ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Integrated Data Mining and Knowledge Discovery Techniques in ERP

[I]Gandhimathi Amirthalingam, [II]Rabia Shaheen, [III]Mohammad Kousar, [IV]Syeda Meraj Bilfaqih
[I,III,IV]Dept. of Computer Science, King Khalid University, Kingdom of Saudi Arabia
[II]Dept. of Computer Science, Iqbal Institute of Technology & Management, Jammu and Kashmir, India

Abstract
Enterprise Resource Planning (ERP) is business process that allows an organization to use a system of integrated applications to manage the business and automate many back office functions related to technology, services and human resources. Data warehouse online analytical processing techniques provided decision makers a set of useful tools to analyze report and graphically represent data of the ERP. It provides an overview of the emerging field machine learning and clarifying how data mining and knowledge discovery in databases are related to each other. This paper provides a comparison of benefits obtained by applying OLAP or data mining techniques and the effect of integrating the both approaches in ERP.

Keywords
Data Mining, ERP, Knowledge Discovery, Machine Learning, Data Warehousing

I. Introduction

Enterprise Resource Planning is a set of applications for core business operations and back-end management that was originally developed for manufacturing and commercial companies. The most significant factor that distinguishes ERP systems from previous generations of information systems such as MRP (Material Requirement Planning) that ERP permits organizations to integrate business processes and optimize the resources available [1].

Machine learning is a mature and well-recognized research area of computer science, mainly concerned with the discovery of models, patterns, and other regularities in data [9]. A learning system uses sample data to generate an updated basis for improved [performance] on subsequent data from the same source and expresses the new basis in intelligible symbolic form Machine learning approaches can be roughly categorized into two different groups:

Symbolic approaches. Inductive learning of symbolic descriptions such as rules trees or logical representations.

Statistical approaches. Statistical or pattern-recognition methods, including k-nearest neighbour or instance-based learning, Bayesian classifiers, neural network learning, and support vector machines.

Research areas related to machine learning and data mining include database technology and data warehouses, pattern recognition and soft computing, text and web mining, visualization, and statistics.

- Database technology and data warehouses are concerned with the efficient storage, access and manipulation of data.
- Pattern recognition and soft computing typically provide techniques for classifying data items.
- Visualization concerns the visualization of data as well as the visualization of data mining results.
- Statistics is a classical data analysis discipline, mainly concerned with the analysis of large collections of numerical data.
- Text and web mining are used for web page analysis, text categorization, as well as filtering and structuring of text documents; natural language processing can provide useful tools for improving the quality of text mining results.

There is a need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data [7, 4]. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). At an abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data [5]. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a short report), more abstract [2] (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases) [6]. At the core of the process is the application of specific data-mining methods for pattern discovery and extraction.

II. Data Mining and KDD

The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase knowledge discovery in databases was coined at the first KDD workshop to emphasize that knowledge is the end product of a data - driven discovery [3]. It has been popularized in the AI and machine-learning fields. Data mining is the application of specific algorithms for extracting patterns from data. Fig. 1 shows an overview of KDD process. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining are essential to ensure that useful knowledge is derived from the data.

The data-mining component of KDD currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data mining step of the KDD process. Database techniques for gaining efficient data access, grouping and ordering operations when accessing data, and optimizing queries constitute the basics for scaling algorithms to larger data sets [11]. Most data-mining algorithms from statistics, pattern recognition, and machine learning assume data are in the main memory and pay no attention to how the algorithm breaks down if only limited views of the data are possible.

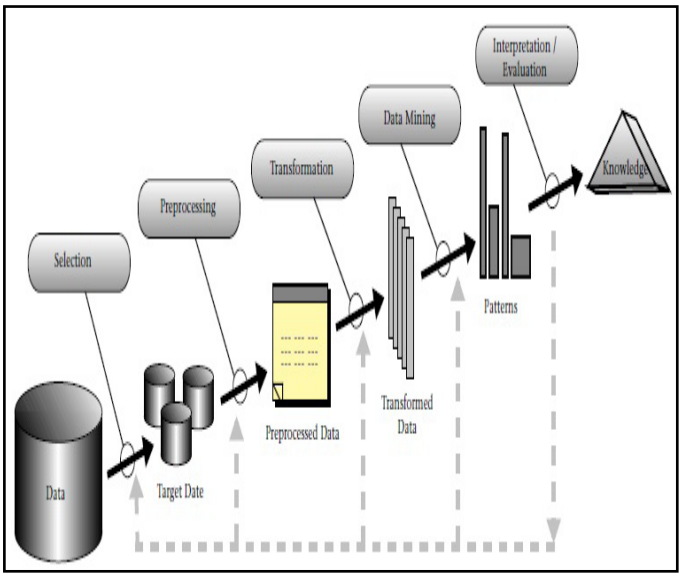

Fig. 1: An Overview of KDD Process

A related field evolving from databases is data warehousing, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps set the stage for KDD in two important ways: (1) data cleaning and (2) data access.

Data cleaning: As organizations are forced to think about a unified logical view of the wide variety of data and databases they possess, they have to address the issues of mapping data to a single naming convention, uniformly representing and handling missing data, and handling noise and errors when possible.

Data access: Uniform and well-defined methods must be created for accessing the data and providing access paths to data that were historically difficult to get to (for example, stored offline).
OLAP tools focus on providing multidimensional data analysis, which is superior to SQL in computing summaries and breakdowns along many dimensions [10]. OLAP tools are targeted toward simplifying and supporting interactive data analysis, but the goal of KDD tools is to automate as much of the process as possible. Thus, KDD is a step beyond what is currently supported by most standard database systems.
The idea of ERP systems is initially started by the development of Material Requirements Planning (MRP) systems to handle the planning and schedule of complex products. On next stages, MRP systems are developed to include some operational areas of the enterprise such as (sales, operation planning and financials). Finally, ERP systems are developed to integrate all business functions of an enterprise to support decision making. Fig. 2 shows modules of a typical ERP.

Fig. 2: Typical ERP Module

III. Machine Learning

The general problem of machine learning is to search a, usually very large, space of potential hypotheses to determine the one that will best fit the data and any prior knowledge. The data may be labelled or unlabelled. If labels are given then the problem is one of supervised learning in that the true answer is known for a given set of data. If the labels are categorical then the problem is one of classification, e.g. predicting the species of a flower given petal and sepal measurements. If the labels are real-valued the problem is one of regression, e.g. predicting property values from crime, pollution, etc. If labels are not given then the problem is one of unsupervised learning and the aim is characterize the structure of the data, e.g. by identifying groups of examples in the data that are collectively similar to each other and distinct from the other data. Fig. 3 shows the machine learning module.
In supervised learning there is necessarily the assumption that the descriptors available are in some related to a quantity of interest. For instance, suppose that a bank wishes to detect fraudulent credit card transactions. In order to do this some domain knowledge is required to identify factors that are likely to be indicative of fraudulent use. These may include frequency of usage, amount of transaction, spending patterns, type of business engaging in the transaction and so forth. These variables are the predictive, or independent, variables **x**. It would be hoped that these were in some way related to the target, or dependent, variable y, deciding which variables to use in a model is a very difficult problem in general; this is known as the problem of feature selection and is NP-complete. Many methods exist for choosing the predictive variables; if domain knowledge is available then this can be very useful in this context. Here we assume that at least some of the predictive variables at least are in fact predictive.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

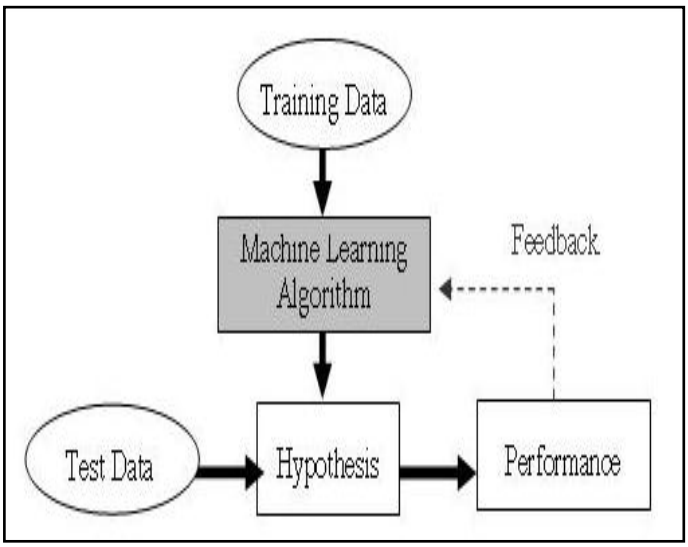

Fig. 3: Machine Learning Module

Assume, then, that the relationship between **x** and y is given by the joint probability density,

$$P(\mathbf{x}, y) = P(\mathbf{x})P(y \mid \mathbf{x}) \quad (1)$$

Eq. (1) formulation allows for y to be either a deterministic or stochastic function of **x** , in reality the available data are generated in the presence of noise so the observed values will be stochastic even if the underlying mechanism is deterministic.

IV. ERP and DATA WareHOUSING

Business Intelligence is providing decision makers with valuable information and knowledge by leveraging a variety of sources of data as well as DFD Structured and unstructured information. The information and data could reside within or outside the organization, could be structured in different ways, and could be either quantitative or qualitative. In some instances, this activity may reduce to calculations of totals and percentages, graphically represented by simple histograms, whereas more elaborate analyses require the development of advanced optimization and learning models [8]. Traditional database systems do not satisfy the requirements of data analysis necessary for BI. They are optimized to support the daily operations of an organization and their primary concern is to ensure the fast access of data in the presence of multiple users, such as ERP system's database.

Data warehouses are used as a data source for On-line analytical processing (OLAP) and machine learning. Fig. 4 shows a typical architecture of business intelligence and data sources to create data warehouse to apply OLAP and data mining. Fig. 5 shows the typical ERP solution. A data warehouse is a collection of subject-oriented, integrated, non-volatile, and time-variant data to support decision making and BI. In large data warehouse environments, many different types of analysis can occur. Data warehouse can be enriched with advance analytics using OLAP (On-Line Analytic Processing) and data mining. Rather than having a separate OLAP or data mining engine, they can also be integrated.

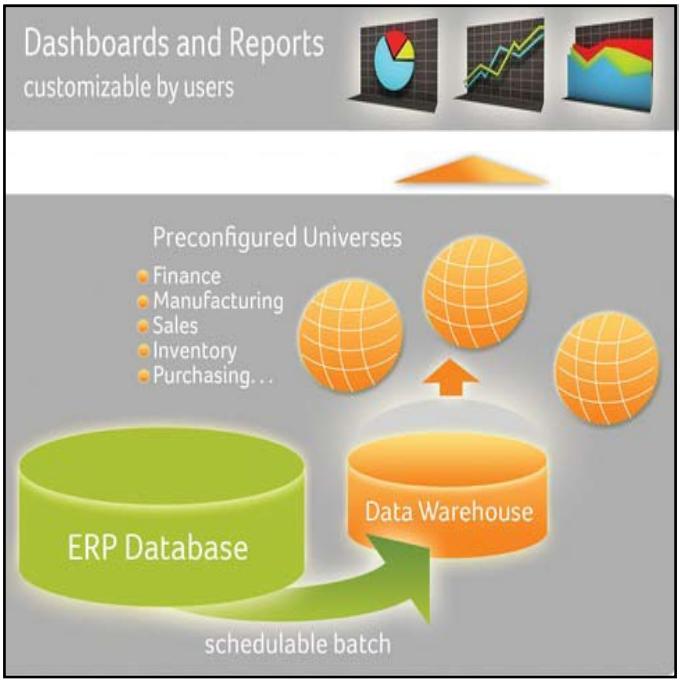

Fig. 4: ERP and Data Warehousing – User portal

V. ERP and Data Mining

Business Data Mining is the process of exploration and analysis to discover meaningful correlations, patterns and trend by sifting through large amounts of data stored in repositories. On the most important applications targeted by data Mining is ERP. Using data mining, businesses may be able to perform effective market analysis, compare customer feedback, identify similar products, retain highly valuable customers and make smart business decisions. Data Mining uses many several predictive and statistical methods in order to explore and analyse data. Such methods include association rule, linear regression, neural networks, regression trees, cluster analysis and classification trees.

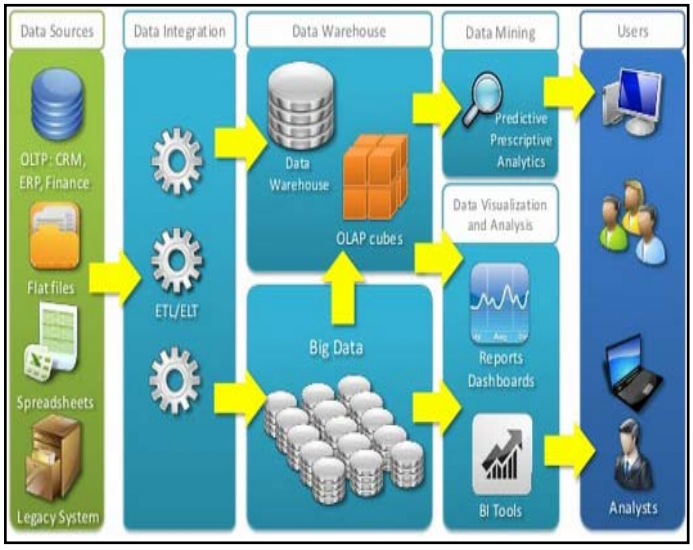

Fig. 5: Typical ERP Solution

VI. Integrating OLAP and Data Mining in ERP

We present for integrating OLAP and data mining so that they can benefit from each other's advances for the ultimate objective of efficiently providing a flexible answer to data mining queries addressed either to a relational or a multi dimensional database. The implementation of OLAP like techniques relies on three

operations on lattices, namely selection, projection and assembly. Table 1 illustrates the scope and benefits of the ERP module with OLAP and Data mining tasks.
Four major steps in integrating multi-dimension module in ERP:

1. Define the data mart of ERP:

Must define the overall methods and scopes during the evaluation stage of the design and determine using Schema or Snowflake in accordance with the requirements. It is an emphasis on the single business activity of the enterprise, such as importing, purchasing or ordering of goods.

2. Choices of fact:

The Cube is built and completed using Fact. Therefore Fact must be able to answer all the possible questions that may occur during the process of decision-making.

Table 1 : Comparing Tasks that can be performed by OLAP vs. DM at ERP Systems

ERP Module	OLAP Tasks	Data Mining
Accounting & Finance Management	Compare and visualize cost / profit of company	Forecast total profit/loss, cash flow of company
Human Resources Management	Compare and visualize salaries, rewards, and evaluation overtime	employee based on historical data
Vendors & Purchase Management	Purchases totals, cost of products' supply chains', Comparing purchases cost of different vendors	Determine best quantities of purchase orders
Production Management	Analyzing cost, scheduled variances, defects, manufacturing line	Applying Classification /Clustering technique
Customer Relationship Management	Analyze sales, customers' response, and issues	Identify customers' behavior patterns
Sales & Distribution Management	Compare total sales, offer, purchase history	Determine items sold, customer behavior, marketing efforts

3. Establishment of Dimension:

Use simple and useful message in words. Codes, abbreviations and Null are all unfitted for dimensions. The explicated time, names or addresses allow more flexibility in inquiries. Each and every item of the Dimension Table carries multiple feedback capabilities in processing the Fact Table. When there are changes over the data, the new data will be added to the "newly added data row". Use the time to tell the track record at certain point. This method allows unlimited times of tracking the changes over data. The deficiency is that it must use time to identify the updated data row and increase the data rows of the dimension table. Use additionally built record column plus the time column to record the changes in time. The good point of it is that there is no need to build additional data row or to change the values architecture of dimension table.

4. The design of aggregation:

Aggregation is the advanced calculated total amount to increase the analysis speed when facing complicated enquiries.

VII. Conclusions

Businesses that use ERP systems can benefit from business intelligence OLAP and data mining (DM) approaches that can apply to ERP's data in order to generate reports, charts and identify new knowledge to support decision makers. OLAP and data mining can perform different tasks on ERP's data, but integrating both approaches with the knowledge discovery is very useful to perform new tasks that may be required by businesses decision makers and ERP users. Integration can help to increase customer's satisfaction, behaviour and ultimately the growth of the organization and provide help for dealing with the customers in more efficient manner.

References

[1] *Abdullah Saad Al-Malaise, "Integration of Automated Decision Support Systems with Data Mining Abstract: A Client Perspective", International Journal of Advanced Computer Science and Applications, Vol.4, No.2, pp. 173-176, 2013.*
[2] *Abdullah Saad Almalaise Alghamdi, " Rules Generation from ERP Datanase" A Successful Implementation of Data Mining", International Journal of Computer Science and Network Security, Vol.12, No.3, pp. 21-29, March 2012.*
[3] *A. Abdullah, Zahid Ullah, "A Framework of an Automated Data Mining Systems Using ERP Model", International Journal of Computer and Electrical Engineering, Vol.1, No. 5, pp. 651-655, 2009.*
[4] *Mohammed K. Kolkas, "Integrated Data Mining Techniques in Enterprise Resource Planning (ERP) Systems", International Journal of Information Science and Intelligent System, Vol. 2, Issue 2, pp. 131-152, 2014.*
[5] *Pang-Ning Tan, Michael Steinbach & Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2005.*
[6] *S.C. Hui, G. Jha, "Data Mining for Customer Service Support", Information & Management, Elsevier 2000.*
[7] *Tamer S. Abdellatif, "Comparing online analytical processing and data mining tasks in enterprise resource planning systems", International Journal of Computer Science Issues, Vol. 8, Issue 6(2), pp. 161-174 , Novemebt 2011.*
[8] *Ruey-Shun Chen, " A Web-based Data Mining System for ERP Decision Making", IEEE SMC, 2002.*
[9] *Sonja Grabner-Kraeuter, Gernot Moedritscher, Martin Waigunyc, Werner Mussnigb, "Performance Monitoring of CRM Initiatives", in Proceedings: IEEE conference on System Sciences, 2007.*
[10] *Sistla Hanumanth Sastry, Prof. M. S. Prasada Babu, " Analysis of Enterprise Material Procurement Leadtime using Techniques of Data Mining.", International Journal of Advanced Research in Computer Science (IJARCS), Vol. 4, Issue 4, pp. 288-301, April 2013.*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

[11] Tiruveedula Gopi Krishna , "Advanced Data Mining Tools in Web Based ERP, ASP Environment". International Journal of Computer Science and Information Technologies, Vol. 5, No. 1, 223 – 226, 2014.