Mining Text Data using Side Information

Gajalakshmi.S[1] Timotta.C[2] Gayathiri.D[3] Vinisha.K[4]
[1,3,4]P.G Scholar [2]Assistant Professor
[1,2,3,4]Department of Computer and Engineering
[1,2,3,4]PPG Institute of Technology, Coimbatore

Abstract— Text mining applications consists of text along with side information which can be also called as attributes ,they are of different kinds such as document links, history of the document, user web logs, access behavior of various users. These side information comes along with the text documents and they can also be non-textual attributes is available with text documents. Such side information embedded with the text document consists of tremendous information which is used in clustering process is embedded into the text document. The estimation of the relative information .Importance of side information may be difficult when some of them are noisy. In that case, it is a risk to combine side information in the mining process; they can even add more noise into the process or improve the quality of text representation for into mining process. Hence a principled method of process is needed to perform a mining process to increase the advantages of side information. The algorithm named COATES is designed with the combination of classical partitioning algorithms under probabilistic models in the form of creating an effective clustering process. The process is then extended to an classification purpose with the algorithm named COLT by presenting experimental results with number of real data sets to show the advantages of using such process.

Key words: Mining Text Data, Clustering

I. INTRODUCTION

Text mining is also called as data mining. Data mining is of deriving high quality of information from a simple text which is referred to as text analytics, the data information is derived in the form of devising patterns which is called as statistical pattern.

Mining text involves the process of parsing with some attribute features and removal of stop words and stemming process with structured data, initial mining consists of text categorization (preprocessing)_text clustering(splitting into groups) Concept evolution production of new class.

Knowledge discovery which means the discovery of the computer of new unknown information by the means of extracting information from simple huge amount of different forms of structured and unstructured textual attributes

Data mining in text forms are the patterns that are extracted from natural text forms rather the database. Information access which is the retrieval of information as then new information is found desired with valid sources. text mining consists of data mining ,information retrieval, web mining, statistics and NLP.

Text mining has the goal of extracting text with the following process, text preprocessing, text transformations, attribute selection, data mining, pattern discovery, interpretation evolution

II. MINING AND CLUSTERING FOUNDATIONS

Data mining with techniques and process with product development for side information in research foundations. The problem in text-clustering is in by the database community. Major focus on work is been with scalable clustering and multidimensional data with different texting forms. Survey with clustering algorithms and classification algorithms are made with COATES and COLT. Clustering problem is made to study in context of text information. The known technique for clustering text is scatter gather technique under the combination of agglomerative and partitioned clustering. The related methods in text clustering are co-clustering methods for text data COLT with expectation Maximization (EM) method with text clustering is processed with Matrix-factorization techniques... Technique of document information based with clustering process, closely related area t of topic-modeling, event tracking, text-categorization is the context, with method in topic-driven clustering in text data was been propose for text clustering the context in keyword extraction. Approach for text clustering with side information is with corpus S in text documents. The total number of documents is N, which are denoted by T1 . . . TN. That is assumed with set of distinct words in entire corpus S will be denoted by W. Associated in each document Ti, a set in side attributes Xi. Attributes Xi is dimensions d that are denoted by (xi1 . . . xid). Such attributes are auxiliary attributes. The notation and analysis, that assume side-attribute xid is binary, hence both numerical with categorical attributes is easily converted to linear type, because different values in categorical attribute is assumed in separate binary attributes, The numerical data is been discredited with binary values of attribute ranges. Examples in side-attributes are:

Web log analysis consists of xir which is of 0-1 variable, that indicates the ith document is been accessed by rth user. The information is used in mean order to then cluster web pages in site with more informative with a technique that is purely based on content of documents. The previous case is of number of pages in site is large; the number of documents in particular user can relatively be small.

- Network application, that xir corresponds to 0-1 variable corresponding can whether be the document Ti is a hyperlink to the next Tr. If desired, it is linkage connectivity with clustering process and then hyperlink graphs is large sparse, and follows number of auxiliary variables that are high and only small fraction take on the value of use inside information
- Documents associated with provenance information, is possible with attributes on large possibilities. Attributes can be a sparsity property. With examples such as auxiliary attributes in real applications. A

All rights reserved by www.ijsrd.com

challenge efficiency perspective, sparsity is taken during the clustering process. The techniques are designed for sparsely and is possible in easily designed way for an approach of non-sparse attributes and information, in treating with attribute values in a clear way

III. COMPLEXITY IN TIME WITH CLUSTERING

Time complexity in this approach is by time need for content and auxiliary based iterations. Running time is determined by means of number in clusters k, the words in text content is lexicon which is dt, number in auxiliary attributes in text is d, with number in documents in N. All iterations of approach, O(k) with cosine computations is performed. The number of N documents is, O(N·k) under cosine computations. cosine computation require O(dt) distance time, to run the time by given O(N·k·dt).

Each iteration computes similarity of auxiliary attributes. The difference in from of content-based in all computation requires O(d) time distance. The total time need in each iterations O(N·k·(d+dt)) with all the running time can made to be obtained in multiplying the value with total number in distance iterations. A little number in iterations within 2 to 9 may be sufficient in convergence for most sessions. The running time can be calculated within as O(N · k · (d + dt)).

IV. COATES

Text clustering with side-information with COATES that it is COntent and Auxiliary attribute based Text clustering algorithm. The inputs of the algorithm are of k clusters. Removable of stop words is made and stemming is also been performed to increase the importance of the attributes. The algorithm consists of two phases,

A. Initialization of Text Documents:

Clustering text approach without side-information. COATES algorithm is used, to provide an quick and efficient clustering approach with the centroids as the starting initial process and the second phase is of auxiliary attribute information.

B. Main Phase with Text Clustering:

After the first phase the main phase algorithm is executed. The phase starts with initial cluster groups and reconstructs clusters with both the text content and auxiliary attribute information. This phase is of alternating iterations with the text content and auxiliary attribute information to improve the quality of clustering. They are called as content iterations and auxiliary iterations. The first phase is to construct an initialization; the second phase is of content and auxiliary information integration. The first phase is just the text clustering with algorithm proposed. The overall approach uses centroids for k clusters created with the phase denoted by L1 . . . Lk, cosine similarity function is used for assignment process. Each auxiliary phase is of probabilistic model, that relates attribute information probabilities to cluster group probabilities, based on clusters that already been created in recent text-based attribute phase. The goal is to create a coherence of text clustering with side-information.

V. COLT

In the case of classification process, the COLT algorithm COntent and auxiLiary attribute-based Text classification algorithm was used against such baseline process: The COLT algorithm consists of supervised clustering approach to partition data into k different clusters groups. The partitioning is then for classification purpose. The processes used in the algorithm are,

A. Feature Selection to Remove Attributes:

First step is of feature selection which is made to remove the attributes that are not related to class label. The process is performed for both text attributes and auxiliary attributes.

B. Initialization with k-means:

The second step is of supervised kmeans process to perform initialization, with purely text content. The membership of class of records in each and every cluster is the main important difference between supervised k-means initialization, and unsupervised initialization hence the k-means clustering algorithm is modeled and so individual clusters consists records only of their particular class.

C. Cluster-Training Model Construction with classifiers:

This phase is the combination of text and side information which is used for the purpose of creating a cluster based model. In case of initialization, purity of clusters is maintained in the phase at once when a supervised cluster are constructed

(1)Tested against Naive Bayes Classifier that cares for Text alone. (2) Testing against SVM classifier only for text (3) test made against supervised clustering method that is embedded with both text and side information. Hence the combination of the algorithms is made to calculate the importance of the approach on both of the process such as text and side information.

One of the major advantage of COLT over COATES is that clusters groups within are class-specific, hence the training document needs to be compared only with the clusters of same class group.

This can require the comparison to O(k) clusters. The COLT classify method performs effective respect to other when the classification accuracy is not changed so much along the whole range of smoothing parameter. Therefore the technique is very robust in the choice of smoothing parameter.

VI. METRICS

The approach is supervised with side information and Clustering attributes and attribute alternatives with noise and the pure text. This can also be with text and side information or the use in both. Each and every data set is with the class labels which are called as the known forms of the data. They are not used in the clustering process. Hence all the classes are computed in the form of purity, for each class, we computed the cluster purity, which is also called as the fraction of dominant clusters in class.

VII. CONCLUSION

The paper is of methods for mining text data As well with side-information. There are large and many amount of side information or Meta data (data about a data) which are

All rights reserved by *www.ijsrd.com*

greatly used to increase the clustering approach are embedded with many and large forms of text databases. To design a clustering method of process, the combinations of iterative partitioning technique with probability estimation process that computes the advantage of different types of side information is used. The general approach is used to design both clustering and classification algorithms. Results on real data sets are illustrating effectiveness of the approach is presented. Results are there to show the use of side-information do greatly enhance quality of text clustering and classification, with the maintenance of high level of efficiency.

REFERENCE

[1] 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. 2012

[2] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A Framework for On-Demand Classification of Evolving Data Streams," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 5, pp. 577-589, May 2011.

[3] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda`, "New Ensemble Methods for Evolving Data Streams," Proc. ACM SIGKDD 15th Int'l Conf. Knowledge Discovery and Data Mining, pp. 139-148, 2010.

[4] S. Chen, H. Wang, S. Zhou, and P. Yu, "Stop Chasing Trends: Discovering High Order Models in Evolving Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 923-932, 2009.

[5] W. Fan, "Systematic Data Selection to Mine Concept-Drifting Data Streams," Proc. ACM SIGKDD 10th Int'l Conf. Knowledge Discovery and Data Mining, pp. 128-137, 2004.

[6] J. Gao, W. Fan, and J. Han, "On Appropriate Assumptions to Mine Data Streams," Proc. IEEE Seventh Int'l Conf. Data Mining (ICDM), pp. 143-152, 2007.

[7] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari, "Adapted One-versus-All Decision Trees for Data Stream Classification," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 5,pp. 624-637, May 2009.

[8] G. Hulten, L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams," Proc. ACM SIGKDD Seventh Int'l Conf. Knowledge Discovery and Data Mining, pp. 97-106, 2001. Masud Et Al.: Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams 1495 Fig 5. Scalability of MCM with number of dimensions

[9] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM), pp. 929-934, 2008.

[10] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Tracking Recurring Contexts Using Ensemble Classifiers: An Application to Email Filtering," Knowledge and Information Systems, vol. 22, pp. 371-391, 2010.

[11] J. Kolter and M. Maloof, "Using Additive Expert Ensembles to Cope with Concept Drift," Proc. 22nd Int'l Conf. Machine Learning (ICML), pp. 449-456, 2005.

[12] D.D. Lewis, Y. Yang, T. Rose, and F. Li, "Rcv1: A New Benchmark Collection for Text Categorization Research," J. Machine Learning Research, vol. 5, pp. 361-397, 2004.

All rights reserved by *www.ijsrd.com*