INTRODUCTION
Currently, most data warehouses are being used for summarization-based, multi-dimensional, online analytical processing (OLAP). However, given the recent developments in data warehouse and online analytical processing technology, together with the rapid progress in data mining research, industry analysts anticipate that organizations will soon be using their data warehouses for sophisticated data analysis. As a result, a tremendous amount of data will be integrated, preprocessed, and stored in large data warehouses.

Online analytical mining (OLAM; also called OLAP mining) is among the many different paradigms and architectures for data mining systems. It integrates online analytical processing with data mining and mining knowledge in multi-dimensional databases. It is a promising direction due to the:

- high quality of data in data warehouses
- available information processing infrastructure surrounding data warehouses
- OLAP-based exploratory data analysis
- online selection of data mining functions

OLAM MINING BENEFITS
Most data mining tools must work on integrated, consistent, and cleaned data. This requires costly preprocessing for data cleaning, data transformation, and data integration. Therefore, a data warehouse constructed by such preprocessing is a

PAYOFF IDEA
Online analytical mining (OLAM) integrates online analytical processing and data mining. It represents a promising direction for mining large databases and data warehouses. By learning about different data mining and data warehousing implementation solutions, technologies, and products, an organization will have a better understanding of how it fits into its overall enterprise and data processes.
valuable source of high-quality data for both OLAP and data mining. Data mining may also serve as a valuable tool for data cleaning and data integration.

Organizations with data warehouses have or will systematically construct comprehensive information processing and data analysis infrastructures surrounding them. The surrounding infrastructures include accessing, integration, consolidation, and transformation of multiple heterogeneous databases; Object-oriented Database Connectivity/Object Linking and Embedding Database (ODBC/OLEDB) connections; Web-accessing and service facilities; and reporting and OLAP analysis tools. Savvy organizations will make the best use of their available infrastructures, rather than constructing everything from scratch.

Effective data mining requires exploratory data analysis. Users often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms. Online analytical mining provides facilities for data mining on different subsets of data and at different levels of abstraction. It does this by drilling, pivoting, filtering, dicing, and slicing on a data cube and on intermediate data mining results. This, together with data/knowledge visualization tools, can greatly enhance the power and flexibility of exploratory data mining.

Users seldom know which kinds of knowledge they wish to mine. By integrating OLAP with multiple data mining functions, online analytical mining provides users with the flexibility to select desired data mining functions, and dynamically swap data mining tasks.

Because data mining functions are usually more computationally expensive than OLAP operations, organizations are challenged to efficiently implement online analytical mining in large data warehouses, and provide fast response. Various implementation methods and ways to perform online analytical mining are discussed below.

OLAM ARCHITECTURE

In a similar manner to how OLAP engine performs online analytical processing, an online analytical mining engine performs analytical mining in data cubes. Therefore, an integrated OLAM and OLAP architecture makes sense, whereby the OLAM and OLAP engines both accept users’ online queries (or commands) via a user graphical user interface (GUI) application programming interface (API). Work with the data cube in the data analysis is performed through a cube API, and a metadata directory guides the data cube access. The data cube can be constructed by accessing or integrating multiple databases, or by filtering a data warehouse via a database API, which may support OLEDB or ODBC connections.

An OLAM engine can perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, and
time-series analysis. Therefore, it usually consists of multiple, integrated data mining modules, making it more sophisticated than an OLAP engine. There is no fundamental difference between the data cube required for OLAP and that for OLAM, although OLAM analysis might require more powerful data cube construction and accessing tools. This is the case when OLAM involves more dimensions with finer granularities, or involves the discovery-driven exploration of multi-feature aggregations on the data cube, thereby requiring more than OLAP analysis.

Moreover, when exploratory data mining identifies interesting spots, an OLAM engine might need to drill through from the data cube into the corresponding relational databases for detailed analysis of particular portions of data; for example, in time-series analysis. Furthermore, a data mining process might disclose that the dimensions or measures of a constructed cube are not appropriate for data analysis. Here, a refined data cube design could improve the quality of data warehouse construction.

**OLAM FEATURES**

A well-thought-out design can help an organization to systematically develop OLAM mechanisms in data warehouses. The following features are important for successful online analytical mining:

- the ability to mine anywhere
- availability and efficient support of multi-feature cubes and cubes with complex dimensions and measures
- cube-based mining methods
- the selection or addition of data mining algorithms
- interaction among multiple data mining functions
- fast response and high-performance mining
- visualization tools
- extensibility

The OLAM process should be exploratory; that is, mining should be performed at different portions of data at multiple levels of abstraction. When using a multi-dimensional database and an OLAP engine, it is easy to carve many portions of data sets at multiple levels of abstraction using OLAP operations such as drilling, dicing/slicing, pivoting, and filtering. Such processes can also be performed during data mining, through interaction with OLAP operations.

Moreover, in some data mining processes, at least some of the data may require exploration in great detail. OLAP engines often provide facilities to drill through the data cube down to the primitive/low-level data stored in the database. The interaction of multiple data mining modules with an OLAP engine can ensure that mining is easily performed anywhere in a data warehouse.
Traditional data cube queries compute simple aggregates at multiple granularities. However, many data mining tasks require discovery-driven exploration of multi-feature cubes, which are complex subqueries involving multiple dependent queries at multiple granularities. This is the case, for example, when a user studies organizations whose growth rate in certain years in the 1990s was less than 60 percent of their average annual growth rate in that decade. The user could then compare the features associated with the poor performance years versus other years at multiple granularities, finding important associations.

Moreover, traditional data cubes support only dimensions of categorical data and measures of numerical data. In practice, the dimensions of a data cube can be of numerical, spatial, and multimedia data. The measures of a cube can also be of spatial and multimedia aggregations, or collections of them. Support of such nontraditional data cubes will enhance the power of data mining.

Cube-based data mining methods should be the foundation of the online analytical mining mechanism. Although there have been studies of concept description, classification, association, prediction, and clustering in relation to cube-based data mining, more research is needed on efficient cube-based mining algorithms.

Different data mining algorithms can generate dramatically different mining results — unlike relational query processing, which generates the same set of answers to a query with different processing efficiency. Therefore, it is important for organizations to provide alternative mining algorithms for a data mining function, giving users a choice.

Moreover, users might wish to develop their own algorithms in order to experiment with or customize a mining task. If they are given standard APIs, and the OLAM system is well modularized, sophisticated users can add or revise data mining algorithms. These user-defined algorithms could make good use of such well-developed system components as data cube accessing, OLAP functionality, and knowledge visualization tools, integrating them with the existing data mining functions.

One OLAM strength is the interaction of multiple data mining and OLAP functions; another is in selecting a set of data mining functions. For example, the steps may be to dice a portion of a cube, to classify the diced portion based on a designated class attribute, to then find association rules for a class of data so classified, and finally to drill down to find association rules at a finer granularity level. In this way, the organization can develop a data mining system that can tour around the selected data space at will, mining knowledge with multiple, integrated mining tools.

Because mining is usually more expensive than OLAP, OLAM may encounter greater challenges for fast response and high-performance processing. While it is highly desirable and productive to interact with the mining process and dynamically explore data spaces, fast response is critical for interactive mining. In fact, miners might choose to trade min-
ing accuracy for fast response, since interactive mining might progressively lead them to focus the search space, and find ever more important patterns. Once users can identify a small search space, they can call up more sophisticated but slower mining algorithms for careful examination.

It is important for organizations to develop a variety of knowledge and data visualization tools, because an OLAM system will integrate OLAP and data mining, and mine various kinds of knowledge from data warehouses. Charts, curves, decision trees, rule graphs, cube views, and box-plot graphs are effective tools to describe data mining results, and can help users interact with the mining process and monitor their data mining progress.

An OLAM system communicates with users and knowledge visualization packages at the top, and data cubes/databases at the bottom. Therefore, it should be carefully designed, systematically developed, and highly modularized.

Moreover, because an OLAM system must integrate with many sub-systems, it should be designed with extensibility in mind. For example, an OLAM system might be integrated with a statistical data analysis package, or be extended for spatial data mining, text mining, financial data analysis, multimedia data mining, or Web mining. Modularity allows easy extension into such new domains.

IMPLEMENTATION OF OLAM MECHANISMS

Because the OLAM mechanism requires efficient implementation, special attention should be paid to:

- modularized design and standard APIs
- support of online analytical mining by high-performance data cube technology
- constraint-based online analytical mining
- progressive refinement of data mining quality
- layer-shared mining with data cubes
- bookmarking and backtracking techniques

An OLAM system might well integrate a variety of data mining modules via different kinds of data cubes and visualization tools. Thus, highly modularized design and standard APIs will be important for the systematic development of OLAM systems, and for developing, testing, and sharing data mining modules across multiple platforms and systems.

In this context, OLEDB for OLAP by Microsoft, and multi-dimensional API (MDAPI) by the OLAP Council, respectively, could be important initiatives toward standardizing data warehouse APIs for both OLAP and mining in data cubes. Shareable visualization tool packages could also prove useful here — in particular, Java-based, platform-independent knowledge visualization tools.
High-performance data cube technology is critical to online analytical mining in data warehouses. There have been many efficient data cube computation techniques developed in recent years that helped in the efficient construction of large data cubes. However, when a mining system must compute the relationships among many dimensions, or examine fine details, it might be necessary to dynamically compute portions of data cubes on-the-fly.

Moreover, effective data mining requires the support of nontraditional data cubes with complex dimensions and measures, in addition to the on-the-fly computation of query-based data cubes and the efficient computation of multi-featured data cubes. This requires further development of data cube technology.

While most data mining requests are query or constraint based, online analytical mining requires fast response to data mining requests. Therefore, the organization must perform mining with a limited scope of data, confined by queries and constraints. In addition, the organization must adopt efficient, constraint-based data mining algorithms. For example, many constraints involving set containments or aggregate functions can be pushed deeply into the association rule mining process. The organization should also explore such constraint-based mining in other data mining tasks.

There is a wide range of data mining algorithms. While some are fast and scalable, higher-quality algorithms generally cost more. Organizations can use a methodology that first applies fast mining algorithms on large data sets to identify the regions/patterns of interest, and then applies costly but more accurate algorithms for detailed analysis of these regions/patterns. For example, in spatial association rule mining, one technique first collects the candidates that potentially pass the roughly determined minimum support threshold, and then further examines only those that pass the rough test, using a more expensive spatial computation algorithm.

Each data cube dimension represents an organized layer of concepts. Therefore, data mining can be performing by first examining the high levels of abstraction, and then progressively deepening the mining process toward lower abstraction levels. This saves the organization from indiscriminately examining all the concepts at a low level.

The OLAM paradigm offers the user freedom to explore and discover knowledge by applying any sequence of data mining algorithms with data cube navigation. Users can often choose from many alternatives when traversing from one data mining state to the next. If users set bookmarks when a discover path proves uninteresting, they can return to a previous state and explore other alternatives. Such marking and backtracking mechanisms can protect users from being lost in the OLAM space.
ANALYTICAL MINING METHODS

There are other efficient and effective online analytical mining techniques. These include the design of a data mining language, incremental and distributed mining of association rules, constrained association mining, mining periodic patterns, a wavelet technique for similarity-based time-series analysis, intelligent query answering with data mining techniques, and a multi-layer database model.

A good data mining query language will support ad hoc and interactive data mining. Such a language can serve as the underlying core for different GUIs in a variety of commercial data mining systems, and facilitate the standardization and wide adoption of the technology.

It is best to update data mining results incrementally, rather than mining from scratch on database updates — especially when a database contains huge amounts of data. And, while it is a straightforward process to work out incremental data mining algorithms for concept description, it is nontrivial to update association rules incrementally.

Ad hoc query-based data mining is best when users wish to examine various data portions with different constraints. Constrained association rule mining supports the constraint-based, human-centered exploratory mining of associations. By this process, too, user-specified constraints can be pushed deeply into the association mining process to reduce the search space.

Many patterns are periodic or approximately periodic in nature; for example, seasons change periodically by year; and temperatures change periodically by day. In some cases, while the whole sequence exhibits no periodicity behavior; some particular points or segments in the sequence could be approximately periodic. For example, while someone might watch a particular TV news show from 7:00 to 7:30 a.m. almost every morning, his TV-watching habit is irregular at other hours.

Using an OLAP-based technique for mining the periodicity of such patterns in large databases can be explored via two cases: (1) with a given period, and (2) with an arbitrary period. For a user-specified given period, such as per day, per week, or per quarter, the organization can aggregate the potential activity patterns for the given period along the time dimension in a data cube. Similar OLAP-based methods apply for mining periodic patterns with arbitrary periods.

Similarity-based time-series analysis is similar to a stock market database. It is used to find similar time-related patterns such as trends and segments in a large, time-series database. In most previous analyses of similarity-based time series, organizations have adopted such traditional trend analysis techniques as Fourier transformation. More recently, wavelet transformation-based similarity mining methods have been used to discover trends or similar curves or curve segments. This method has proven efficient and effective at mining large time-series databases.
Database queries can be answered intelligently using concept hierarchies, data mining results, or online data mining techniques. For example, instead of bulky answers, a summary of answers can be presented, allowing users to manipulate the summary by drilling or dicing. Alternatively, related answers or rules can be presented in the form of associations or correlations, based on association mining results.

The autonomy and semantic heterogeneity among different databases present a major challenge for cooperating multiple databases. Tools to handle this problem use methods for schema analysis, transformation, integration, and mediation. However, because schema level analysis may be too general to solve the problem, the organization should consider data-level analysis, whereby the database contents are analyzed.

The organization can construct a multi-layer database model utilizing a common data access API, and using generalization-based data mining to generalize the database contents from a primitive level to multiple higher levels. A multi-layer database provides a useful architecture for intelligent query answering, and helps in information exchange and interoperability among heterogeneous databases. This is because the low-level heterogeneous data is transformed into high-level, relatively homogeneous information that can then be used for effective communication and query/information transformation among multiple databases.

OLAM AND COMPLEX DATA TYPES

It is challenging to extend the online analytical mining method to complex types of data. These include complex data objects, spatial data, and text and multimedia data.

Object-oriented and object-relational databases introduce advanced concepts into database systems, including object identity, complex structured objects, methods, and class/subclass hierarchies. A generalization-based data mining method can generalize complex objects, construct a multi-dimensional object cube, and perform analytical mining in such an object cube.

Here, objects with complex structures can be generalized to high-level data with relatively simple structures. For example, an object identifier can be generalized to the class identifier of the lowest class where the object resides. In addition, an object with a sophisticated structure can be generalized into several dimensions of data that reflect the structure, the generalized value, or other features of the object.

A spatial database stores nonspatial data representing other properties of spatial objects and their nonspatial relationships, as well as spatial data representing points, lines, and regions. A spatial data cube consists of both spatial and nonspatial dimensions and measures, and can be modeled by the star or snowflake schema, resembling its relational counterpart. Spatial data mining can be performed in a spatial database as well as in a spatial data cube.
Text analysis methods and content-based image retrieval techniques play an important role in mining text and multimedia data. By one method of online analytical mining of text and multimedia data, text/multimedia data cubes are built, whereupon the cube-based relational and spatial mining techniques are extended toward mining text and multimedia data.

Organizations can also mine the Web access patterns stored in Web log records. Web log records are preprocessed and cleaned, and multiple dimensions are built, based on such Web access information as page start time, duration, user, server, URL, next page, and page type. The process includes construction of a WebLog, and the performance of time-related, multi-dimensional data analysis and data mining.

CONCLUSION

The rapid development of data warehouse and OLAP technology has paved the way toward effective online analytical mining. Analysts anticipate that OLAM will become a natural addition to OLAP technology, which enhances data analysis in data warehouses.

Notes

1. The OLAP Council’s proposed multi-dimensional API, now on version 2.0. The earlier, abortive version was called the MD-API (with a hyphen). Very few or no vendors are likely to support even the 2.0 version (which was released in January 1998), and no vendor has even announced a date for supporting it.
2. DBMiner is an intelligent data mining and data warehousing system. This educational release allows university teachers and research institutes to have a comprehensive data mining system to teach students and researchers the concepts and skills of data mining and data warehousing. Its first professional version has been released with enhanced data mining capabilities for professional users.
3. The traditional periodicity detection methods, such as Fast Fourier Transformation, find the periodicity of the whole sequence, but not the periodicity of a particular point/segment in the sequence.

John Vacca is an information technology consultant and internationally known author based in Pomeroy, Ohio.