Comparative Analysis of Issues Related to Centralized and Distributed Warehouse Architectures

Preeti Baser¹, Dr. Jatinderkumar R. Saini²

¹Preeti Baser, Assistant Professor, SJPIBMCA, Gandhinagar, Gujarat, India – 382 007
Research Scholar, R. K. University, Rajkot

²Director (I/C) & Associate Professor,
Narmada College of Computer Application, Bharuch, Gujarat, India – 392 011.
Research Guide, R. K. University, Rajkot

ABSTRACT

Data mining normally collects the data from large databases at central warehouse and applies mining techniques on collected data and discovers patterns and generates useful information which is used for further decision making process. Data mining has a lot of importance because of its vast applicability. It is being used progressively more in business applications for understanding and predicting valuable data, like consumer buying actions and buying tendency, profiles of customers, industry analysis and financial services. Distributed data mining is concerned with discovering patterns and trends from large distributed databases. This paper presents an overview of general architecture of centralized warehouse and distributed data warehouse. It also presents various issues related to both the approaches. The paper ends with our views on comparative analysis of centralized warehouse and distributed warehouse.

Keywords: Centralized Warehouse (CW), Distributed Warehouse (DW), Knowledge Discovery (KD)

1. INTRODUCTION

Knowledge discovery (KD) is a process aiming at the extraction of previously unknown and implicit knowledge out of large databases which may potentially be of added value for some given application [8]. The automated extraction of unknown patterns, or data mining (DM), is a central element of the KD process. Data Mining is used in several applications like market research, consumer behavior, direct marketing, genetics, text analysis, customer relationship management, intrusion detection [6] and financial services. Mining techniques can help companies to provide better, customized services and support decision making. However, there are many challenges of data mining posed by very large and complex data set. One of the main challenges in data mining is the development of efficient techniques that scale up to large and possibly physically distributed data sets. Distributed Data Mining (DDM) is one solution of this challenge. Distributed Data Mining (DDM) [3], [7] is concerned with the application of the classical Data Mining procedure in a distributed computing environment trying to make the best of the available resources (communication network, computing units and databases). Distributed data mining mines data sources regardless of their physical locations.

2. RELATED WORK

Virtual data warehouses provide end-users with direct access to multiple operational databases and files through middleware tools [1]. The virtual data warehouse uses a network to allow end-users on terminals or client workstations direct access to operational system. The virtual warehouse is a subject oriented, current-valued and detailed-only collection of data in support of an organization’s need for up-to-the-second operation information. An Operational Data Store (ODS) represents a collective, integrated view of the current operations of the organization. It is fed by the operational databases and serves the role of integrating data within the operational system. The operational data store is a subject-oriented, integrated, volatile, current-valued and detailed-only collection of data in support of an organization’s need for up-to-the-second, integrated and collective information [10]-[12]. The data warehouse contains consolidated data extracted from multiple operational systems and/or merged with data from other external information systems. It supports long to medium-term decisions. It is a subject-oriented, integrated, non-volatile, time-variant, summarized and detailed collection of data supporting managerial decision making for the entire organization. Data
marts are subject-oriented, integrated, non-volatile, time variant summarized and detailed collection of data in support of divisional decision making.

There are two approaches for distributed data warehouses. Inmon’s approach assumes the existence of both local and global data warehouses with data stored in each being mutually exclusive. The local data warehouse contains data of interest to the local site and includes historical data in addition to local DSS functions. The global data warehouse contains data common across the corporation and data integrated from various local staging areas for inclusion into central location. Inmon’s approach assumes that data found in any local site are not stored in the global site and vice versa [11]. White’s approach, also known as “Two-Tier Data Warehouse” is combination of a centralized data warehouse and decentralized data mart. The data mart or decentralized data mart contains de-normalized and summarized data that is of value to specific user or user group [2].

3. GENERAL ARCHITECTURE AND ISSUES OF CENTRALIZED DATA WAREHOUSE

Figure-1 depicts the schematic diagram of centralized warehouse based data mining. In this model, data mining works by regularly uploading data in the warehouse for subsequent centralized data mining application. This centralized approach is fundamentally inappropriate for most of distributed data mining applications. The long response time, lack of proper use of distributed resources, and the fundamental characteristics of centralized data mining algorithm do not work well in distributed environments.

![General architecture of centralized data warehouse](image)

The traditional warehouse-based architectures for data mining have centralized data repository [4]. However, this is usually either ineffective or infeasible because of following issues:

(a) **Storage limitation**: Another problem arises with need to scale up to massive data sets which are distributed over large number of sites. For example, the NASA Earth Observing System (EOS) is a data collector for satellites producing 1450 data sets of about 350GB per day and pair of satellites at a very high rate which are stored and managed by different systems geographically located all over the USA [5]. Any online mining of such huge and distributed data sets needs to transfer all data to central site, where data mining is done; the central storage of the data would require a huge data warehouse of enormous cost.

(b) **Privacy Constraints**: The privacy issue is playing an increasingly important role in the emerging data mining applications. There are many popular data mining applications that deal with sensitive data, such as people’s medical and financial records. The central collection of such data is not desirable as it puts their privacy into risk. In certain cases (e.g. banking, telecommunication) the data might belong to different, perhaps competing, organizations that want to exchange knowledge without the exchange of raw private data.

(c) **Communication cost**: In fact, various wired and wireless networks such as internet, intranets, local area networks, ad hoc wireless networks and sensor networks etc. produce many distributed resources of data. However, even if we have enough capacity to handle the data storage and data mining at a central site, the transfer of huge data volumes over network might take extremely much time and also require an unbearable financial cost. Even a small volume of data might create problems in wireless network environments with limited bandwidth. Note also that communication may be a continuous overhead, as distributed databases are not always constant and unchangeable. On the contrary, it is common to have databases that are frequently updated with new data or data streams that are constantly record information (e.g. remote sensing, sports statistics, etc.).

(d) **Unstructured data**: In enterprise applications, data is distributed in heterogeneous sources coupling in either a tight or loose manner. Distributed data sources associated with business line are often complex, for instance, some is of high frequency or density, mixing static and dynamic data, mixing multiple structure (like natural language text, images, time series, continuous data streams, multi-relational and object data types etc.) of data. For this data, data integration and data matching are difficult to conduct. It is not possible to store them in centralized storage and not feasible for processing in a centralized manner.

(e) **Computational cost**: The computational cost of mining a central data warehouse is much bigger than the sum of the cost of analyzing smaller parts of the data that could also be done in parallel.
(f) **Load balancing:** Data are distributed on cluster nodes. If the data are quickly centralized using the relatively fast network, proper balancing of computational load among a cluster of nodes is needed.

(g) **Multi-organizational applications:** Many data mining application scenarios involve multi-organizational systems. In the simplest case, the owner of the database (the data-owner) may delegate the responsibility of mining information to another party (the data-miner). Another possible scenario is cross-domain network threat management. These systems are inherently distributed, and hence require specialized techniques for efficient data mining. These requirements have led to the development of the Distributed Data Mining (DDM) approach. DDM is an effective and scalable solution for mining huge and distributed data sets in distributed computing environments.

### 4. GENERAL ARCHITECTURE AND ISSUES OF DISTRIBUTED DATA WAREHOUSE

In some distributed environments, central collection of data from every site may create heavy traffic over limited bandwidth and this may also drain a lot of power from the devices. A distributed architecture for data mining is likely to reduce the communication load and also reduce network cost across different sites. We need data mining architectures that pay careful attention to the distributed data, computing and communication in order to consume them in a near optimal fashion. A general architecture of a DDM is depicted in Figure 2. The first phase normally involves the analysis of the local database at each distributed site. Mining algorithm is applied on each local site and generates local model. Then, the discovered knowledge is usually transmitted to a merger site to generate global model, where the integration of the distributed local models is performed.

![Figure 2 General architecture of distributed data warehouse](image)

There are various issues with Distributed Data warehouse are as follows:

(a) **Homogeneous and Heterogeneous data:** One of the main issues of DDM is that data is heterogeneous, complex and noisy. In homogeneous DDM, the local databases have same attributes and in the same format, while heterogeneous DDM, the attributes at each site are different or in different format. Heterogeneous data is more complex than homogeneous data for data mining tasks.

(b) **Integration of Results:** The integration is not simple putting together results from all sites. An interesting pattern in a local database may not be interesting globally. For example, a frequent item set at a local site may be infrequent globally. Since the goal of distributed data mining is to generate globally interesting patterns, the patterns and their properties should be collected from all sites and verified globally for their interestingness.

(c) **Data skewness:** The statistical distributions of data, such as attribute values and class memberships, are usually different among local databases. A local model is obtained by mining a local database is unavoidably affected by such distribution. Such data skewness can make the local models inaccurate, sometimes even useless.

(d) **Duplicate data:** Some or all data in a local database may be duplicated at other sites. Duplication of data makes it harder to maintain data consistency.

(e) **Communication cost:** In the centralized data mining, the main concern for the efficiency of a data mining algorithm is its I/O and CPU time. In a distributed environment, one has to consider the communication cost. For a slow network, the communication cost will dominate the overall cost. The communication cost is determined by the network bandwidth and the number of messages that are sent across the network. A cost model is different from that in centralized data mining is needed for distributed data mining.

### 5. RESULTS AND FINDINGS

Table 1 has demonstrated that distributed data mining techniques are scalable. A data mining technique is scalable when its performance does not degrade much with the increase of data set size. In distributed data mining the number of sites involved in a task can grow or shrink with the data set size. The overall performance of such a system remains steady.
Table: 1 Tradeoff amongst various costs with data mining strategies

<table>
<thead>
<tr>
<th>No.</th>
<th>Data Mining Strategies/Cost</th>
<th>Centralized Data Warehouse</th>
<th>Distributed Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Communication Cost</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>2</td>
<td>Storage Cost</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>3</td>
<td>Computation Cost</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>4</td>
<td>Abstraction Level</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>5</td>
<td>Privacy</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>6</td>
<td>Accuracy</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

6. CONCLUSION

As compared to centralized data mining, distributed data mining techniques are faster, efficient and cost effective. Of course, compared with centralized data mining, the techniques in distributed data mining are more complex. Careful design is required for a distributed data mining task.

REFERENCES


AUTHOR

Preeti Baser received MCA and M.Sc in Mathematics degrees from Gujarat University in 2001 and 2004 respectively. She has over 7 years of experience in teaching. She is Ph.D research scholar in Computer Science. Her area of research is Data Mining, Web Technology and Database Management Systems.

Dr. Jatinderkumar Saini received Doctor of Philosophy in Computer Science. He is currently working as Director (I/C) & Associate Professor at Narmada College of Computer Applications, Bharuch, Gujarat. He has more than 20 research publications in national and international journals.