

1.1 D2.1.1 — DBMiner

Abstract. DBMiner is an on-line analytical mining system, developed for interactive mining of multiple-level knowledge in large relational databases and data warehouses. The distinct feature of the system is its tight integration of on-line analytical processing (OLAP) with a wide spectrum of data mining functions, including characterization, comparison, association, classification, prediction, and clustering. The system facilitates query-based, interactive mining of multi-dimensional databases by implementing a set of advanced data mining techniques, including OLAP-based induction, multi-dimensional statistical analysis, progressive deepening for mining refined knowledge, meta-rule guided mining, and data and knowledge visualization. DBMiner integrates smoothly with commercial relational database and data warehouse system, and provides a user-friendly, interactive data mining environment with high performance. With extensions to the DBMiner system, several specialized data mining system prototypes, including *GeoMiner*, *MultiMediaMiner*, and *WeblogMiner*, have been designed and developed for mining complex types of data with interesting applications.

1.1.1 Introduction

DBMiner (<http://db.cs.sfu.ca/DBMiner>) is a data mining system, originated from the Intelligent Database Systems Research Laboratory, Simon Fraser University, Canada, with the address: *Intelligent Database Systems Research Laboratory, School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6*, and the Web address: <http://db.cs.sfu.ca>.

With years of efforts on research into data mining and knowledge discovery in databases, Dr. Jiawei Han and his graduate students, predominantly Jenny Y. Chiang, Sonny H.S. Chee, Yongjian Fu, Hua Zhu, Eddie Kim, and several others, have designed and developed the DBMiner system which integrates on-line analytical processing (OLAP) with a wide spectrum of data mining functions and performs on-line analytical mining in relational databases and data warehouses.

The first research paper on data mining from the research laboratory was published in 1989 [1], and the first comprehensive description of the DBMiner system was published in 1996 [6]. A good set of published research papers related to the system can be fetched from the Web: <http://db.cs.sfu.ca>. The system has been demonstrated in major conferences on database systems, artificial intelligence, and data mining and knowledge discovery. The educational version of the system, *DBMiner E1.0*, has been developed based on the research system prototype and was released in 1998 by *DBMiner Technology Inc.* (<http://www.dbminer.com>).

The major distinct feature of the DBMiner system is its tight integration of on-line analytical processing (OLAP) with data mining [2, 3, 5]. This integration leads to a promising data mining methodology, called *on-line analytical mining* (OLAM), where the system provides a multidimensional view of its data and creates an interactive data mining environment: users can dynamically select data mining and OLAP functions, perform OLAP operations, such as drilling, dicing/slicing, and pivoting, on data mining results and/or perform mining operation on OLAP results, i.e., mining different portions of data at multiple levels of abstraction.

1.1.2 Software architecture

The software architecture of the DBMiner system is shown in Figure 1, which takes data from a relational database and/or a data warehouse, integrates and transforms them into a multidimensional database (portions or all of which could be consolidated into data cube(s)), and then performs multi-dimensional on-line analytical processing and on-line analytical mining, based on user's data mining or on-line analytical processing requests.

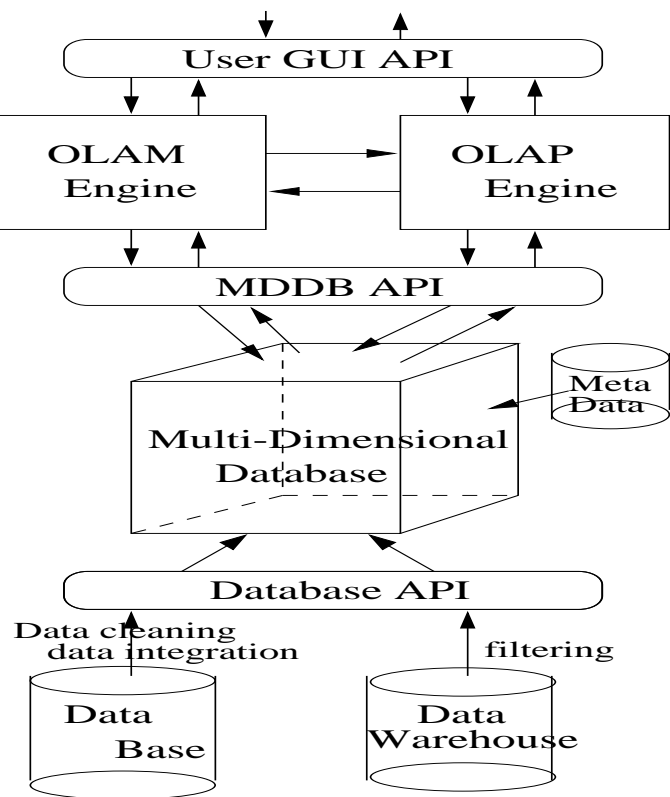


Figure 1: Software architecture of DBMiner: An integrated OLAM and OLAP architecture

The core module of the architecture is an OLAM engine which performs on-line analytical mining in multi-dimensional databases in a similar manner as an OLAP engine performs on-line analytical processing. The OLAM engine in the DBMiner system performs multiple data mining tasks, including concept description, association, classification, prediction, clustering, and time-series analysis.

More importantly, the system integrates OLAM and OLAP engines, both of which accept users' on-line queries (or commands) via a User_GUI_API and work on the multidimensional database via an MDDDB_API. Notice in the architecture, OLAM and OLAP engines interact each other because the former may take the output of the latter and perform mining on the OLAP results; whereas the latter may take the output of the former and perform OLAP on the mining results. A meta data directory, which stores (1) database schema, (2) data warehouse schema, and (3) concept hierarchy information, is used to guide the access of the multidimensional database and the execution of dimension-related OLAP operations, such as drilling and slicing. The multi-dimensional database can be constructed by accessing database(s), filtering data warehouse(s), and/or integrating multiple such sources, via a Database_API which may support OLEDB or ODBC connections. To facilitate OLAP and OLAM operations, it is often useful to consolidate the multi-dimensional database by materializing portions or all of it into data cube(s).

1.1.3 Input data and knowledge

The DBMiner system takes data from relational database system, which may contain single or multiple relational tables, and/or data warehouse system, which may contain multidimensional database(s) or materialized data cube(s) and represent cleansed, integrated, and consolidated data. It also takes other forms of data, such as spreadsheets, via ODBC connections.

Another kind of input is meta data which consists of relational database and data warehouse schema definitions and background knowledge in the form of concept hierarchies. A concept hierarchy can be specified based on the relationships among database attributes (called *schema hierarchy*) or by set groupings (called *set-grouping hierarchy*) and be stored in the form of relations in the database. Moreover, they can be generated automatically for both numerical and categorical attributes by data distribution analysis.

1.1.4 Output knowledge: manipulation and visual presentation

There are many forms of visual presentation of knowledge generated by the DBMiner system, depending on the data mining tasks and user preference: Data summarization, characterization, and comparison generate cross-tabulation tables, generalized rules, bar charts, pie charts, curves, or other forms of graphical outputs. Classification generates visual display of decision trees or decision tables. Association generates association rule tables, association planes, and association rule graphs. Prediction generates prediction curves for numerical data or predicted value distributions in the form of pie charts for categorical data. Clustering generates maps (for two-dimensional analysis) with different clusters contoured by its silhouette and painted in different colors.

Moreover, the system provides facilities to view concept hierarchies and data cube contents. Concept hierarchies are presented in a tree form similar to the directory/subdirectory structures. Data cube contents are presented in two forms: (1) a 3-dimensional cube form, where the size and color of each cuboid in the 3-D cube represent the summarization of corresponding selected measures within a set of 3-dimensional intervals; and (2) a 3-dimensional boxplot form, where each boxplot represents the data dispersion view (including median, first quarter, third quarter, whiskers, and outliers) of the corresponding intervals.

An important feature of the system is its flexible manipulation, such as performing drilling, dicing, and/or other transformations, on the output knowledge. For example, after classification on a combination of dimensions and levels, drilling can be performed on both classified and classifying dimensions to derive classification tree over the new set of data. Furthermore, one can drill through a node of the decision tree to view the set of detailed data forming the current class.

1.1.5 Data mining tasks supported by the system

The DBMiner system supports the following data mining tasks [4]:

- **Characterization.** This function generalizes a set of task-relevant data into a *generalized data cube* which can then be viewed at multiple levels of abstraction from different angles, be presented in various visual/graphical forms, or be used for extraction of characteristic rules. OLAP operations can be performed on generalized data to drill or dice on the regions of interest for further analysis.
- **Comparison.** This function extracts the generalized features that distinguish the class being examined (the *target class*) from other classes (called *contrasting classes*) and presents them in visual/graphical forms for comparison or in the form of **discriminant rules**. When there are many attributes involved in the analysis, it is often necessary to perform attribute relevance analysis and rank the attributes based on their importance to the comparison task.
- **Classification.** This function analyzes a set of training data (i.e., a set of objects whose class label is known), constructs a model for each class based on the features in the data, and adjust the model based on the test data. The model so constructed is presented in the form of decision trees or classification rules, and is used to classify future data and develop a better understanding of data in the database.

- **Association.** This function mines a set of association rules in a multiple dimensional database. The rules so mined can be used for cross-market analysis, correlation analysis, etc. A user may specify the rules to be mined in the form of *meta-patterns*, such as “*major(s:student, x) ∧ P(s, y) → grade(s, y, z)*” to confine the search for desired rules. One can also drill along any dimension to mine rules at multiple levels of abstraction.
- **Prediction.** This function predicts the values or value distributions for certain missing or unknown data in a selected set of objects. This involves finding the set of attributes relevant to the attribute of interest (by some statistical analysis) and predicting the value distribution based on the set of data similar to the selected object(s). For example, an employee’s potential salary can be predicted based on the salary distribution of similar employees in the company.
- **Clustering.** This function groups a selected set of data objects into a set of clusters to ensure that its inter-cluster similarity is low and its intra-cluster similarity is high. Two-dimensional clusters can be presented in a map, with different clusters painted in different colors. High-dimensional clustering can also be performed in multidimensional databases.
- **Time-series analysis.** This module contains several time-series data analysis functions, including similarity analysis, periodicity analysis, sequential pattern analysis, and trend and deviation analysis.

The DBMiner E1.0 release consists of the modules: *characterization*, *comparison*, *classification*, *association*, and *prediction*. The *clustering* and *time series analysis* modules will be released in future versions.

1.1.6 Support for task and method selection

The DBMiner system supports task and method selection via a window-based graphical user interface, where users may choose different mining tasks using a mining wizard or interact with the data mining results for mining at alternative dimensions and levels. Moreover, an SQL-like data mining query language, DMQL, has been designed, and a user-composed mining query is presented in the form of DMQL for examination before submitting for execution.

1.1.7 Support of KDD process

Since the DBMiner system works with a data warehouse, the two essential preprocessing tasks for knowledge discovery, *data cleaning* and *data integration*, are performed, when necessary, by the (supported) underlying data warehouse system. However, data consolidation (aggregation grouped by multiple dimensions and levels), data selection, and concept hierarchy construction can be performed in the DBMiner system. Moreover, concept hierarchies, multi-dimensional data, and intermediate data mining results are all stored in database relations, and their indexing and accessing are supported by commercial database system products.

In DBMiner, most of the postprocessing process on the discovered patterns is integrated with the data mining process since a data mining query provides not only the selection of task-relevant data and the mining task but also the interestingness measure (such as mining thresholds, including support, confidence, noise, etc.) and desired rule patterns (such as in meta-pattern guided mining of associations). Such an integration of mining and pattern evaluation not only reduces the search space but also facilitates the focus of mining process.

1.1.8 Main applications

The DBMiner system can be used as a general purpose, on-line analytical mining system for both on-line analytical processing (OLAP) and data mining in relational databases and data warehouses.

The system has been tested in several medium to large relational databases, including NSERC (Natural Science and Engineering Research Council of Canada) research grant information system, U.S. City-County Data Book, and several industry proprietary databases, with excellent performance.

Several specialized data mining system prototypes, including GeoMiner, MultiMediaMiner, and WeblogMiner, have been developed by extensions to the DBMiner system. A special purpose data mining system for banking industry is being developed based on the DBMiner system.

1.1.9 Current status

The DBMiner system has been evolving from a research system prototype to an industry product. However, its research innovation and new technology progress is still tightly connected with the university research laboratory.

The minimum hardware requirement for the DBMiner system is a Pentium-166 machine with 64 MB RAM. The system runs on Windows/NT and Windows-95. The system can be directly linked to Microsoft SQLServer, or communicate with various relational database systems, including Microsoft Access, Oracle, and others, via its ODBC connections.

A mini-version of DBMiner E1.0 can be freely downloaded at <http://db.cs.sfu.ca/DBMiner>. A scalable version of DBMiner E1.0 (Educational) can be purchased via <http://www.dbminer.com> for educational and research usage.

References

- [1] Y. Cai, N. Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 213–228. AAAI/MIT Press, 1991.
- [2] J. Han. Towards on-line analytical mining in large databases. *ACM SIGMOD Record*, 27:97–107, 1998.
- [3] J. Han, S. Chee, and J. Y. Chiang. Issues for on-line analytical mining of data warehouses. In *Proc. 1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'98)*, pages 2:1–2:5, Seattle, Washington, June 1998.
- [4] J. Han, J. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, W. Gong, M. Kamber, G. Liu, K. Koperski, Y. Lu, N. Stefanovic, L. Winstone, B. Xia, O. R. Zaïane, S. Zhang, and H. Zhu. DBMiner: A system for data mining in relational databases and data warehouses. In *Proc. CASCON'97: Meeting of Minds*, pages 249–260, Toronto, Canada, November 1997.
- [5] J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 399–421. AAAI/MIT Press, 1996.
- [6] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. R. Zaïane. DBMiner: A system for mining knowledge in large relational databases. In *Proc. 1996 Int'l Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 250–255, Portland, Oregon, August 1996.
- [7] K. A. Ross, D. Srivastava, and D. Chatziantoniou. Complex aggregation at multiple granularities. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98)*, March 1998.
- [8] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98)*, March 1998.