ISSN No. 0976-5697

Volume 4, No. 7, June 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Review: The Role of Data Mining over Peer-to-Peer Networks

Bharat Kumar
Research Scholar
MJRP University,
Jaipur, India.
bharatbcv@gmail.com

Shilpi Gupta
Dept. of Computer Science,
School of Engineering and System Science,
M.D.S. University,
Ajmer (Raj.), India.
guptashilpi2489@gmail.com

Kamal Kumar Jyotiyana,
Dept. of Computer Science,
School of Engineering and System Science,
M.D.S. University,
Ajmer (Raj.), India.
kamal.jyotiyana26@gmail.com

Abstract: Data Mining is a prominent field in Information Technology. Related to Data (or Information), it simply means Knowledge Discovery from Data Source. The main objective of this research is to study data mining over peer-to-peer network. As Data Mining is a process to turn raw data into meaningful information, it has great advantage over peer to peer network. The Peer-to-peer networks are using in many applications such as file sharing, e-commerce, and social networking. All of these applications require useful data to secure and place from source to destination. The advantage of this paper is better understanding of data mining principles and working of peer to peer network. In the future, it can be deeply research and develop.

Keywords: Data Mining, Peer-to-Peer Network, E-Mule, Kazaa.

I. INTRODUCTION

Data mining is the process of searching, cleaning, collecting analyzing data from various sources of databases for the purpose of evaluation. In simpler words, it is automatic analysis of large chunk of online files in order to discover patterns that may otherwise go unexplored. Data mining known as "knowledge discovery," refers to computer assisted tools and techniques for sifting through and analyzing these vast data stores in order to find trends, patterns, correlations that can guide decision making and increase understanding. Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can be analyzed to predict future trends. Because data mining tools predict future trends and behaviours by reading through databases for hidden patterns, it allows organizations to make proactive and knowledge-driven decisions. It provides solution to the problem those were time consuming earlier. [1] [2] [3]

Peer-to-peer computing is the sharing of computer resources and services by direct exchange between the systems. These resources include the exchange of information, processing cycles, cache storage, and disk storage for files. Peer-to-peer networks are typically used for connecting nodes via largely ad hoc networks. It has no central administrator or coordinator. Peers simultaneously function as both "clients" and "servers". [4] [5]

Data mining application areas over the network:

1). Marketing/Retailing: Data mining can aid direct marketers by providing them with useful and accurate trends about their customers purchasing behaviour, as well as it provides knowledge-sharing over the network.

2). Banking/Crediting: Data mining can assist financial institutions in areas such as credit reporting and loan information. It provides flexibility to the Online-banking operations.

3). Law enforcement: Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviours. It helps security alert system to generate and traverse information within the network.

4). Researchers: Data mining can assist researchers by speeding up their data analysing process. It help researcher to focus on the significant data and provide functionality to change data at real time over the network.

5). Call Detail Record Analysis: Data mining can mine incoming data to see use patterns, build customer profiles from these patterns and then construct a tiered pricing structure to maximize profit. Mobile Phone communication based organization use data mining to discover meaningful data within the range of network.

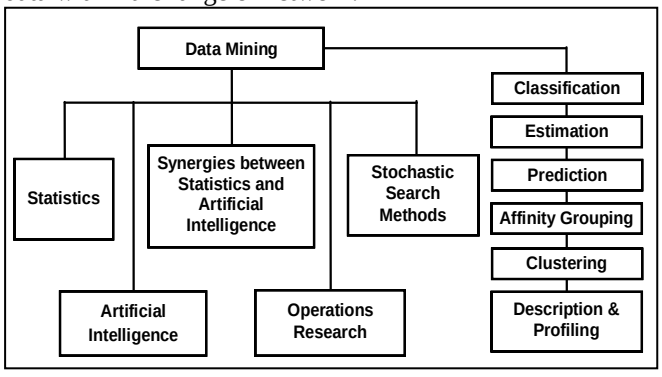

Figure 1. Process of Data Mining

6). Segmentation: Data mining is used to segment large data. It helps to manage information with communication to computer node.

7). Sales forecasting: To examine time-based patterns retailers make stocking decisions. Actually, it uses internet (the network of networks) to get the information.

8). Manufacturing: By applying data mining in operational data, manufacturers can get the information about faulty equipments and determine optimal control parameters.

9). Government: Data mining helps government agency to analyse and extract information about the policies. Network makes it possible to cover all the country location and their records [6]. Characteristics of Peer-to-peer network:

1). Efficient use of resources:
- Unused bandwidth, storage, processing power at the edge of the network.

2). Scalability:
- Consumers of resources also donate resources
- Aggregate resources grow naturally with utilization.

3). Reliability:
- Replicas
- Geographic distribution
- No single point of failure.

4). Ease of administration:
- Self organization of Nodes.
- No need to deploy servers to satisfy demand
- Built in fault tolerance, replication, and load balancing.

5). No central point of failure:
- E.g. Internet and the web do not have a central point of failure.
- Most internet and web services use the client-server model (e.g. HTTP), so a specific service does have a central point of failure.

6). Anonymity-Privacy:
- Not easy in a centralize system

7). Dynamism:
- Highly dynamic environment.
- Ad hoc communication and collaboration. [7]

II. PEER-TO-PEER DATA MINING

Peer-to-peer network is useful file sharing, instant messaging, voice communication, sensor nets, collaboration, backup, distributed computing, and defence. Peer-to-peer network such as the *e-Mule* and *Kazaa* file sharing networks, which are based on point-to-point connection without central server. Data integration applications such as peer-to-peer web mining uses data stored in the Web Browser's cache of different machines connected via a peer-to-peer network, which revolutionize the business of internet search engine.

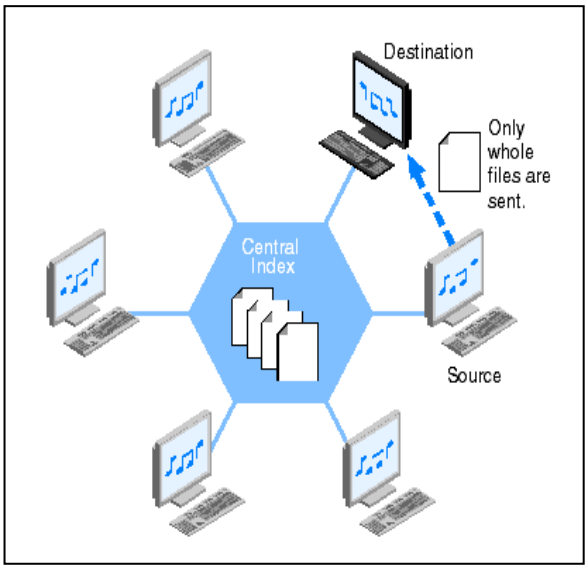

Figure 2. Peer-to-Peer Network

A peer-to-peer data clustering algorithm groups the URL's visited by each user corresponding to different subject by exchanging information with other peers. The common example is Google Reader, which is used to store meaningful information on server. This web-base information can be very useful to the individuals. This may help characterizing each user based on their browsing pattern, and forming communities of peers with similar interests. There are many other similar interesting information integration and knowledge discovering applications involving in data distribution on the peer-to-peer network. Data analysis plays an important role in most non-trivial information integration and retrieval applications. However, most of the mining techniques are designed for centralized applications where all the data are stored at central place. It requires developing some distributed data mining algorithms that are fundamentally decentralized, asynchronous, communication efficient, and scalable.

In the Below figure, it illustrates the application categories by visited URLs according to three subjects (movies, baseball, hurricanes) as well as exchanging information with other peers. Clearly, maintaining the user's privacy is an important issue in such applications.

Although most current peer-to-peer networks deal primarily with file sharing applications (for example music and movies). This research is based on peer-to-peer network while considering wide server-less network with point-to-point connections. This opens up other potential application areas for peer-to-peer data mining, including mobile ad hoc networks (MANET), sensor networks, and federated databases without central coordinator sites. [1] [5] [4]

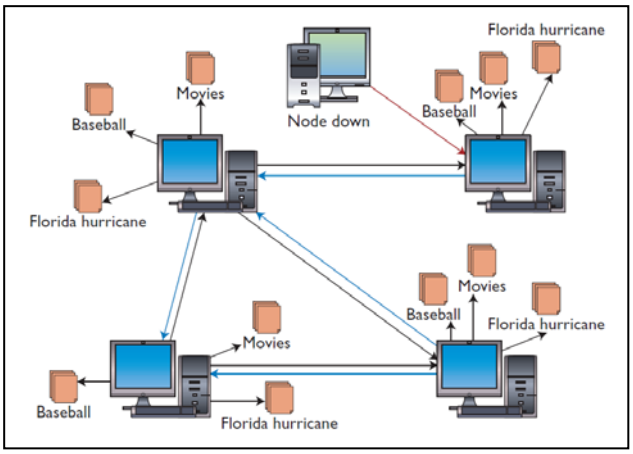

Figure 3. Peer-to-Peer Network using Web Mining application

III. CHALLENGES OF PEER-TO-PEER NETWORK'S DATA MINING

There are various issues and challenges associated with mining on peer-to-peer network. The issues like limited bandwidth, limited memory, CPU capacity and limited battery power are very crucial in the peer-to-peer computing environment since it reduces the efficiency of mining.

High communication cost, non scalability are some of the challenges in peer-to-peer mining.

1). Scalability: Peer-to-peer systems can be very large scale such as millions of nodes, which typically join and leave continuously. These properties are very challenging.

2). Fault Tolerance: Peer-to-peer system must be robust enough to recover from loosing data and partial results as well as peer failures. Peer-to-peer data mining algorithms must be able to work well even if some peer fails.

3). Privacy and Security: Peer-to-peer system must be secure from spread of virus, malware, spyware and adware. Avoid use of steganography. Peers can be assigned reputation values. Secure system from pseudo spoofing and shilling attacks.

4). Decentralization: For task involving large volume of data, centralised system lead to huge data transfer and hence reduce the efficiency and quality of data mining.

5). Asynchronism: The number of nodes in the peer-to-peer system available in the large amount, usually in the range of million. There are a lot of factors like limited bandwidth, connection latency that prevents successful synchronization between the entire peer-to-peer networks.

6). Communication Efficiency: Peer-to-peer data mining algorithms should be communication efficient. A data mining algorithm that is designed to analyze the large volumes of data stored in peer-to-peer systems must be able to work well with less exchange of data among the nodes. In contrast the communication overhead, it should be minimized as much as possible while performing distributed data analysis.

7). Any-time-ness (or Anytimeness): The data mining algorithms for peer-to-peer systems should be incremental. Since, the data at the peers change very frequently, it is not suitable to design algorithms that need to begin from the scratch at every data change. In some applications the rate of data changing will be more than the computation rate. [1]

IV. PEER-TO-PEER DATA MINING: APPLICATION AREAS

There are various applications of peer-to-peer data mining.

1). File Sharing Networks: This is the most popular application area where maximum number of peer-to-peer networks exists today. Peer-to-peer media file sharing networks (such as Napster, Kazaa) have already gained by popularity. These networks allow users to freely join, provided they allow files to be stored on their computers. Data mining application like clustering information for search, personalization and recommendation are very useful for such networks.

2). Sensor Networks: Light-weight, inexpensive sensor with wireless communication capabilities can be easily deployed in large numbers to form sensor networks. The potential of such networks is already widely recognized. The sensor may need to operate in a router less environment with no global IPs and communicate with each other in peer-to-peer fashion.

3). Mobile Ad hoc Networks: Mobile Ad hoc Networks are increasing attention in many wireless application domains. Such ad hoc networks may play a key role in defining how we communicate at work and social environment in the future. Several data rich environments are emerging and they are likely to need data analytic supports for efficient and personalized services.

4). Peer-to-Peer Scientific Computing Platform: Peer-to-peer scientific computing platform is another application area where peer-to-peer data mining algorithms can be useful. One of the most used peer-to-peer service is the 'Chinook' platform for peer-to-peer informatics service. The Chinook platform facilitates exchange of analysis techniques worldwide.[4][8]

V. EVOLUTION OF PEER-TO-PEER DATA MINING

Peer-to-peer data mining is a new field that has grown out of distributed data mining, which itself is a fairly new research area. Distributed data mining has evolved over the past 5-10 year as an effort to introduce distributed versions of many standard data mining algorithms, such as association rule mining, Bayesian network learning and clustering. However, most such efforts assume a stable network and data, and so they can't be applied directly to peer-to-peer network. Researchers have develop various complex algorithms by applying efficient primitives, many peer-to-peer data mining efforts have focussed on developing primitive operations (such as average, sum, max, random sampling and so on), laying foundation for more sophisticated data analysis and mining algorithms.

Wojtek kowalczyk and his colleagues developed the model which can calculate mean of data distributed in peer-to-peer network. In contrast to these approaches, which all require resources that scale directly with the system size, local algorithms can compute results and make definite claims regarding correctness using information from just a handful of nearby neighbours in a peer-to-peer system. The

resources required by such algorithms are often independent of system size, which presents obvious benefits for scalability and fault tolerance. Ran Wolff, Kanishka Bhaduri, and Hillol Kargupta have also proposed algorithms for monitoring k-means clustering in peer-to-peer networks. [1] [5]

VI. CONCLUSION

Peer-to-peer systems are increasingly being popular for many applications that go beyond sharing media files. Huge amount of data are stored in various types of networks which is based on communication between peer to peer. As discussed above, the approaches of data mining make information to scale over network without loss of data. The study about data mining and its application concludes that it is the best way to fetch meaningful information from the available databases (at peer) over the network.

VII. ACKNOWLEDGEMENT

I take this opportunity to express my profound gratitude and deep regards to my guide (Associate Professor Dr. Neeraj Bhargava) for his exemplary guidance, monitoring and constant encouragement throughout the article.

VIII. REFERENCES

[1]. Souptik Datta, Kanishka Bhaduri, Chris Giannella, Ran Wolff, Hillol Kargupta "Distributed Data Mining in Peer-to-Peer Networks".
[2]. Wojtek Kowalczyk, Mark Jelasity, A.E. Eiben "Towards Data Mining in Large and Fully Distributed Peer-to-peer Overlay Networks".
[3]. Joyce Jackson "Data Mining: A Conceptual Overview", Communication of the Association for Information Systems, Volume 8, 2002.
[4]. "Peer-to-Peer Computing: Overview, Significance and Impact, E-learning, and Future Trends".
[5]. E. Anupriya, N.Ch.S.N.Iyengar "A framework for optimizing the performance of peer-to-peer distributed data mining algorithms", International Journal of Computing Science and Communication Technologies, Volume 3, No. 1, July 2010.
[6]. Kanishka Bhaduri "Efficient Local Algorithms for Distributed Data Mining in Large Scale Peer-to-Peer Environments: A Deterministic Approach".
[7]. Kien A. Hua, Duc A. Tran, and Tai Do, "ZIGZAG: An Efficient Peer-to-Peer Scheme for Media Streaming", INFOCOM 2003.
[8]. Haimonti Dutta, Ananda Matthur "Distributed Optimization Stratergies for Mining on Peer-to-Peer Networks".
[9]. Ping Luo, Hui Xiong, Kevin Lu, Zhongzhi Shi "Distributed Classification in Peer-to-Peer Networks".