**Q.2a. Explain the architecture of data mining system?**

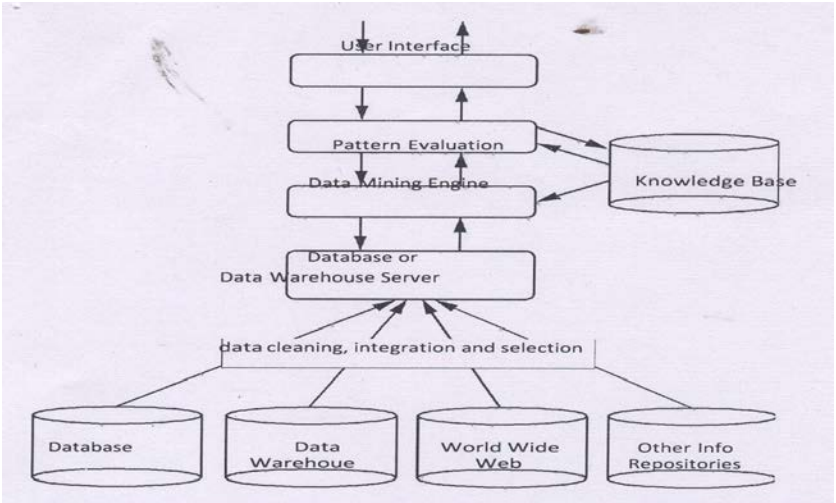
**Ans 2 (a):** The architecture of a typical data mining system may have the following major components:

**Database, data warehouse, World Wide Web, or other information repository:** This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

**Database or data warehouse server:** The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request

**Knowledge base:** This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

**Data mining engine:** This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.




**Pattern evaluation module:** This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

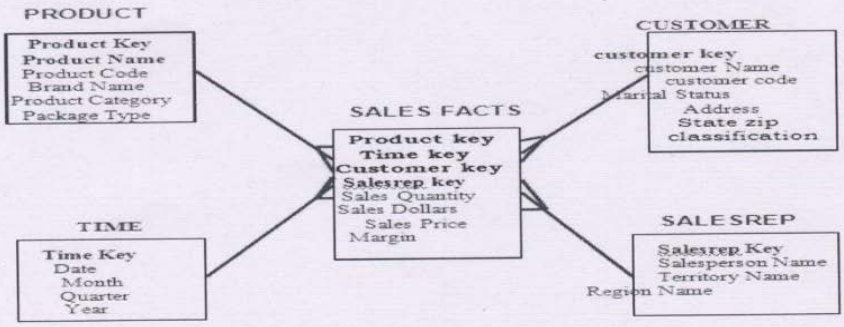
**User interface:** This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

**b.Why are decision tree classifiers so popular? Explain.**

**Ans 2(b):** The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule induction systems.
During tree construction, *attribute selection measures* are used to select the attribute that best partitions the tuples into distinct classes. When decision trees are built, many of the branches may reflect noise or outliers in the training data. *Tree pruning* attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

**Q.3a. Explain the snowflake schema with the help of examples.**

**Q3(a) "Snowflaking"** is a method of normalizing the dimension tables in a STAR schema. When you completely normalize all the dimension tables, the resultant structure resembles a snowflake with the fact table in the middle.
For example
The sales fact table contains quantity, price, and other relevant metrics. Sales rep, cus- tomer, product, and time are the dimension tables. This is a classic STAR schema, denor- malized for optimal query access involving all or most of the dimensions. The model is not in the third normal form.




Assume that there are 500,000 product dimension rows. These products fall under 500 product brands and these product brands fall under 10 product categories. Now suppose one of your users runs a query constraining just on product category. If the product dimension table is not indexed on product category, the query will have to search through 500,000 rows.
"Snowflaking" or normalization of the dimension tables can be achieved in a few different ways.
The following options indicate the different ways you may want to consider for normalization of the dimension tables:

- Partially normalize only a few dimension tables, leaving the others intact
- Partially or fully normalize only a few dimension tables, leaving the rest intact
- Partially normalize every dimension table
- Fully normalize every dimension table

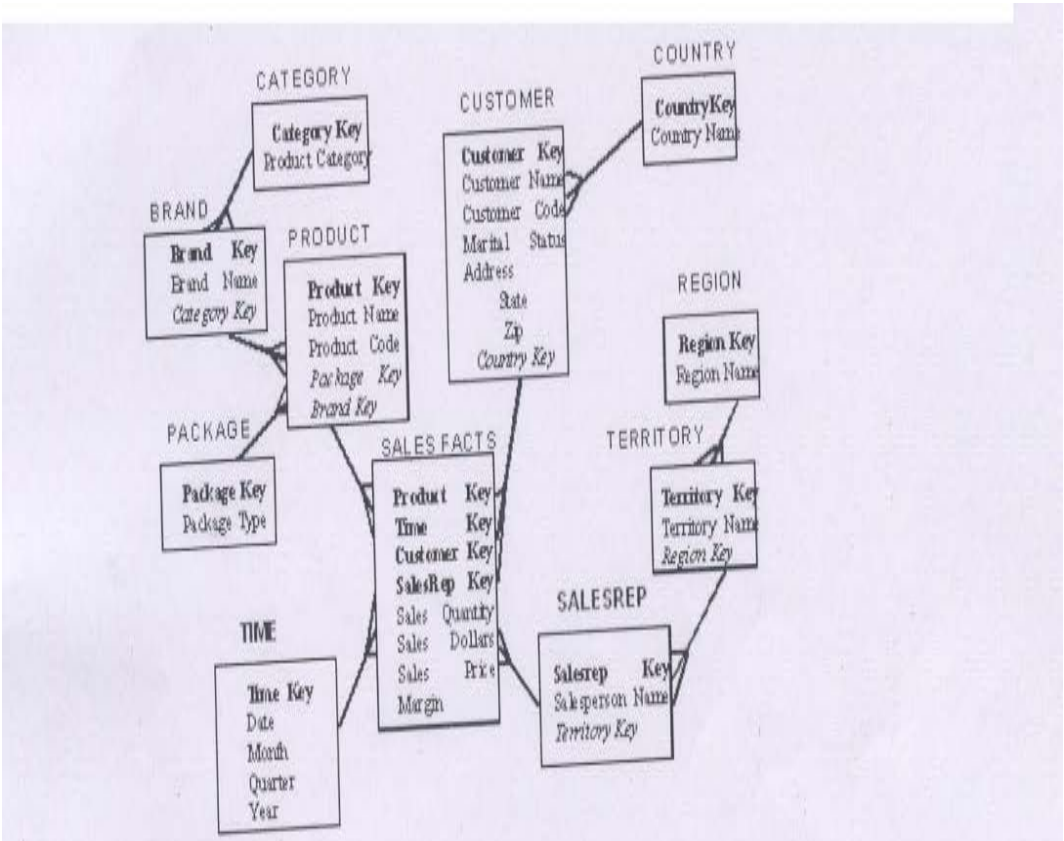



Figure shows the version of the snowflake schema for sales in which every dimension table is partially or fully normalized

The original STAR schema for sales as shown in Figure contains only five tables, whereas the normalized version now extends to eleven tables. In the snowflake schema, the attributes with low cardinality in each original dimension table are removed to form separate tables. These new tables are linked back to the original dimension table through artificial keys.

**b.What are various steps of data preprocessing?**

**Ans 3 (b):** Data preprocessing is an important issue for both data warehousing and data mining, as real-world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation, and data reduction.

**Data cleaning** routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Data cleaning is usually performed as an iterative two-step process consisting of discrepancy detection and data transformation.

**Data integration** combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute toward smooth data integration.

**Data transformation** routines convert the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between small ranges, such as 0.0 to 1.0.

**Data reduction** techniques such as data cube aggregation attribute subset selection, dimensionality reduction, numerosity reduction, and discretization can be used to obtain a reduced representation of the data while minimizing the loss of information content.

**Q.4 a. Explain how the data warehouse acts as a basis for EIS.**

**Ans 4 (a):** It is in the EIS environment that the data warehouse operates in its most effective state. The data warehouse is tailor-made for the needs of the EIS analyst. Once the data warehouse has been built, the job of the EIS is infinitely easier than when there is no foundation of data on which the EIS analyst can operate. The figure shows how the data warehouse supports the need for EIS data.

With a data warehouse, the EIS analyst does not have to worry about the following:

1. Searching for the definitive source of data.
2. Creating special extract programs from existing systems.
3. Dealing with unintegrated data.
4. Compiling and linking detailed and summary data and the linkage between the two.
5. Finding an appropriate time basis of data (finding historical data).
6. Management constantly changing its mind about what needs to be looked at next

In addition, the EIS analyst has a rich supply of summary data available

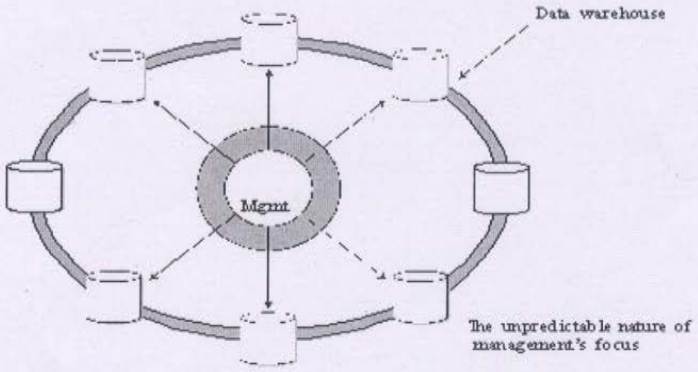



**Figure** The data warehouse supports management's need for EIS data.

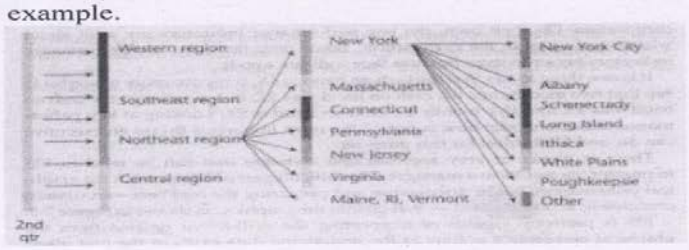
The important reason why the data warehouse serves the needs of the world of EIS is this: The data warehouse operates at a low level of granularity. The data warehouse contains — for lack of a better word — *atomic data*. The atomic data can be shaped one-way, and then another. When management has a new set of needs for information that has never before been encountered in the corporation, the very detailed data found in the data warehouse sits, waiting to be shaped in a manner suited to management's needs. Because of the granular atomic data that resides in the data warehouse, analysis is flexible and responsive. The detailed data in the data warehouse sits and waits for future unknown needs for information. This is why the data warehouse turns an organization from a reactive stance to a proactive stance

**b. Explain drill-down analysis and event mapping in context of EIS.**

**Ans 4(b): Dill down analysis**

Dilling down refers to the ability to start at a summary number and to break that summary into a successfully finer set of summarizations. By being able to get at the detail beneath a summary number, the manager can get a feel for what is happening, especially where the summary number is surprising.

example.




The manager has seen second quarter summary results and wants to explore them further. The manager then looks at the regions that have contributed to the summary analysis. The figures analyzed are those of the Western region, the southern region , the northern region and the the central region.

In looking at the numbers of each regions , the manager decides to look more closely at the Northeast region's numbers. The Northeast's numbers are made up of totals from New York, Massachusetts, Connectricts and many more. Of the states , the manager then decided to look more closely at the number for New York.

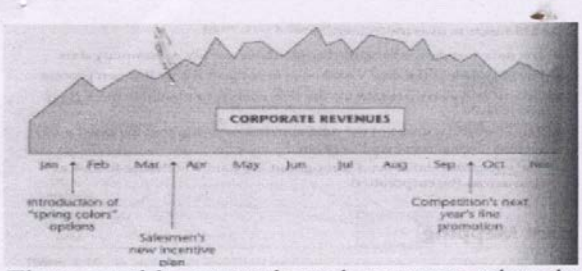
In each case , the manager has selected a path going from summary to detail, and then successively lower level. In such a fashion the manager can determine where the troublesome results are.

EIS is perfectly capable of supporting the drill down process from the graphical perspective as long as the underlying data exists in the first place. If the data to analyze does not exist, the drill down process becomes very tedious and awkward.

**Event mapping**

A useful technique in using data warehouse for EIS processing is event mapping. The simplest way to depict event mapping is to start with a simple trend line.


Shows that corporate revenues have varied by months, as expected. The trend has been calculated from data found in the data warehouse. The trend of revenue s in and of itself is interesting but gives only superficial view of what is going on with the corporation. To enhance the view, events are mapped onto the trend line.




Three notable events have been mapped to the corporate revenue trend line- the introduction of a " spring colors" line of products, the advent of a sales incentives program and the introduction of competition. Now the relationship between corporate revenues and significant events begins to take on a different perspective.

For some sorts of events, event mapping is the only way to measure the results. Some events and activities cannot be measured directly and have to be measured in a correlative fashion. Cost justification and actual cost benefit cannot be measured any other way for some types of events.

**Q.5a. Discuss the problem related to use and storage of unstructured data in data warehouse? What are the methods to capture and store external information?**

**Ans 5(a):** Several issues relate to the use and storage of external data in the data ware- house.

**1. Frequency of availability**

Unlike internally appearing data, there is no real fixed pattern of appearance for external data. This irregularity is a problem because constant monitoring must be set up to ensure that the right external data is captured. For some environments (such as the Internet), monitoring programs can be created and used to build automated alerts and automated loads.

**2. Totally undisciplined**

To be useful, and for placement in the warehouse, a certain amount of reformatting and structuring of external data is needed to transform it into an internally acceptable and usable form in the data warehouse. A common practice is to convert the external data as it enters the data warehouse environment. External key data is converted to internal key data. Or, external data is passed through simple edits, such as a domain check. In addition, the data is often restructured so that it is compatible with internal data.

In some cases, the level of granularity of the external data will not match that of the internal systems of the corporation.

3. **Unpredictability**

External data may come from practically any source at almost any time.

There are many **methods** to capture and store external information. One of the best places to locate external data if it is voluminous is on a bulk storage medium such as near-line storage. With near-line storage, the external data is still accessible, but it doesn't cost huge amounts of money to store. Of course, extensive indexing of the external data can be done, and those indexes can be stored in both disk storage and near-line storage. In such a manner, many requests about the external data can be managed without actually having to go to the external data. In addition, some requests can be handled entirely inside the index of the external data itself. Also, if an extensive index of external data is created, the external data can be tied to structured data and the data ware- house. The index can then be used to determine what external data to bring to disk storage. In this case, only external data that is prequalified and preselected would be brought to disk storage.

Another technique for handling external data that is sometimes effective is to create two stores of external data. One store contains all of the external data, and another, much smaller

**b. Explain multiway array aggregation scheme for full cube computation.**

Ans Page 164 of Text book.

**Q.6 a. Discuss the criteria used to compare and evaluation of the classification and prediction method.**

**Ans 6 (a):** Classification and prediction methods can be compared and evaluated according to the following criteria:

i. **Accuracy:** The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data (i.e., tuples without class label information). Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

ii. **Speed:** This refers to the computational costs involved in generating and using the given classifier or predictor.

iii. **Robustness:** This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.

iv. **Scalability:** This refers to the ability to construct the classifier or predictor efficiently given large amounts of data.

v. **Interpretability:** This refers to the level of understanding and insight that is provided by the classifier or predictor. Interpretability is subjective and therefore more difficult to assess.

**b.How classification is done by back-propagation. Give an example of a general multilayered feed-forward neural network.**

**Ans6(b)**
ns:

Backpropagation learns by iteratively processing a set of training samples, comparing the network's prediction for each sample with the actual known class label. For each training sample, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual class. These modifications are made in the "backwards" direction, that is, from the output layer, through each hidden layer down to the first hidden layer (hence the name *backpropagation*). Although it is not guaranteed, in general the weights will eventually converge, and the learning process stops.

The backpropagation algorithm performs learning on a **multilayer feed-forward** neural network. An example of such a network is shown in Figure . The inputs correspond to the attributes measured for each training sample. The inputs are fed simultaneously into a layer of units making up the **input layer**. The weighted outputs of these units are, in turn, fed simultaneously to a second layer of "neuronlike" units, known as a **hidden layer**. The hidden layer's weighted outputs can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the **output layer**, which emits the network's prediction for given samples.

The units in the hidden layers and output layer are sometimes referred to neurodes, due to their symbolic biological basis, or as output units. The mu

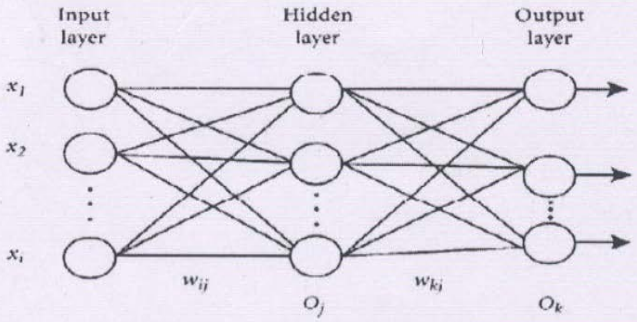



**Figure** A multilayer feed-forward neural network: A training sample, $X = (x_1, x_2, \ldots, x_i)$, is fed to the input layer. Weighted connections exist between each layer, where $w_{ij}$ denotes the weight from a unit $j$ in one layer to a unit $i$ in the previous layer.

layer neural network shown in Figure has two layers of output units. Therefore, we say that it is a **two-layer** neural network. Similarly, a network containing two hidden layers is called a *three-layer* neural network, and so on. The network is **feed-forward** in that none of the weights cycles back to an input unit or to an output unit of a previous layer. It is **fully connected** in that each unit provides input to each unit in the next forward layer.

Multilayer feed-forward networks of linear threshold functions, given enough hidden units, can closely approximate any function.

**c.Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules.**

**Ans 6(c):** There are several approaches to reduce the number of rules. Here we list a few:
•Mine only closed frequent patterns to reduce the number of redundant rules.
•Use multilevel rule mining and generate lower-level rules only when they are non-redundant given the high-level rules. For example, we may find rules at the product category level first. If we find that milk→cheese[support =0.1,conf =0.9] and at the lower level we get milk 2%→provolone[support =0.01,conf =0.92], this may be redundant, i.e., this would be the expected support and confidence given the high-level rule.
•Use domain knowledge to define templates for the rules to be mined and define minimum support, confidence, and correlation measures.

**Q.7a. Describe Market Basket Analysis with suitable example.**

**Ans 7 (a):** Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision making processes, such as catalog design, cross-marketing, and customer shopping behavior analysis.

A typical example of frequent itemset mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their \shopping baskets" (Figure 5.1). The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

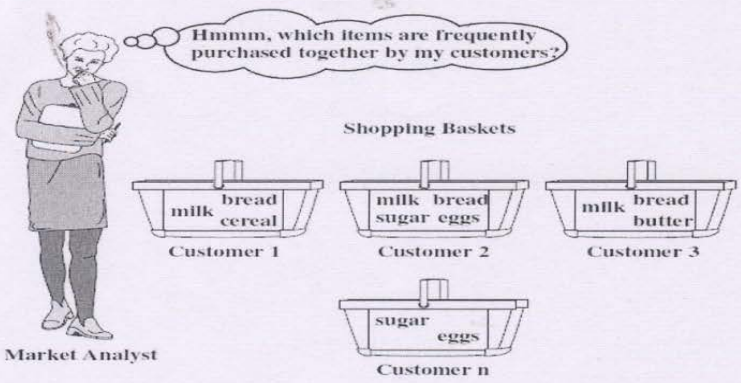



Figure 5.1: Market basket analysis.

**Example** Market basket analysis: Suppose, as manager of an All Electronics branch, you would like to learn more about the buying habits of your customers. Specifically, you wonder, \Which groups or sets of items are customers likely to purchase on a given trip to the store?" To answer your question, market basket analysis may be performed on the retail data of customer transactions at your store. You can then make use of the results to plan marketing or advertising strategies, or in the design of a new catalog. For instance, market basket analysis may help you design different store layouts. In one strategy, items that are frequently purchased together can be placed in close proximity in order to further encourage the sale of such items together. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help to increase the sales of both of these items. In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading towards the software display to purchase antivirus software and may decide to purchase a home security system as well. Market basket analysis can also help retailers to plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers.

If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules. For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule (5.1) below:

Computer → antivirus software [support = 2%; confidence = 60%] (5.1)

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule (5.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both **a minimum support threshold** and **a minimum confidence threshold.** Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlations between associated items.

**b. For class characterization, what are the major differences between a data cube-based implementation and a relational implementation such as attribute-oriented induction? Discuss**

**which method is most efficient and under what conditions this is so.**

**Ans 7 (b):** For class characterization, the major differences between a data cube-based implementation and a relational based implementation such as attribute-oriented induction include the following:

•Process control: Under a data cube-based approach, the process is user-controlled at every step. This includes the selection of the relevant dimensions to be used as well as the application of OLAP operations such as roll-up, roll-down, slicing and dicing. A relational approach does not require user interaction at every step, however, as attribute relevance and ranking is performed automatically.

•Supported data types and measures: The relational approach supports complex data types and measures, which restrictions in current OLAP technology do not allow. Thus, OLAP implementations are limited to a more simplified model for data analysis.

•Precomputation: An OLAP-based implementation allows for the pre-computation of measures at different levels of aggregation analysis.

•Precomputation: An OLAP-based implementation allows for the pre-computation of measures at different levels of aggregation Based upon these differences, it is clear that a relational approach is more efficient when there are complex data types and measures being used, as well as when there are a very large number of attributes to be considered. This is due to the advantage that automation provides over the efforts that would be required by a user to perform the same tasks. However, when the data set being mined consists of regular data types and measures that are well supported by OLAP technology, and then the OLAP-based implementation provides an advantage in decency. This results from the time saved by using pre-computed measures, as well as the flexibility in investigating mining results provided by OLAP functions.

**Q.8 a. Differentiate between partitioning methods and density based methods of cluster analysis.**

Ans. Page401, 418 of text book

**b. State why, for the integration of multiple heterogeneous information source, many companies in industry prefer the update-driven approach (which constructs and uses data warehouse), rather than the query-driven approach (which applies wrappers and integrators)? Describe situation where the query-driven approach is preferable over the update-driven approach.**

**Ans 8 (b):** For decision-making queries and frequently asked queries, the update-driven approach is more preferable. This is because expensive data integration and aggregate computation are done before query processing time. For the data collected in multiple heterogeneous databases to be used in decision-making processes, any semantic heterogeneity problems among multiple databases must be analyzed and solved so that the data can be integrated and summarized. If the query-driven approach is employed, these queries will be translated into multiple (often complex) queries for each individual database. The translated queries will compete for resources with the activities at the local sites, thus degrading their performance. In addition, these queries will generate a complex answer set, which will require further filtering and integration. Thus, the query-driven approach is, in general, inefficient and expensive. The update-driven approach employed in data warehousing is faster and more efficient since most of the queries needed could be done off-line. For queries that either are used rarely, reference the most current data, and/or do not require aggregations, the query-driven approach is preferable over the update-driven approach. In this case, it may not be justifiable for an organization to pay the heavy expenses of building and maintaining a data warehouse if only a small number and/or relatively small-sized databases are used. This is also the case if the queriesrely on the current data because data warehouses do not contain the most current information

**Q.9 a. What is web mining? Explain the types of web mining. (8)**

**Ans 9 (a): Web Mining: Web mining** - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are **Web usage mining**, **Web content mining** and **Web structure mining**.

**A. Web Content Mining** Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of multimedia documents such as images, video, and audio, which are embedded in or linked to the web pages.

It is also quite different from Data mining because Web data are mainly semi-structured and/or unstructured, while Data mining deals primarily with structured data. Web content mining is also different from Text mining because of the semi-structure nature of the Web, while Text mining focuses on unstructured texts.

**B. Web Structure Mining**

Web Structure Mining focuses on analysis of the link structure of the web and one of its purposes is to identify more preferable documents. The different objects are linked in some way. The intuition is that a hyperlink from document A to document B implies that the author of document. A thinks document B contains worthwhile information. Web structure mining helps in discovering similarities between web sites or discovering important sites for a particular topic or discipline or in discovering web communities.

**C. Web Usage Mining**

Web Usage Mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining, discover user navigation patterns from web data, tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. Web usage mining collects the data from Web log records to discover user access patterns of web pages. There are several available research projects and commercial tools that analyze those patterns for different purposes. The

insight knowledge could be utilized in personalization, system improvement, site modification, business intelligence and usage characterization.

**b.What are the major challenges faced in bringing data mining research to market? Illustrate one data mining research issue that, in your view, may have a strong impact on the market and on society. Discuss how to approach such a research issue.**

**Ans 9 (b):** Due to the high demand for transforming huge amounts of data found in databases and other information repositories into useful knowledge, it is likely that data mining will become a thriving market. There are, however, several bottlenecks remaining for data mining research and development. These include:

The handling of increasingly complex data: Such data include unstructured data from hypertext, documents, spatial and multimedia data, as well as from legacy databases, active databases, and the Internet.

•Visualization and data mining: The visualization of database contents would help users comprehend mining results and redirect miners in the search for promising patterns. This requires the development of easy-to-use and "easy-to-see" tools.

• The integration of mined knowledge into a knowledge-base, an expert system, a decision support system, or even a query optimizer.

• Market or domain-specific in-depth data mining with the goal of providing business-specific data mining solutions.

• Invisible data mining, where systems make implicit use of built-in data mining functions

Many may believe that the current approach to data mining has not yet won a large share of the market for system applications owing to the fact that the importance and usefulness of this kind of knowledge has not completely been made aware to the public and the market. Currently, not every university offers undergraduate courses on this topic in computing science departments. Offering more courses on data mining may be a good start. Furthermore, success stories regarding the use of data mining could be featured more prominently in the media.

**Text book**

**I. Data Mining, Concepts and Techniques, Jiawei Han and Micheline Kamber, Elsevier, Second Edition, 2006**