ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Mining Text Document on the Utilization of Side Information

Shekhar S. Patil[1], R. B. Wagh[2]

P.G. Student, Department of Computer Engineering, R.C.Patel Institute of Technology, Shirpur, Maharashtra, India[1]

Associate Professor, Department of Computer Engineering, R.C.Patel Institute of Technology, Shirpur, Maharashtra, India[2]

ABSTRACT:In many text mining applications, side information is accessible forth with the text documents. Such side information mightbe of various types, like document provenance information, the hyperlinks inside the document, user get right of entry to behavior from web logs, or added non-textual attributes which are embedded into the text document. Such attributes might contain an large amount ofinformation for clustering purposes. However, the about accent of this side-information is also troublesome to estimate, abnormally if some of the information is noisy. It will be problematic to include side information into the mining process, because it can either advance the superior of the representation for the mining process, or will add noise to the process. Therefore, we need to proper way to perform the mining process,therefore on maximize the benefits from using this side information. Here we have presented the improved clustering techniques, where they combined an iterative partitioning technique..

KEYWORDS:Data Mining, Clustering, Side Information

I. INTRODUCTION

The problem of text clustering arises within the context ofmany application domains like the social networks,online, and different digital collections.The speedily increasingamounts of text information within the context of those large on-linecollections has led to associate in nursing interest in making effective mining andscalable algorithms.An incredible quantity of workhas been done in recent years on the matter of clusteringin text collections [7], [15], [17], [18] within theinformation retrieval communities and database.However,this work is primarily designed for the matter of pure textclustering, within the absence of different kinds of attributes. Inmany application domains, an incredible quantity of side informationis also associated along side the documents.This is as a result of text documents usually occur within the contextof a range of applications in which there could also be alarge amount of different kinds of meta information or database attributes which may be helpful to the clustering process. Following are the example of such side information:
In an application within which we have a tendency to track user access behavior of internet documents, the user-access behaviormay be captured within the type of web logs. for every document, the meta information might correspond tothe browsing behavior of the various users. Such logs may be wont to enhance the standard of the mining process in an exceedingly means that is a lot of significant to the user, and additionally application sensitive. This is often as a result ofthe logs will typically obtain refined correlations in content that cannot be picked up by the raw textalone.

Many text documents contain links among them,which can even be treated as attributes. Such linkscontain a lot of helpful data for mining purposes. As within the previous case, such attributes mightoften give insights regarding the correlations amongdocuments in a very approach which cannot be simply accessiblefrom raw content.

Many web documents have meta-data related to them that correspond to totally different kind of attributes just like the source or alternative data concerning the origin of the document. In othercases, data like possession, location, or

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

may be temporalinformation is also informative for miningpurposes. During a variety of network and user-sharingapplications, documents are also related touser-tags, which can even be quite informative.

While such side information will generally be helpful in improving the standard of the cluster method, it is a risky approach once the side information is noisy. In such cases, it will truly worsen the standard of the mining process. Therefore, we'll use an approach that carefully ascertains the coherence of the cluster characteristics of the aspect information thereupon of the text content. This helps in magnifying the cluster effects of each form of data. The core of the approach is to see a clustering in which the text attributes and side information offer similar hints regarding the character of the underlying clusters and at a similar time ignore those aspects within which conflicting hints square measure provided.

In order to realize this goal, we are going to combine a partitioning approach with a probabilistic estimation process that determines the coherence of the side attributes within the clustering process. A probabilistic model on the side information uses the partitioning data (from text attributes) for the aim of estimating the coherence of various clusters with side attributes.This helps in abstracting out the noise among the membership behavior of various attributes.The partitioning approach is specifically designed to be very efficient for big data sets. This may be necessary in situations during which the data sets are very large. We are going to present experimental results on real data set and illustrate the effectiveness and efficiency of the approach.
While our primary goal during this paper is to study the clustering problem, we tend to note that such an approach willalso be extended in essence to alternative data mining issues during which auxiliary information is available with text. Such scenarios are very common during a wide variety of data domains. Therefore, we'll additionally propose a way during this paper so as to extend the approach to the problem classification. We will show that the extension of the approach to the classification problem provides superior results due to the incorporation of side information. Our goal is to indicate that the benefits of using side information extend beyond a pure clustering task, and can give competitive benefits for a wider variety of problem scenarios.

II. RELATED WORK

The Algorithm design which combine combines classical partitioning algorithms with probabilistic models [2]. Shekhar S. Patil and R. B. Wagh are presented a survey of mining text data using side information may be found in [1]. The problem of text content clustering has been studied extensively by using the database community [9], [16], [21]. The main focus of this work has been on scalable clustering of multi- dimensional data of different kinds [9], [10],[16], [21]. A. Jain and R. Dubes are perform a standard survey of clustering algorithms may be observed in [11]. The trouble of clustering has additionally been studied quite extensively inside the context of text document. C. C. Aggarwal and C. X. Zhai are shows a survey of text content clustering methods may be found in [5]. D. Cutting, D. Karger, J., et. al are focused on one of the most well-known techniques for text clustering is the scatter gather method [15], which uses a mixture of agglomerative and partition clustering. H. Schutze and C. Silverstein are studied on different related strategies for text clustering which use similar strategies are discussed in [17]. T. Liu, S. Liu, Z. Chen, et.al has been proposed an Expectation Maximization (EM) method for text clustering in [12]. W. Xu, X. Liu, and Y. Gong are proposed Matrix Factorization techniques for text clustering in [20]. This technique selects phrases from the report based on their relevance to the clustering technique, and makes use of an iterative EM method with the intention to refine the clusters. A. Banerjee and S. Basu are studied on a intently associated area is that of topic modeling and occasion or event tracking [14]. C. C. Aggarwal, S. C. Gates, and P. S. Yu are studied on which may be in textual categorization[13].H. Frigui and O. Nasraoui are discussed the methods for text clustering within the text context of keyword extraction in [8]. M. Steinbach, G. Karypis, et. al can be discovered a comparative take a look at of different clustering methods in [18].
C. C. Aggarwal and P. S. Yu are focused on the problem of text clustering has additionally been studied in context of scalability in [7]. However, all of these techniques are designed for the case of pure text data, and do now not work for instances in which blended with different varieties of data. A few restrained works has been performed on clustering textual content inside the context of network based linkage information [3], [4], [19] although this work isn't applicable to the case of general side information attributes. On the basis of survey, this can provide a first technique to the use of

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

different kind of attributes along with text clustering. We will show the blessings of using such a method over pure textual content primarily based clustering. Such an technique is in particular useful, while the auxiliary information is enormously informative, and provides powerful guidance in developing greater coherent clusters. We are able to also increase the approach to the problem of textual content classification, which has been studied appreciably inside the literature. Distinctive surveys on text classification can be located in [6].

III. PREVIOUS METHODOLOGY

There was presented methods for mining text data with the use of side-information. Some of the types of text databases contain a large amount of side information or Meta information, which can be used in order to improve the clustering process. In order to design the clustering method, there were combined an iterative partitioning technique with a probability estimation method that computes the importance of different kinds of side information. This general approach is used in order to design both clustering and classification algorithms. That was also present results on real data sets illustrating the effectiveness.
We discuss with this algorithm as COATES throughout the paper, which corresponds to the truth that it's far a COntent and Auxiliary based TExtual content cluStering algorithm. We anticipate that an enter to the algorithm is the range of clusters k. As within the case of all text clustering algorithms, it's miles assumed that stop word were removed, and stemming has been achieved in order to improve the discriminatory power of the attributes. This algorithm has working into 2 phases:

1. Initialisation Phase

We use a lightweight initialization section wherein a fashionable text content clustering technique is used without any side-information. For this motive, we use the algorithm defined in [15]. The motive that this algorithm is used, because it's far easy algorithm that can quick and efficiently provide an affordable preliminary start line. Centroids and the partitioning created with the aid of the clusters shaped inside the first section provide an initial starting point for the second phase. We are aware that the first segment is based on text content only, and does no longer use the auxiliary data.

2. Main Phase

The main phase of the algorithm is executed after the primary phase. This phase starts off evolved off with those preliminary companies, and iteratively reconstructs those clusters with the usage of each the text content and the auxiliary information. This phase performs alternating iterations which use the textual content and auxiliary attribute information in an effort to improve efficiency of the clustering. We call these iterations as content iterations and auxiliary iterations respectively. The combination of the 2 iterations is known as a major iteration. Each main iteration hence carries two minor iterations, similar to the auxiliary and text based methods respectively.

The initialization phase and main phase are working in the fig. 1. The initialization phase is lightweight phase. Main phase of algorithm are execute after completion of initialization phase.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

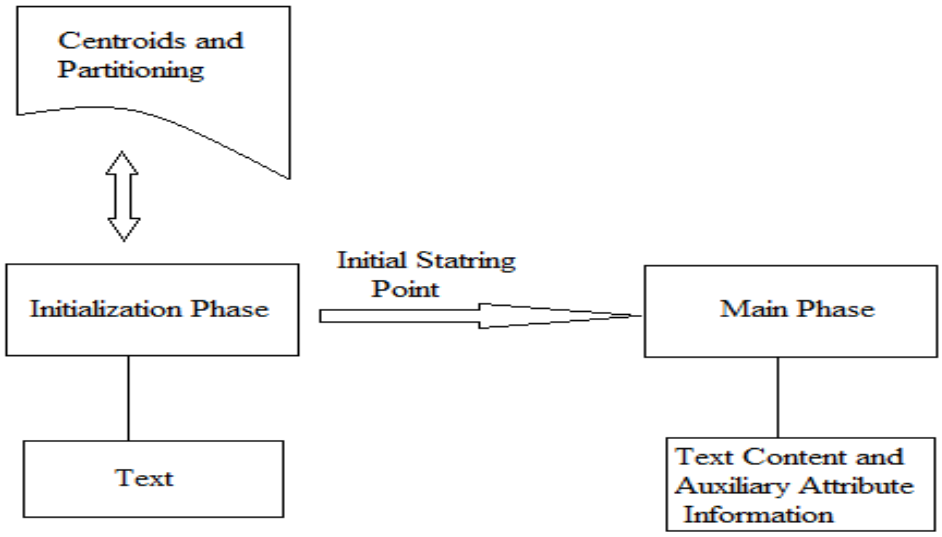

Fig. 1 Working Flow of Algorithm

IV. ADABOOST ALGORITHM

To improve the efficiency of proposed technique we are going to presented new and improved classification technique which is known as Adaboost algorithm. In existing system they have used Naive Bayes or SVM algorithm as a baseline algorithm. Existing results shows that Adaboost is more efficient than SVM and Naive Bayes hence we are proposing this in our paper with COLT algorithm.

It is utilized in conjunction with learning algorithms to improve their performance. The output of the COATES and COLT is combined into a weighted sum that represents the ultimate output of the boosted classifier AdaBoost is sensitive to noisy data and outliers.

Adaboost Algorithm

INPUT: training set of *m* examples $((x_1, y_1),\ldots\ldots, (x_m, y_m))$

INITIALIZE: $D_1(i) = 1/m$ for all i

FOR $t = 1, \ldots, T$ **Do**

Call Weak Learner, Providing it with the distribution D_t.

Call back a hypothesis $h_{t: X \to Y}$

Calculate the error of $h_t: \epsilon_t = \sum_{i:ht(xi)\neq yi} Dt(i)$.

IF $\epsilon_t > 1/2$ **THEN**

Set $T = t - 1$ and abort loop

END IF

Update distribution D_t: $D_{t+1}(i) = \frac{Dt(i)}{Zt} \times \begin{cases} Bt, & if\ ht(xi) = yi \\ 1, & Otherwise \end{cases}$ Where Z_t is the normalization constant

END FOR

OUTPUT: The final hypothesis: $h_{fin} = arg\ max\ (y \in Y) \sum_{t:ht(x)=y} log \frac{1}{\beta t}$

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

These are the steps which are combine clustering and classification algorithm. By using this Adaboost algorithm we can get the proper result about the review if it will be positive or negative.

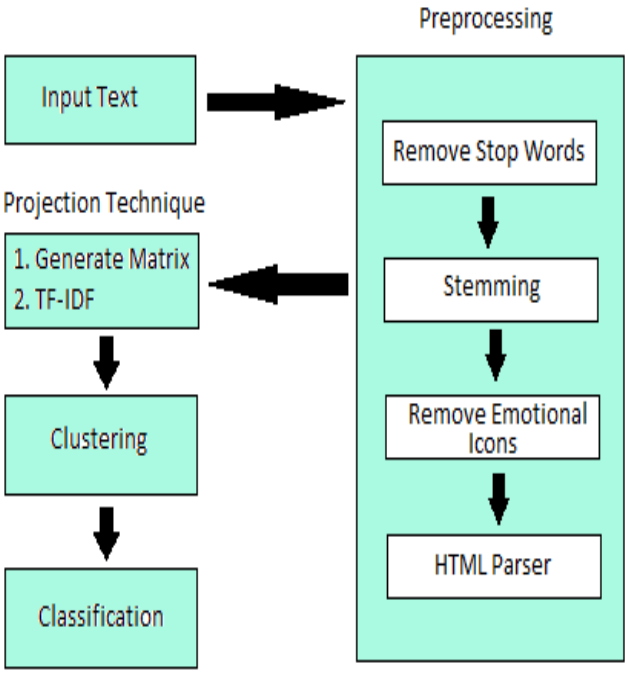

Fig. 2 Flow diagram of the proposed workflow for mining text data

I. User Authentication

The system is required user authentication module for security purpose. Security is just like a registration form and login id, password. Whenever system is running time then user are first logged it.

II. Input Text

After user login then we needed input text data. This file is data which part of IMDB dataset. The multiple file are merged and perform the preprocessing on text data.

III. Preprocessing

In pre-processing the following activity are performed in the system.

Remove Stop Words

Stop words are removed in the available text (like as a, an, the, am, is, are etc. or repeated words).This stop words are repeated in sentence. That means preposition, auxiliary verb (am, is, are)are removed for the resizing text.

Stemming-

Stemming is the process of remove suffix s, es, ing, ed etc. which are joined to the some of the meaningful word.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Remove Emotional Icons -

Sometimes text has uses the symbols of smile e.g. ☺,☹ which is meaningful so that time system also removes that smile symbol.

HTML Parser-

The online web data is available with html tag. In mining process the html tag has no proper meaning. So this meaningless data is also removed.

IV. Clustering

The clustering is the process of classification of data into group. This classification is possible by using COATES algorithm in to the earlier system.

V Classification

Classification is the arrangement of data by user required label. In our system is taking the four labels in classification drama, comedy, short and documenrty. The data are divided into that label. This classification is possible using COLT algorithm in the earlier system.

VI. RESULT ANALYSIS

For the performance evaluation of mining text data using side information and system, the system is run on configuration having Windows 8 pro with 3 GB RAM. This method is implemented in java. For this system, java works on the front end and MySQL on the back end. MySQL is used to store all user databases and updated text file. For this, the Net Beans IDE 8.0.2 is used, which contains the Java development tool. For this system, we used IMDB (Internet Movie DataBase) dataset. We extracted movies from the top four genres in IMDB which were labeled by Short, Drama, Comedy, and Documentary.

I Effectiveness of System

The following table and graph shows the effectiveness of existing and proposed system.

Table 1. Effectiveness of System

Sr. No.	Data Size	Existing System	Proposed System
1	1000	0.80	0.89
2	2000	0.83	0.90
3	3000	0.85	0.92
4	4000	0.89	0.94

As we can see in the above the graph we know that x-axis shows the different data sizes and Y-axis represents the purity level for every data size. We can see that purity level increases as data size grows. Here in this graph it clearly represents that proposed system achieves higher purity than existing system.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

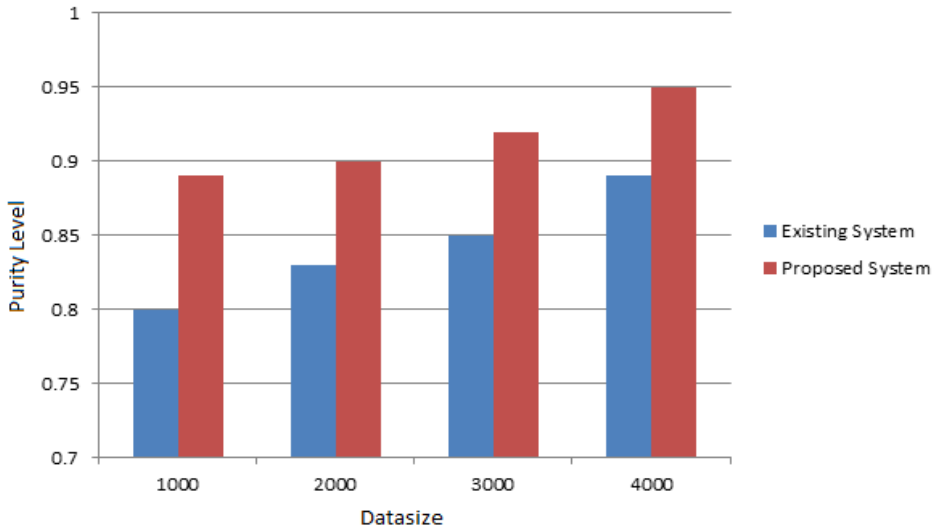

Fig.3 The Graph of Effectiveness of proposed system.

The above table and graph shows the Time complexity of existing and proposed system.

VII. CONCLUSION

In this paper, we have studied different methods for mining text data with the use of side-information. Here we have presented the improved clustering techniques, where they combined an iterative partitioning technique. We have presented new algorithm for classification technique known as Adaboost algorithm which we combined with the COLT algorithm. Our result shows that our proposed technique is more efficient in classification criteria.

REFERENCES

[1] Shekhar S. Patil ,R.B.Wagh, "Mining Text Data Using Side Information: A Survey", in Proc. IJMTER Conf.,Nasik, India Volume 3, Issue 4, April 2016 Special Issue of ICRTET'2016.
[2] C. C. Aggarwal,Philip S. Yu and Yuchen Zhao, "On the Use of Side Information for Mining Text Data" IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp. 1415-1429, June 2014.
[3] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data. New York, NY, USA: Springer, 2010.
[4] C. C. Aggarwal, Social Network Data Analytics. New York, NY, USA: Springer, 2011.
[5] C. C. Aggarwal and C.-X. Zhai, Mining Text Data. New York, NY, USA: Springer, 2012.
[6] C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in Mining Text Data. New York, NY, USA: Springer, 2012.
[7] 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.
[8] H. Frigui and O. Nasraoui, "Simultaneous clustering and dynamic keyword weighting for text documents," in Survey of Text Mining, M. Berry, Ed. New York, NY, USA: Springer, 2004, pp. 45–70.
[9] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. ACM SIMOD Conf., New York, NY, USA, 1998, pp. 73–84.
[10] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," Inf. Syst., vol. 25, no. 5, pp. 345–366, 2000.
[11] A. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.
[12] 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. 488–495.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

[13] C. C. Aggarwal, S. C. Gates, and P. S. Yu, "On using partial supervision for text categorization," IEEE Trans. Knowl. Data Eng., vol. 16, no. 2, pp. 245–255, Feb. 2004.
[14] A. Banerjee and S. Basu, "Topic models over text streams: A study of batch and online unsupervised learning," in Proc. SDM Conf., 2007, pp. 437–442.
[15] 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.
[16] R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in Proc. VLDB Conf., San Francisco, CA, USA, 1994, pp. 144–155.
[17] H. Schutze and C. Silverstein, "Projections for efficient document clustering," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 74–81.
[18] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD, 2000, pp. 109 –110.
[19] Y. Sun, J. Han, J. Gao, and Y. Yu, "TopicModel: Information network integrated topic modeling," in Proc. ICDM Conf., Miami, FL, USA, 2009, pp. 493–502.
[20] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in Proc. ACM SIGIR Conf., New York, NY, USA, 2003, pp. 267–273.
[21] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACMSIGMOD Conf., New York, NY, USA, 1996, pp. 103–114