Online Data Mining

Héctor Oscar Nigro
Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina

Sandra Elizabeth González Císaro
Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina

INTRODUCTION

Several approaches for intelligent data analysis are not only available but also tried and tested. Online analytical processing (OLAP) and data mining represent two of the most important approaches. They mainly emphasize different aspects of the data and allow deriving of different kinds of information. So far, these approaches have mainly been used in isolation (Schwarz, 2002).

OLAP is a powerful data analysis method for multidimensional analysis of data warehouses. Data mining is the extraction of interesting (i.e., nontrivial, implicit, previously unknown, and potentially useful) information or patterns from data in large databases (Fayad, Piatesky-Shapiro, Smyth, & Utthurasamy, 1996; Han & Kamber, 2001).

OLAP is more frequently used for verification of the existing hypothesis. Data mining tries to generate such a hypothesis by uncovering hidden patterns. It is essentially an inductive process. It means that OLAP and data mining should be used together to complement each other (Schwarz, 2002).

Motivated by the popularity of OLAP technology, Han developed an online analytical mining (OLAM) mechanism for multidimensional data mining.

This paper is organized in the following way: First, the background section, in which OLAP and data mining are introduced. Second, the main section is divided into the following subsections: Definition, Online Mining—Expected Characteristics, Olam Architecture, Olam and Complex Data Types, and OLAM System—Essential Functionality. Then, Future Trends in the area. Finally the Conclusions, References, Key Terms.

BACKGROUND

Basically, an OLAP system must show multidimensional data with dimensions, measures, and hierarchies through the cube’s paradigm. The multidimensional model simplifies for users the process of making complex queries, modifying information in a report, swapping between aggregated and detailed data, selecting part of the data, and so forth (Han & Kamber, 2001; Lenz & Shoshani, 1997).

Selected functionalities include the following:

- drilling down
- rolling up
- dice & slice
- pivoting
- cubing

There are three types of OLAP systems:

- **Rolap**: Relational OLAP works on normalized data; the classic models are star and snowflake. Its characteristic is having few storing spaces but with the disadvantage of presenting the information more slowly.
- **Molap**: Multidimensional OLAP works with precalculated data for the optimization of response time. It is very fast but it uses more space than Rolap.
- **Hybrid**: Hybrid OLAP combines the benefits of both.

Data mining is an especially interesting and very complex task, as defined in the previous section, and it is also a confluence of multiple disciplines (see Table 1).

A good data mining query language will support ad hoc and interactive data mining. Ad hoc query-based data mining is better when users want to examine various data portions with different constraints. Database queries can be answered intelligently using concept hierarchies, data mining results, or data mining techniques.

The major functions of data mining follow (Han, Chee Sonny, & Chiang Jenny, 1999; Han & Kamber, 2001):

1. **Characterization**: Generalizes a set of task-relevant data into a generalized data cube, which can then be used to extract different kinds of rules or be viewed at multiple levels of abstraction from different angles. In particular, it derives a set of characteristic rules, which summarize the general characteristics of a set of user’s specified data (called the target class).
2. **Comparison**: Mines a set of discriminant rules that summarize the features, which distinguish the
class being examined (the target class) from other classes (called contrasting classes).

3. **Classification**: Analyzes a set of training data (i.e., a set of objects whose class label is known) and constructs a model for each class based on the features in the data. A set of classification rules is generated by such a classification process, which can be used to classify future data and develop a better understanding of each class in the database.

4. **Association**: Discovers a set of association rules at multiple levels of abstraction from the relevant set(s) of data in a database.

5. **Prediction**: Predicts the possible values of some missing data or the value distribution of certain attributes in a 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).

6. **Cluster analysis**: Groups a selected set of data in the database or data warehouse into a set of clusters to ensure the interclass similarity is low and the intra-class similarity is high.

7. **Time series analysis**: Performs data analyses for time-related data in databases or data warehouses, including similarity analysis, periodicity analysis, sequential pattern analysis, and trend and deviation analysis.

### Online Data Mining: Main Features

**Definition**

*Online analytical mining (OLAM; also called OLAP mining) is among the many different paradigms and architectures for data mining systems. It integrates OLAP with data mining and mining knowledge in multidimensional databases.* (Han, 1997)

<table>
<thead>
<tr>
<th>Table 1. A list of topics related to data mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Database systems, data warehouse, and OLAP</td>
</tr>
<tr>
<td>• Statistics</td>
</tr>
<tr>
<td>• Machine learning</td>
</tr>
<tr>
<td>• Visualization</td>
</tr>
<tr>
<td>• Information science</td>
</tr>
<tr>
<td>• High performance computing</td>
</tr>
<tr>
<td>• Business and application domain knowledge expertise</td>
</tr>
<tr>
<td>• Other disciplines: neuronal networks, mathematical modeling, information retrieval, pattern recognition, and others (Han, 2000)</td>
</tr>
</tbody>
</table>

---

**Online Mining: Expected Characteristics**

The following features are important for successful OLAM (Han, 1998a):

- Ability to mine anywhere
- Carve any portion of a database by drilling, dicing, filtering, and so forth, for mining
- Drill through to raw data and mining
- Support of multifeature cubes and cubes with complex dimensions and measures
- Discovery-driven exploration of multifeature cubes
- Spatial and multimedia dimensions/measures
- Cube-based mining method
- Multimining tasks and multilevel mining
- Selection and addition of data mining algorithms
- Different algorithms may generate different results
- Standard APIs allow users to develop their own algorithms
- Integration of data mining functions
- First dicing, then classification, then association, and so forth
- Ad-hoc query-based (constraint-based) mining
- Fast response and high performance online
- Visualization tools
- Data and knowledge visualization tools
- Extensibility