Visual Data Mining Framework Based on a Peer-to-Peer Architecture

Hamadi Abdelkrim, Nader Fahima
Ecole nationale Supérieure d'Informatique, ESI
Algiers, Algeria
e-mail : {a_hamadi, f_nader@}esi.dz

Abstract—Nowadays, the continuous hardware evolution is driving computer systems to be able to store large amounts of data; therefore, exploring and analyzing this huge volume of data is becoming more and more difficult to manage. Peer-to-Peer (P2P) network produce more computing power in terms of processing, communication systems and large storage. Visual Data Mining (VDM) using P2P infrastructure can be more beneficial to improve both performance and quality of the selected data. This paper aims at describing a prototype of our framework Visual Data Mining Distributed (VDMD) based on VDM algorithms in a P2P architecture.

Keywords-Visual Data Mining; Peer-to-Peer Architecture; JXTA specification

I. INTRODUCTION

Internet is a network consisting of millions of computers connected at any given time. All the computers are theoretically connected to one another, and information stored on any of these systems can be accessed. The topology of computers on the Internet is a group of machines spread out in various locations. Computers within each group or subnet are visible to each other and sometimes visible to other subnets on the Internet.

Advances in computing and communication over networks, such as Internet, intranets, and wireless networks, have resulted in various pervasive distributed environments. Many of these environments have to deal with massive data collections in terabyte scale maintained over geographically distributed sites. The data is collected as potential source of valuable information, providing a new competitive advantage. However, finding the valuable hidden information is a difficult task. Thus, some VDM algorithms are applied to find classifiers, associations, clusters and other patterns in large and distributed data sets. The purpose of VDM is either to help explain the past, or try to predict the future based on past data. Data mining techniques help identify patterns in a vast data store, and then build models that concisely represent those patterns [1].

Distributed computing plays an important role in the VDM process for several reasons. First, VDM often requires a huge amount of resources in terms of storage space and computation time. Second, data are often inherently distributed into several databases, making the centralized processing of this data not very inefficient and prone to security risks [1].

In this paper, we describe our efforts to create a framework to implement VDM in P2P architecture. The remainder of this paper is organized as follows. Section II presents a review of VDM. Section III introduces the P2P architecture. Section IV discusses the proposed VDMD framework and concludes with future work.

II. VISUAL DATA MINING

For data mining to be effective, it is important to include the human in the data exploration process and combine mainly creativity and pattern recognition abilities of human with the enormous storage capacity and the computational power of today's computers. The basic idea of data visualization is to present the data in some visual form, allowing the human to get insight into data. Visualization becomes useful as soon as the data analysis starts and the exploration goals are still vague.

Data Exploration usually follows a three step process: Overview first, zoom and filter, and then details-on-demand [2]. The Data Mining (DM) expert always needs to get an overview of the data what helps him to identify interesting patterns in this data. This corresponds to Data understanding phase of CRISP-DM methodology [3].

The techniques of VDM can be classified based on three criteria (see Figure 1): the data to be visualized, the visualization technique, and the interaction and distortion technique used [4].

The data, usually, consists of a large number of records each consisting of a list of values calling in data mining attributes or in visualization dimensions. We call the number of variables the dimensionality of the data set. Data sets may be one-dimensional, such as temporal data; two-dimensional, such as geographical maps; multidimensional, such as tables from relational database.

The next data types are text/hypertext or hierarchies/graphs. Text data type is distinguished by the fact that it cannot be easily described directly by numbers and therefore text has to be firstly transformed into describing vectors, for example word counts. Graphs types are widely used to represent relations between data, not only data alone. The last types are algorithms and software.

 ISBN: 978-1-61208-381-0

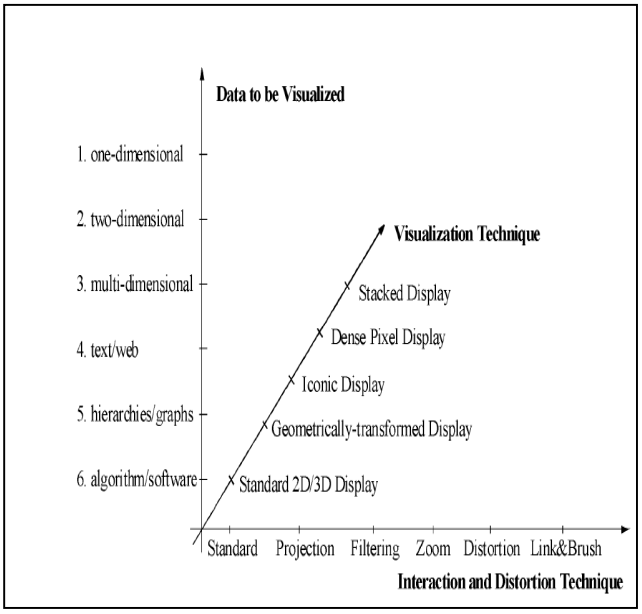

Figure 1. Classification of information visualization techniques.

Note that our framework is designed to support different data types and that it can use a combination of multiple visualization and interaction techniques.

III. DISTRIBUTED ARCHITECTURES

The client-server architecture, P2P architecture and hybrid architecture try to achieve scalability through various means. Scalability can be achieved either by increasing the resources or by reducing the consumption.

A. The Client-Server architecture

The client–server model of computing [5] is a distributed application structure that partitions tasks between the providers of a resource or service, called servers, and service requesters, called clients. Often clients and servers communicate over a computer network on separate hardware, but both client and server may reside in the same system. A server host runs one or more server programs which share their resources with clients. A client does not share any of its resources, but requests a server's content or service function. Clients, therefore, initiate communication sessions with servers which await incoming requests.

B. P2P architecture

P2P computing or networking [5] is a distributed application architecture that partitions tasks between peers. Peers are equally privileged, equipotent participants in the application. Peers make a portion of their resources, such as processing power, disk storage or network bandwidth, directly available to other network participants, without the need for central coordination by servers or stable hosts. Peers are both suppliers and consumers of resources, in contrast to the traditional client-server model in which the consumption and supply of resources is divided. Emerging collaborative P2P systems are going beyond the era of peers doing similar things while sharing resources, and are looking for diverse peers that can bring in unique resources and capabilities to a virtual community thereby empowering it to engage in greater tasks beyond those that can be accomplished by individual peers, yet that are beneficial to all the peers.

C. The Hybrid architecture

It is possible to combine P2P architecture with a server-based architecture. Hybrid models are a combination of peer-to-peer and client-server models. A common hybrid model is to have a central server that helps peers find each other. There are a variety of hybrid models, all of which make trade-offs between the centralized functionality provided by a structured server/client network and the node equality afforded by the pure peer-to-peer unstructured networks. Currently, hybrid models have better performance than either pure unstructured networks or pure structured networks because certain functions, such as searching, do require a centralized functionality but benefit from the decentralized aggregation of nodes provided by unstructured networks.

IV. OUR FRAMEWORK VDMD

Guedes et al. [6] described a service-oriented architecture (SOA) that offers simple abstractions for users and supports computationally intensive applications for data mining. Zhang et al. [7] adopted Web Services Description Language (WSDL) for specifying the interfacing of the data mining components, and Business Process Execution Language for Web Services (BPEL2WS) for specifying the execution flow.

The proposed VDMD framework is based on the hybrid topology where computers can be both client and server. The VDMD framework offers the following features:

- Huge amount of resources in storage space and computation time.
- Make the information from mass of data discovery and display to be better.
- A fragmentation of the databases.

Figure 2 shows the VDMD framework architecture. It is composed of nine main components described as follows:

- **DatabasePeer**: This peer is responsible for all database access and control. Data is received by the peer and placed in one database associated with the peer. Depending on the needs, the database can be either on the same node as the peer or on a different one as: VDM Server A, VDM Server B, Server A, Server B, Server C. The tables and indexes of the databases can be partitioned.
- **DateminingPeer**: This peer is charged to execute the data mining programs.

 ISBN: 978-1-61208-381-0

- **GatheringPeer**: This peer is responsible for gathering any data result from "DM Peer Group" and saving that data to a businessPeer. This peer could be a spider that looks at the result of data mining programs for data.
- **OverviewPeer:** This peer is responsible to get the user a global overview of the data.
- **FilterPeer:** This peer is responsible to execute all filters on the data.
- **DetailPeer:** This peer is responsible for executing the "on-demand details" requests.
- **ZoomPeer:** This peer can display more information about the selected data.
- **BusinessPeer:** This peer is responsible for acting as a buffer between the GUIClientPeer and the "VDM Peer Group" (OverviewPeer, FilterPeer, DetailPeer and ZoomPeer). This peer simply receives a packet and forwards it to the "VDM Peer Group", but additional logic could be incorporated.
- **GUIClientPeer:** This peer is responsible for requesting an image from the database to be displayed. The GUIClientPeer will typically be a GUI-based application that a person will use to request data from the data. The peer will interact with a BusinessPeer, which will in turn attempt to communicate with "VDM Peer Group".

Our protytype can support any VDM technique. Bellow, a working flow:

- We supposed that the datasource are fragmented on three servers : VDMServerA and ServerB and ServerC.
- Implementing the program to access to data on three machines "DatabasePeer".
- Choosing the circle segments technique [8] from the dense pixel techniques, it maps each dimension value to a coloured pixel and group these pixels belonging to each dimension into adjacent areas. To arrange the pixels on the screen, there are two techniques: recursive pattern technique and circle segments technique.
- Implementing the program of circle segments technique on three machines "Datamining Peer" so each machine can perform this program on the data of the three servers: ServerA, ServerB and ServerC.
- Imlplementing the interaction and distorition techniques(Dynamic projections, interactive filtering, intercative zomming and interactive distortion) on four machines as following: OverviewPeer, FilterPeer, DetailPeer and ZoomPeer.
- Implementing a GUI application which offers the services and displays the result to the final users.

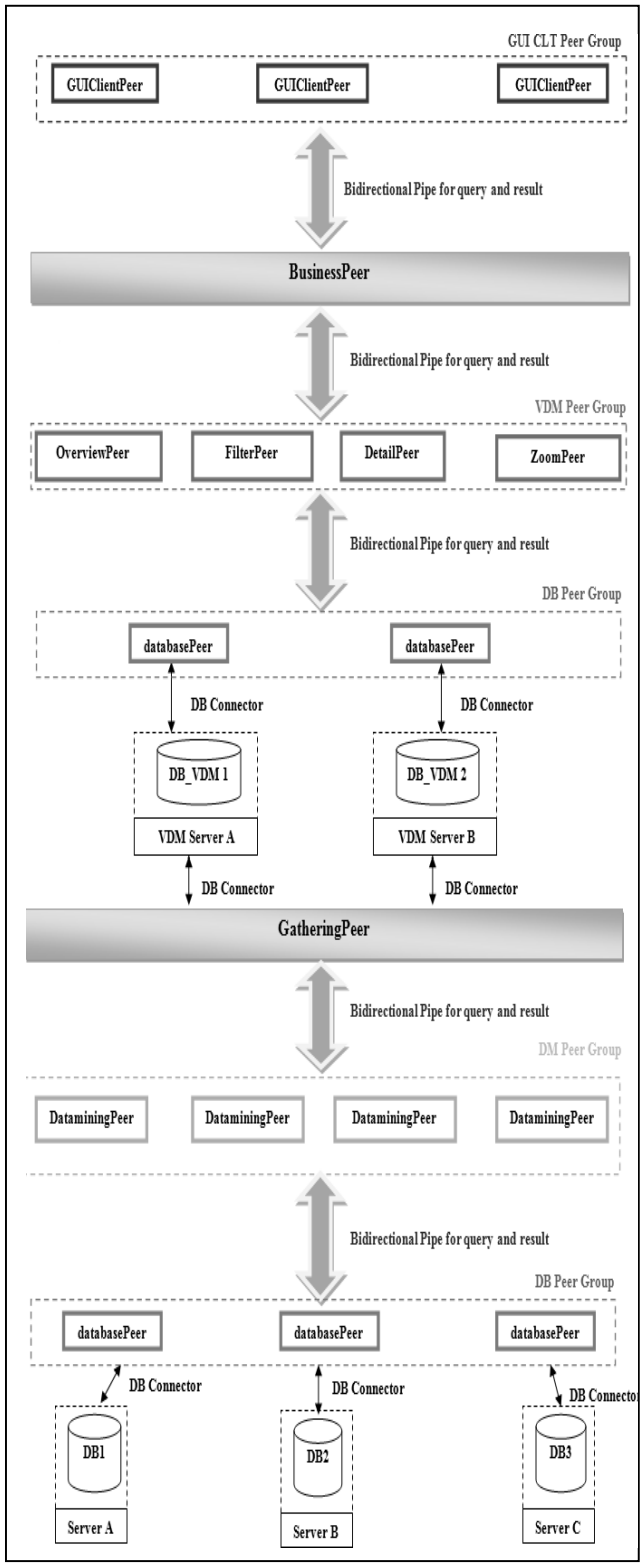

Figure 2. The basic architecture of VDMD framework.

Until now, we did not implement the proposed framework. Our plan is to test it with different data sets and data mining techniques then compare it to client-server and grid architectures.

V. CONCLUSIONS

The proposed Visual Data Mining framework based on Peer-to-Peer architecture has a modular, extensible design. Our architecture can use various data mining programs [9] and handle a huge amount of data [10]. In our future work, we plan to continue our research as bellow:

- Test and evaluate the P2P architecture with Peersim simulator [11] then compare it with a Client-Server architecture and a grid architecture.
- Implement the P2P architecture with JXTA specifications.

There are major futures challenges developing this framework :

- Handle various structures of data.
- Compare the performances of our architecture with the grid architecture.
- Improve the displaying of the data to the final user.
- Introduce other peer group.

REFERENCES

[1] I. Niskanen and J. Kantorovitch,"Ontology driven data mining and information visualization for the networked home", Research Challenges Infromation Science (RCIS), Fourth International Conference, May 2010, pp. 147-156.

[2] B. Shneiderman, "The Eye Have It: A Task by Data Type Taxonomy for Information Visualizations", Visual Languages, 1996.

[3] CRISP-DM: Crosss Industry Standard Process for Data-Mining, 1999.http://www.crisp-dm.org/.

[4] A. Keim Daniel, " Information Visualization and Visual Data Mining", IEEE Transactions on Visualization and Computer Graphics, January-March 2002, vol. 8, no. 1.

[5] A. Yahyavi and B. Kemme, "Peer-to-Peer Architectures for Massively Multiplayer Online Games: A Survey", ACM Computing Surveys (CSUR), vol 46, iss. 1, October 2013.

[6] D. Guedes, W. Meira and R. Ferreira, "A Service-Oriented Architecture for High-Performance DataMining", IEEE Internet Computing , , July-August 2006, vol 10, No 4, pp. 36 – 43.

[7] X. Zhang H-F. Wong, and W.Cheung, "A Privacy-Aware Service-oriented Platform for Distributed Data Mining", Proceedings of 3rd IEEE Conference on Entreprise Computing, E-Commerce and E-Services (EEE'06), San Francisco, California, June 26-29 2006, pp. 44-48

[8] M. Ankerst, D.A. Keim, and H. Kriegel, "'Circle Segments': A Technique for Visually Exploring Large Multidimensional Data Sets", Proc. Visualization '96, Hot Topic Session, San Francisco, CA, 1996.

[9] E. Kandogan, "Visualizing Multi-Dimensional Clusters, Trends, and Outliers using Star Coordinates", Proc. ACM SIGKDD '01, 2001, pp. 107-116.

[10] U. Fayyad, U. Piatetsky-Shapir, and G. Smyth, "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, vol. 39, iss. 11, Nov 1996, pp. 27-34.

[11] http://peersim.sourceforge.net/, retrieved: October, 2014.

 ISBN: 978-1-61208-381-0