Intelligent Data Analysis of E-Learning System Based on Data Warehouse, OLAP and Data Mining Technologies

JELENA MAMCENKO, IRMA SILEIKIENE
Information Technology Department
Vilnius Gediminas Technical University
Saulėtekio al. 11, Vilnius
LITHUANIA

Abstract: The data warehouse provides the data foundation for the data - the place where the data that goes into the process of knowledge discovery is stored. Data mining may be used to automatically perform knowledge discovery by giving the mining algorithm loose cues about potential relationships and letting the algorithm work on the data to discover the relationships and items to focus on further. OLAP is complimentary to data mining and is most likely the first, and most preferred, manner of discovering knowledge. OLAP works through a user performing specific, rather than general, interactive analysis with the data. If a data warehouse is present in the environment, either it or a data mart, would be the database used by OLAP.
In this article we tried to introduce possible ways of intelligent E-learning system's data analysis using data warehouse, OLAP and data mining technologies. Some aspects and suggestion of using these intelligent analysis components are presented.

Keywords: Intelligent Data Analysis, E-Learning, OLAP, Data Mining, Data Warehouse

1 Introduction

Business intelligence (BI) applications and their underlying data warehouses have been used in corporations for many years for enterprise reporting and business planning. Data warehousing has enabled IT organizations to extract and integrate data from diverse transactional systems into warehouse databases containing clean and consistent information stored in a form that business users can understand [1].

Data warehousing by itself, however, is not a complete answer to the information needs of business users. IT organizations have discovered that unless the information in the data warehouse is thoroughly documented and easy to access, complexity still limits data warehouse usage. To help solve this problem, vendors introduced business intelligence tools that focused on improving access to information in a data warehouse, and on offering a rich set of analytical processing capabilities. The emergence of these tools saw the introduction of online analytical processing (OLAP) and data mining technologies, and the use of the Web as a cost-effective way to deliver business information to a large user audience [1].

To reduce development time and to improve ease of use, vendors have also begun offering packaged analytical applications that exploit existing business intelligence and data warehouse technologies. These application packages are out-of-the box solutions whose business domain-specific analytics can be customized as needed to suit an organization's business operations.

2 Data Warehouse, OLAP and Data Mining Technologies

As the amount of the data gathered in relational databases kept growing exponentially with time, the ability to manually process the data was getting more and more limited. This led to the invention of a new model of data processing called On-Line Analytical Processing (OLAP). OLAP offers new ways to perceive both the physical and logical data organization, as well as new ways to process transactions. OLAP tools are not used to support the day-to-day operations of a database system; they are used mainly for decision support, hypothesis analysis, and statistical processing. Typical OLAP operations, often called cubing operations, include rollup (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) along one or more dimensions hierarchies, slicing (the extraction from a data cube of summarized data for a given dimension value), dicing (the extraction of "sub-cube" or intersection of several slices) and pivot (rotating of the axes of a data cube so that one may examine a cube from different angles) [2]. A database server implementing the OLAP functionality is called a data warehouse. According to W. Inmon [3], [4], widely considered the father of the modern data warehouse, a data warehouse is a "subject-oriented, integrated, time-variant, non-volatile collection of data in support of decision making". Subject orientation means that the data stored in the data warehouse are organized around selected subjects which are crucial from the point of view of a given organization. The data warehouse usually integrates data from several data sources e.g. operational databases, spreadsheet files, legacy database systems, the Internet, and operational enterprise database systems (Fig. 1). Contrary to traditional database systems, the data warehouse contains snapshots of the data from different time moments, thus enabling a history-driven analysis of the data. Finally, data stored in the data warehouse are non-volatile, which means that once loaded into the data warehouse, the data never change or evolve. Therefore, the data warehouse can be regarded as a huge repository of integrated data, where the only permitted operations are loading of new data and querying of existing data. Data warehouses are used as the source for advanced analysis, e.g. OLAP reporting tools or data mining [5].

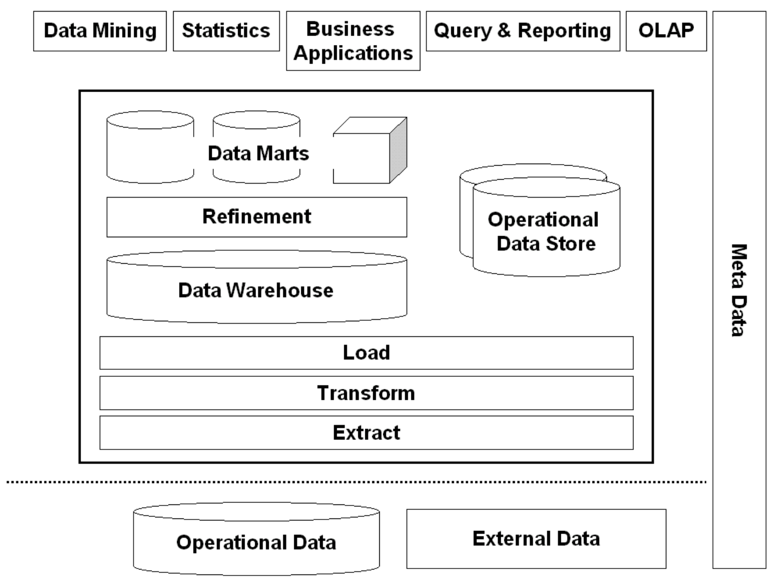

Fig. 1. Data Warehouse components

Other technology that can be used for querying the warehouse is data mining. The amount of data to be processed usually discards the possibility to process the data manually. Recent advances in data capture and data storage (e.g. the widespread use of bar code technology and digital cards) further increase the amount of data which are continuously loaded into databases. Unfortunately, the progress in data gathering techniques is not accompanied by the increased ability to process and utilize the data. Companies are gathering gigabytes of business data, without being able to actually put these data to work. Many public and private organizations are harvesting scientific and medical data, multimedia resources, web logs [6], text documents, or sensor measurements, yet current database and data warehouse technologies are not sufficient to efficiently process the gathered data and discover useful knowledge from it. Therefore, advanced systems are required to automatically process very large amounts of data and acquire useful knowledge from the data self-reliantly. The process of discovering novel, previously unknown, potentially useful, understandable and valid patterns from a large amount of data is called data mining or knowledge discovery in databases [7]. Data mining is a discipline which brings together databases, decision support systems, machine learning, artificial intelligence, statistics, data visualization, and several other disciplines [8]. Data mining uses different models of knowledge to present patterns discovered in data. These models include, but are not limited to association rules, decision trees, sequential patterns, clusters and others. In parallel, numerous algorithms have been developed to discover and maintain patterns [9]. The knowledge discovery process executes on the data that are provided by the data warehouse. These data are pre-processed, transformed, cleaned, and enriched with a background and domain knowledge. Most data mining algorithms heavily depend on the performance of the underlying database management system. Unfortunately, contemporary database systems offer very little or no support for data mining tasks. In addition, most of data mining methods are expensive in terms of time, computing power, and memory utilization.

The recommendations for collaboration Data Warehouse, OLAP and Data Mining are the next [10]:

- To build an ODS (Operational Data Store) where will be collected and cleaning data from OLTP systems;
- To build a star schema data warehouse with fact table and dimensions table;
- To use data in the data warehouse to build an OLAP data mart.

Then the recommended method for building data warehouse and data mining data marts could be quite the same [10]:

- To build an ODS where will be collected and cleaning data from OLTP systems;
- To build a star schema data warehouse with fact table and dimensions table;
- To pick the dimension which is of main interest, for instance, customers – to use aggregation and pivot on the fact table and maybe one or two other dimensions in order to build a flat record schema or a data mart for the mining techniques.

As a star schema model or multidimensional model, a data warehouse should be a prerequisite for OLAP data marts, even if it is not a prerequisite for a data mining project; it may help as a design guideline. OLAP and data mining projects could use the same infrastructure. The construction of the star schema and extracting/transforming/loading steps to build the data warehouse are the responsibilities of the IT department. An IT department should of course take into account the business users' requirements on OLAP as cubes or multidimensional databases, reports and also data mining models to design the data warehouse.

OLAP and data mining can use the same data, the same concepts, the same metadata and also the same tools, perform in synergy, and benefit from each other by integrating their results in the data warehouse [11].

3 E-learning System in Vilnius Gediminas Technical University

Today's universities and education institutions temporizing with the modern day circumstances should at most approach to students and assure the potential for them to study, contribute up-to-date knowledge for everyone by all available ways. Thus, part-time and extramural studies are coming apace back to study programs, and distant learning is ploughing its way.

3.1 E-Learning IS

Already since 2003 VGTU Information Technologies Department jointly with Kaunas University of Technology offer distant learning study program "Distant Learning in Information Technologies" at Master level next to traditional full-time Engineering Informatics study program. Starting with this academic year part-time Engineering Informatics basic studies and extramural studies were offered for college graduates.

As this new approach to studies has occurred, teachers were motivated to develop and offer students distant learning course on their subject. Thus, a need to develop and implement distant learning system has occurred, in order to satisfy all students', teachers' and administration needs.

An information system enabling to realise the blended learning method has been designed and implemented at the Information Technologies Department of VGTU. However, for the analysis of student data only the asynchronous learning portion of the information system will be used, since this portion is already used in master's studies for a considerable time, and a sufficient amount of data has been accumulated. The synchronous learning portion is still being tested and is not used with students yet for practical purposes.

Asynchronous courses are created with the technology of Lotus IBM LearningSpace, which is implemented on the irma.vtu.lt server. The information on users and their log-in rights is stored on a LDAP server, which is implemented on the gama.vtu.lt server. Shared data on user activities in the e-learning system are recorded and stored in the files of the Domino Web Servers Log. The schema of users' data workflow is presented in Fig. 2.

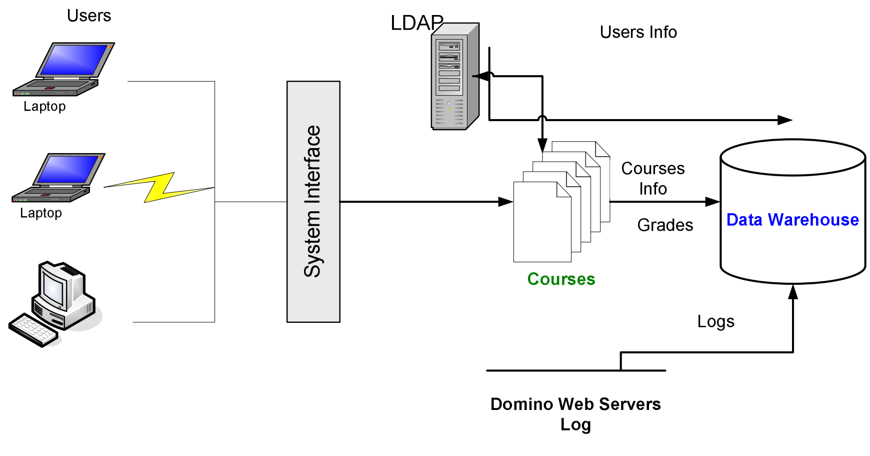

Fig. 2. Users' data workflow schema of e-learning information system

The information on users from the LDAP server, the course information and student results from LearningSpace courses, and the logs from the Domino Web Server are converted and stored in the Data Warehouse. The data from the Data Warehouse will be used in data analysis.

At this time there are 11 e-courses in the LearningSpace course catalogue including four courses for master's studies: ODL (Open Distance Learning) Infrastructure, Multimedia, Organising Information Activities, and Virtual Learning Environment. The e-courses have been used in master's studies for three years; about 70 master students who are identified in the system have studied these courses. Other courses are meant for bachelor students in Informatics as supplementary material or lecture notes. Students are not registered for these courses; they log in using a shared student account. Data analysis can be performed on the data of master students and bachelor students as well.

3.2 Intelligent Data Analysis of E-Learning system

E-Learning system's Data Warehouse contains several months' data from several data sources (Fig. 2).

The implementation of data mining techniques to blended delivery courses can help to understand visitors needs, for instance, modification of web page to better fit for the user, Web page creation that are unique per user or using the desires of a user to determine what documents to retrieve. These pieces of knowledge could lead to better course location on web site and education strategies. Likewise, an analyst could perform OLAP on the data warehouse to determine what kind of courses are the most popular and what students attend the most. For instance, as a dimension could be student, event, Date, complementary or teacher. As a measure we can choose a degree of learning, satisfaction, complementary or other. But there is one requirement for analyzed data: data should be organized and stored so that it can be analyzed from any dimension or at any level of the hierarchy.

According to all above and to the result which were received in previous works [12], data mining would enable to help the learners who are interested in certain areas by suggesting relevant or complimentary courses of which they might not be aware, providing with a personalized registration Web page [13]. For the learning provider, they will have the chance to view data of learners and courses from different angles in order to have full picture, enabling them to make the most profitable decision via targeting the class of users of interest to them and investing more in courses that are highly required by their targeted classes of students [14].

4 Conclusions and future work

OLAP and data warehousing are complementary fields. A data warehouse stores and manages data. OLAP subsets and transforms data warehouse data into strategic information. OLAP ranges from basic navigation and browsing ("slice and dice"), to calculations, to more serious analysis such as time series and complex modelling. As decision-makers exercise more advanced OLAP capabilities, they can use data access to gain knowledge rather than information only.

The final result of OLAP techniques can be very simple (e.g., frequency tables, descriptive statistics, simple cross-tabulations) or more complex (e.g., they may involve seasonal adjustments, removal of outliers, and other forms of cleaning the data). Although Data Mining techniques can operate on any kind of unprocessed or even unstructured information, they can also be applied to the data views and summaries generated by OLAP to provide more in-depth and often more multidimensional knowledge. In this sense, Data Mining techniques could be considered to represent either a different analytic approach (serving different purposes than OLAP) or as an analytic extension of OLAP. Online analytical processing (OLAP) and data mining are integral parts of any decision support process.

In future work we will consider on performing analysis of student activity and behaviour as well, will intend to research the relationship between student activity and study results; i.e., to research how intensely students study in e-courses, and whether their study results depend on the intensity. In addition to that it will be interesting to classify them by particular characteristics. Having the university domain, is that once we perform the clustering we might end up observing a cluster being full of students coming from a certain postcode, it might be interesting to investigate it further.

Performing analysis using OLAP technology it is possible to build data cube and aggregate the number of visited pages, the number or error pages, the generated traffic and visit estimation time according to possible dimensions such as date and time, courses, students and maybe others.

References:

[1] White C. J. Corporate Performance Optimization Guide. Intelligent Business Strategies, Sponsored by Oracle Corporation, 2002.

[2] Naydenova I., Kaloyanova K., Basic Approaches of Integration between Data Warehouse And Data Mining. Is available on http://isdg.fmi.uni-sofia.bg/reports/DMDWIntegrationLast.doc.

[3] Inmon W.H., *Building the Data warehouse, Fourth Edition*, Wiley Publishing, Inc, 2005.

[4] Inmon, W.H. *Building the Data warehouse, Fourth Edition*, Wiley Publishing, Inc, 2005.

[5] Jarke M., Lenzerini M., Vassiliadis P., Vassiliou Y., *Fundamentals of Data Warehouses*. Springer Verlag, 2003.

[6] Cooley R., Mobasher B., Srivastava J., Web Mining: Information and Pattern Discovery on the World Wide Web.

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

[8] Fayyad U. M., Piatetsky-Shapiro G., Smyth P., From Data Mining to Knowledge Discovery in Databases, *AI Magazine* 17(3), p.37-54, 1996.

[9] Witten I. H., Frank E., *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 2000.

[10] Baragoin C., Ch. M. Andersen., S. Bayerl and oth., Mining your Own Business in Telecoms Using DB2 Intelligent Miner for Data. © *International Business Machine Corporation*, First Edition, 2001.

[11] Forcht K.A., Cochran K., Using Data Mining and Data Warehousing Techniques. *Industrial Management & Data System*s, Volume 5, 1999, pp.189-196.

[12] Kulvietis G., Mamcenko J., Sileikiene I., Data Mining Application for Distance education Information System. *WSEAS Transactions on Information Science and Applications,* Issue 8, Volume 3, p.1482-1488, August 2006.

[13] Monk D., Using Data Mining for E-Learning Decision Making. *The Electronic Journal of e-Learning*, Volume 2, Issue 1, pp. 41-54, 2005. Is available online at www.ejel.org.

[14] Hanna M., Data Mining in the e-learning domain. *Campus-Wide Information Systems*, Vol. 21 Nr.1, 2004, pp. 29-34.