Towards On-Line Analytical Mining in Large Databases *

Jiawei Han

Intelligent Database Systems Research Laboratory

School of Computing Science, Simon Fraser University, British Columbia, Canada V5A 1S6

URL: http://db.cs.sfu.ca/ (for research group) http://db.cs.sfu.ca/DBMiner (for system)

Abstract

Great efforts have been paid in the Intelligent Database Systems Research Lab for the research and development of efficient data mining methods and construction of on-line analytical data mining systems.

Our work has been focused on the integration of data mining and OLAP technologies and the development of scalable, integrated, and multiple data mining functions. A data mining system, `DBMiner`, has been developed for interactive mining of multiple-level knowledge in large relational databases and data warehouses. The system implements a wide spectrum of data mining functions, including characterization, comparison, association, classification, prediction, and clustering. It also builds up a user-friendly, interactive data mining environment and a set of knowledge visualization tools. In-depth research has been performed on the efficiency and scalability of data mining methods. Moreover, the research has been extended to spatial data mining, multimedia data mining, text mining, and Web mining with several new data mining system prototypes constructed or under construction, including `GeoMiner`, `MultiMediaMiner`, and `WebLogMiner`.

This article summarizes our research and development activities in the last several years and shares our experiences and lessons with the readers.

1 Introduction

The research into data mining in our lab started in early 1989, when we proposed an efficient knowledge discovery method, *attribute-oriented induction* [4]. Since then, we have investigated a set of interesting data mining methods for mining relational data, data warehouse data, spatial data, data formed with complex objects, text data, and multimedia data. These include enhancement of attribute-oriented induction [13, 16], automatic generation and adjustment of concept hierarchies [16], mining multi-level association rules [15], meta-rule guided mining of associations [22], incremental and distributed mining of associations [8, 7], constraint pushing in association mining [10, 27], mining periodicity and similarity in time-series data [11, 30], multi-level classification and prediction [23, 6], spatial data cube construction [21], spatial association rule mining [24], OLAP mining [12], Weblog mining [31], etc.

A data mining system, `DBMiner` [16, 14], has been constructed with our years of research and development. The system integrates data mining with on-line analytical processing (OLAP) and implements a spectrum of data mining functions, including characterization, comparison, association, classification, prediction, and clustering. An important goal of the system is to perform *multiple functional, on-line analytical mining* in large databases and data warehouses, where the *on-line analytical mining* implies that data mining is performed in a way similar to on-line analytical processing (OLAP) in multi-dimensional databases, i.e., mining can be performed, *interactively* (i.e., by mouse clicking and with quick response) when possible, in different portions of a multi-dimensional database and at different levels of abstraction.

This paper summarizes our work related to the research and development of on-line analytical mining mechanisms. The remaining of the paper is organized as follows. In Section 2, we present the on-line analytical mining mechanisms designed and implemented in the DBMiner system. In Section 3, we introduce our additional research into analytical mining methods. In Section 4, we present our work on mining complex types of data, including spatial data, complex data objects, text data, multimedia data, and Web data. Finally, we summarize our study and point out some future research directions in Section 5.

2 OLAP + Data Mining → On-Line Analytical Mining

On-line analytical processing (OLAP) is a powerful data analysis method for multi-dimensional analysis of data

*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 Hughes Research Laboratories.

warehouses [5]. Motivated by the popularity of OLAP technology, we develop an On-Line Analytical Mining (OLAM) mechanism for multi-dimensional data mining in large databases and data warehouses. We believe this is a promising direction to pursue based on the following observations.

1. 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 [9]. A data warehouse constructed by such preprocessing serves as a valuable source of cleaned and integrated data for OLAP as well as for data mining.

2. Effective data mining needs exploratory data analysis. A users often likes to traverse flexibly through a database, select any portions of relevant data, analyze data 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.

3. It is often difficult for a user to predict what kinds of knowledge to be mined beforehand. By integration of OLAP with multiple data mining functions, on-line analytical mining provides flexibility for users to select desired data mining functions and swap data mining tasks dynamically.

However, data mining functions usually cost more than simple OLAP operations. Efficient implementation and fast response is the major challenge in the realization of on-line analytical mining in large databases or data warehouses. Therefore, our study has been focused on the efficient implementation of the on-line analytical mining mechanism. The methods that we developed include the efficient computation of data cubes by integration of MOLAP and ROLAP techniques, the integration of data cube methods with dimension relevance analysis and data dispersion analysis for concept description, and data cube-based multi-level association, classification, prediction and clustering techniques. These methods will be discussed in detail in the following subsections.

2.1 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, it is suggested to have an integrated OLAM and OLAP architecture as shown in Figure 1, where the OLAM and OLAP engines both accept users' on-line queries (instructions)

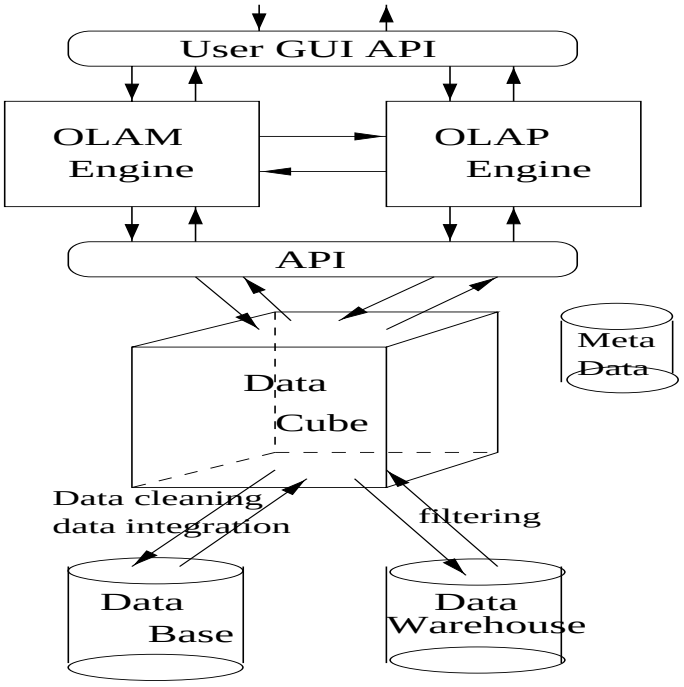

Figure 1: An integrated OLAM and OLAP architecture

and work with the data cube in the analysis. Furthermore, an OLAM engine may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, etc. Therefore, an OLAM engine is more sophisticated than an OLAP engine since it usually consists of multiple mining modules which may interact with each other for effective mining.

Since some requirements in OLAM, such as the construction of numerical dimensions, may not be readily available in the commercial OLAP products, we have chosen to construct our own data cube and build the mining modules on such data cubes. 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 a larger number of dimensions with finer granularities, and thus require more powerful data cube construction and accessing tools than OLAP analyses. Since OLAM engines are constructed either on customized data cubes which often work with relational database systems, or on top of the data cubes provided by the OLAP products, it is suggested to build on-line analytical mining systems on top of the existing OLAP and relational database systems, rather than from the ground up.

2.2 Data cube construction

Data cube technology is essential for efficient on-line analytical mining. There have been many studies on efficient computation and access of multidimensional databases, such as [1, 5, 33].

Our early development of attribute-oriented induction method [13] adopts two generalization techniques: (1) *attribute removal*, which removes attributes which represent low-level data in a hierarchy, and (2) *attribute generalization*, which generalizes attribute values to their corresponding high level ones. Such generalization leads to a new, compressed generalized relation with *count* and/or other aggregate values accumulated. This is similar to the relational OLAP (ROLAP) implementation of the roll-up operation.

For fast response in OLAP and data mining, our later implementation has adopted data cube technology as follows: when data cube contains a small number of dimensions, or when it is generalized to a high level, the cube is structured as compressed sparse array but is still stored in a relational database (to reduce the cost of construction and indexing of different data structures). The cube is precomputed using a chunk-based multiway array aggregation technique similar to [33]. However, when the cube has a large number of dimensions, it becomes very sparse with a huge number of chunks. In this case, a relational structure is adopted to store and compute the data cube, similar to the ROLAP implementation. We believe such a dual data structure technique represents a balance between multidimensional OLAP (MOLAP) and relational OLAP (ROLAP) implementations. It ensures fast response time when handling medium-sized cubes/cuboids and high scalability when handling large databases with high dimensionality.

Notice that even adopting the ROLAP technique, it is still unrealistic to materialize all the possible cuboids for large databases with high dimensionality due to the huge number of cuboids. It is wise to materialize more of the generalized, low dimensionality cuboids besides considering other factors, such as accessing patterns and the sharing among different cuboids.

A 3-D data cube/cuboid can be selected from a high-dimensional data cube and be browsed conveniently using the DBMiner 3-D cube browser as shown in Figure 2, where the size of a cell (displayed as a tiny cube) represents the entry *count* in the corresponding cell, and the brightness of the cell represents another measure of the cell. Pivoting, drilling, and slicing/dicing operations can be performed on the data cube browser with mouse clicking.

2.3 Concept description

Concept/class description plays an important role in descriptive data mining. It consists of two major func-

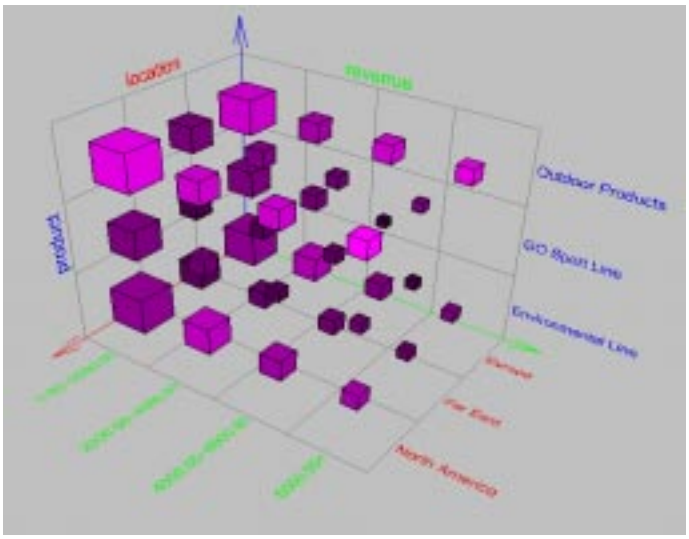

Figure 2: Browsing of a 3-dimensional data cube in DBMiner