Issues for On-Line Analytical Mining of Data Warehouses *

(Extended Abstract)

Jiawei Han, **Sonny H.S. Chee** and **Jenny Y. Chiang**

Intelligent Database Systems Research Laboratory
School of Computing Science, Simon Fraser University, British Columbia, Canada V5A 1S6
{ han, schee, ychiang }@cs.sfu.ca
URL: http://db.cs.sfu.ca/

Abstract

Data warehouses and OLAP engines are expected to be widely available in the near future. The data in data warehouses has been cleansed, integrated, and preprocessed, and infrastructures have been built surrounding data warehouses for efficient data analysis. Therefore, data warehouses or OLAP databases are expected to be a major platform for data mining in the future. We discuss the issues related to efficient and effective data mining in large data warehouses and/or data marts, including the desired architectures for an integrated on-line analytical processing (OLAP) and on-line analytical mining (OLAM) system, the expected features of OLAM, and how to implement such a system effectively.

1 Introduction

With recent developments of data warehouse and OLAP technology [5], it is expected that a tremendous amount of data will be integrated, preprocessed, and stored in large data warehouses. Currently, most of the data warehouses are being used for summarization-based, multi-dimensional, on-line analytical processing. However, with the rapid progress in data mining research [8], it is expected that powerful data mining tools will be introduced step-by-step for sophisticated data analysis in data warehouses.

In the field of data mining, substantial research has been performed for data mining at various platforms, including transaction databases, relational databases, spatial databases, text databases, time-series databases, flat files, data warehouses, etc.

Among many different paradigms and architectures of data mining systems, **On-Line Analytical Mining (OLAM)** (also called **OLAP mining**), which integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multi-dimensional databases, is a promising direction [9] for the following reasons.

1. High quality of data in data warehouses. Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data transformation, and data integration as preprocessing steps [8]. A data warehouse constructed by such preprocessing serves as a valuable source of high quality of data for OLAP as well as for data mining. Notice that data mining may also serve as a valuable tool for data cleaning and data integration as well [8].
2. Available information processing infrastructure surrounding data warehouses. Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple, heterogeneous databases, ODBC/OLEDB connections, Web-accessing and service facilities, reporting and OLAP analysis tools, etc. It is prudent to make the best use of the available infrastructures rather than constructing everything from scratch.
3. OLAP-based exploratory data analysis. Effective data mining needs exploratory data analysis [14]. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms. On-line analytical mining provides facilities for data mining on different subsets of data and at different levels of abstraction, by drilling, pivoting, filtering, dicing and slicing on a data cube and on some intermediate data mining results. This, together with data/knowledge visualization tools, will greatly enhance the power and flexibility of exploratory data mining.
4. On-line selection of data mining functions. Often a user may not know what kinds of knowledge that she wants to mine. By integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

*Research was supported in part by a research grant and a CRD grant from the Natural Sciences and Engineering Research Council of Canada, a grant NCE:IRIS/Precarn from the Networks of Centres of Excellence of Canada, and grants from B.C. Advanced Systems Institute, MPR Teltech Ltd., National Research Council of Canada, and Seagate Software.

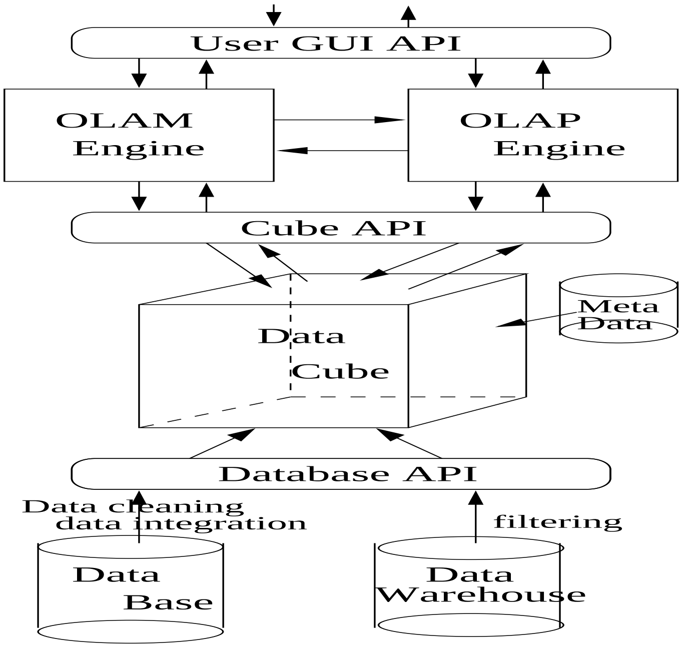

Figure 1: An integrated OLAM and OLAP architecture

The above observations motivate us to study the ways to perform on-line analytical mining and their implementation methods. Since data mining functions are usually computationally more expensive than OLAP operations, efficient implementation and fast response is the major challenge in the realization of on-line analytical mining in large data warehouses. This paper presents our view on the architecture, expected features, and methods for on-line analytical mining mechanisms in data warehouses and discusses some issues related to efficient implementations of on-line analytical mining.

The remaining of the paper is organized as follows. In Section 2, we present a system architecture for on-line analytical mining. In Section 3, we present the expected features of analytical mining. In Section 4, we discuss the methods for implementation of on-line analytical mining. Finally, we summarize our study and point out some future research directions in Section 5.

2 Architecture for on-line analytical mining

An OLAM engine performs analytical mining in data cubes in a similar manner as an OLAP engine performs on-line analytical processing. Therefore, we suggest an integrated OLAM and OLAP architecture, as shown in Figure 1, where the OLAM and OLAP engines both accept users' on-line queries (or commands) via a User_GUI_API and work with the data cube in the data analysis via a Cube_API. A meta data directory is used to guide the access of the data cube, and the data cube can be constructed by accessing and/or integrating multiple databases and/or by filtering a data warehouse via a Database_API which may support OLEDB or ODBC connections. Since an OLAM engine may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, etc., it usually consists of multiple, integrated data mining modules and is more sophisticated than an OLAP engine.

With many OLAP products available on the market, it is important to develop on-line analytical mining mechanisms directly on top of the constructed data cubes and OLAP engines. Based on our analysis, there is no fundamental difference between the data cube required for OLAP and that for OLAM, although OLAM analysis may often involve the analysis of more dimensions with finer granularities, or involve the discovery-driven exploration of multi-feature aggregations on the data cube [21, 22], and thus require more powerful data cube construction and accessing tools than OLAP analysis. Moreover, when exploratory data mining identifies interesting spots, an OLAM engine may sometimes (such as in time-series analysis) need to drill through from the data cube into the corresponding relational databases for detailed analysis of particular portions of data. Furthermore, a data mining process may disclose that the dimensions and/or measures of a constructed cube may not be appropriate for data analysis and thus suggest a refined design of the data cube, which may in turn improve the quality of data warehouse construction.

3 Expected features of on-line analytical mining

On-line analytical mining is at its infancy, and consequently it is important to discuss what should be its desired features. A thoughtful design will help systematic development of OLAM mechanisms in data warehouses.

We believe the following features are important for successful on-line analytical mining.

1. Ability to mine anywhere. With a multidimensional database and an OLAP engine, it is easy to carve any portions of data sets at multiple levels of abstraction using OLAP operations, such as drilling, dicing/slicing, pivoting, filtering, etc. This greatly facilitates the on-line analytical mining process since such a process should be exploratory in nature, that is, mining should be performed at different portions of data at multiple levels of abstraction. By interaction with OLAP operations, one can perform drilling, dicing/slicing, and pivoting during data mining as well. Moreover, some data mining process may need to explore at least some of the data in great detail. An OLAP engine often provides 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 will ensure that mining can be easily performed anywhere in a data warehouse.

2. Availability and efficient support of multi-feature cubes and cubes with complex dimensions and measures. Traditional data cube queries compute simple aggregates at multiple granularities. Many data mining tasks need discovery-driven exploration of *multi-feature cubes* [21, 22], which are complex subqueries involving multiple dependent queries at multiple granularities. For example, one may examine the companies whose growth rate in certain year(s) in 1990s is much weaker, e.g., less than 50% of their average annual growth rate in 1990s, compare the features associated with the poor performance years versus other years at multiple granularities, and find 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 the collections of such object pointers. Support of such non-traditional data cubes will enhance the power of data mining.

3. Cube-based mining methods. Cube-based data mining methods should be the core of the on-line analytical mining mechanism. There have been many studies on cube-based data mining including concept description [12], classification [16], association [15], prediction [6], clustering [2], etc. Cube-based mining may inherit the spirit of relational or transactional data mining methods, such as [3, 8], but may explore many distinct features of cube-based multi-dimensional optimization. More studies are needed on efficient cube-based mining algorithms.

4. Selection or addition of data mining algorithms. Unlike relational query processing which generates the same set of answers to a query with different processing efficiency, different data mining algorithms may generate dramatically different mining results. Therefore, it is important to provide alternative mining algorithms for a data mining function and give user a choice of data mining algorithms.

   Moreover, a user may even like to develop her own algorithms for experimentation or for customizing a mining task. If standard open APIs are provided and the OLAM system is well-modularized, it is possible for sophisticated users to add or revise data mining algorithms. By doing so, user-defined data mining algorithms can make the best use of some well-developed system components, such as data cube accessing, OLAP function, and knowledge visualization tools, and be integrated with the existing data mining functions.

5. Interaction among multiple data mining functions. The strength of OLAM should be not only at the selection of a set of data mining functions but also at the interaction among multiple data mining and OLAP functions. For example, one may like to first dice a portion of a cube, classify the diced portion based on a designate class attribute, and find association rules for a class of data so classified, and drill down to find association rules at a finer granularity level. This will lead to a data mining system which may tour around the selected data space at will and mine knowledge with multiple, integrated mining tools.

6. Fast response and high performance mining. OLAM may encounter greater challenges for fast response and high performance processing than OLAP because mining is usually more expensive than OLAP. It is highly desirable and productive to interact with the mining process and dynamically explore data spaces. However, fast response is critical for interactive mining. Sometimes one may even like to trade mining accuracy for fast response since interactive mining may progressively lead miners to focus the search space and find more and more important patterns. Once a user can identify a small search space, more sophisticated but slower mining algorithms can be called up for careful examination.

7. Visualization tools. Since an OLAM system will integrate OLAP and data mining and mine various kinds of knowledge from data warehouses, it is important to develop a variety of knowledge and data visualization tools [4]. Charts, curves, decision trees, rule graphs, cube views, boxplot graphs, etc. are effective tools to describe data mining results and help users monitor the progress of data mining and interact with the mining process.

8. Extensibility. An OLAM system is a system which will communicate with users and knowledge visualization packages at the top and data cubes/databases at the bottom. Thus, it should be highly modularized with careful design and systematic development. Moreover, an OLAM system should be designed with extensibility in consideration since an OLAM system will be expected to be integrated with many subsystems or be extended in many ways. For example, an OLAM system may be integrated with a statistical data analysis package, or be extended for spatial data mining, text mining, financial data analysis, multimedia data mining, Web mining, etc. A modularized design may lead to easy extension towards new domains.

4 A discussion on the implementation of OLAM mechanisms

Although there are many studies and proposals for efficient and effective data mining algorithms, OLAM mechanism requires fast response due to its nature of interactive mining. Thus, it poses new challenges on efficient implementation. We believe that special attention should be paid to the following implementation considerations.

1. Modularized design and standard APIs. It is expected that an OLAM system may integrate a variety of data mining modules with different kinds of data cubes and visualization tools. Therefore, highly modularized design and standard APIs will be important for systematic development of OLAM systems and for developing, testing, and sharing data mining modules across multiple platforms and systems. For example, *OLEDB for OLAP* [18] and *MDAPI* [7], by Microsoft and OLAP Council respectively, could be important initiatives towards standardization of data warehouse APIs not only for OLAP but also for mining in data cubes. It is also interesting to develop sharable visualization tool packages, especially some Java-based, platform-independent knowledge visualization tools.

2. Support of on-line analytical mining by high performance data cube technology. High performance data cube technology is critical to on-line analytical mining in data warehouses. There have been many efficient data cube computation techniques developed in recent years, such as [1, 13, 23, 20], which helps efficient construction of large data cubes. However, since a mining system may need to compute the relationships among a good number of dimensions or examine the fine details, but such data may not always be materialized beforehand, it will be necessary to dynamically compute portions of data cubes on the fly. Moreover, besides on-the-fly computation of query-based data cubes, the efficient computation of multi-featured data cubes, and the support of nontraditional data cubes with complex dimensions and measures, are crucial to effective data mining. Therefore, further development of data cube technology will provide enhanced support of on-line analytical mining.

3. Constraint-based on-line analytical mining. On-line analytical mining requires fast response upon data mining requests whereas most data mining requests are query-based, or constraint-based. This requires mining not only be performed with a limited scope of data, confined by queries and/or constraints, but also adopt efficient, constraint-based data mining algorithms. For example, many constraints involving set containments and/or aggregate functions can be pushed deeply into the association rule mining process [19]. Such constraint-based mining should be explored in many other data mining tasks.

4. Progressive refinement of data mining quality. There is a wide spectrum of data mining algorithms: some are fast and scalable but may not be of as high quality as some relatively expensive ones. An interesting methodology for on-line analytical mining is to apply fast mining algorithms first on large data sets to identify the regions/patterns of interest, and then apply costly but more accurate algorithms for detailed analysis of these regions/patterns. For example, in spatial association rule mining, a technique has been developed to first collect the candidates which "potentially" pass the minimum support threshold by certain rough spatial evaluation, such as using minimum bounding rectangles (MBRs), and only those which pass the rough test will be further examined using a more expensive spatial computation algorithm [17].

5. Layer-shared mining with data cubes. Since each dimension of a data cube represents an organized layer of concepts, data mining can be performing by first examining the high levels of abstraction and then progressively deepening the mining process towards lower levels of abstraction. This will save the efforts of indiscriminative examination of all the concepts at low level. Layered mining of association rules has been studied in [11, 15]. It is important to explore this optimization at mining other kinds of knowledge.

6. Bookmarking and backtracking techniques. The OLAM paradigm offers the user complete freedom to explore and discover knowledge by applying any sequence of data mining algorithms with data cube navigation. Often a user has a choice of many alternatives when traversing from one data mining state to the next. It would be useful if she can set bookmarks: if a discover path proves uninteresting, she can return to a previous state and explore other alternatives. Efficient support of such marking and backtracking mechanisms will protect users from being "*lost in the OLAM space*".

5 Conclusions

We discussed the motivation, architecture, expected features, and implementation issues of on-line analytical mining (OLAM) (or OLAP mining). On-line analytical mining (OLAM) integrates on-line analytical processing and data mining. It represents a promising direction for mining large databases and data warehouses.

The rapid development of data warehouse and OLAP technology has paved the way towards effective on-line analytical mining. It is expected that OLAM will be a natural addition to OLAP technology to enhance the power of data analysis in data warehouses.

We have been studying and implementing OLAM mechanisms in the DBMiner system [10]. Many expected features and methods discussed in this paper are being added incrementally and tested in the system. However, significant work is still required for a thorough implementation of the methods discussed in this abstract. We will continue working in this direction and report our progress towards that goal in the future.

References

[1] S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proc. 1996 Int. Conf. Very Large Data Bases*, pages 506–521, Bombay, India, Sept. 1996.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data*, Seattle, Washington, June 1998.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 1994 Int. Conf. Very Large Data Bases*, pages 487–499, Santiago, Chile, September 1994.

[4] C. Brunk, J. Kelly, and R. Kohavi. MineSet: An integrated system for data mining. In *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, pages 135–138, Newport Beach, California, August 1997.

[5] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26:65–74, 1997.

[6] S. Cheng. *Statistical Approaches to Predictive Modeling in Large Databases.* M.Sc. Thesis, Simon Fraser University, Canada, January 1998.

[7] OLAP council. MDAPI specification version 2.0. In *http://www.olapcouncil.org/research/apily.htm*, 1998.

[8] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

[9] J. Han. Towards on-line analytical mining in large databases. *ACM SIGMOD Record*, 27:97–107, 1998.

[10] 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.

[11] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. 1995 Int. Conf. Very Large Data Bases*, pages 420–431, Zurich, Switzerland, Sept. 1995.

[12] 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.

[13] V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, pages 205–216, Montreal, Canada, June 1996.

[14] J. Elder IV and D. Pregibon. A statistical perspective on knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 83–115. AAAI/MIT Press, 1996.

[15] M. Kamber, J. Han, and J. Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. In *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, pages 207–210, Newport Beach, California, August 1997.

[16] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han. Generalization and decision tree induction: Efficient classification in data mining. In *Proc. of 1997 Int. Workshop Research Issues on Data Engineering (RIDE'97)*, pages 111–120, Birmingham, England, April 1997.

[17] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int'l Symp. Large Spatial Databases (SSD'95)*, pages 47–66, Portland, Maine, Aug. 1995.

[18] Miscrosoft. OLEDB for OLAP programmer's reference version 1.0. In *http://www.microsoft.com/data/oledb/olap*, 1998.

[19] R. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data*, Seattle, Washington, June 1998.

[20] K. Ross and D. Srivastava. Fast computation of sparse datacubes. In *Proc. 1997 Int. Conf. Very Large Data Bases*, pages 116–125, Athens, Greece, Aug. 1997.

[21] 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.

[22] 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.

[23] Y. Zhao, P. M. Deshpande, and J. F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. In *Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data*, pages 159–170, Tucson, Arizona, May 1997.