Data Mining with Cloud Computing: - An Overview

Miss. Rohini A. Dhote
Department of Computer Science & Information Technology
HVPM, COET Amravati, Maharashtra, India.

Dr. S. P. Deshpande
HOD at Computer Science Technology Department,
HVPM, COET Amravati, Maharashtra, India

Abstract: Data mining is a process of extracting potentially useful information from data, so as to improve the quality of information service. The integration of data mining techniques with Cloud computing allows the users to extract useful information from a data warehouse that reduces the costs of infrastructure and storage. Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent Years Wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge Market analysis, fraud detection, and customer retention, production control and science exploration. Data mining can be viewed as a result of the natural evolution of information technology. Security and privacy of user's data is a big concern when data mining is used with cloud computing. This paper introduces the basic concept of cloud computing and data mining firstly, and sketches out how data mining is used in cloud computing;

Keywords: Cloud Computing, Data mining, data mining in cloud computing

1. INTRODUCTION

We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS). Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence.

Data mining has been an effective tool to analyze data from different angles and getting useful information from data. Classification of data, categorization of data, and to find correlation of data patterns from the dataset. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. The cloud makes it possible for you to access your information from anywhere at any time. The cloud removes the need for you to be in the same physical location as the hardware that stores your data. The use of cloud computing is gaining popularity due to its mobility, huge availability and low cost.

2. DATA MINING

Data Mining involves the use of sophisticated data and tool to discover previously unknown, valid patterns and Relationship in large data sets. These tools can include statistical models, mathematical algorithms and machine learning methods (algorithms that improve their performance automatically through experience, such as neural network or decision trees). Consequently, data mining consist of more than collecting and managing data; it also includes analysis and prediction. Data mining is becoming common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly used data mining to reduce cost, enhance research, and increase sales. In public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purpose such as measuring and improving program performance. The main objective of Data Mining is to find patterns automatically with minimal user input and efforts. Data Mining is a powerful tool capable of handling decision making and for forecasting Future trends of market. Data Mining tools and techniques can be successfully applied in various fields in various forms.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure shows data mining as a step in an iterative knowledge discovery process.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw

data collections to some form of new knowledge. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, *data cleaning* and *data integration* can be performed together as a pre-processing phase to generate a data warehouse. *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called “gold mining” and not “rock mining”, thus by analogy, data mining should have been called “knowledge mining” instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

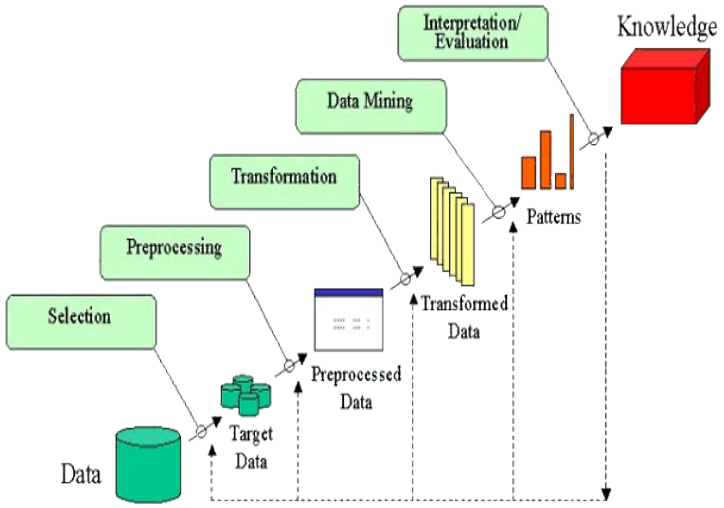

Figure 2.1 KDD Process

2.1 Data Mining Techniques

Characterization: Data characterization is a summarization of general features of objects in a target class, and produces what is called *characteristic rules*. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

Discrimination: Data discrimination produces what are called *discriminate rules* and is basically the comparison of the general features of objects between two classes referred to as the *target class* and the *contrasting class*.

Association analysis: Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*,

Identifies the frequent item sets. Another threshold, *confidence*, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules.

Classification: Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection.

Prediction: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending

trends, or predict a class label for some data. The latter is tied to classification.

Clustering: Useful for exploring data and finding natural groupings. Members of a cluster are more like Each other than they are like members of a different cluster. Common examples include finding new customer segments and life sciences discovery

3. CLOUD COMPUTING

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.

Cloud computing is receiving a great deal of attention, both in publications and among users, from individuals at home to the U.S. government. Yet it is not always clearly defined.Cloud computing is a subscription-based service where you can obtain networked storage space and computer resources. One way to think of cloud computing is to consider your experience with email. Your email client, if it is Yahoo!, Gmail, Hotmail, and so on, takes care of housing all of the hardware and software necessary to support your personal email account. When you want to access your email you open your web browser, go to the email client, and log in. The most important part of the equation is having internet access. Your email is not housed on your physical computer; you access it through an internet connection, and you can access it anywhere. If you are on a trip, at work, or down the street getting coffee, you can check your email as long as you have access to the internet. Your email is different than software installed on your computer, such as a word processing program. An email client is similar to how cloud computing works. Except instead of accessing just your email, you can choose what information you have access to within the cloud.

3.1Types of Cloud

Public cloud is offered over the Internet and are owned and operated by a cloud provider. Some examples include services aimed at the general public, such as online photo storage services, e-mail services, or social networking sites.

Private cloud, the cloud infrastructure is operated solely for a specific organization, and is managed by the organization or a third party.

Community cloud, the service is shared by several organizations and made available only to those groups. The infrastructure may be owned and operated by the organizations or by a cloud service provider.

Hybrid cloud is a combination of different methods of resource pooling (for example, combining public and community clouds).

3.2 Cloud Computing Models

1. **Software as a Service (SaaS):** In this model, a complete application is offered to the customer, as a service on demand. A single instance of the service runs on the cloud & multiple end users are serviced. On the customers‟ side, there is no need for upfront investment in servers or software licenses, while for the provider, the costs are lowered, since only a single application needs to be hosted & maintained. Today SaaS is offered by companies such as Google, Sales force, Microsoft, Zoho, etc.

2. **Platform as a Service (Paas):** Here, a layer of software, or development environment is encapsulated & offered as a service, upon which other higher levels of service can be built. The customer has the freedom to build his own applications, which run on the provider's infrastructure.

3. Infrastructure as a Service (Iaas): IaaS provides basic storage and computing capabilities as standardized service over the network. Servers, storage systems, networking equipment, data centre space etc. are pooled and made available to handle workloads.

Figure 3.2.1 Cloud Computing Models

3.3 Security

The information housed on the cloud is often seen as valuable to individuals with malicious intent. There is a lot of personal information and potentially secure data that people store on their computers, and this information is now being transferred to the cloud. This makes it critical for you to understand the security measures that your cloud provider has in place, and it is equally important to take personal precautions to secure your data.

The major issue of Cloud is represented by security. Before adopting this technology, beneficiaries should know that they will be surrendering all their company's sensitive information to a third-party Cloud service provider. This could potentially impose a great risk to the company. Hence, businesses need to make sure that they choose the most reliable service provider, who will keep their information totally secure. Switching to the cloud can actually improve security for a small business, as mentioned by Michael Redding, managing director of Accenture Technology Labs. "Because large cloud computing companies have more resources, he says, they are often able to offer levels of security an average small business may not be able to afford implementing on its own servers" (Outsource IT Headaches to the Cloud(The Globe and Mail)).

4. DATA MINING IN CLOUD COMPUTING

"Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users."

Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on your own hard drive or updating applications for your needs, you use a service over the Internet, at another location, to store your information or use its applications. Doing so may give rise to certain privacy implications.

The Microsoft suite of cloud-based administrations presents another specialized sneak peak of Data Mining in the Cloud as "DMCloud". The data mining tasks include:

- Analyze Key Influencers
- Detect Categories
- Fill From Example
- Forecast

5. CONCLUSION

This paper provides an overview of the necessity and utility of data mining in cloud computing. Cloud computing offers benefits for organizations and individuals. There are also privacy and security concerns. If you are considering a cloud service, you should think about how your personal information, and that of your customers, can best be protected. Actually we are discussing the cloud computing data mining for the advance use of security in data loss purpose. While the data we are storing in cloud is being separated in different servers for a security but the hackers using the cheap and raw cloud computing for the misuse of the software.

REFERENCES

[1] Chappell, D., A short introduction to cloud Platforms: An enterprise-oriented view, White Paper, 13 Pages, San Francisco, Chappell and Associates, 2008
[2] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
[3] G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.
[4] Fayyad, U.M. "Data Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert 11(5)*, 1996.
[5] Dunham, M.H. (2003). Data mining introductory and Advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.
[6] Information on http://en.wikipedia.org/wiki/Cloud_computing.

Authors Details:-
Miss. Rohini A. Dhote student of ME 1st Year (CSIT) at HVPM COET Amravati, Maharashtra (India).
Dr. S. P. Deshpande is Head of Department in PG Department of Computer Science & Technology, HVPM COET Amravati, Maharashtra

