ISSN 2395-1621

Clustering Approach for Mining Text Data

[#1]Akshay A Bhujugade, [#2]Amol K.Dammewar, [#3]Ravindra J.Bondre, [#4]Prof.H.V.Kumbhar

[1]akshaybhujugade65@gmail.com
[2]amoldammewar77@gmail.com
[3]ravindrabondre101@gmail.com

[#1234]Department of Computer Engineering PVPIT,PUNE.

ABSTRACT

Any text mining application may contain side information. This side information may be any links in the document, web logs which contain user access behaviour, provenance information, the links for any document or any other non-textual attributes which are embedded into the text document. All these attributes may contain a huge amount of information for clustering purposes. But it is difficult to count the concerned importance of this side information especially when some of the data is noisy. In that matter, it is dangerous to merge side-information into the mining process because it can upgrade the quality of the representation for the mining process or can add noise in this system. Thus, there should be a right way to do this mining process so that it will make use of side information to maximize their advantages. Therefore, it is suggested to design an efficient algorithm which makes combination of classical portioning algorithm with probabilistic models in order to create an effective clustering approach. Afterwards, extension to the classification problem is also shown.

Keywords: Data Mining, K-means, Document Clustering, COATES algorithm, COLT algorithm.

ARTICLE INFO

Article History

Received :8th March 2016

Received in revised form :

10th March 2016

Accepted : 13th March 2016

Published online :

18th March 2016

I. INTRODUCTION

Side information contains meta data which is present with text document in several text mining domains. Document important information, the links in the document, other non-textual attributes which are contained no of hits to web logs or types of user from web logs these are the different kind of side information. During text clustering different problem arises in various application such as web, social networks and also some other digital data. The web is collection of large amount of text so in order to create efficient and more scalable algorithms for text mining approach. Meta data is present at origin of documents of the web. The user access behaviour is computed in the form of web logs. A document contains the links which may contain more information for mining process.The process of grouping the objects with similarity is termed as Clustering. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, information retrieval, image analysis, pattern recognition, and bioinformatics.

In order to achieve this goal, we will combine a partitioning approach with a probabilistic estimation process, which will determine the coherence in side-attributes for clustering process. A probabilistic model on the side information will uses partitioning information for the purpose of estimating the coherence of different types of clusters with the side attributes. This will helps us in abstracting out the noise in the membership behaviour of different types attributes. The partitioning approach is specially designed for the efficient of large data sets. This can be useful in many scenarios were the data sets are large. We will present experimental results on a number of real data sets, and illustrate the effectiveness and efficiency of the approach.

II. LITERATURE SURVEY

Paper presented by Charu C. Aggarwal, Philip S. Yu demonstrates [1] that real time clustering and segmentation

of text data records is required in many applications such as news group filtering, text crawling, and document organization. The categorical data stream clustering problem also has a number of applications to the problems of customer segmentation and real time trend analysis. By making the use of a statistical summarization methodology, an online approach for clustering massive text and categorical data streams is presented here.

Paper presented by S. Zhong demonstrates [3] that clustering data streams has been a new research topic, recently used in many real data mining applications, and has attracted a lot of research attention. However, there is not much work on clustering high-dimensional streaming text data. This paper merges an efficient online spherical k-means algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams. The OSKM algorithm modifies the spherical k-means algorithm, using online update based on the well known. Winner Take All competitive learning. It has been shown to be as efficient as SPKM, but much superior in clustering quality. The scalable clustering strategy was previously developed to deal with very large data bases that cannot fit into a limited memory and that are too expensive to read/scan multiple times. Using this method, one keeps only sufficient statistics for history data to retain (part of) the contribution of history data and to accommodate the limited memory. To make the proposed clustering algorithm adaptive to data streams, a forgetting factor is introduced here that applies exponential decay to the importance of history data. The older a set of text documents, the less weight they carry. The experimental results demonstrate the efficiency of the proposed algorithm and reveal an intuitive and an interesting fact for clustering text streams—one needs to forget to be adaptive.

Paper presented by Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey demonstrates [11] that for information retrieval, document clustering has not been well used. There are two main categories for its objection: first, for large corporation clustering is too slow and second, that retrieval is not improved by clustering. When clustering is used in an to improve conventional search techniques then only such problems are coming. However, t clustering as an information access tool in its own right obviates these objections, and provides a powerful new access paradigm. Document clustering is presented as primary operation in document browsing technique. Fast clustering algorithms are also presented which support this interactive browsing paradigm.

Paper presented by S. Guha, R. Rastogi, and K. Shim demonstrates [6] that for discovering groups and identifying interesting distributions in the underlying data clustering is used in data mining. Traditional clustering algorithms either favor clusters with spherical shapes and similar sizes. In this paper a clustering algorithm is presented which is called CURE that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction. Having more than one representative point per cluster allows CURE to adjust well to the geometry of non-spherical shapes and the shrinking helps to dampen the effects of outliers. CURE employs a combination of random sampling and partitioning to handle large databases. A random sample drawn from the data set is first partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to gain the desired clusters. In this paper experimental results shows that the quality of clusters produced by CURE is much better than those found by existing algorithms. Further, in this paper it is demonstrated that random sampling and partitioning enable CURE to not only outperform existing algorithms but also to scale well for large databases without sacrificing clustering quality.

III.PROPOSED SYSTEM

In this system, we are storing documents of different subject on server where each user or guest can read that document. He can represent his view opinion through comments on that document. We are applying COLT algorithm for clustering the documents. Documents will be clustered according to content in the document. For each comment gini index will be calculated which is used to calculate inequality of the comment with document. Finally review for that document is generated depending on the comments.

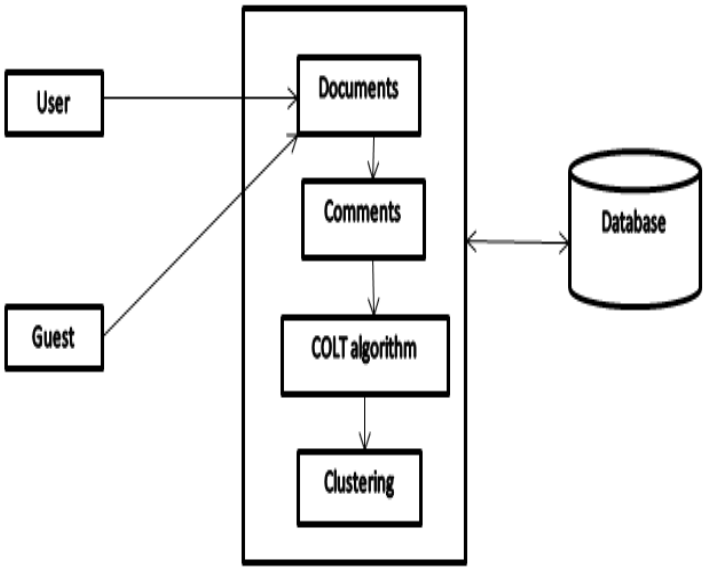

Fig. System Architecture

IV. ALGORITHM

Algorithm COATES(NumClusters k,
Corpus:T1…TN,Auxilary Attribute:X1…XN)
Begin
Use content-based algorithm in [27] to crete initial set of custers c1…..ck;
Let centroids of c1…..ck be
Denoted by L1….Lk;
T=1;
While not (termination_criterion**) do**
begin
{First minor iteration}
Use cosine-similarity of each document Ti to centroids L1…Lk in order to determine
the closest cluster to Ti and update the cluster assignments C1…Ck;
Denote assigned cluster index for document Ti by qc(i,t);
Update cluster centroids L1….Lk to the centroids of updated clusters C1….Ck;
{Second Minor iteration}
Compute gini-index of Gr for each auxiliary attribute r with respect to current
Clusters C1….Ck;
Mark attributes with gini-index which is standared-deviations below the mean
as non-discriminatory;
{ for document Ti let Ri be the set ofattributes which take on the value of 1,
And for which gini-index is discriminatory;}
for each document Ti use the method discussed in section 2 to determine
the posterior probability p(Ti)
Denote qa(I,t) as the cluster-index with highest posterior probability of assignment for document Ti;
Update clustr-centroids L1…..Lk with the use of posterior probabilities as discussed in section 2;
T=t+1;
end
end

Algorithm COLT(NumClusters: k,Corpus:T_1…..T_N,
Auxilary Attributes:X_1…..X_N);
Labels l_1…l_N);
begin
Perform feature selection on text and auxiliary attributes with the use of class labels and gini index as explained in section 3;
Use supervised version of algorithm in [27] to create initial set of k clusters denoted by
C_1…C_K , so that each cluster Ci contains only records of a particular class;
Let centroids of C_1…C_k be denoted by L_1…L_k;
t=1;
while not (termination_criterion) **do**
begin
{First minor iteration**}**
Use cosine-similarity of each document T_i to centroids L_1…L_k in order to determine
the closest cluster to T_i(which belongs to same class)and update the cluster
assignment C_1…C_k;
Denote assigned cluster index for Document T_i by q_c(i,t);
Update clustr centroids L_1….L_k to the centroids of updated cluster C_1…C_k;
{Second Minor Iteratoin}
Compute gini–index of G_r for each auxiliary attribute r with respect to current
cluster C_1…C_k;
Mark attribute with gini-index which is γ standard-deviation below the mean an
non- discriminatory;
{ for document T_i let R_i be the set of attributes which take on the value of 1, and for
which gini-index is discriminatory;}
For each document T_i use the method discussed
in section 2 to determine the posterior probability $P^n(T_i \in C_j | R_i)$;
Denote q_a (i,t) as the cluster-index with highest posterior probability of assignment
for document T_i which also belongs to the same class;
Update cluster-centroids L_1…L_k with theuse of posterior probabilities as discussed in
section 2;
t = t + 1;
end
end

V. CONCLUSION

In this paper, methods are discussed for mining text data with making use of side information. Side information may be presented in many forms of text database which are used to enhance the clustering process. Iterative portioning technique is combined with a estimation process to design the clustering method which gives the importance of different kinds of side information. This general method is used to design both clustering and classification algorithms. COATES and COLT approach can greatly enhance the quality of text clustering and classification, while maintaining a high level of efficiency.

VI. ACKNOWLEDGEMENT

We are thankful to our guide Prof.Mrs.H.V.Kumbhar for her proper guidance and valuable suggestions. I am also thankful to Prof. Mr. N. D. Kale the Head Of Computer Engineering Department and other faculty members. We also thankful to our Family members and friends for support. We once again extend our since thanks to all of them and very much thankful to library staff & Computer Engineering Department staff for their kind co-operation.

REFRENCES

[1] 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.

[2] C. C. Aggarwal and H. Wang, *Managing and Mining Graph Data*. New York, NY, USA: Springer, 2010.

[3] S. Zhong, ―Efficient streaming text clustering,‖ *Neural Netw.*,vol. 18, no. 5–6, pp. 790–798, 2005.

[4] C. C. Aggarwal and C.-X. Zhai, *Mining Text Data*. New York, NY,USA: Springer, 2012.

[5] C. C. Aggarwal and C.-X. Zhai, "A survey of text classificationalgorithms," in *Mining Text Data.* New York, NY, USA: Springer,2012.

[6] 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.

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

[8] 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.

[9] 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.

[10] J. Chang and D. Blei, "Relational topic models for document networks," in *Proc. AISTASIS*, Clearwater, FL, USA, 2009, pp. 81–88.

[11] 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.