ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

An Efficient Way to Mine Text Data with Side Information

M Mintu

M.Tech P.G. Scholar, Department of CSE, MIT Anjarakandy, Kannur, Kerala, India

ABSTRACT: We need to find correlatios among dozens of fields in large relational databases, for this we use data mining techniques.Retrieving information from text is referred to as text data mining. So many text documents contain side information. Links in the document, user access behavior from web logs etc. are examples of side information. Such side information which are embedded in text document may contain vast amount of information about clustering. It may be hard to determine the importance of side information when some of the data are erroneous. So this side information must be used in a systematic way in mining process so as to maximize the benefit from using this side information. It can enhance the efficiency of clustering. After effective clustering, this approach can be extended to classification. In this paper we propose a technique to solve problems in existing clustering techniques based on side information.

KEYWORDS: Data Mining, Text Mining, Side information, Clustering, Classification.

I. INTRODUCTION

A vast majority of data are stored in document and they are increasing day by day. So text mining is a technique used to extract information from text documents. The text clustering becomes an issue in many type of application domains such as web, social network etc. However in many applications a vast amount of side information is embedded with the documents. Side information provide a huge amount of information that can enhance clustering process. The problem of text clustering has been studied in [6], [7]. Clustering is a technique for automatically organizing a large collection of text. In the context of a number of applications text document contains a large amount of meta information which may support to the clustering process.
Some examples of such side information are:

- User access behaviour
  Keep a web log based on user access behaviour and based on that behaviour data mining process can be implemented. Such logs can help to enhance the quality of the mining process.
- Links in text documents
  Various text documents having links in them. It can also be treated as attributes. It contains information of data mining process.

Charu C Aggarwal et al. refers examples of side attributes [1] as:

- In a web log analysis application, x_{ir}maps to the 0-1 variable, which identifies whether or not the i^{th} document has been accessed by the r^{th} user.
- In a network application, x_{ir}maps to the 0-1 variable which identifies whether or not the i^{th} document $T_i$has a hyperlink to the r^{th} page T_r.

In network application they used BAA (Binary Auxiliary Attribute). So through this they can attain only yes or no values. Also the purity of cluster is below 50% in CORA and IMDB data set. So in this paper, we are proposing solution for these problems.

II. RELATED WORK

A general survey of clustering algorithm is specified by Jain and Dubes [4]. The problem of clustering has also been found in the context of text data. Cluster analysis is the study of algorithms and methods for classifying objects. The cluster analysis aims to find a convenient and valid organization of the data. Clustering algorithms are geared toward finding structure in the data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Guha et al. [5] studied clustering algorithms for data with boolean and categorical attributes. They proposed a novel method of links to measure the similarity between a pair of data points. They developed a robust hierarchical clustering algorithm ROCK (RObust Clustering using linKs) that employs links and not distances when merging clusters.

Detailed surveys on text classification has been defined by Aggarwal and Zhai [3]. The problem of classification has been included in data mining, machine learning, database, and information retrieval communities. In the case of text data, the problem is similar to that of discrete set-valued classification attributes, when the frequencies of the words are ignored. Therefore, text mining techniques must be designed to efficiently manage large numbers of elements with varying frequencies. Almost all the common techniques for classification such as decision trees, rules, Bayes methods, have been extended to the case of text data.

A survey of text clustering methods has been identified by Aggarwal and Zhai [2]. The clustering problem is defined as finding groups of similar objects in the data. Similarity function is used to measure similarity between the objects. The clustering problem can be very helpful in the text domain, where the objects to be clusters can be of different partitions such as documents, paragraphs, sentences or terms. A good text clustering requires effective feature selection and a proper choice of the algorithm.

One of the most well-known methods for text-clustering called the Scatter/Gather technique defined by Cutting et al. [7], uses a collection of agglomerative and partitional clustering. First of all the system scatters the collection into a small number of document clusters. Based on these, the user selects one or more of the clusters. The selected clusters are clustered together to form a subcollection. It is clustered again to scatter subcollection into a small collection of document groups which are again presented to the user. The groups become smaller for each successive iteration. Other similar methods for text-clustering which use same methods are discussed by Schutze and Silverstein [6]. An easy method to speed up clustering is to speed up the distance calculations at the clustering routines. They have shown that projecting documents via LSI and truncation offers a dramatic advantage over full profile clustering in terms of time efficiency. The improved efficiency, is not accompanied by a loss of cluster quality.

While studying the existing paper [1], we happen to find that side attributes are BAA. So we attain only yes or no values.

III. PROPOSED SYSTEM

A. *Methodology:*

While studying the existing paper, we happen to find that side attributes are BAA. So we attain only yes or no values. So we try to propose a new method to retrieve access count also.

Also we find that purity of cluster is below 50% in Cora and IMDB data sets. So we analyse the equation to findout purity,

$$P=\frac{\sum_{i=1}^{k} c_i}{\sum_{i=1}^{k} n_i} \qquad \text{eq. (1)}$$

Where c_i is the input clusters and n_i is the number of data points.

So from the equation, we identify that purity is proportional to sum of input clusters and inversely proportional to sum of number of data points.

B. *Problem Definition:*

1. Charu C Aggarwal et al. in [1] completed their work based on binary attribute values. In web log analysis application, they accept that x_{ir} relates to the 0-1 variable, which shows whether the i^{th} document has been accessed by the r^{th} user.
2. In a network application, they assumed that x_{ir} corresponds to the 0-1 variable corresponding to whether or not the i^{th} document T_i has a hyperlink to the r^{th} page T_r.
3. The cluster purity always lies between 0 and 1 [1]. Cluster purity 1 represents perfect clustering whereas a poor clustering will provide very low value of the cluster purity.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

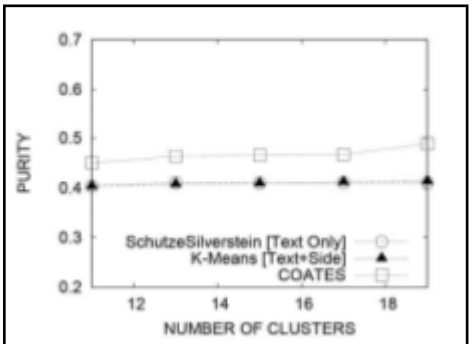


Fig.1. Effectiveness Vs. Number Of Clusters (CORA)

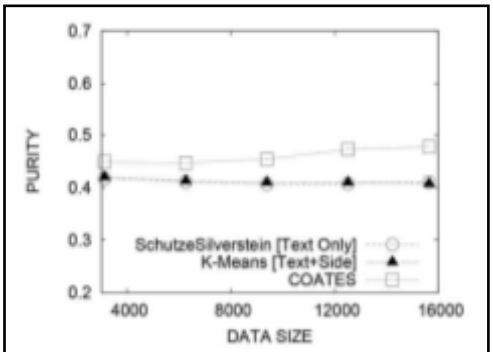
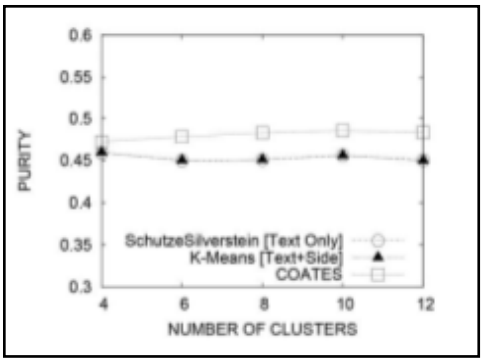


Fig.2. Effectiveness vs. data size (CORA)



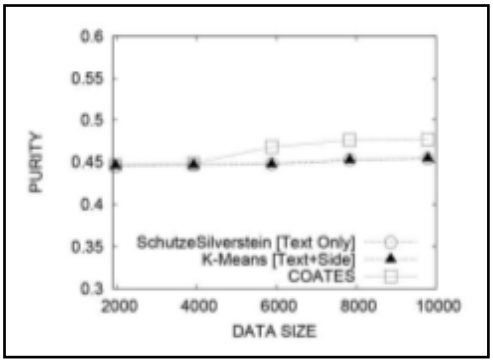


Fig.3. Effectiveness Vs. Number Of Clusters (IMDB) Fig.4. Effectiveness vs. data size (IMDB)

While analysing the graph, we can find purity for CORA and IMDB lies between 0.4 to 0.5. There lies only below 50% purity.

C. *Proposed System*

1. By considering numerical auxiliary attribute value instead of BAA. In web log analysis application, 0's of BAA can be replaced by 0 and 1's by value greater than 0 in numerical auxiliary attribute.
2. In the case of network application, 0's of BAA can be replaced by -1, 1's of BAA can be spited as values 0(for links that are not accessed) and values greater than 0(for links that are accessed) denotes access count. Here -1 and 0's are considered as outliers.
3. Purity P is the ratio of sum of input clusters to the sum of number of data points(eq. (1)). We can increase the purity by reducing data points and increasing sum of input clusters. The number of data points can be reduced by moving more points to the outliers as from above solution.

IV. EXPECTED RESULTS

In proposed system, the side attribute were taken as numerical auxiliary attribute value. We can cluster our data more efficiently using numerical auxiliary attribute value. The existing issues were solved and we get:

- Decrease in data points

The number of data points can be reduced by moving more points to the outliers.

- Increase in cluster purity

Purity P is proportional to sum of input clusters and inversely proportional to the sum of number of data points (eq. (1)). We can increase the purity by reducing data points and increasing sum of input clusters.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

V. CONCLUSION

We identified two problems in existing system. First relates to the user access behavior in network application. Second relates to the purity of clustering. In this paper we proposed solution for these problems. By implementing the proposed solution we can improve efficiency and performance of clustering process.

REFERENCES

1. Aggarwal, C. C., Zhao, Y., and Yu, P. S., "On the Use of Side Information for Mining Text Data" , IEEE Transactions on knowledge and data engineering vol 26,no.6 pp 1415-1429, 2014
2. Aggarwal, C. C., and Zhai, C. X., "Mining Text Data ". New York, NY, USA: Springer, 2012.
3. Aggarwal, C. C., andZhai, C. X., "A survey of text classification algorithms,"in Mining Text Data. New York, NY, USA: Springer, 2012.
4. Jain,A., and Dubes, R., "Algorithms for Clustering Data". Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.
5. Guha, S., Rastogi, R., and Shim, K., "ROCK: A robust clustering algorithm for categorical attributes", Inf. Syst., vol. 25, no. 5,pp. 345-366, 2000.
6. SchutzeH., and Silverstein, C., "Projections for efficient document clustering", in Proc. ACM SIGIR Conf., New York, NY, USA, pp. 74-81, 1997.
7. Cutting, D., Karger,D., Pedersen, J., and Tukey, J., "Scatter/Gather: A cluster- based approach to browsing large document collections", in Proc. ACM SIGIR Conf., New York, NY, USA, pp. 318-329, 1992.

BIOGRAPHY

M Mintuis an M Tech PG Scholar in Department of Computer Science and Engineering, Malabar Institute of Technology, Anjarakandy, Kannur University. Her research interests is in Data Mining (Text Mining).