ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 408- 410 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

An efficient Privacy Preserving Data Clustering over Distributed Networks

[1]G. Kananka Maha Laksmi, [2]V.Visweswara Rao
[1]Final M.Tech Student, [2]Assistant Professor
[1,2]Dept of Computer Science and Engineering
[1,2]Pydah College of Engineering, Visakhapatnam, AP, India

Abstract: Privacy preserving over data mining in distributed networks is still an important research issue in the field of Knowledge and data engineering or community based clustering approaches, privacy is an important factor while data sets or data integrates from different data holders or players for mining. Secure mining of data is required in open network. In this paper we are proposing an efficient privacy preserving data clustering technique in distributed networks.

INTRODUCTION

In distributed networks or open environments nodes communicates with each other openly for transmission of data, there is a rapid research going on secure mining. Research work on privacy preserving techniques while mining of data either in classification, association rule mining or clustering.

Randomization and perturbation approach available for privacy preserving process and it can be maintained in two ways, one is cryptographic approach here real data sets can be converted to unrealized datasets by encoding the real datasets and the second one imputation methods, here some fake values imputed between there real dataset and extracted while mining with some rules[1][2].

Clustering is a process of grouping similar type of objects based on distance (for numerical data) or similarity (for categorical data) between data objects. In distributed environment data holders or players maintains individual data sets and every node or vertex is connected with each other by an edge along with their quasi identifiers [3].

Most existing text clustering algorithms are designed for central execution. They require that clustering is performed on a dedicated node, and are not suitable for deployment over large scale distributed networks. Therefore, specialized algorithms for distributed and P2P clustering have been developed, such as [6], [7], [8], [9]. However, these approaches are either limited to a small number of nodes, or they focus on low dimensional data only. In distributed environment, nodes are represented by Privacy preserving Distributed clustering algorithm proposed by "S. Jha, L. Kruger, and P. McDaniel", here data can be clustered by grouping the similar type of objects and secure transmission through protocols [4].Perturbation Method of string transformation proposed for privacy preserving clustering technique by using geometric techniques [10].

RELATED WORK

In social network, nodes can be represented as vertices and those vertices V (v_1,v_2…v_n) connected through set of edges E in a undirected graph G (V,E) and non-identifying attribute to describe node is known as quasi-identifier. Clustering can be performed on the quasi identifiers like age and gender, distributed clustering groups similar type of objects based on minimum distance between the nodes.

In distributed networks, data can be numerical data or categorical .Numerical data can be compared with respect to quasi identifier difference and Categorical data can be compared with similarity between the data objects. Text Clustering methods are divided into three types. They are partitioning clustering, Hierarchal clustering, fuzzy clustering. In partitioning algorithm, randomly select *k* objects and define them as *k* clusters. Then calculate cluster centroids and make clusters as per the centroids. It calculates the similarities between the text and the centroids. It repeats this process until some criteria specified by the user.

In this paper we are proposing architecture, every data holder or player clusters the documents itself after preprocessing and computes the local and global frequencies of the documents for calculation of file Weight or document weights. Distributed k means algorithm is one of the efficient distributed clustering algorithm. In our work we are working towards optimized clustering in distributed networks by improving traditional clustering algorithm. In our approach if a new dataset placed at data holder, it requests the other data holders to forward the relevant features from other data holders instead of entire datasets to cluster the documents.

In distributed clustering nodes can be clustered based on the common edges which are connected through vertices and should have similar set of weights between the edges, but weight does not leads to optimal solution because it is not optimal measure to consider it.

Every individual data holder or player maintains their transactions or patterns, in horizontal

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 408- 410 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

partitioning, every data holder forwards their patterns to centralized server after encryption of patterns which are at individual end, At centralized server received pattern can be decrypted with decoder and forwarded to Boolean Matrix to extract frequent pattern from the received patterns. For experimental purpose we establish connection between the nodes and Central location (Key generation center) through network or socket programming, Key can be generated by using improved LaGrange's polynomial equation and key can be distributed to user [8].

Even though various horizontal and vertical partition mechanisms available, privacy is the major concern while transmission of data from data holders securely. Centralized server performs required data mining operations over received data.

PROPOSED SYSTEM

In this paper we are proposing an efficient and secure data mining technique with optimized k-means and cryptographic approach, for cluster the similar type of information, initially data points need to be share the information which is at the individual data holders or players. In this paper we are emphasizing on mining approach not on cryptographic technique, For secure transmission of data various cryptographic algorithms and key exchange protocols available.

Individual peers at data holders initially preprocess raw data by eliminating the unnecessary features from datasets after the preprocessing of datasets, compute the file Weight of the datasets or preprocessed feature set in terms of term frequency and inverse document frequencies and computes the file relevance matrix to

reduce the time complexity while clustering datasets. We are using a most widely used similarity measurement i.e cosine similarity

$Cos(d_m,d_n)= (d_m * d_n)/Math.sqrt(d_m * d_n)$

Where

d_m is centroid (Document weight)

d_n is document weight or file Weight

In the following example diagram shows a simple way to retrieve similarity between documents in at individual data holders by computing cosine similarity prior clustering as follows.

	d1	d2	d3	d4	d5
d1	1.0	0.77	0.45	0.32	0.67
d2	0.48	1.0	0.9	0.47	0.55
d3	0.66	0.88	1.0	0.77	0.79
d4	0.89	0.67	0.67	1.0	0.89
d5	0.45	0.88	0.34	0.34	1.0

In the above table $D(d_1,d_2….d_n)$ represents set of documents at data holder or player and their respective cosine similarities, it reduces the time complexity while computing the similarity between the centroids and documents while clustering. In our approach we are enhancing K Means algorithm with recentoird computation instead of single random selection at every iteration, the following algorithm shows the optimized k-means algorithm as follows

Algorithm:

1: Select K points as initial centroids for initial iteration

2: until Termination condition is (user specified maximum no of iterations)

3: get_relevance(dm,dn)

Where dm is the document M Weight from relevance matrix

$d_n$is the document N Weight from relevance matrix

4: Assign each point to its closest centroid to form K clusters

5: Recompute the centroid with intra cluster data points (i.e average of any k data points in the individual cluster).

Ex: $(P_{11}+P_{12}+….P_{1k}) / K$

All points from the same cluster

6. Compute new centorid for merged cluster

In the traditional approach of k means algorithm it randomly selects a new centroid.In our approach we are enhacing by prior construction of relevance matrix and by considering the average k random selection of document Weight for new centroid calculation .

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 408- 410 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

CONCLUSION

We are concluding our current research work with efficient privacy preserving data clustering over distributed networks. Quality of the clustering mechanism enhanced with preprocessing, relevance matrix and centroid computation in k-means algorithm and cryptographic technique solves the secure transmission of data between data holders or players and saves the privacy preserving cost by forward the relevant features of the dataset instead of raw datasets. We can enhance security by establishing an efficient key exchange protocol and cryptographic techniques while transmission of data between data holders or players.

REFERENCES

[1] Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases FoscaGiannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang.

[2] Privacy Preserving Decision Tree Learning Using Unrealized Data Sets Pui K. Fong and Jens H. Weber-Jahnke, Senior Member, IEEE Computer Society.

[3] Anonymization of Centralized and Distributed Social Networks by Sequential ClusteringTamirTassa and Dror J. Cohen.

[4] Privacy Preserving ClusteringS. Jha, L. Kruger, and P. McDaniel.

[5] Tools for Privacy Preserving Distributed Data Mining **,** Chris Clifton, Murat Kantarcioglu, Xiaodong Lin, Michael Y. Zhu

[6] S. Datta, C. R. Giannella, and H. Kargupta, "Approximate distributed K-Means clustering over a peer-to-peer network," *IEEE TKDE*, vol. 21, no. 10, pp. 1372–1388, 2009.

[7] M. Eisenhardt, W. M¨ uller, and A. Henrich, "Classifying documents by distributed P2P clustering." in *INFORMATIK*, 2003.

[8] K. M. Hammouda and M. S. Kamel, "Hierarchically distributed peer-to-peer document clustering and cluster summarization," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 5, pp. 681–698, 2009.

[9] H.-C. Hsiao and C.-T. King, "Similarity discovery in structured P2P overlays," in *ICPP*, 2003.

[10] Privacy Preserving Clustering By Data Transformation Stanley R. M. Oliveira ,Osmar R. Za¨ıane