# Data Mining Architecture – Data Mining Types and Techniques

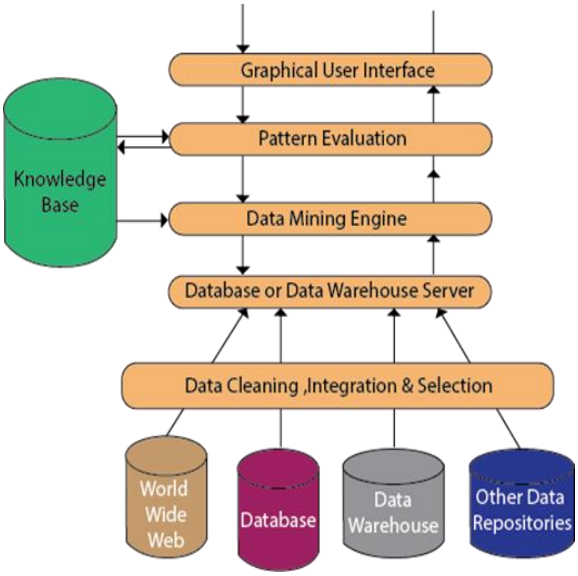
Dr. Akhilesh Saini
*Associate Professor (Computer Science), Ch. K.R.Godara Memorial College, Bashir, Tibbi, India*

***Abstract*** **- Data mining architecture is for memory-based data mining system. We can say it is a process of extracting interesting knowledge from large amounts of data. That is stored in many data sources. Such as file systems, databases, data warehouses. Also, knowledge used to contributes a lot of benefits to business and individual. Data mining offers tools for the discovery of relationship, patterns and knowledge from a massive database in order to guide decisions about future activities. Applications from various domains have adopted this technique to perform data analysis efficiently. Several issues need to be addressed when such techniques apply on data these are bulk at size and geographically distributed at various sites. In this paper we describe system architecture for a scalable and a portable distributed data mining application. The system contains modules for secure distributed communication, database connectivity, organized data management and efficient data analysis for generating a global mining model. Performance evaluation of the system is also carried out and presented. Certified Data Mining and Warehousing. Data and architecture design. Data architecture in Information Technology is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations.**

## INTRODUCTION

In this architecture, data mining system uses a database for data retrieval. In loose coupling, data mining architecture, data mining system retrieves data from a database. And it stores the result in those systems.

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.




Types of Data Mining Architecture

Data is collected through business transactions and stored in relational database systems. Also, these business processes have been built to provide analytical reports. That is for business users to make decisions. As also data is now stored in database or data warehouse system? So data mining system should be designed to decouple.

This question leads to four possible architectures:

a. No-coupling Data Mining

In this architecture, data mining system does not use any functionality of a database. A no-coupling data mining system retrieves data from a particular data sources.

The no-coupling data mining architecture does not take any advantages of a database. That is already very efficient in organizing, storing, accessing and retrieving data. The no-coupling architecture is considered a poor architecture for data mining system. But it is used for simple data mining processes.

b. Loose Coupling Data Mining
In this architecture, data mining system uses a database for data retrieval. In loose coupling, data mining architecture, data mining system retrieves data from a database. And it stores the result in those systems.
Data mining architecture is for memory-based data mining system. That does not must high scalability and high performance.

c. Semi-Tight Coupling Data Mining
In semi-tight coupling, data mining system uses several features of data warehouse systems. That is to perform some data mining tasks. That includes sorting, indexing, aggregation. In this, some intermediate result can be stored in a database for better performance.

d. Tight Coupling Data Mining
In tight coupling, a data warehouse is treated as an information retrieval component. All the features of database or data warehouse are used to perform data mining tasks. This architecture provides system scalability, high performance, and integrated information.

There are three tiers in the tight-coupling data mining architecture:

i. Data Layer
We can define data layer as a database or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in the data layer. Thus, we can present to end-user in form of reports or another kind of visualization.

ii. Data mining application layer
It is to retrieve data from a database. Some transformation routine has to perform here. That is to transform data into the desired format. Then we have to process data using various data mining algorithms.

iii. Front-end layer
It provides the intuitive and friendly user interface for end-user. That is to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

Data Mining Techniques

There are several data mining techniques present, mentioned below:-

a. Decision Trees:-
It's the most common technique, we use for data mining. As because of its simplest structure. The root of decision tree act as a condition. Each answer leads to specific data that help us to determine final decision based upon it.

b. Sequential Patterns:-
As we use this to discover regular events, similar patterns in transaction data. The historical data of customers helps us to identify the past transactions in a year.

c. Clustering:-
Having similar characteristics clusters objects have to form, by using automatic method. We use clustering, to define classes. Then suitable objects have to place in each class.

d. Prediction:-
We use this method defines the relationship between independent and dependent instances.

e. Association:-
It is also known as relation technique. Also, in this, we have to recognize a pattern. That it is based upon the relationship of items in a single transaction.
Also, we can suggest the technique for market basket analysis. That is to explore the products that customer frequently demands.

f. Classification:-
This is based on machine learning. We use this to classify each item in a particular set into predefined groups. Although, this method adopts mathematical techniques. Such as neural networks, linear programming, and decision trees and so on.

Required Technological Drivers
As data mining applications are present for all size machines. Such as mainframe, workstations, clouds, client, and server. The size of enterprise applications varies from 10 Gb to 100 Tb. NCR systems are preferring for deliver the applications exceeding 100

Tb. The technological drivers are as:-

a. Database size
As for maintaining and processing the huge amount of data, we need powerful systems.
b. Query Complexity

To analyze the complex and large number of queries, we need a more powerful system.
Data Source:-
The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes:-
Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server:-
The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:-
The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.
In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:-
The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.
This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

Graphical User Interface:-
The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

Knowledge Base:-
The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.

Data Mining Process:-
Data Mining is a process of discovering various models, summaries, and derived values from a given collection of data. The general experimental procedure adapted to data-mining problems involves the following steps:

State the problem and formulate the hypothesis :-
Most data-based modeling studies are performed in a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies tend to focus on the data-mining technique at the expense of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypotheses formulated for a single problem at this stage. The first step requires the combined expertise of an application domain and a data-mining model. In practice, it usually means a close interaction between the data-mining expert and the application expert. In successful data-mining applications, this cooperation does not stop in the initial phase; it continues during the entire data-mining process.
1.Collect the data:-
This step is concerned with how the data are generated and collected. In general, there are two distinct possibilities. The first is when the data-generation process is under the control of an expert (modeler): this approach is known as a designed experiment. The second possibility is when the expert cannot influence the data- generation process: this is known as the observational approach. An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, the sampling distribution is completely unknown after data are collected, or it is partially and implicitly given in the data-collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. Also, it is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. If this is not the case, the estimated model cannot be successfully used in a final application of the results.
Preprocessing the data:-
In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts. Data preprocessing usually includes at least two common tasks:

1. Outlier detection (and removal) – Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such no representative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:
a. Detect and eventually remove outliers as a part of the preprocessing phase
b. Develop robust modeling methods that are insensitive to outliers.

2. Scaling, encoding, and selecting features : – Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [−100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently. Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling. These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a data-mining process. Data-preprocessing steps should not be considered completely independent from other data-mining phases. In every iteration of the data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a priori knowledge in the form of application-specific scaling and encoding.

4. Estimate the model :-
The selection and implementation of the appropriate data-mining technique is the main task in this phase. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task. The basic principles of learning and discovery and analyze specific techniques that are applied to perform a successful learning process from data and to develop an appropriate model.

5. Interpret the model and draw conclusions :-
In most cases, data-mining models should help in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex "black-box" models. Note that the goals of accuracy of the model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models. The problem of interpreting these models, also very important, is considered a separate task, with specific techniques to validate the results. A user does not want hundreds of pages of numeric results. He does not understand them , he cannot summarize, interpret, and use them for successful decision making.

Classification of Data mining Systems: -
The data mining system can be classified according to the following criteria: Database Technology Statistics Machine Learning Information Science Visualization Other Disciplines

Some Other Classification Criteria: -
1. Classification according to kind of databases mined
2. Classification according to kind of knowledge mined
3. Classification according to kinds of techniques utilized Classification according to applications adapted

Classification according to kind of databases mined :-
We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object-relational, or data warehouse mining system.

Classification according to kind of knowledge mined:-
We can classify the data mining system according to kind of knowledge mined. It is means data mining system are classified on the basis of functionalities such as:

- Characterization
- Discrimination Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Classification according to kinds of techniques utilized:-
We can classify the data mining system according to kind of techniques used. We can describes these techniques according to degree of user interaction involved or the methods of analysis employed.

Classification according to applications adapted:-
We can classify the data mining system according to application adapted. These applications are as follows:

- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail

Major Issues In Data Mining: -
Mining different kinds of knowledge in databases The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover broad range of knowledge discovery task.

Interactive mining of knowledge at multiple levels of abstraction:- The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

Incorporation of background knowledge: - To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple level of abstraction.

Data mining query languages and ad hoc data mining: - Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

Presentation and visualization of data mining results: - Once the patterns are discovered it needs to be expressed in high level languages, visual representations. This representations should be easily understandable by the users.

Handling noisy or incomplete data : - The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

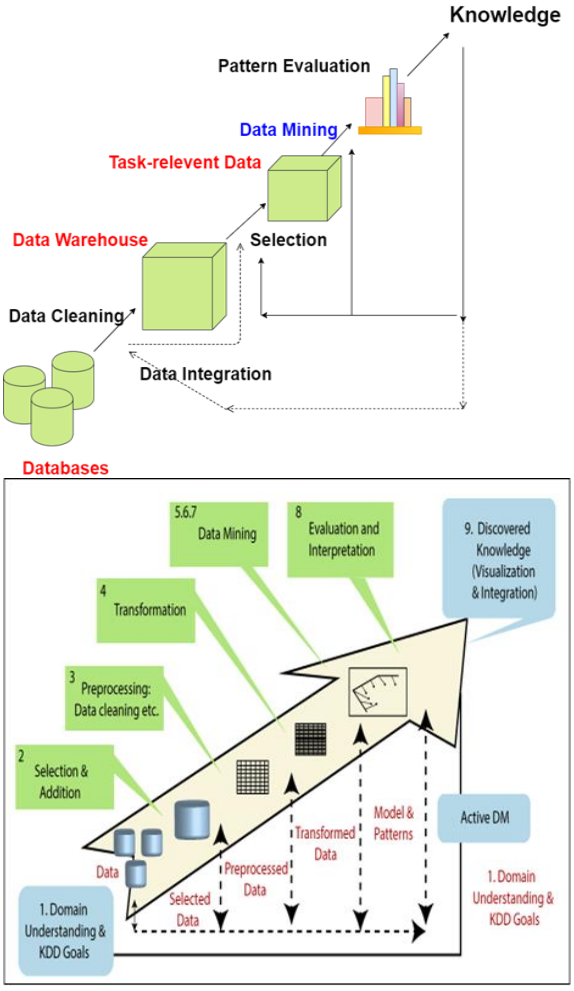
Pattern evaluation: - It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Efficiency and scalability of data mining algorithms.
In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

Parallel, distributed, and incremental mining algorithms: –
The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithm divide the data into partitions which is further processed parallel. Then the results from the partitions is merged. The incremental algorithms, updates databases without having mine the data again from scratch.

KDD- Knowledge Discovery in Databases
The term KDD stands for Knowledge Discovery in Databases. It refers to the broad procedure of discovering knowledge in data and emphasizes the high-level applications of specific Data Mining techniques. It is a field of interest to researchers in various fields, including artificial intelligence, machine learning, pattern recognition, databases, statistics, knowledge acquisition for expert systems, and data visualization.
The main objective of the KDD process is to extract information from data in the context of large databases. It does this by using Data Mining algorithms to identify what is deemed knowledge.
The Knowledge Discovery in Databases is considered as a programmed, exploratory analysis and modeling of vast data repositories. KDD is the organized procedure of recognizing valid, useful, and understandable patterns from huge and complex data sets. Data Mining is the root of the KDD procedure, including the inferring of algorithms that investigate the data, develop the model, and find previously unknown patterns. The model is used for extracting the knowledge from the data, analyze the data, and predict the data.
The availability and abundance of data today make knowledge discovery and Data Mining a matter of impressive significance and need. In the recent development of the field, it isn't surprising that a wide variety of techniques is presently accessible to specialists and experts.

The KDD Process:-
The knowledge discovery process(illustrates in the given figure) is iterative and interactive, comprises of nine steps. The process is iterative at each stage, implying that moving back to the previous actions might be required. The process has many imaginative aspects in the sense that one cant presents one formula or make a complete scientific categorization for the correct decisions for each step and application type. Thus, it is needed to understand the process and the different requirements and possibilities in each stage.
The process begins with determining the KDD objectives and ends with the implementation of the discovered knowledge. At that point, the loop is closed, and the Active Data Mining starts. Subsequently, changes would need to be made in the application domain. For example, offering various features to cell phone users in order to reduce churn. This closes the loop, and the impacts are then measured on the new data repositories, and the KDD process again. Following is a concise description of the nine-step KDD process, Beginning with a managerial step:




1. Building up an understanding of the application domain:-

This is the initial preliminary step. It develops the scene for understanding what should be done with the various decisions like transformation, algorithms, representation, etc. The individuals who are in charge of a KDD venture need to understand and characterize the objectives of the end-user and the environment in which the knowledge discovery process will occur (involves relevant prior knowledge).

2. Choosing and creating a data set on which discovery will be performed:-

Once defined the objectives, the data that will be utilized for the knowledge discovery process should be determined. This incorporates discovering what data is accessible, obtaining important data, and afterward integrating all the data for knowledge discovery onto one set involves the qualities that will be considered for the process. This process is important because of Data Mining learns and discovers from the accessible data. This is the evidence base for building the models. If some significant attributes are missing, at that point, then the entire study may be unsuccessful from this respect, the more attributes are considered. On the other hand, to organize, collect, and operate advanced data repositories is expensive, and there is an arrangement with the opportunity for best understanding the phenomena. This arrangement refers to an aspect where the interactive and iterative aspect of the KDD is taking place. This begins with the best available data sets and later expands and observes the impact in terms of knowledge discovery and modeling.

3. Preprocessing and cleansing:- In this step, data reliability is improved. It incorporates data clearing, for example, Handling the missing quantities and removal of noise or outliers. It might include complex statistical techniques or use a Data Mining algorithm in this context. For example, when one suspects that a specific attribute of lacking reliability or has many missing data, at this point, this attribute could turn into the objective of the Data Mining supervised algorithm. A prediction model for these attributes will be created, and after that, missing data can be predicted. The expansion to which one pays attention to this level relies upon numerous factors. Regardless, studying the aspects is significant and regularly revealing by itself, to enterprise data frameworks.

4. Data Transformation:-

In this stage, the creation of appropriate data for Data Mining is prepared and developed. Techniques here incorporate dimension reduction( for example, feature selection and extraction and record sampling), also attribute transformation(for example, discretization of numerical attributes and functional transformation). This step can be essential for the success of the entire KDD project, and it is typically very project-specific. For example, in medical assessments, the quotient of attributes may often be the most significant factor and not each one by itself. In business, we may need to think about impacts beyond our control as well as efforts and transient

issues. For example, studying the impact of advertising accumulation. However, if we do not utilize the right transformation at the starting, then we may acquire an amazing effect that insights to us about the transformation required in the next iteration. Thus, the KDD process follows upon itself and prompts an understanding of the transformation required.

5. Prediction and description:-
We are now prepared to decide on which kind of Data Mining to use, for example, classification, regression, clustering, etc. This mainly relies on the KDD objectives, and also on the previous steps. There are two significant objectives in Data Mining, the first one is a prediction, and the second one is the description. Prediction is usually referred to as supervised Data Mining, while descriptive Data Mining incorporates the unsupervised and visualization aspects of Data Mining. Most Data Mining techniques depend on inductive learning, where a model is built explicitly or implicitly by generalizing from an adequate number of preparing models. The fundamental assumption of the inductive approach is that the prepared model applies to future cases. The technique also takes into account the level of meta-learning for the specific set of accessible data.

6. Selecting the Data Mining algorithm:-
Having the technique, we now decide on the strategies. This stage incorporates choosing a particular technique to be used for searching patterns that include multiple inducers. For example, considering precision versus understandability, the previous is better with neural networks, while the latter is better with decision trees. For each system of meta-learning, there are several possibilities of how it can be succeeded. Meta-learning focuses on clarifying what causes a Data Mining algorithm to be fruitful or not in a specific issue. Thus, this methodology attempts to understand the situation under which a Data Mining algorithm is most suitable. Each algorithm has parameters and strategies of leaning, such as ten folds cross-validation or another division for training and testing.

7. Utilizing the Data Mining algorithm:-
At last, the implementation of the Data Mining algorithm is reached. In this stage, we may need to utilize the algorithm several times until a satisfying outcome is obtained. For example, by turning the algorithms control parameters, such as the minimum number of instances in a single leaf of a decision tree.

8. Evaluation:-
In this step, we assess and interpret the mined patterns, rules, and reliability to the objective characterized in the first step. Here we consider the preprocessing steps as for their impact on the Data Mining algorithm results. For example, including a feature in step 4, and repeat from there. This step focuses on the comprehensibility and utility of the induced model. In this step, the identified knowledge is also recorded for further use. The last step is the use, and overall feedback and discovery results acquire by Data Mining.

9. Using the discovered knowledge:-
Now, we are prepared to include the knowledge into another system for further activity. The knowledge becomes effective in the sense that we may make changes to the system and measure the impacts. The accomplishment of this step decides the effectiveness of the whole KDD process. There are numerous challenges in this step, such as losing the "laboratory conditions" under which we have worked. For example, the knowledge was discovered from a certain static depiction, it is usually a set of data, but now the data becomes dynamic. Data structures may change certain quantities that become unavailable, and the data domain might be modified, such as an attribute that may have a value that was not expected previously.

## CONCLUSION

Data mining is a very powerful and useful methodology and technology for generating information for decision making. Future developments are expected to make data mining even more powerful and useful. Despite this, data mining is not without limitations. Before highlighting the limitations of data mining and discussing some future directions.

## REFERENCES

[1] Anonymous. (1999). “Data mining”, Chain Store Age, October, p. 42.
[2] Anonymous. (2002). “Data mining digs deep to improve on quality”, Professional Engineering, Vol. 15 No. 11, p. 45.
[3] Anonymous. (2004). “‘Top 10 technologies’ confirms interest in information security, spam control”, The CPA Journal, Vol. 74 No. 4, p. 15.
[4] Baker, S. and Baker, K. (1998). “Mine over matter”, The Journal of Business Strategy , Vol. 19 No. 4, pp. 22-26. Biggs, M. (2000). “Resurgent text-mining technology can greatly increase your firm’s ‘intelligence’ factor”, InfoWorld, Vol. 22 No. 2, p. 52.
[5] Conclusion 243 Bloemer, J., Brijs, T., Swinnen, G. and Vanhoof, K. (2002), “Identifying latently dissatisfied customers and measures for dissatisfaction management”, The International Journal of Bank Marketing, Vol. 20 No. 1, pp. 27-27.
[6] Chopoorian, J. A., Witherell, R., Khalil, O. E. M. and Ahmed, M. (2001). “Mind your own business by mining your data”, SAM Advanced Management Journal, Vol. 66 No. 2, pp. 45-51. Cody, W. F., Kreulen, J. T., Krishna, V. and Spangler, W. S. (2002). “The integration of business intelligence and knowledge management”, IBM Systems Journal, Vol. 41 No. 4, pp. 697-713.
[7] Deck, S. (1999). “Data mining”, Computerworld, Vol. 33 No. 13, p. 76. Fayyad, U. and Uthurusamy, R. (2002). “Evolving data mining into solutions for insight”, Association for Computing Machinery, Vol. 45 No. 8, pp. 28-31.
[8] Gertosio, C. and Dussauchoy, A. (2004), “Knowledge discovery from industrial databases”, Journal of Intelligent Manufacturing, Vol. 15 No. 1, pp. 29-37. Gillespie, G., (2000). “There’s gold in them thar’ databases”, Health Data Management, Vol. 8 No. 11, pp. 40-52.