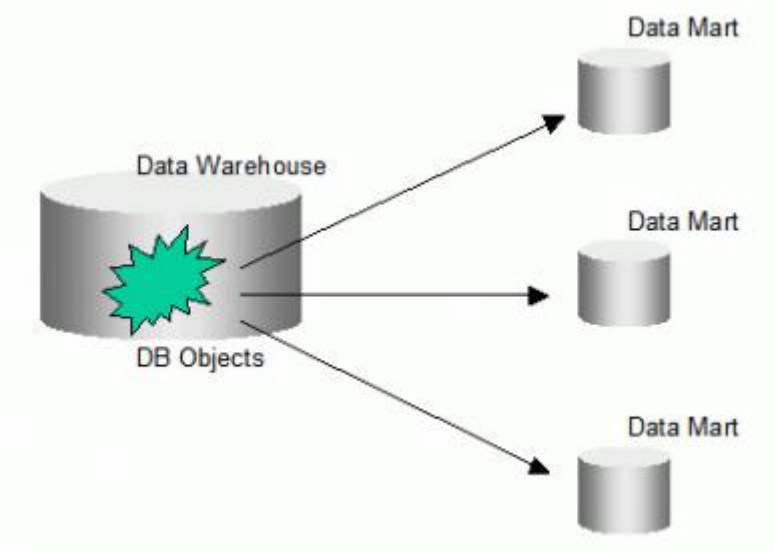
Q.1 Explain the use of frequent item set generation process.

Frequent sets play an essential role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters. The mining of association rules is one of the most popular problems of all these. The identification of sets of items, products, symptoms and characteristics, which often occur together in the given database, can be seen as one of the most basic tasks in Data Mining.

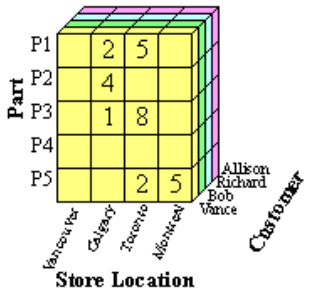
Frequent itemset (Agrawal et al., 1993, 1996) are a form of frequent pattern. Given examples that are sets of items and a minimum frequency, any set of items that occurs at least in the minimum number of examples is a frequent itemset. For instance, customers of an on-line bookstore could be considered examples, each represented by the set of books he or she has purchased. A set of books, such as {"*Machine Learning*," "*The Elements of Statistical Learning*," "*Pattern Classification*,"} is a frequent itemset if it has been bought by sufficiently many customers. Given a frequency threshold, perhaps only 0.1 or 0.01% for an on-line store, *all* sets of books that have been bought by at least that many customers are called frequent. Discovery of all frequent itemset is a typical data mining task. The original use has been as part of association rule discovery. Apriori is a classical algorithm for finding frequent itemset.

Q.2 Differentiate between data marts and data cubes.

A data mart is a *concept*, whereas a cube is an *implementation option*. A data mart is a problem-specific data store, designed to hold information for reporting / analysis / insight around a specific organizational function (typically), and contains a subset of data from a data warehouse. A Data Mart is the staging area for data that serves the needs of a particular segment or business unit. It is a subset of the data in the data warehouse that focuses the information to a particular subject or operational department, fitted to the purpose of the users without redundancy.



The 'cube' metaphor refers to a non-relational data store which represents many dimensions of related data. Cube is normally a shortening of 'hypercube' - a reference to a multidimensional concept from geometry (and wider mathematics) - and highlights that a 'cube' can have more than 3 dimensions.



Data marts are often implemented using cubes, although some data marts are implemented using relational databases (using star and snowflake schemas). More recently, data marts are being implemented in columnar databases and on non-relational / non-cube technologies such as Hadoop.

Q.3 Explain the OLAP operation with example.

A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. The OLAP operation can be categories into following types:

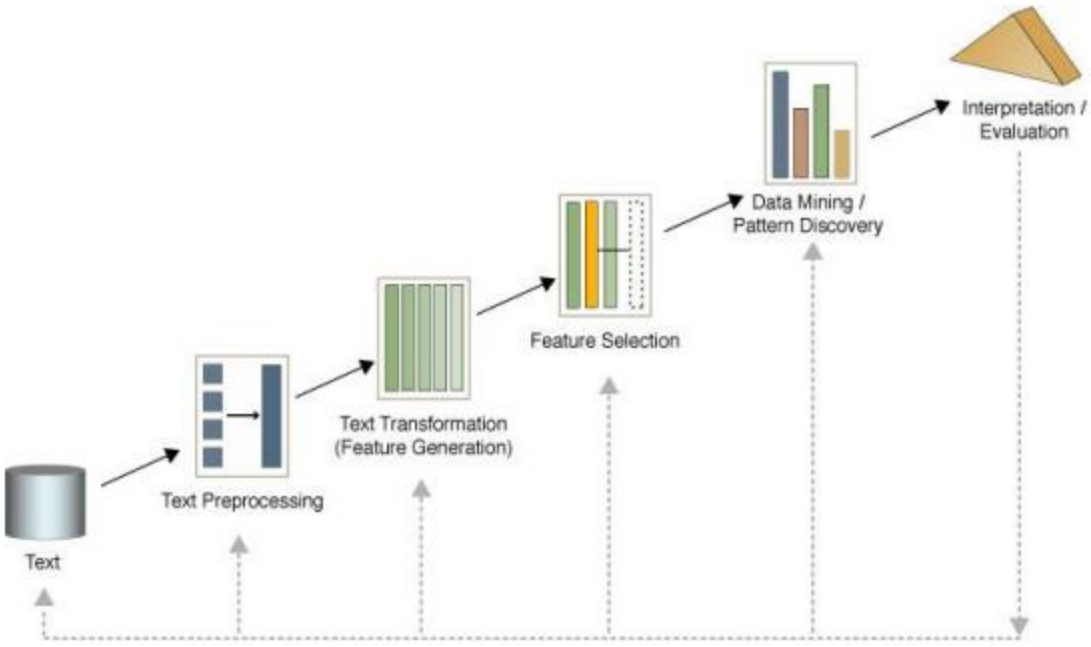
Rollup :
The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. Rather than grouping the data by city, the resulting cube groups the data by country. When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.
Drill-Down:
The Drill-down is the reverse of roll-up. It navigates from less details data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. In data warehousing, generally, Drill-through is the ability to navigate from data/reporting of a data source to data found in another data source. A drill-down adds more detail to the given data, it can also be performed by adding new dimensions to a cube.
Slice and Dice:
The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. Example, you could slice a cube by using a particular product and view all sales of that product across all dates and customers. The dice operation defines a subcube by performing a selection on two or more dimensions. Example, you could slice a cube by using a particular product and view all sales of that product across all dates and customers.
Pivot (rotate):
Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

Q.4 List the drawbacks of ID3 algorithm with over-fitting and its remedy technique.

Q.5 What is text mining? Explain the text indexing techniques.

Text mining is the procedure of synthesizing information, by analyzing relations, patterns, and rules among textual data. This procedure contains text summarization, text categorization, and text clustering.

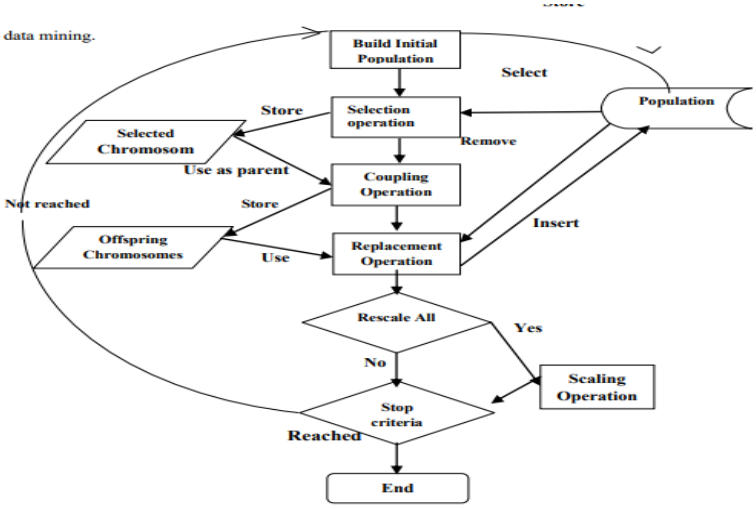
Text summarization is the procedure to extract its partial content reflecting its whole contents automatically.
2. Text categorization is the procedure of assigning a category to the text among categories predefined by users
3. Text clustering is the procedure of segmenting texts into several clusters, depending on the substantial relevance.



Q.6 Describe genetic algorithm using as problem solving technique in data mining .

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic algorithms find application in bioinformatics, phylogenetic, computational science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields.

Keywords
GA, Classifier, data mining.



Flowchart of a genetic algorithm

Genetic Programming is very efficient in problem solving compared to other proposals but its performance is very slow when the size of the data increases. This paper proposes a model for multi-threaded Genetic Programming classification evaluation using a NVIDIA CUDA GPUs programming model to parallelize the evaluation phase and reduce computational time. Three different well-known Genetic Programming

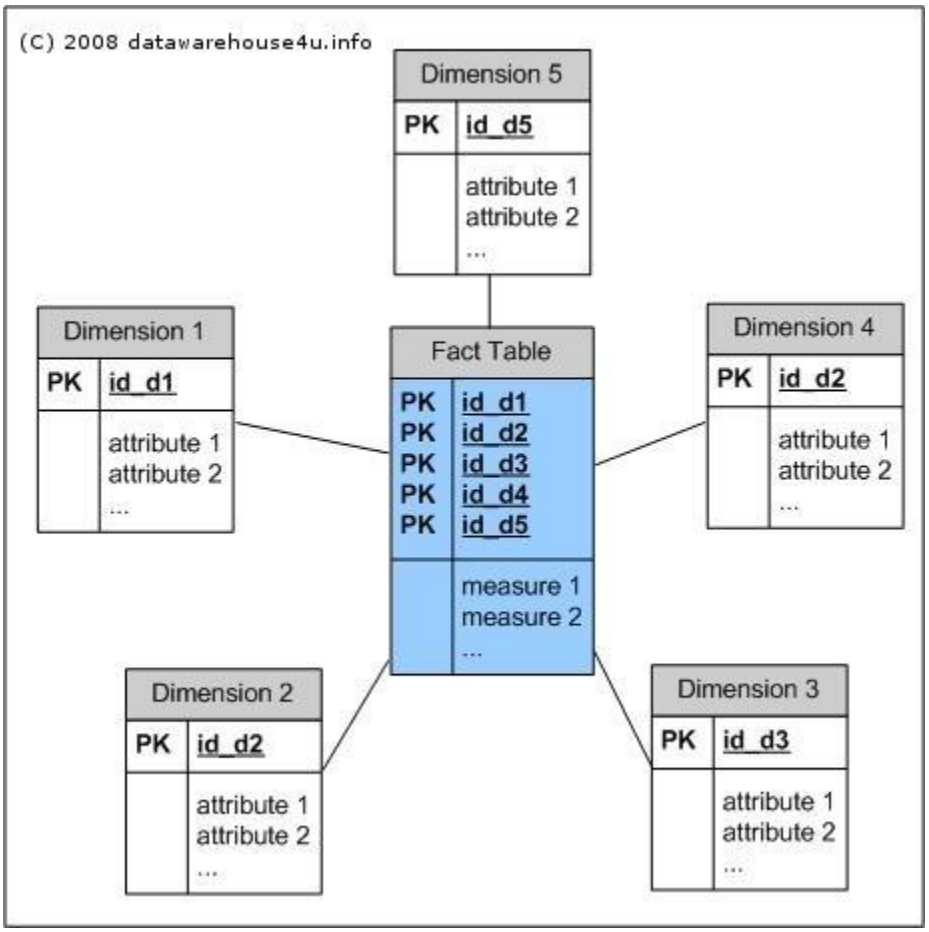
classification algorithms are evaluated using the parallel evaluation model proposed. Experimental results using UCI Machine Learning data sets compare the performance of the three classification algorithms in single and multithreaded Java, C and CUDA GPU code. Results show that our proposal is much more efficient.

Q.7 What do you mean by www mining? Explain WWW mining techniques.

The term **Web Mining** was coined by Orem Etzioni (1996) to denote the use of data mining techniques to automatically discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web. The World Wide Web is a rich, enormous knowledge base that can be useful to many applications. The WWW is huge, widely distributed, global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce, hyperlink information, access and usage information. The Web's large size and its unstructured and dynamic content, as well as its multilingual nature make extracting useful knowledge from it a challenging research problem.

Q.8 What is DMQL? How do you define Star Schema using DMQL?

The Data Mining Query Language (DMQL) was proposed by Han, Fu, Wang, et al. for the DBMiner data mining system. The Data Mining Query Language is actually based on the Structured Query Language (SQL). Data Mining Query Languages can be designed to support ad hoc and interactive data mining. This DMQL provides commands for specifying primitives. The DMQL can work with databases and data warehouses as well. DMQL can be used to define data mining tasks. Particularly we examine how to define data warehouses and data marts in DMQL.



The star schema architecture is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of fact table and the points of the star are the dimension tables. Usually the fact tables in a star schema are in third normal form(3NF) whereas

dimensional tables are de-normalized. Despite the fact that the star schema is the simplest architecture, it is most commonly used nowadays and is recommended by Oracle.

The main characteristics of star schema:

Simple structure - easy to understand schema

Great query effectives - small number of tables to join

Relatively long time of loading data into dimension tables - de-normalization, redundancy data caused that size of the table could be large.

The most commonly used in the data warehouse implementations - widely supported by a large number of business intelligence tools