CHAPTER-29

Data Mining, System Products and Research Prototypes

29. Data Mining, System Products and Research Prototypes

Although data mining is a young field with many issues that still need to be researched in depth, there are already great many off-the-shelf data mining system products and domain-specific data mining application software available. As a young discipline, data mining has a relatively short history and are constantly evolving-new data mining systems appear on the market every year; new functions, features, and visualization tools are added to existing systems on a constant basis; and efforts toward the standardization of data mining language have only just begun. Therefore, it is not our intention in this book to provide a detailed description of commercial data mining systems. Instead, we describe the features to consider when selecting a data mining product and offer a quick introduction to a few typical data mining systems.

29.1 How to Choose a Data Mining System

With many data mining system products available on the market, you may ask, "What kind of system should J choose?" Some people may be under the impression that data mining systems, like many commercial relational database systems, share the same well-defined operations and a standard query language, and behave similarly on common functionalities. If such were the case, the choice would depend more on the systems' hardware platform, compatibility, robustness, scalability, price, and service. Unfortunately, this is far from reality. Many commercial data mining systems have little in common with respect to data mining functionality or methodology and may even work with completely different kinds of data sets.

To choose a data mining system that is appropriate for your task, it is important to have a multiple dimensional view of data mining systems. In general, data mining systems should be assessed based on the following multiple dimensional features.

Data types: - -

Most data mining systems that are available on the market handle formatted, record-based, relational-like data with numerical, categorical, and symbolic attributes. The data could be in the form of ASCII text, relational database data, or data warehouse data. It is important to check what exact format(s) each system you are considering can handle. Some kinds of data or applications may require specialized algorithms to search for patterns, and so their requirements may not be handled by off-the-shelf, generic data mining systems. Instead, specialized data mining systems may be used, which mine either text documents, geo-spatial data, multimedia data, time-series data, DNA sequences, Weblog records or other Web data, or are dedicated to specific applications (such as finance, the retail industry, or telecommunications). Moreover, many data mining companies offer customized data mining solutions that in corporate essential data mining functions or methodologies. _ .

System issues:

A given data mining system may run on only one operating system, or on several. The most popular operating systems that host data mining software are UNIX and Microsoft Windows (including 95, 98, 2000, and NT). There are also data mining systems that run on OS/2, Macintosh, and Linux. Large

industry-oriented data mining systems should ideally adopt a client/server architecture, where the client could be a personal computer running on Microsoft Windows, and the server could be a set of powerful parallel computers running on UNIX. A recent trend has data mining systems providing Web-based interfaces and allowing XML data as input and/or output.

Data sources:

This refers to the specific data formats on which the data mining system will operate. Some systems work only on ASCII text files, whereas many others work on relational data, accessing multiple relational data source. It is important that a data mining system supports ODBC connections or OLE DB for ODBC connections. These ensure open database connections, that is, the ability to access any relational data (including those in DB2, Informix, Microsoft SQL Server, Microsoft Access, Microsoft Excel, Oracle, Sybase, etc.), as well as formatted ASCII text data. A data mining system that operates with a data warehouse should follow the OLE DB for OLAP standard, since this helps ensure that the system is able to access the warehouse data provided not only by Microsoft SQL Server 7.0 but also by other data warehouse products supporting the standard.

29.2 Data, mining functions and methodologies:

Data mining functions form the core of a data mining system. Some data mining systems provide only one data mining function, such as classification. Others may support multiple data mining functions, such as description, discovery-driven OLAP analysis, association, classification, prediction, clustering, outlier analysis, similarity search, sequential pattern analysis, and visual data mining. For a given data mining function (such as classification), some systems may support only one method, while others may support a wide variety of methods (such as decision tree analysis, Bayesian networks, neural networks, genetic algorithms, case-based reasoning etc.). Data mining systems that support multiple data mining functions and multiple methods per function provide the user with greater flexibility and analysis power. Many problems may require users to try a few different mining functions or incorporate several together and different methods can be shown to be more effective than others for different kinds of data. In order to take advantage of the added flexibility, however, users may require further training and experience. Thus such systems should also provide novice users with convenient access to the most popular function and method, or to default settings.

29.3 Coupling data mining with database anti/or data warehouse systems:

A data mining system should be coupled with a database and/or data warehouse system, where the coupled components are seamlessly integrated into a uniform information processing environment. In general, there are four forms of such coupling: no coupling, loose coupling, semi tight coupling, and tight coupling. Some data mining systems work only with ASCII data files and are not coupled with database or data warehouse systems at all. Such systems have difficulties handling large data sets and using the data stored in database systems. In data mining systems that are loosely coupled with database and data warehouse systems, the data are retrieved into a buffer or, main memory by database or warehouse operations, and then mining functions are applied to analyze the retrieved data. These systems tend to vive poor scalability and may be inefficient when executing some data mining queries. The coupling of a data mining system with a database or data warehouse system may be semi-tight, providing the efficient

implementation of only a few essential data mining primitives (such as sorting, indexing, aggregation, histogram analysis, multiway join, and! the pre-computation of some statistical measures). Ideally, a data mining system should be tightly coupled with a database system in the sense that the data mining and data retrieval processes are integrated by optimizing data mining queries deep into the iterative mining and retrieval process. Tight coupling of data mining with OLAP-based data warehouse systems is also desirable so that data mining and OLAP operations can be integrated to provide OLAP-mining features.

Scalability:

Data mining has two kinds of salability issues: row (or database size) scalability and column (or dimension) scalability. A data mining system is considered row scalable if, when the number of rows is enlarged 10 times, it takes no more than 10 times to execute the same data mining queries. A data mining system is considered column scalable if the mining query execution time increases linearly with the number of columns (or attributes or dimensions). Due to the curse of dimensionality, it is much more challenging to make a system column scalable than row scalable.

Visualization tools:

"A picture is worth a thousand words"-this is very true in data mining. Visualization in data mining can be categorized into data visualization, mining result visualization, mining process visualization, and visual data mining. The variety, quality, and flexibility of visualization tools may strongly influence the usability, interpretability, and attractiveness of a data mining system.

29.4 Data ruining query language and graphical user interface:

Data mining is an exploratory process. An easy-to-use and high-quality graphical user interface is essential in order to promote user-guided, highly interactive data mining. Most data mining systems provide user-friendly interfaces for mining. However, unlike relational database systems, where most graphical user interfaces are constructed on top of SQL, most data mining systems do not share any underlying data mining query language. Lack of a standard data mining language makes it difficult to standardize data mining products and to ensure the interoperability of data mining systems. Recent efforts at defining and standardizing daft mining query languages, one such language are Microsoft's OLE DB for DM.

Examples of Commercial Data Mining Systems

As mentioned earlier, due to the infancy and rapid evolution of the data mining market, it is not our intention in this book to describe any particular commercial data mining system in detail. Instead, we briefly outline a few typical data mining systems in order to help the reader get an idea of what can be done with current data mining products.

Many data mining systems specialize in one data mining function, such as classification, or just one approach of a data mining function, such as decision tree classification. Other systems provide a broad spectrum of data mining functions. Here we introduce a few systems that provide multiple data mining functions and explore multiple knowledge discovery techniques.

Intelligent Miner is an IBM data-mining product that provides a wide range of data mining algorithms including association, classification, regression, predictive modeling, deviation detection, sequential pattern analysis, and clustering. It also provides an application toolkit containing neural

network algorithms, statistical methods, data preparation tools, and data visualization tools. Distinctive features of Intelligent Miner include the scalability of its mining algorithms and its tight integration with IBM's DB2 relational database system.

SAS Institute Inc. developed Enterprise Miner who provides multiple data mining algorithms including regression, classification, and statistical analysis packages. A distinctive feature of Enterprise Miner is its variety of statistical analysis tools, which are built based on the long history of SAS in the market of statistical analysis.

Silicon Graphics Inc. (SGI) developed MineSet. It also provides multiple data mining algorithms including association and classification, as well as advanced statistics and advanced visualization tools. A distinguishing feature of MineSet is its set of robust graphics tools (using powerful graphics features of SGI computers), including rule visualize, tree visualizer, map visualizer, and (multidimensional data) scatter visualize, for the visualization of data and data mining results.

Integral Solutions Ltd. (ISL) developed Clementine. It provides an integrated data mining development environment for end users and developers. Multiple data mining algorithms, including rule induction, neural nets, classification, and visualization tools, are incorporated in the system. A distinguishing feature of Clementine is its object-oriented, extended module interface, which allows users' algorithms and utilities to be added to Clementine's visual programming environment. Clementine has been acquired by SPSS Inc.

DBMiner Technology Inc developed DBMiner. It provides multiple data mining algorithms including discovery-driven OLAP analysis, association, classification, and clustering. A distinct feature of DBMiner is its data-cube based on-line analytical mining, which includes efficient frequent-pattern mining functions, and integrated visual classification methods

There are many other commercial data mining products, systems, and research prototypes that are also fast evolving. Interested readers may wish to consult tin surveys on data warehousing and data mining products.

29.6 Review Questions

1 How to Choose a Data Mining System

2Explain about Data, mining functions and methodologies:

3Explain about Coupling data mining with database anti/or data warehouse systems:

4 Explain about Data ruining query language and graphical user interface

29.7 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition