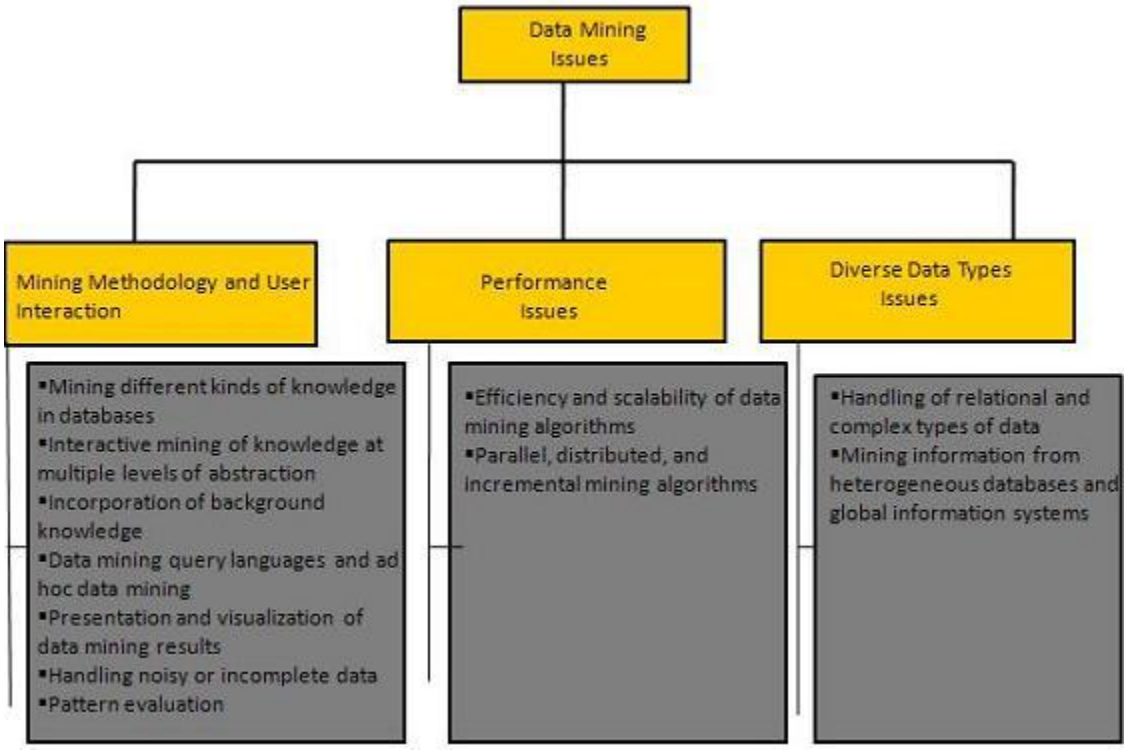
Data Mining - Issues

Introduction

Data mining is not that easy. The algorithm used are very complex. The data is not available at one place it needs to be integrated form the various heterogeneous data sources. These factors also creates some issues. Here in this tutorial we will discuss the major issues regarding:

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kind of issues:

- **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.

Performance Issues

It refers to the following issues:

- **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.

Diverse Data Types Issues

- **Handling of relational and complex types of data.** - The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.

- **Mining information from heterogeneous databases and global information systems.** - The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining knowledge from them adds challenges to data mining.

Source:

http://www.tutorialspoint.com/data_mining/dm_issues.htm