International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)

International Conference on Information Systems and Computing (ICISC-2013), INDIA.

A STRIDE TOWARDS DEVELOPING A DISTRIBUTED SYSTEM FOR THE IDENTIFICATION OF TOP-L INNER PRODUCT ELEMENTS

S.VEENA[1], DR.P.RANGARAJAN[2]

[1]*RESEARCH SCHOLAR, Sathyabama University, Chennai*
[2] *PROFESSOR & HEAD, R.M.D Engineering College, Chennai.*

Email : djohnveena@gmail.com

Abstract

A distributed scenario can be of two types: (1) homogeneous – where only a fraction of each feature is observed at every site or (2) heterogeneous – where only some of the features are observed at each site. For either scenario centralizing all the data in order to build a global model is not an appropriate solution due to the high cost of centralizing and storage requirement at the central node. Therefore, distributed algorithms are required to solve most data mining problems in p2p networks. In general, a distributed algorithm in this setting should (1) not require global synchronization, (2) be communication efficient, and (3) be resilient to moderate changes in the network topology. We proposed Gossip based probabilistic approximate algorithm. This algorithm relies on properties of random walk on network to provide estimates for various estimates for various statistics of data stored in network. The computation result of this algorithm is exponentially fast. The most important quality of Gossip based probabilistic approximate algorithm is that they provide probabilistic guarantees for the accuracy of result.

***Keywords*-- Distributed data mining, inner product, peer-to-peer network**

I. INTRODUCTION

The inner product between two vectors measures how similar or close they are to each other. Inner product is an important primitive for many ma-chine learning and data mining applications such as classifier learning and clustering [1]. These are used for various kinds of tasks such as information retrieval from the web, text processing, predictive modeling and the like. Traditional data mining techniques assume that all data is available at a central location. However there exist situations in which the data is inherently distributed over a large, dynamic network containing no special servers or clients, for example, peer-to-peer (p2p) networks . In many application scenarios, it is often desirable to know only the top inner products. Such a need is often felt even in emerging large-scale peer-to-peer (P2P) applications such as the formation of interest-based online communities [2]. P2P networks are large, dynamic, and asynchronous and have little central control. Peer-to-Peer (P2P) systems are distributed systems in which nodes of equal roles and capabilities exchange information and services directly with each other. In recent years, P2P has emerged as a popular way to share huge volumes of data. The key to the usability of a data-sharing P2P system, and one of the most challenging design aspects, is efficient techniques for search, route queries and retrieval of data [5].

P2P applications and networks are taking foot by the day, and new systems are proposed continuously with ever novel features and better performances [4].It is very difficult, if not impossible, to transfer all the data to a single peer to do the computation since no one would have such extensive storage and computational capabilities, let alone the enormous communication overhead. In the online community formation example, each peer may be associated with a feature vector describing its Web surfing patterns, and the goal is to find peers having similar interest (browsing patterns). This helps in routing queries to peers with relevant interests, resulting in better network-search results. In most cases, each peer may be interested in finding only a few peers with similar interest and not all of them.

II. RELATED WORK

Many other applications such as network intrusion detection over data streams [2], query routing in sensor networks, and efficient decision tree construction in distributed environments demonstrate the same needs. If the entire data can be conveniently accessed, it is easy to compute the inner product matrix and determine the top ones. However, much of the world's data is distributed over a multitude of systems connected by communication channels of varying capacity.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

This calls for new techniques to perform data mining in a distributed environment.

2.1 AN ORDER STATISTICS-BASED APPROXIMATE LOCAL ALGORITHM

In this paper, we consider the problem of identifying the global top-l inner products (attribute wise) from distributed data. We assume that data is scattered among a large number of peers such that each peer has exactly the same set of attributes (or features). In the data mining literature, this is often referred to as a horizontally partitioned (homogeneously distributed) data scenario. We propose an order statistics-based approximate local algorithm for solving the problem. Here, the local algorithm is one where a peer communicates only with its neighbors (a formal definition will be given later). At the heart of our algorithm is the ordinal approximation based on theories from order statistics.

III. INNER PRODUCT COMPUTATION

The curse of dimensionality makes data analysis significantly difficult [7] . It even dictates the cost of centralization, since the latter increases with increasing dimension. There exists a number of techniques such as principal components analysis (PCA), singular value decomposition (SVD), etc. for dimensionality reduction in the centralized setting. These can be applied to reduce the dimensionality and then the inner product entries of this reduced space can be centralized to find the significant entries in the new space. However these off-the-shelf techniques do not scale well in large scale peer-to-peer networks with respect to communication, computation and storage. In many cases such as the Internet, distributed file sharing networks (e.g. Gnutella, Bit Torrents), local area networks, sensor net-works and peer-to-peer networks the data is inherently distributed. Thus there exists great scope for development of distributed algorithms for performing a wide variety of tasks that are otherwise quite easily solvable, in a distributed scenario.

One such task is inner product computation. Inner product computation is a very powerful primitive in machine learning and data mining that can be used for computing Euclidean distance (clustering), information gain (classifiers, bayes net) and correlation between vectors. Inner product can also be used for computing the angle between two vectors. Now, if we consider each feature as a column vector, then the inner product between two feature vectors measures the "similarity" between them in terms of the angle between them. In other words, higher the inner product value more is the similarity between two features and vice versa.

Efficiency of local algorithms is the major issue. Thus we will develop an algorithm which will increase the efficiency, accuracy of local algorithms.

3.1 PEER TO PEER DATA MINING

A lot of works have been proposed by researchers regarding peer-to-peer network. A brief review of some of the recent researches is presented here:

Kanishka Bhaduri *et al.* [2] proposed a new algorithm for efficiently identifying some user specified l entries of the inner product matrix that belong to the top $1 - p$ percentile of the population. In order to achieve low communication complexity for our distributed algorithm, it used an ordinal statistics based approach together with cardinal sampling Ordinal statistics provides a general framework for estimating distribution free confidence intervals for population percentiles. What this means is that, for any data distribution, it can use the same theory developed here in order to estimate the top-l elements. Using simple cardinal approximation is more communication intensive. Similarly, it cannot use only ordinal sampling since the inner product entries are distributed among the peers. Thus, using both, we can achieve good results. In this work we bounded both the message complexity of our algorithm and the error in our decision making. It provided experimental results that substantiate our claims regarding accuracy and message complexity of our algorithm[12].

IV. IDENTIFYING TOP-L INNER PRODUCTS

Kamalika Das *et al.* [2] developed a distributed algorithm for efficiently identifying top-l inner products from horizontally partitioned data. To achieve low communication overhead, it use an order statistics-based approach together with cardinal sampling. Ordinal statistics provides a general framework for estimating distribution-free confidence intervals for population percentiles. Cardinal sampling helps to combine the inner product values that are distributed among the peers. Local algorithms can be exact or approximate. However, the class of exact local algorithms that currently exists in the literature work for simple primitives such as average and L2-norm. For solving more complicated distributed problems, researchers have developed approximate solutions. The ordinal analysis technique developed in this paper belongs to this genre of approximate local algorithms.

Kanishka Bhaduri *et al.* [11] offers a local distributed algorithm for multivariate regression in large peer-to-peer environments.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

The algorithm can be used for data compaction, data modeling and classification tasks in many emerging peer-to-peer applications for bioinformatics, astronomy, social networking, sensor networks and web mining. Computing a global regression model from data available at the different peer-nodes using a traditional centralized algorithm for regression can be very costly and impractical because of the large number of data sources, the asynchronous nature of the peer-to-peer networks, and dynamic nature of the data/network. This paper proposes a two-step approach to deal with this problem. First, it offers an efficient local distributed algorithm that monitors the "quality" of the current regression model. If the model is outdated, it uses this algorithm as a feedback mechanism for rebuilding the model. The local nature of the monitoring algorithm guarantees low monitoring cost. Experimental results presented in this paper strongly support the theoretical claims.

Kanishka Bhaduri *et al.* [10] offers a scalable and robust distributed algorithm for decision-tree induction in large peer-to-peer (P2P) environments. Computing a decision tree in such large distributed systems using standard centralized algorithms can be very communication-expensive and impractical because of the synchronization requirements. The problem becomes even more challenging in the distributed stream monitoring scenario where the decision tree needs to be updated in response to changes in the data distribution. This paper presents an alternate solution that works in a completely asynchronous manner in distributed environments and offers low communication overhead, a necessity for scalability. It also seamlessly handles changes in data and peer failures.

4.1 IDENTIFYING MOST FREQUENT ITEMSET

Souptik Datta *et al.* [11] data intensive large-scale distributed systems like peer-to-peer (P2P) networks are becoming increasingly popular where centralization of data is impossible for mining and analysis. Unfortunately, most of the existing data mining algorithms work only when data can be accessed in its entirety. Finding all the network-wide frequent itemsets is computationally difficult and usually has large communication overhead in such environment. This paper focuses on developing a communication efficient algorithm for discovering frequent itemsets from a P2P network. A sampling-based approach is adopted to find approximate solution instead of an exact solution with probabilistic guarantee. The benefit of approximation technique is reflected in the low communication overhead in discovering majority of frequent itemsets with probabilistic guarantee.

The main principal followed by the algorithm assumes that an independent and identically distributed (id) sample of the entire data is available at one location to generate a set of candidate item sets. Collecting id sample from a P2P network is a challenging problem because of varying degrees of connectivity and sizes of data shared. The paper first addresses this issue and shows how an id sample of nodes and data can be collected from a P2P network using random walk. It applies the proposed sampling technique to identify most of the frequent itemsets from a P2P network. Theoretical analysis shows how to decide about optimum sample size and minimize communication to compute the results.

V. CONCLUSION AND FUTUREWORK

There are so many local algorithms are available for identifying top-l inner product in peer-to-peer network. The efficiency of those algorithm and quality of result is the major issue. There are two types of local algorithms in terms of accuracy exact and approximate. In an exact local algorithm, once the computation terminates, the result computed by each peer is the same as that compared to a centralized execution. For more complicated tasks, researchers have proposed approximate local algorithms using probabilistic techniques. Here we proposed Gossip based probabilistic approximate algorithm. This algorithm relies on properties of random walk on network to provide estimates for various estimates for various statistics of data stored in network. The computation result of this algorithm is exponentially fast. The most important quality of Gossip based probabilistic approximate algorithm is that they provide probabilistic guarantees for the accuracy of result.

Acknowledments

I would like to specially thank God , who guided me through the way and made all things possible. I would like to acknowledge and extend my heartfelt gratitude to my supervisor **Prof. P.RANGARAJAN** whose help, stimulating suggestions, knowledge, experience and encouragement helped me in all the times of study and analysis of the research. I am deeply indebted to **Mr.D.John Aravindhar**, Associate Professor, HITS for his vital encouragement, support and constant reminders in carrying out my research work. I would like to express my gratitude to Director, **Thiru P.Venkatesh Raja**, S.A.Engineering College, Chennai for his inspiration and support towards my research work and also to the Principal Dr.S.Suyambazhahan for his motivation in doing my research work.

I would also like to thank my friends of Information Technology, S.A.Engineering College for their understanding and assistance given to work in this area and also to my family members who were a great source of support and encouragement.

REFERENCES

[1] K. Liu, H. Kargupta, and J. Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 1, pp. 92-106, Jan 2006.

[2] Kanishka Bhaduri, Kamalika Das, Kun Liu, Hillol Kargupta, "Distributed Identification of Top-l Inner Product Elements and its Application in a Peer-to-Peer Network", CIKM , 2006.

[3] B. Babcock and C. Olston, "Distributed Top-k Monitoring," Proc. ACM SIGMOD '03,pp. 28-39, 2003.

[4] Renato Lo Cigno, Alessandro Russo, "On Some Fundamental Properties of P2P Push/Pull Protocols ," Italian Ministry of University and Research, 2009.

[5] Anis ISMAIL, Aziz BARBAR, "P2PDOMAIN CLASSIFICATION USING DECISION TREE," International Journal of Peer to Peer Networks (IJP2P) Vol.2, No.3, July 2011.

[6] H.A. David, Order Statistics. John Wiley & Sons, 1970.

[7] J. H. Friedman, "On bias, variance, 0/1-loss, and the curse-of-dimensionality," Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 55–77, 1997.

[8] C. Aggarwal, "On k-anonymity and the curse of dimensionality," in Proceedings of the 31st VLDB Conference, 2005.

[9] Kanishka Bhaduri , Hillol Kargupta, "A Scalable Local Algorithm for Distributed Multivariate Regression," SIAM Data Mining Conference, 2008.

[10] Kanishka Bhaduri ,Ran Wolff, Chris Giannella, Hillol Kargupta, "Distributed Decision Tree Induction in Peer-to-Peer Systems," Article first published online: 20 MAY 2008.

[11] Souptik Datta, Hillol Kargupta, "A communication efficient probabilistic algorithm for mining frequent itemsets from a peer-to-peer network," Article first published online: 16 JUN 2009.

[12] A.-M. Kermarrec, L. Massoulie´, and A.J. Ganesh, "Probabilistic Reliable Dissemination in Large-Scale Systems," IEEE Trans. Parallel and Distributed Systems, March 2003 (vol. 14 no. 3), pp. 248-258