xi. Such attributes are referred as auxiliary attributes. 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.

Text Pre-processing

Mining from a pre-processed text is easy as compare to natural languages documents. So, it is important to pre-process the text before clustering. To reduce the dimensionally of the documents words, special methods such as filtering andstemming are applied. Filtering methods remove those words from the set of all words which are irrelevant. Stop word filtering or stop words removal is a standard filtering method. Words like conjunctions, prepositions, articles, etc. are removed. Stemming is a technique used to find out the root/stem of a word. Stems are thought to be useful for improving retrieval performance because they reduce variants of the same root word to a common concept. For example, consider the words user, users, used, and using. The stem of these words is use. Similarly the stem of words engineering, engineered, and engineer is engineer. Furthermore, stemming has the secondary effect of reducing the size of the indexing structure because the number of distinct index terms is reduced. This can be done as follows.

- if a word ends with a consonant other than *s*, followed by an *s*, then delete *s*.
- if a word ends in *es*, drop the *s*.
- if a word ends in *ing*, delete the *ing*unless the remaining word consists only of one letter or of *th*.
- If a word ends with *ed*, preceded by a consonant, delete the *ed*unless this leaves only a single letter.

Many of the most frequently used words in English are worthless in text mining. These words are called stop words. For example the, of, and, to, etc. in stop word removal the stop words such as the , to etc. are removed.

Partitioning Clustering

The simplest and most fundamental version of cluster analysis is

partitioning method. It organizes the objects of a set into several exclusive groups or clusters. All objects are partitioned so that no hierarchy exists among the clusters. For a given dataset D, of n objects and k, the number of clusters to form. Then a partitioning algorithm organizes the objects into k partitions (k<=n) where each partition represents a cluster. There are two methods used in partitioning approach. K means and k medoids approach. K-Means is probably the most widely used clustering technique. K means can be done using the following steps. Input: K (number of clusters). Output: clusters. The steps involved are as follows.

1. The algorithm randomly selects k points as initial cluster centres.
2. Each point from the dataset is assigned to the closest cluster based upon the cosine similarity among each point and each cluster center.
3. Each cluster center is then recomputed as the average of points in that cluster.
4. Step ii and iii are repeated until the clusters are formed.

Probabilistic Model Based Clustering

For using the probabilistic method in this approach here it evaluates the conditional probability of X given Y. This probability tells us how likely it is to observe concept X provided that we already observed instance Y. The output of this is the calculation of the probability that an observed instance Y belongs in concept X, which is also known as the posterior probability and denoted by P(X/Y). The calculation is performed based on the following formula:

P(X/Y) = (P(Y/X).P(X)).P(Y) (1)

COATES Clustering Process

COATES means COntent and Auxiliary attribute based TExtcluStering algorithm. It is assumed that the input to this algorithm is the number of clusters k. The process is started after stop-words are removed and stemming has been performed.This is done to improve the discriminatory power of attributes. Once all the pre-processing is per-formed then retrieval of side information from the documents takes place. The links in the documents are considered to be side information in the documents. For the clustering process, the algorithm requires two phases.They are as follows:

- First phase: In the first phase the clustering is performed on the pure text data from the documents without making any use of the side information.
- Second phase: After the completion of the first phase the next phase started. This is the main phase of the algorithm. In the second phase the re-clustering is done with the help of text content as well as the side information.

In the first phase clusters are formed with the use of content based iteration. This phase simply construct an initialization which provides a good starting point for the clustering. This can be done by using k-means algorithm in the partitioning method. The next is auxiliary based iteration. Before starting auxiliary iteration calculation of gini index is

Table 1: Precision and Recall for document processing

Files	tp	tn	fp	fn	Precision	Recall	Accuracy
50	37	7	4	2	0.902	0.948	0.88
100	70	20	6	4	0.921	0.945	0.90
150	108	33	6	4	0.947	0.964	0.94
200	160	33	2	5	0.988	0.968	0.96

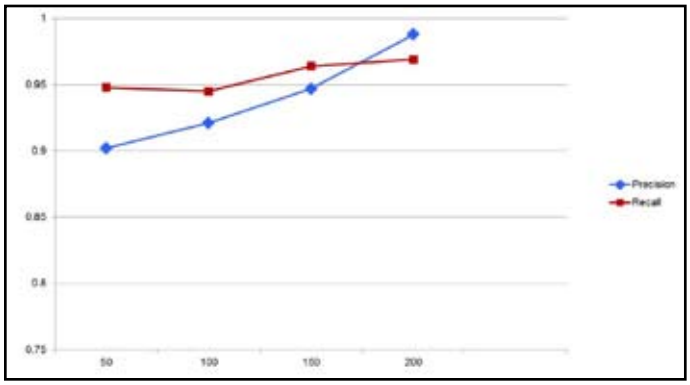

Fig. 1: Graph showing precision and recall.

needed ie; gini index of each attribute based on the clusters created by the last content based iteration is to be calculated.

$$Gr = \sum_{j=1}^{k} prj^2 \qquad (2)$$

To construct a probabilistic model of membership of the data to the clusters, prior and posterior probabilities has to be computed. The prior probability that the document Ti belongs to the cluster Cj can be denoted P(Ti Cj).To compute the posterior probabilities P(Ti Cj/ Xi) of membership of a record at the end of the auxiliary iteration, use the auxiliary attributes Xi which are associated with Ti.
The output of this process after all this steps is clusters based on side information. Then feedback and time frequency concept is used.The count of search and download is taken for faster retrieval of information Here it uses an indirect feedback. Based on the download made by users the documents are prioritized. This is done by using a counter. Then at next time of search, the most downloaded documents are prioritized as high in listing.

V. Result

The dataset used is the 20 News-groups dataset. This dataset comprise of 20000 text documents taken from 20 news-groups, this news gathering is put away in subdirectory, with each one article put away as a different document. To evaluate the accuracy of our clustering algorithm, Recall and Precision performance metrics can be calculated. The value of precision and recall can be calculated as:

Precision = tp/(tp+fp) (3)
Recal l= tp/(tp+fn) (4)

tn/True Negative :case was negative and predicted negative.
tp/True Positive : case was positiveand predicted positive.
fn / False Negative : case waspositive but predicted negative.
fp / False Positive: case was negativeand predicted positive.

VI. Conclusion

The increasing amount of text data in large collections led to the

ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

creation of scalable and efficient mining algorithms. Many works has been done by considering the text clustering problem. However all these works deals with pure text clustering. It does not consider other kinds ofattributes. In many applications tremendous amount of side information can be seen. Sometimes this side- information may be noisy and it worsens the quality of clustering. So this work needs a way for performing the mining process, so that the advantages from using this side information can be maximized. This work needs to use an approach which carefully discovers the connection of the clustering characteristics of the side information with that of text content. In order to design the clustering method, combined an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information.

VII. Acknowledgement

I have taken efforts in this review of clustering for mining using side information. However, it would not have been possible without the kind support and help of many individuals. I am highly indebted to Associate Prof. Smitha ES for her guidance and constant supervision as well as for providing necessary information regarding this approach.

References

[1] *C. C. Aggarwal and C.-X.Zhai, "Mining Text Data", New York, NY, USA: Springer, 2012.*
[2] *J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd ed., Elsevier, Morgan Kaufmann, 2006.*
[3] *A. Jain and R. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.*
[4] *G. Salton, "An Introduction to Modern Information Retrieval. London", U.K.: McGraw Hill, 1983.*
[5] *C. C. Aggarwal and C.-X.Zhai, "A survey of text classification algorithms," in Mining Text Data.New York, NY, USA: Springer, 2012.*
[6] *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.*
[7] *I. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in Proc. ACM KDD Conf., New York, NY, USA, 2001, pp. 269–274*
[8] *T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering," in Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488–49*
[9] *C. C. Aggarwal, S. C. Gates, and P. S. Yu, "On using partial supervision for text categorization," IEEE Trans. Knowledge. Data Eng., vol. 16, no. 2, pp. 245–255, Feb. 2004.*
[10] *H. Schutze and C. Silverstein, "Projections for efficient document clustering," in Proc. ACM SIGIR Conf., New York, USA, 1997, pp. 74–81.*
[11] *Domingos, P. and M.J. Pazzani, 1997. "On the optimality of the simple Bayesian classifier under zero-one loss," Mach. Learn., 29(2-3): 103-130.*