ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(32), DOI: 10.17485/ijst/2015/v8i32/87597, November 2015* ISSN (Online) : 0974-5645

Utilizing the Amenities of Unstructured Peer to Peer Framework in Extraction of Large Scale Data Clusters

A. R. Arunachalam*

Department of CSE, Bharath University Chennai - 600073, Tamil Nadu, India;
ararunachalam78@gmail.com

Abstract

Historical data are commonly used in large scale industry for efficient manipulation and retrieval of datasets. Managing a historical data in multidimensional data cubes is a complicated process. This paper aims to utilize the amenities of a peer to peer framework in extraction and management of large scale data clusters. The data clusters are mined by finding interesting patterns in specialized domains.

Keywords:

1. Introduction

In a peer to peer environment process and workload is shared by all the peers. All the peers are equipotent in a peer to peer scenario. In other words there is no prioritization for sharing the workload or other system resources such as CPU time and memory.

2. Structured P2P Network

In a structured peer to peer network every peer is equipotent and share the resources, process and workload equally among the all the peers. The resources include computing time, CPU registers and secondary hard disc storage spaces. The peers are connected to one another in an equivalent connection. The overlay network defines whether a peer to peer environment is structured or unstructured. In a structured overlay network, the standard set of algorithms and routines are used for process sharing among the peers. Data retrieval and management among the peers are easier and not complicated as in the case of unstructured peer to peer network.

2.1 Unstructured P2P Network

An unstructured P2P network is formed when the overlay links are established arbitrarily. Such networks can be easily constructed as a new peer that wants to join the network can copy existing links of another node and then form its own links over time. In an unstructured P2P network, if a peer wants to find a desired piece of data in the network, the query has to be flooded through the network to find as many peers as possible that share the data. Aggregate data describes data combined from several measurements.

Domain Driven Data mining (d3m) proved to be an emerging concept in managing large scale industry specific data. In majority of industry specific database systems, the traditional data mining techniques proved to be inefficient in generating data patterns. d3m involves specialized mining techniques based on the domain

* *Author for correspondence*

specific knowledge. The Domain specific knowledge discovery involves analysis of the domain environment and the classification of data present in that domain. The classification and clustering of data is used to identify and discover interesting patterns. These interesting patterns are used to retrieve multidimensional aggregate data based on specific domain.

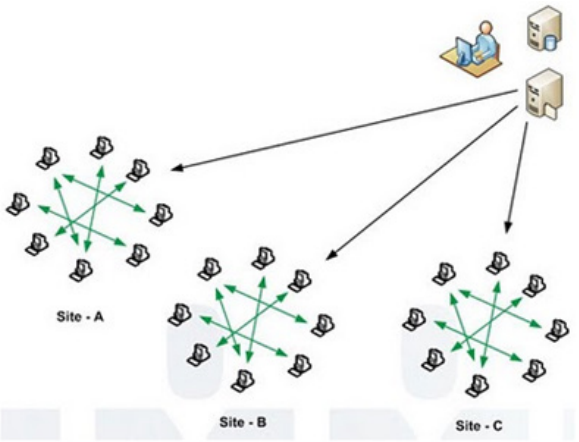

Figure 1.

3. Multidimensional Historical Data

In complex business environments, individuals who analyze demand data and use forecast results are normally interested in seeing the information presented in a way that helps them better manage certain aspects of their business, such as sales territories, major customers or product lines. As part of that process, they need an easy, efficient way to focus in and segment the data stored in their database systems.

Consequently, in today's business world, there is usually no such thing as a simple, stand-alone item that is managed and forecasted in isolation. Each of your product items may be characterized by multiple dimensions, such as SKU number, sales region, market segment, customer, product line, etc. Demand data for each item will be measured and collected across all dimensions. In addition, other information, such as conversion factors, forecast parameter values and promotional.

3.1 Domain Driven Approach

Database administrators find it difficult in adapting the traditional data mining techniques in real time, knowledge centered environment. Industrial domains need a powerful, efficient, deliverable, problem solving routines that identifies patterns based on the imparted intelligence. Actionable Knowledge Discovery (AKD) is the key factor to take in consideration. The Actionable knowledge is obtained by imparting specialized knowledge about the domain to the framework.

Knowledge Discovery in Databases (KDD) refers to the process of extraction of Actionable, Deliverable knowledge from the raw data obtained from the database. Knowledge refers to relationships and patterns that provide problem solving solutions between the data elements.

Actionable Knowledge Discovery (AKD) involves multiple dimensions of requirements from both macro level and micro level from the real world applications. There exists a difference between intrinsic academic thinking and business expectations.

Any Domain driven approach had to consider the following key features. Environment-factors surrounding data mining models and systems.

- **Human Intelligence**-Human centered and imparting human thought process to the framework is crucial for solving complex problems.
- **Process**-In the domain driven approach, the mining process involves dynamic and iterative involvements of environmental elements.
- **Pattern**-Interesting data patterns are generated by balancing the technical and business perspective based on the subjective aspects.
- **Adaptation**-The pattern generated must be dynamically adaptable to the domain.it must cater to the changing business needs and user needs and expectations.
- **Actionability**-Several metrics to be developed to test the actionability of the discovered patterns in the business domain.
- **Deliverable**-The solutions obtained should be presented to the end user in a business friendly way that is compatible with business operating systems and rules.

These are the key elements in a specific user domain.

- The purpose of domain specific data mining is to convert effective data mining algorithms in to problem solving algorithms.
- Domain driven approach converts the abstract, synthetic refined data in to real life data.
- Domain driven approach is a multistep iterative process.
- D3m is a human centered and human mining methodology.

- D3m involves closed loop problem solving methods in open environment.
- It uses ad-hoc, dynamic and customizable models.
- It's a trade-off between technical significance and business expectation.

4. Imparting Ubiquitous Intelligence

Ubiquitous intelligence makes d3m deliver business friendly, decision making rules and actions with technical significance. Ubiquitous intelligence involves imparting social intelligence, the network domain where the domain being adopted.

4.1 Imparting Network Intelligence to the Framework

Network intelligence involves both web intelligence and information distribution among different networks. The information surrounding the network is used to obtain deliverable knowledge for the business problem. Network intelligence in some way or the other caters for the Actionable Knowledge Discovery (AKD) process.

In our project, a framework is being proposed that utilises the amenities of peer to peer (p2p) networks for the effective extraction and management of aggregate data. Imparting network intelligence to the domain in order to mine and perform analytical processing of data to obtain deliverable, data patterns is the idea proposed.

The primary aspects of imparting network intelligence provide a

- Possibility of discovering business intelligence in aggregate data's distributed among different peers to find a solution to a business problem.
- Discovering data clusters and classification among the peers.
- Imparting unstructured peer to peer network information in discovering pattern mining on target data.
- Utilising peer to peer network facilities to pursue information search process and to enhance Actionable Knowledge Discovery (AKD) process.

4.2 Social Intelligence

Social intelligence refers to intelligence behind group interactions, behaviour and regulations of the user. Human social intelligence refers to human social cognition, emotional intelligence, consensus construction and group decision.

In mining patterns in complex environment and peer based networks, social intelligence plays a main aspect.

Social intelligence can be used in

- Use of human intelligence in supervised data mining and range query evaluation.
- Multiagent data mining and warehousing to facilitate human model interaction to facilitate group decision making and autonomous auto selection.
- Developing performance evaluation models to evaluate and maintain quality of data mining software.
- Project management, business process management and analysing the delivery of data from the analyst to the operational department.

4.3 Human Intelligence

Human intelligence may refer to direct involvement of human empirical knowledge, belief, intention, expectation, run time supervision, and experts group into Actionable Knowledge Discovery (AKD).

Implicit or indirect involvement of human intelligence includes imaginary thinking, emotional intelligence, inspiration and brainstorming. Data mining in an enterprise involves both intrapersonal and interpersonal human intelligence.

Interactive system for mining and understanding cross market trade behavior is a typical example of interpersonal human intelligence.

To involve hum an intelligence into AKD many issues needs to be addressed,

- Interactive data mining.
- Human centered interactive data mining deal with interface design and major roles played by humans in pattern mining and generation.
- Dynamic user modeling and to input the characteristics to the data mining systems.
- Online user interaction which support online users to interact with the data mining system.
- Group decision making.
- Adaptive mining for the clients to adapt to the pattern mining process.
- Distributed interaction catering with multiple clients to interact with the data mining servers.

- Consensus building which facilitates the users and provider clients to find mutually agreeable findings and pattern generation.

5. Intelligence Meta Synthesis and Peer to Peer Environment

Experts often find it difficult to impart ubiquitous intelligence into complex data mining framework. Intelligence meta synthesis proposes to introduce meta synthesis (m space) supporting the interaction of metadata and meta computing (m computing).

It integrates the ubiquitous computing techniques including the knowledge engineering process, human centered computing, social computing and behavioral analysis.

The domain driven data mining (d3m) concepts and methodologies such as social intelligence, human intelligence and combining all the features in meta synthesis process is combined together to form a single, unified justified algorithm.

This single, unified justified algorithm combining the domain driven modules is induced in unstructured peer to peer environment.

5.1 Data Compression

Data compression is used just about everywhere. Data compression is useful because it helps reduce the consumption of expensive resources, such as hard disk space or transmission bandwidth. On the downside, compressed data must be decompressed to be used, and this extra processing may be detrimental to some applications. For instance, a compression scheme for video may require expensive hardware for the video to be decompressed fast enough to be viewed as its being decompressed (the option of decompressing the video in full before watching it may be inconvenient, and requires storage space for the decompressed video). The design of data compression schemes therefore involves trade-offs among various factors, including the degree of compression, the amount of distortion introduced (if using a lossy compression scheme) and the computational resources required to compress and uncompress the data.

Compression is useful because it helps reduce the consumption of expensive resources, such as hard disk space or transmission bandwidth. On the downside, compressed data must be decompressed to be used, and this extra processing may be detrimental to some applications. For instance, a compression scheme for video may require expensive hardware for the video to be decompressed fast enough to be viewed as it is being decompressed (the option of decompressing the video in full before watching it may be inconvenient, and requires storage space for the decompressed video). The design of data compression schemes therefore involves trade-offs among various factors, including the degree of compression, the amount of distortion introduced (using a lossy compression) and the computational resources required to compress and uncompress the data.

6. General Framework of the Project

- Creating justified algorithm to automatically abstract patterns of data in the distributed unstructured peer to peer network.
- Client side scripting for designing the receiver client to obtain the distributed data obtained from different peers.
- Performing lossy compression technique to compress the data to be distributed among the peers.
- Indexing the distributed data by using R tree for creating synopsis and sub synopsis for the indexed data's.
- Creating the justifying algorithm combining all the other modules.

7. Modules

- Adopting domain specific data mining methodologies to facilitate intelligent retrieval of relevant data distributed among the peers.
- Designing receiver client to search for the required data among the peers.
- Compressing the data in the client side.
- Indexing the values by creating synopsis and sub synopsis.
- Creating the justifying algorithm to facilitate the identification of the peer in which the relevant data is distributed.

7.1 Module Description

Domain specific mining methodologies

In an unstructured peer to peer environment, data's are distributed among different set of peers. There is no standard set of algorithms or routines for efficient management and usage of hardware and software resources. By using the amenities of peer to peer based systems, retrieval of data distributed among the peers

may be facilitated. The drawback discovered in this IEEE paper is that it extracts only frequently used datasets. The probability of routing a peer which holds the most frequent datasets is more when compared to the peer which holds the less frequently used datasets.

Our paper propose to solve the problem by adopting domain driven data mining techniques and creation of range queries that intelligently transforms data specific query results to solution derivable Actionable, deliverable patterns of data.

8. Designing the Receiver Client

In this module we design a client which is giving request to the server for the desired file which may be present among the peers in the network. The client itself must have the view of server to select the files from the server. While sending a request it transfers its own Internet Protocol (IP) address so that the server comes to know whether the request is obtained from a valid user or not. After checking it has to take care by the provider client.

9. Compressing the Data in Provider Side

In this module compression of data in the provider side is concentrated by clustering the data's by their similar types so that the data's are compressed under common clusters. By compressing the data by their similar types facilitates easy distribution of data among the peers and easy manipulation of the data by creating synopsis and sub synopsis.

9.1 Indexing the Values

Indexing is done based on the R-tree concept. The sub-synopsis is created which was manifested based on its two contents i.e., the description and the indexing structure. The indexing structure must be efficient enough for the client to find the provider in the network and then able to look for the file it's searching for.

9.2 Justifying Algorithm

CHIST algorithm is used for the compression of data. The data sets are compressed individually and then scattered in the network. Replication is the major hindrance so that it has to be governed. Duplication of data must be done on demand only.

10. Compressing Data in Provider Side

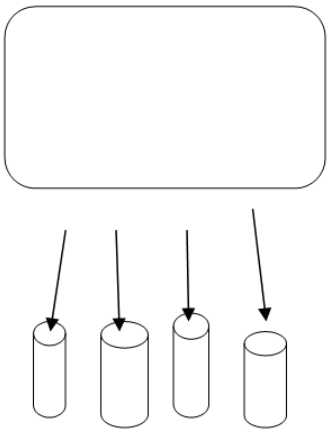

Figure 2. Provider client.

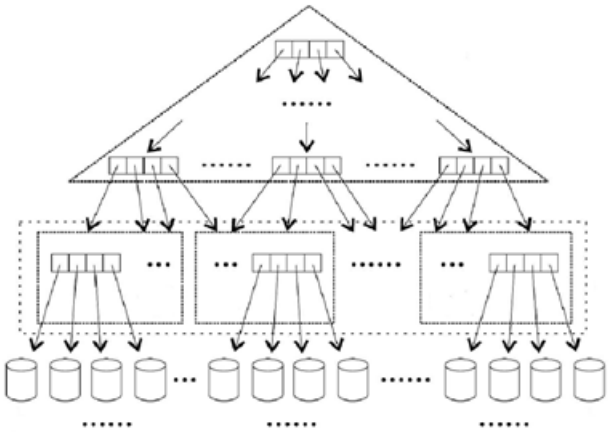

Figure 3. Indexing the values.

11. Existing Project

Multidimensional historical aggregate data is mined from large, complex databases to extract useful patterns to solve complex business problems. The data is compressed and scattered among the set of equipotent peers. When a data set is searched the possible set of peers which may hold the relevant data is mapped to the receiver client.

12. Drawback in Existing Project

This framework provides efficient data retrieval for most commonly retrieved data sets, or the previously retrieved data sets. It's not efficient in mapping the peers which contain the data clusters which is not most commonly

retrieved. The data sets are routed based on the synopsis and the sub synopsis that is created. When a receiver client request for a data it is mapped to the corresponding peer only if the data matches the sub synopsis indexed on the provider. It does not perform data mining and classification for in frequently used data sets.

13. Proposed Framework

The proposed framework generates useful patterns of data based on behavioral analysis of the client. The data sets are clustered by their similar types. This facilitates relevant mapping of peers to the respective data's. Based on the specific domains, the range queries are evaluated to predict the peers which share the data among them. This ensures mapping of infrequent data distributed among the peers to the requesting client.

14. References

1. Gkantsidis C, Mihail M, Saberi A. Random walks in peer to peer networks. INFOCOM. IEEE 23rd Annual Joint Conference of the Computer and Communications Societies; 2004.
2. Kaliyamurthie KP, Udayakumar R, Parameswari D, Mugunthan SN. Highly secured online voting system over network. Indian Journal of Science and Technology. 2013; 6(S6):4831-6.
3. Kiran Kumar TVU, Karthik B. Improving network life time using static cluster routing for wireless sensor networks. Indian Journal of Science and Technology. 2013; 6(S5):4642-7.
4. Langeswaran K, Revathy R, Kumar SG, Vijayaprakash S, Balasubramanian MP. Kaempferol ameliorates Aflatoxin B1 (AFB1) induced hepatocellular carcinoma through modifying metabolizing enzymes, membrane bound ATPases and mitochondrial TCA cycle enzymes. Asian Pacific Journal of Tropical Biomedicine. 2012; 2(S3):S1653-9.
5. Gupta A, Agrawal D, Amr El A. Approximate range selection queries in peer to peer systems. Proceedings of the CIDR Conference; 2003.
6. Muruganantham S, Srivastha PK, Khanna. Object based middleware for grid computing. Journal of Computer Science. 2010; 6(3):336-40.
7. Khanna V, Thooyamani KP, Saravanan T. Simulation of an all optical full adder using optical switch. Indian Journal of Science and Technology. 2013; 6(S6):4733-6.
8. Guttman A. R-tree dynamic index structure for spatial searching. International Conference on Management of Data; 1984.
9. Jagadish HV, Ooi BC, Vu QH. BATON: Balanced tree structure for peer to peer networks.
10. Cao L. Domain driven data mining-challenges and prospects. IEEE Transactions on Knowledge and Data Engineering; 2010.