Journal of Advances in Science and Technology

Vol. IV, No. VIII, February-2013, ISSN 2230-9659

DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASES

www.ignited.in

Journal of Advances in Science and Technology
Vol. IV, No. VIII, February-2013, ISSN 2230-9659

Data Mining and Knowledge Discovery in Databases

Rishipal Bangarh

Astt. Prof., DAV College, Pehowa –Kurukshetra (India), Pin. No. -136128

Abstract – There are many approaches for data cleaning. Some of them are: parsing[3], data transformation, duplicate elimination and statistical method. A large variety of tools is available in the market to support data transformation and data cleaning tasks, in particular for data warehousing. Some tools concentrate on a specific domain, such as cleaning name and address data, or a specific cleaning phase, such as data analysis or duplicate elimination. Due to their restricted domain, specialized tools typically perform very well but must be complemented by other tools to address the broad spectrum of transformation and cleaning problems.

----------♦----------

INTRODUCTION

Knowledge discovery process transforms data into knowledge [Cios K.J., 2000]. The key issue in KDD is to realize that there is more information hidden in the data than what could be made out at first sight. Before one attempts to extract useful knowledge from data, it is important to understand the overall approach. The process of knowledge discovery in databases is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process generalizes to non-database sources of data, although it emphasizes databases as a primary source of data. It consists of many steps (one of them is DM), each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyses massive datasets, how to interpret and visualize the results, and how to model and support the interaction between humans and machines. It also concerns support for learning and analyzing the application domain.

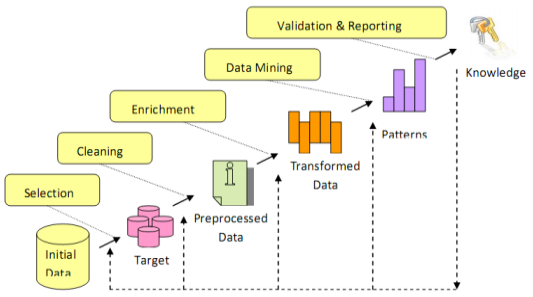

Figure : The process of knowledge discovery.

The starting point for any data mining activity is the formation of information requirement related to a specific action, that is, what do you want to know and what do you want to do with this knowledge? In optimal situation, data mining is an ongoing process. Organizations should continually work on their data, constantly identifying new information needs and trying to improve the data to make it match the goals better. In this way, any organization will become a learning system [Adriaans P. et.al., 2003]. However, the data mining begins with the requirement for knowledge to start with. Given below are the steps of the knowledge discovery process that are to be performed in order to search novel and useful patterns in the organization's data-store.

DATA SELECTION

After the formation of information requirements, the first logical step is to collect and select the target data. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain some patterns while remaining concise enough to be mined in an acceptable timeframe [Singh P.K., 2009]. Selection is the process of selecting the right data from the data stored in operational databases, data files and data warehouses. Gathering this information in a large organization is not an easy task since it may involve low level conversion of data such as from flat files to relational table or from hierarchical system to relational system. If the organization has built up a data warehouse[1], then it presents a stable and reliable environment to collect the relevant data for knowledge discovery. A common source for data is a data-mart[2] or data warehouse.

DATA CLEANING

Before data mining algorithms can be used, a target data set must be cleaned and transformed into uni-format data. Data cleaning is very important step of knowledge discovery as many discrepancies may exist in the selected data that may affect the final results. The old saying "garbage-in-garbage-out" is particularly applicable to the typical data mining projects where large data sets collected via some automatic methods (e.g., via the Web) serve as the input into the data analyses [Adriaans P. et.al., 1999]. Often, the method by which the data is gathered, are not tightly controlled, and so the data may contain out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), and the like. It is therefore a good idea to put in some efforts in examining the data in order to clean the polluted data. Data cleaning deals with eliminating incorrect, invalid or unknown data, depending on data quality and the algorithms to be used. Data quality problems are present in data collections, such as files and databases due to misspellings during data entry, missing information or other invalid data. High quality data need to pass a set of quality criteria such as accuracy, integrity, validity, completeness, consistency, uniformity, density and uniqueness etc.

There are many approaches for data cleaning. Some of them are: parsing[3], data transformation[4], duplicate elimination[5] and statistical method[6]. A large variety of tools is available in the market to support data transformation and data cleaning tasks, in particular for data warehousing. Some tools concentrate on a specific domain, such as cleaning name and address data, or a specific cleaning phase, such as data analysis or duplicate elimination. Due to their restricted domain, specialized tools typically perform very well but must be complemented by other tools to address the broad spectrum of transformation and cleaning problems [Rahm E. et.al., 2000].

DATA ENRICHMENT

Data enrichment is done so that the analysis can be performed more rapidly and it may give more accurate knowledge from the data. At this stage some new attributes may be added in the existing tables to make it more detailed regarding the facts. Extra information can be purchased to add the items describing the individuals to enrich the database. Matching the information from bought-in databases with the original database can be difficult because family relationship has to be established in the database.

Sometimes data reduction is also applied to projects where the goal is to aggregate or amalgamate the information contained in large datasets into manageable (smaller) information nuggets. Data reduction methods can include simple tabulation, aggregation or more sophisticated techniques like clustering, principal component analysis, etc. Redefinition of variables for the purpose of reducing the complexity or the range of values (e.g., by rounding, clustering, binarisation, etc.) may also be used for conditioning the data for mining. The data may undergo a lot of transformations to obtain a lean target dataset.

DATA MINING

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. [Adriaans P. et.al., 1999]. The discovery process is a combination of human involvement and autonomous methods of discovery. Autonomous methods may include automated task integration, for instance, integration of variable selection, knowledge mining and result optimization. One might wonder that there is a relationship between different types of problems and certain machine learning techniques. For instance, neural networks are somewhat better at classification tasks whereas genetic algorithms perform better at problem solving tasks. Inductive logic programming has a high score at problem solving tasks. A KDD environment therefore must support these different types of techniques to encompass different areas of knowledge engineering, classification and problem solving [Adriaans P. et.al., 2003]. Through data mining techniques, a knowledge model is obtained representing behaviour patterns in relevant problem variables or relations between them. Several algorithms are frequently tested generating different models e.g. IDT (Induction of Decision Trees), Neural Nets, Genetic Algorithms, Fuzzy Techniques (fuzzy logic, fuzzy sets, etc.), Rule induction, VSM (Vector Support Machines), Bayesian Networks etc. In section 1.5 data mining techniques are discussed in detail.

VALIDATION AND REPORTING

The objective of this phase is to evaluate the model with respect to problem solving perspective. The step of validation and reporting basically combines two different functions:

- Analysis of the model that describes the behaviour of data.

- Application of the results of the data mining model to the problem in hand.

Modeling is the act of building a model[7] in one situation where you know the answer and then applying it to another situation that you don't. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Before a model could be used to explain the behaviour of data in similar situations, it needs to be validated. In case several models were developed through separate algorithms/methods, they must be compared based on performance and/or error rates. Statistical and experimental approaches (e.g.

significance tests, cross validation, etc.) are used for this phase.

Reporting the results of data mining can take many forms. In many cases, reporting can be done using traditional database query tools; however, several new data visualization techniques are emerging, ranging from simple scatter diagrams[8] to complex interactive environments that enable us to fly over landscapes containing information about data sets [Fayyad U. M. et.al.,1996]. Herein the domain user may want to view the data mining model using different visualization techniques from different angles before they can be applied to real data. It is advisable to test models on real problems. If the results fulfill the problem objectives, the project can move to its conclusive phase i.e. deployment, otherwise more iteration involving data preparation, modeling and evaluation with changed parameters has to be done.

REFRANCES

1. Data warehouse is a repository of an organization's electronically stored data, designed to facilitate reporting and analysis. Data warehousing arises with an organization's need for reliable, consolidated, unique and integrated reporting and analysis of its data, at different levels of aggregation.

2. Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization. Data marts are often derived from subsets of data in a data warehouse.

3. Parsing is a technique that is performed for the detection of syntax errors.

4. Data transformation is the mapping of the data from their given format into the format expected by the appropriate application.

5. Duplicate elimination algorithms determine whether data contains duplicate representations of the same entity.

6. Statistical methods helps in analyzing the data using the values of mean,