Advances in Engineering Software 47 (2012) 164–169

Contents lists available at SciVerse ScienceDirect

Advances in Engineering Software

journal homepage: www.elsevier.com/locate/advengsoft

Analysis the effect of data mining techniques on database

Niyati Aggarwal [a], Amit Kumar [a,*], Harsh Khatter [b], Vaishali Aggarwal [b]

[a] *Department of Computer Science & Engineering, Jaypee University of Engineering & Technology University, Guna, M.P., India*
[b] *Ajay Kumar Garg Engineering College, Ghaziabad, India*

ARTICLE INFO

Article history:
Received 13 September 2011
Received in revised form 19 December 2011
Accepted 31 December 2011
Available online 2 February 2012

Keywords:
Data acquisition
Databases
Data warehouse
Data mining
Data miner
Intelligent miner

ABSTRACT

In today's information society, we witness an explosive growth of the amount of information becoming available in electronic form and stored in large databases. Data mining can help in discovering knowledge. Data mining can dig out valuable information from databases in approaching knowledge discovery and improving business intelligence. In this paper, we have discussed the involvement and effect of data mining techniques on relational database systems, and how its services are accessible in databases, which tool we require to use it, with its major pros and cons in various databases. Through all this discussion we have presented how database technology can be integrated to data mining techniques.

1. Introduction

In recent years data mining has become a very popular technique for extracting information from the database in different areas due to its flexibility of working on any kind of databases and also due to the surprising results [1].

Data mining is the search for valuable information in large volumes of data [1]. With the increase of availability of databases containing structures, mining techniques especially designed for this type of data are becoming more and more important [2]. To improve both software productivity and quality, software engineers are increasingly applying data mining algorithms to various Software Engineering tasks [3].

The progress in data acquisition and successful development of storage technology at cheaper rates, along with limited human capabilities in analyzing and understanding big databases have tempted scientists and researchers to move forward towards the specific field of knowledge discovery in databases (KDD). The huge amount of available data, jointly with a poor understanding of the processes that have generated them, enforces the use of data mining techniques to extract frequent structural patterns that may convey important information [2].

Databases are too big, and data mining can help to extract interesting knowledge from data in large collections. But still we are not completely aware about how to use data mining, and with which database it works well. This paper discuss about data mining technique with respect to different databases. It is just like a comparative study of various databases regarding how we can use data mining techniques on database technology.

Despite the potential effectiveness of data mining to significantly enhance data analysis, this technology is destined be a niche technology unless an effort is made to integrate this technology with traditional database systems. This is because data analysis needs to be consolidated at the warehouse for data integrity and management concerns. Therefore, one of the key challenges is to enable integration of data mining technology seamlessly within the framework of traditional database systems.

Till now, researches are going on in this technique to use it in an efficient manner to get desirable results in database technology. In this paper, we introduce IBM DB2, Microsoft SQL Server, MYSQL, and ORACLE for Data mining. Data mining techniques, based on statistics and machine learning can significantly boost the ability to analyze data.

2. Representing mining models in databases

Data mining is defined as the process of discovering hidden and potentially useful information from very large databases [4]. The progress in data mining research has made it possible to implement several data mining operations efficiently on large databases. While this is surely an important contribution, we should not lose sight of the final goal of data mining – it is to enable database application writers to construct *data mining models* (e.g., a decision tree

* Corresponding author.
E-mail addresses: niyati.aggrawal@gmail.com (N. Aggarwal), amitrathi10@yahoo.co.in (A. Kumar), harsh_khatter@yahoo.com (H. Khatter), sweetvish08@gmail.com (V. Aggarwal).

0965-9978/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advengsoft.2011.12.013

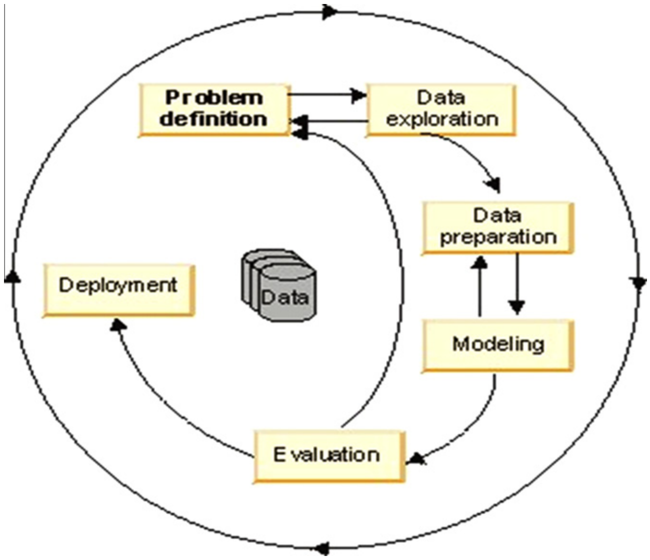


Fig. 1. Representing data mining processes in databases.

classifier, regression model, segmentation) from their databases, to use these models for a variety of predictive and analytic tasks, and to share these models with other applications. Recognizing the above fact, it is obvious that a key aspect of integration with database systems that needs to be looked into is how to treat data mining models as first class objects in databases. Unfortunately, in that respect, data mining still remains an island of analysis that is poorly integrated with database systems. Recall that a data mining model (e.g., classifier) is obtained via applying a data mining algorithm on a training data set [5]. How the data mining model works, this is represented in Fig. 1, this figure shows the processes of data mining in a systematic way.

3. IBM DB2

IBM gives data mining functionality in higher versions of DB2 database. The data mining technologies from IBM help you to detect fraud, prevent customer churn, segment your customers, and simplify market basket analysis. Intelligent Data Miner is present here to provide the data mining functionality. A Fig. 2 shows the simple interaction model of IBM DB2 Intelligence Miner.

IBM DB2, Data Warehouse Edition [6] includes the following data mining features:

- Data mining functions in DB2 Data Warehouse Edition Design Studio.
- Data mining functions in DB2 Data Warehouse Edition Administration console.
- Intelligent Miner Easy Mining.
- Intelligent Miner Modeling.
- Intelligent Miner Scoring.
- Intelligent Miner Visualization.

In DB2 for data mining, a special miner is present, named, Intelligent miner. This miner performs:

3.1. DB2 Intelligent Miner Modeling

Intelligent Miner Modeling provides Intelligent Miner Modeling technology as DB2 extenders. It enables SQL application programs (SQL API) to call Associations discovery, Clustering, Classification, and Transform Regression operations to develop analytic models based on data accessed by DB2 Universal Database Version 8.2 or Version 8 SQL. Using the SQL API, you can build Associations, Distribution-based Clustering, Tree Classification, and Transform Regression PMML models that are stored in DB2 tables [6].

3.2. DB2 Intelligent Miner Scoring

Intelligent Miner Scoring provides scoring technology as DB2 extenders. It enables application programs to apply Predictive Model Markup Language (PMML) models to large databases, subsets of databases, or single rows or cases. Application programs use the SQL API, which consists of user-defined functions (UDFs) and user-defined methods (UDMs), to perform the scoring operation. Intelligent Miner Scoring includes Intelligent Miner Scoring Java Beans, which enables you to score a single data record in a Java (TM) application given a PMML model. This can be used to integrate scoring in e-business applications, for example, for real-time scoring in customer relationship management (CRM) systems. Fig. 3 shows the Intelligent Miner features in business environment.

3.3. DB2 Intelligent Miner Visualization (IMV)

IMV provides association visualize, classification visualize, regression visualize, and clustering visualize presenting and analyzing data modeling results.

3.3.1. IM for data models

Intelligent Miner (IM) for data provides mining functions and algorithms to create and store mining models in DB2 databases. We can migrate the mining models that are created with IM for Data by exporting them to PMML and importing the PMML file into the database. Using IM Scoring includes the following steps:

- Importing the mining model into a DB2 table, where it is stored as a large object.
- Applying the model to data that is stored in DB2 tables.
- Storing the scoring results in DB2 tables.
- Extracting information about the results.

4. Microsoft SQL server

Microsoft provides a variety of services in its SQL server for data mining support. SQL Server Data Mining is a collection of machine learning algorithms that explore your data for patterns. Once discovered, these patterns can be browsed for greater insight into your data, or they can be applied to new data to create "predictions" – which allow you to determine unknown facts about data based on data the algorithms' have seen before. SQL Server 2005 comes with many services. This describes what components are necessary to perform data mining once you have the SQL Server 2005 [7]. These services are:

4.1. Analysis Services

Analysis Services is the only required component to install on the server. If you want to do data mining against existing SQL Server 2000 databases or other data sources (DB2, Oracle, Access, etc.), this is the component you need to install.

4.2. Reporting Services

Install Reporting Services if you want to be able to create reports that work against your data mining models.

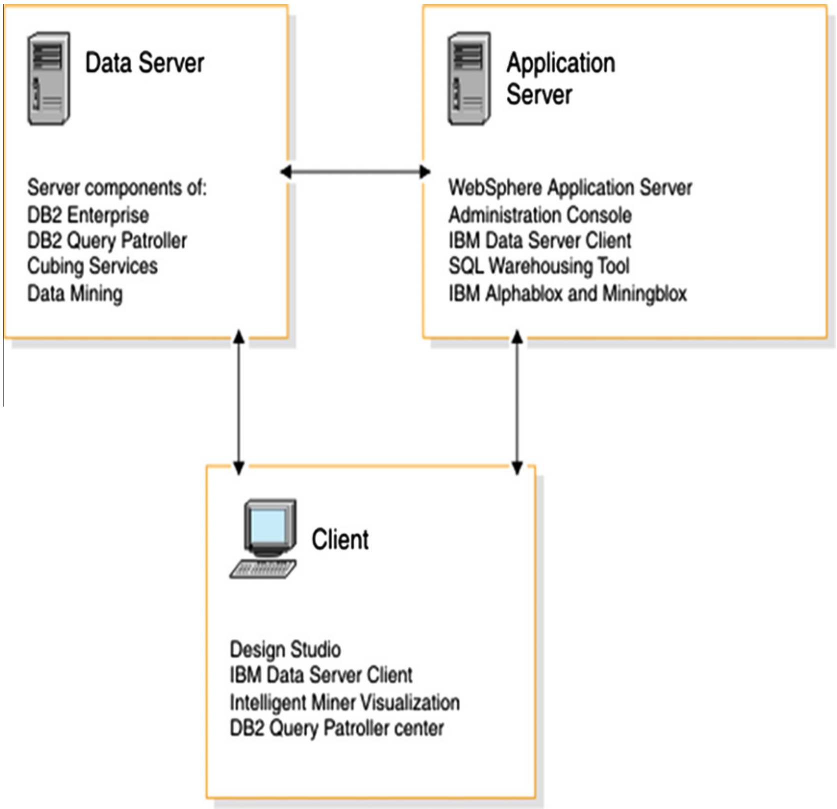


Fig. 2. IBM DB2 Intelligence Miner.

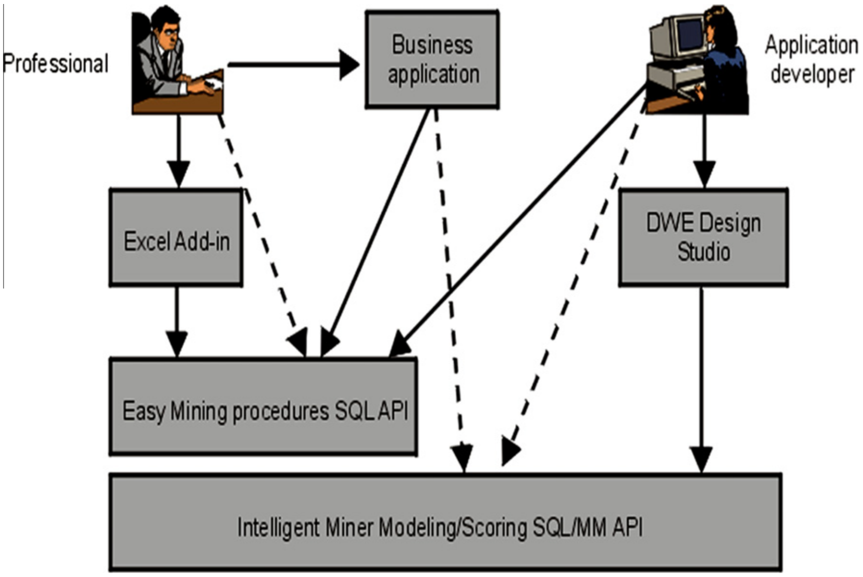


Fig. 3. Intelligent Miner features in business environment.

4.3. SQL Server Database Services

You only need to install the SQL Server relational engine if you want to use it as a data source, or if you want to use the data mining samples and tutorials (to know about how to use data mining models and services in SQL server database).

4.4. SQL Server Integration Services

Installing SSIS causes the SSIS service to be installed on your server, allowing the running of scheduled packages. Install this if you want to use the integrated data mining tasks and transforms on your server.

4.5. Workstation Components

Install the Workstation Components on any client machine that will be creating mining models, authoring reports and SSIS packages, or managing Analysis Services. The Workstation Components work equally well when installed on the same machine as the server.

4.6. Advanced

You just need to click the Advanced button to install the samples and sample databases. All the samples are located under Client Components/Documentation and Samples.

SQL Server 2005 Analysis Services includes a rich set of nine algorithms developed by the SQL Server Data Mining product team in collaboration with Microsoft Research: Microsoft Decision Trees, Microsoft Clustering, Microsoft Time Series, Microsoft Association Rules, Microsoft Sequence Clustering, Microsoft Naive Bayes, Microsoft Neural Network, Microsoft Linear Regression, Microsoft Logistic Regression.

5. MYSQL

MySQL is developed by a subsidiary of Oracle. MySQL being open source allows customization which allows users to add their own requirements. MySQL also supports large number of embedded applications. These features add to the scalability characteristic.

MySQL does not have data mining services in-built [8]. We need an external tool for using data mining services. Any tool that can tap an ODBC–JDBC data source would be able to use a MySQL database.

Here we are giving names of some data mining tools in MySQL [9]:

5.1. DBMyne 2008

DBMyne is a data mining software. In the wide area of data mining DBMyne addresses the field of decision cube analysis. DBMyne is an easy to use software that helps you to analyze and present data stored in computer databases.

5.2. The Query Tool 2005 6.0.1

The Query Tool is a powerful data mining application. It allows you to perform data analysis on any SQL database. It has been developed predominately for the non technical user. No knowledge of SQL is required, most actions are data driven.

5.3. MyLobEditor 1.5

MyLobEditor is a database tool that helps DBA and Database Programmer edit MySQL LOB (text, blob) data directly and import/export LOB data. Main features:

- Edit LOB data directly.
- Batch and automatic import/export LOB data.

5.4. DBF-to-MySQL 1.1

DBF-to-MySQL is a program to move DBF databases to MySQL server. Key features:

- all DBF data types and attributes are supported,
- works with all versions of Unix and Windows MySQL servers,
- merges DBF data into an existing MySQL database.

5.5. MySQLToAccess 1.2

MySQLToAccess is a data conversion tool that helps DBA and database programmer convert MySQL data to Access.

The MySQL database has become the world's most popular open source database because of its high performance, high reliability and ease of use. It is also the database of choice for a new generation of applications built on the LAMP stack (Linux, Apache, MySQL, PHP/Perl/Python.) Many of the world's largest and fastest-growing organizations including Facebook, Google, Adobe, Alcatel Lucent and Zappos rely on MySQL to save time and money powering their high-volume Web sites, business-critical systems and packaged software.

6. Oracle database

Oracle provides the most complete, open, and unified enterprise content management (ECM) platform that enables you to build in a single repository both high volume imaging applications such as accounts payable and claims processes as well as high performance delivery applications, such as consumer-based Web sites. In Oracle, there is a Oracle Exadata Database Machine is the only database machine that provides extreme performance for both data warehousing and online transaction processing (OLTP) applications, making it the ideal platform for consolidating onto grids or private clouds. It is a complete package of servers, storage, networking, and software that is massively scalable, secure, and redundant.

Oracle Data Mining (ODM) embeds data mining within the Oracle database. There is no need to move data out of the database into files for analysis and then back from files into the database for storing. The data never leaves the database – the data, data preparation, model building, and model scoring results all remain in the database. This enables Oracle to provide an infrastructure for application developers to integrate data mining seamlessly with database applications.

ODM is designed to support production data mining in the Oracle database. Production data mining is most appropriate for creating applications to solve problems such as customer relationship management, churn, etc., that is, any data mining problem for which you want to develop an application.

ODM provides single-user milt-session access to models. Model building is either synchronous in the PL/SQL interface or asynchronous in the Java interface.

6.1. Oracle Data Mining programming interfaces

ODM integrates data mining with the Oracle data base and exposes data mining through the following interfaces [10]:

(1) Java interface: Allows users to embed data mining in Java applications.
(2) dbms_data_mining & dbms_data_mining_transform: Allow users to embed data mining in PL/SQL applications.

6.2. ODM data mining functions

Data mining functions are based on two kinds of learning: *supervised* (directed) and *unsupervised* (undirected). Supervised learning functions are typically used to predict a value, and are sometimes referred to as *predictive models* which includes classification, regression, attribute importance. Unsupervised learning functions are typically used to find the intrinsic structure, relations, or affinities in data but no classes or labels are assigned aprioi. These are sometimes referred to as *descriptive models* which includes clustering, association models and, feature extraction. Table 1 shows some of these functions and algorithms.

Table 1
DBMS_DM summary of functions and algorithms.

Mining function	Mining algorithm
Classification	Naive Bayes (NB) – default algorithm Adaptive Bayes Network (ABN) Support vector machine (SVM)
Regression	Support vector machine (SVM)
Association	Association rules (AR)
Clustering	k-Means (KM)
Feature extraction	Non-negative matrix factorization (NMF)

6.3. Oracle Data Miner

Oracle Data Miner is a powerful tool bundled with the Oracle RDBMS. It provides fast and easy way to develop BI applications based on the data stored in the Oracle RDBMS. ODM SQL and Java APIs can be used to develop and deploy applications. It allows data stored in the Oracle RDBMS to be mined automatically. Real time results can be generated as the data, models and results all resided in the database and no additional data movement is required. Fig. 4 shows the server window of Oracle Data Mining.

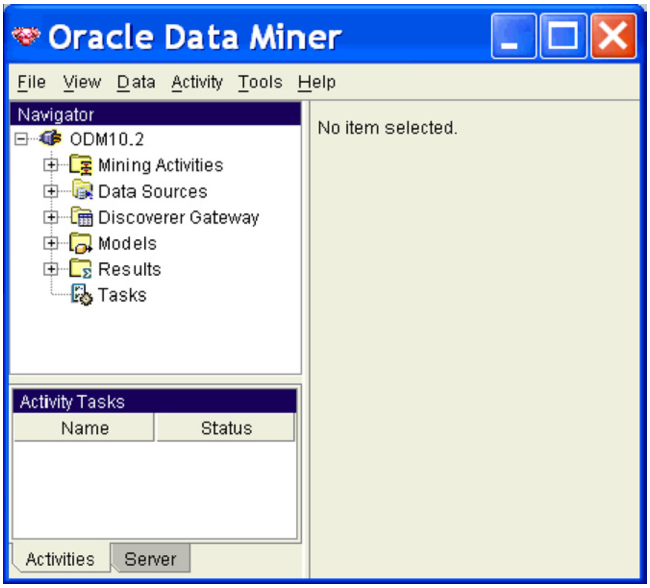
Oracle provides two tools to assist analysts in data mining activities: Oracle Data Miner and Oracle Spreadsheet Add-In for Predictive Analytics.

Oracle Data Miner is a graphical user interface for Oracle Data Mining. Oracle Data Miner's easy-to-use wizards guide you through the data preparation, data mining, model evaluation, and model scoring process. Oracle Data Miner is supported on Windows 2000, Windows XP Professional Edition, and Linux.

Oracle Data Miner, shown in Fig. 5, is designed to support data mining in the Oracle database. It provides single-user multi-session access to models. Model building is either synchronous in the PL/SQL interface or asynchronous in the Java interface.

7. Comparative study

On the basis of major parameters, the effect of data mining techniques on different databases is discussed here.



No item selected.

Fig. 4. Oracle Data Miner.

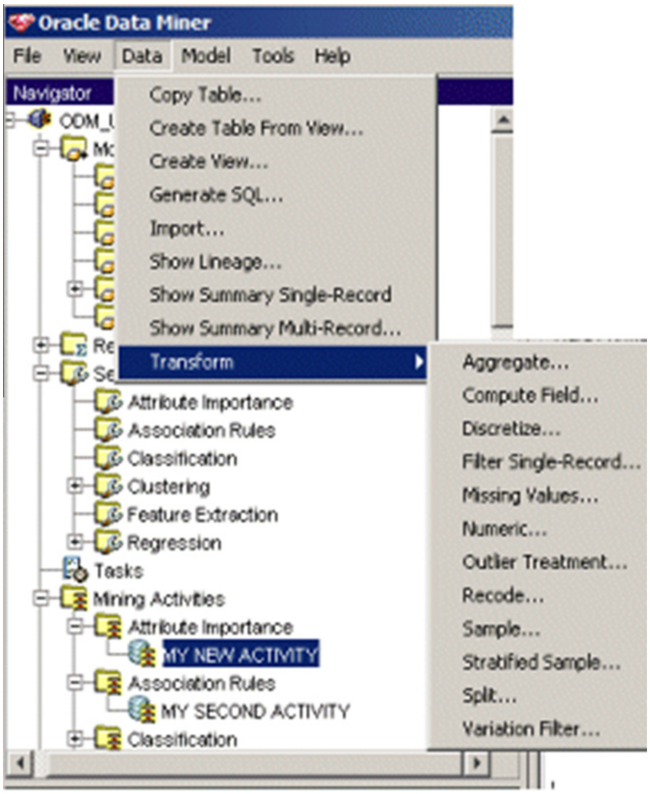


Fig. 5. Oracle Data Miner.

- The very first point which comes in mind of user before using any database is how much space it will take. Disk space (for windows) required for MS SQL Server 2008 is around 2.048 GB, for IBM DB2 is 1.5 GB, for Oracle 11g is 3.47 GB, and for MySQL is 260 MB excluding its tool and other services.
- Second parameter is tools and services. In MS SQL server, there are two important services, i.e. SSIS and SSAS. In DB2, IBM main focused on its special tool, An Intelligent Miner. In ORACLE, it differs from version to version but presently its ODM, Oracle Data Miner (or, Oracle Data Mining). In MYSQL, we have seen that some add-ons tools are used to give support of data mining.
- The third most important parameter is *Security*. In MS SQL Server, the security manager is present. Although, Microsoft SQL Server 2008 offers security feature enhancements that help provide effective management of security feature configuration, strong authentication and access control, powerful encryption and key management capabilities, and enhanced auditing. Whereas IBM DB2 is a less secure database, more vulnerable to users or hackers subverting the security due to the security model that adds security after the fact. Oracle has an excellent, long-standing reputation in security, as witnessed by Oracle's dominant market share among the most security-conscious customers in the world. MySQL uses security based on Access Control Lists (ACLs) for all connections, queries, and other operations that users can attempt to perform. There is also support for SSL-encrypted connections between MySQL clients and servers.

8. Conclusion

In this paper, we have discussed the data mining in various databases. But after seeing the various parameters as discussed in above section we conclude that no database is individually well suited for data mining in all aspects. Selection of particular database depends on user's requirement. Say if he has less disk space available in hand then it is better to choose MySQL, but as there is no inbuilt tool available for data mining so Add-ons are needed.

But as per the services, MS SQL server gives better results and users can easily access the features. And in DB2 and in ORACLE, a user has to use the ''Intelligent Miner'' and ''Oracle Data Miner''. Similarly user can consider other parameters which are already discussed in comparative study section.

References

[1] Muhammad Shahbaz, Syed Athar Masood, Muhammad Shaheen, Ayaz Khan. Data mining methodology in perspective of manufacturing databases. J Am Sci; 2010.

[2] Giovanni Da San Martino. Mining structured data. IEEE computational intelligence magazine, Alessandro Sperduti Università di Padova, Italy; February 2010.

[3] Tao Xie, Suresh Thummalapenta, David Lo, Chao Liu. Data mining for software engineering. IEEE Computer Society in August 2009.

[4] Sung SY, Wang K, Chua BL. Data mining in a large database environment. National University of Singapore (1996 IEEE).

[5] Amir Netz, Surajit Chaudhuri, Jeff Bernhardt, Usama Fayyad. Integration of data mining and relational databases. USA: Microsoft.

[6] Overview of DB2 Data Warehouse Edition. <http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.dwe.tutorial.doc/tutgenabswelcome.htm>.

[7] Services of MS SQL Server. <http://www.sqlserverdatamining.com/ssdm/Home/FAQ/tabid/55/Default.aspx>.

[8] Data Mining and analysis tool for MySQL. <http://forums.mysql.com/read.php?32,52672,64786>.

[9] Introduction to Data Mining Tool MySQL. <http://www.smartcode.com/downloads/data-mining-tool-MySQL.html>.

[10] Introduction to Oracle Data Mining. <http://download.oracle.com/docs/html/B10698_01/1intro.htm>.