Knowledge Discovery in Data-Mining

Shivali¹, Joni Birla², Gurpreet³
1,2,3Department of Computer Science &Engineering, Ganga Institute of Technology and Management, Kablana, Jhajjar, Haryana, India

Abstract- Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

A warehouse is a commercial building for storage of goods. It is manufacturers, importers, exporters, wholesalers, transport businesses, customs, etc. They are usually large plain buildings in industrial areas of cities and towns and villages.

I. INTRODUCTION
Advances in data gathering storage and distribution have created a need for computational tools and techniques to aid in data analysis. Data Mining and Knowledge Discovery in Databases (KDD) is a rapidly growing area of research and application that builds on techniques and theories from many fields including statistics databases pattern recognition and learning data visualization uncertainty modelling data warehousing and OLAP optimization and high performance computing. KDD is concerned with issues of scalability the multi-step knowledge discovery process for extracting useful patterns and models from raw data stores (including data cleaning and noise modelling) and issues of making discovered patterns understandable. Data Mining and Knowledge Discovery is intended to be the premier technical publication in the field providing a resource collecting relevant common methods and techniques and a forum for unifying the diverse constituent research communities. The journal publishes original technical papers in both the research and practice of DMKD surveys and tutorials of important areas and techniques and detailed descriptions of significant applications. Short application summaries are published in a special section. The journal accepts paper submissions of any work relevant to DMKD. A summary of the scope of Data Mining and Knowledge Discovery includes Theory and Foundational Issues: Data and knowledge representation; modelling of structured textual and multimedia data; uncertainty management; metrics of interestingness and utility of discovered knowledge; algorithmic complexity efficiency and scalability issues in data mining; statistics over massive data sets. Data Mining Methods: including classification clustering probabilistic modelling prediction and estimation dependency analysis search and optimization. Algorithms for data mining including spatial textual and multimedia data (e.g. the Web) scalability to large databases parallel and distributed data mining techniques and automated discovery agents. Knowledge Discovery Process: Data pre-processing for data mining including data cleaning selection efficient sampling and data reduction methods; evaluating consolidating and explaining discovered knowledge; data and knowledge visualization; interactive data exploration and discovery. Application Issues: Application case studies; data mining systems and tools; details of successes and failures of KDD; resource/knowledge discovery on the Web; privacy and security issues.

II. WHAT DOES KNOWLEDGE DISCOVERY IN DATABASES (KDD) MEAN?
Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. Major KDD application areas include marketing, fraud detection, telecommunication and manufacturing.

III. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING
It brings together the latest research—in statistics, databases, machine learning, and artificial intelligence—that are part of the exciting and rapidly growing field of Knowledge Discovery and Data Mining. Topics covered include fundamental issues, classification and clustering, trend and deviation analysis, dependency modeling, integrated discovery systems, next generation database systems, and application case studies. The contributors include leading researchers and practitioners from academia, government laboratories, and private industry.

The last decade has seen an explosive growth in the generation and collection of data. Advances in data collection, widespread use of bar codes for most commercial products, and the computerization of many business and government transactions have flooded us with data and generated an urgent need for new techniques and tools that can intelligently and automatically assist in transforming this data into useful knowledge. This book is a timely and comprehensive overview of the new generation of techniques and tools for knowledge discovery in data.
IV. WHAT IS THE KNOWLEDGE DISCOVERY PROCESS?

There is some confusion about the terms data mining, knowledge discovery, and knowledge discovery in databases, so we first define them. Note, however, that many researchers and practitioners use DM as a synonym for knowledge discovery; DM is also just one step of the KDP.

Data mining was defined in just add here that DM is also known under many other names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing.

The knowledge discovery process (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It 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 analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain.

This defines the term knowledge extraction in a narrow sense. While the acknowledge that extracting knowledge from data can be accomplished through a variety of methods — some not even requiring the use of a computer — uses the term to refer to knowledge obtained from a database or from textual data via the knowledge discovery process.

V. STEPS OF THE KNOWLEDGE DISCOVERY IN DATABASES PROCESS

Data mining is actually the core step in Knowledge Discovery in Databases (KDD) process. Though KDD is used synonymously to represent data mining, both these are actually different. Some preprocessing steps before data mining and post processing steps after data mining are to be completed to transform the raw data as useful knowledge. Thus, data mining alone might not give you what you actually look for.

KDD is an iterative process that transforms raw data into useful information. Different steps of Knowledge Discovery in Databases are:

1. **Understanding**: The first step is understanding the requirements. You need to have a clear understanding about the application domain and your objectives, whether it is to improve your sales, predict stock market etc. It should also know whether you are going to describe your data or predict information.

2. **Selection of data set**: Data mining is done on your current or past records. Thus, you should select a data set or subset of data, in other words data samples, on which you need to perform data analysis and get useful knowledge.

   - **A. Data cleaning**
     Data cleaning is the step where noise and irrelevant data are removed from the large data set. This is a very important preprocessing step because your outcome would be dependent on the quality of selected data. As part of data cleaning, you might have to remove duplicate records, enter logically correct values for missing records, remove unnecessary data fields, standardize data format, update data in a timely manner and so on.

   - **B. Data transformation**
     With the help of dimensionality reduction or transformation methods, the number of effective variables is reduced and only useful features are selected to depict data more efficiently based on the goal of the task. In short, data is transformed into appropriate form making it ready for data mining step.

   - **C. Selection of data mining task**
     Based on the objective of data mining, appropriate task is selected. Some common data mining tasks are classification, clustering, association rule discovery, sequential pattern discovery, regression and deviation detection. We can choose any of these tasks based on whether we need to predict information or describe information.

   - **D. Selection of data mining algorithm**
     Appropriate method(s) is to be selected for looking for patterns from the data. You need to decide the model and parameters that might be appropriate for the method. Some
popular data mining methods are decision trees and rules, relational learning models, example based methods etc.

E. Data mining
Data mining is the actual search for patterns from the data available using the selected data mining method.

F. Pattern evaluation
This is a post processing step in KDD which interprets mined patterns and relationships. If the pattern evaluated is not useful, then the process might again start from any of the previous steps, thus making KDD an iterative process.

G. Knowledge consolidation:
This is the final step in Knowledge Discovery in Databases (KDD). The knowledge discovered is consolidated and represented to the user in a simple and easy to understand format. Mostly, visualization techniques are being used to make users understand and interpret information.

Though these are the main steps in any KDD process, some of the steps could be done combined during the actual process. For example, considering the convenience, data selection and data transformation can be combined together. Even after presenting knowledge to the user, new data can be added to the data set or mining can be further refined or a different data mining method can be chosen to get more accurate results. Thus, KDD is completely an iterative process.

When we analyze different steps of KDD process, we could understand that we are mining data to get useful information or knowledge. Thus, knowledge mining would be the more appropriate term rather than data mining.

VI. KNOWLEDGE DISCOVERY PROCESS MODELS
Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa. We restrict our discussion to those models that have been popularized in the literature and have been used in real knowledge discovery projects.

VII. ACADEMIC RESEARCH MODELS
The efforts to establish a KDP model were initiated in academia. In the mid-1990s, when DM field was being shaped, researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain. The two process models developed in 1996 and 1998 are the nine-step model by Fayyad et al. and the eight-step model by Anand and Buchner. Below we introduce the first of these, which is perceived as the leading research model. The second model is summarized.

VIII. KNOWLEDGE DISCOVERY PROCESS MODELS
The Fayyad et al. KDP model consists of nine steps, which are outlined as follows:

- Developing and understanding the application domain. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.
- Creating a target data set. Here the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.
- Data cleaning and preprocessing. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.
- Data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.
- Choosing the data mining task. Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.
- Choosing the data mining algorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.
- Data mining. This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.
- Interpreting mined patterns. Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.
- Consolidating discovered knowledge. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge.

Notes: This process is iterative. The authors of this model declare that a number of loops between any two steps are usually executed, but they give no specific details. The model provides a detailed technical description with respect to data analysis but lacks a description of business aspects. This model has become a cornerstone of later models.

Major Applications: The nine-step model has been incorporated into a commercial knowledge discovery system called MinedSet (for details, see Purple Insight Ltd). The model has been used in a number of different domains, including engineering, medicine, production, e-business, and software development.
IX. INDUSTRIAL MODELS

Industrial models quickly followed academic efforts. Several different approaches were undertaken, ranging from models proposed by individuals with extensive industrial experience to models proposed by large industrial consortia. Two representative industrial models are the five-step model by Cabena et al., with support from IBM and the industrial six-step CRISP-DM model, developed by a large consortium of European companies. The latter has become the leading industrial model, and is described in detail next.

The CRISP-DM (Cross-Industry Standard Process for Data Mining) was first established in the late 1990s by four companies: Integral Solutions Ltd. (a provider of commercial data mining solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company). The last two companies served as data and case study sources.

The development of this process model enjoys strong industrial support. It has also been supported by the ESPRIT program funded by the European Commission. The CRISP-DM Special Interest Group was created with the goal of supporting the developed process model. Currently, it includes over 300 users and tool and service providers.

The CRISP-DM KDP model consists of six steps, which are summarized below:

- Business understanding. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into a DM problem definition, and designs preliminary project plans to achieve the objectives. It is further broken into several substeps, namely,
  1) determination of business objectives,
  2) assessment of the situation,
  3) determination of DM goals, and
  4) generation of a project plan.

- Data understanding. This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets. Data understanding is further broken down into—collection of initial data,
  1) description of data,
  2) exploration of data, and
  3) verification of data quality.

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data understanding. This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets. Data understanding is further broken down into—collection of initial data,
  1) description of data,
  2) exploration of data, and
  3) verification of data quality.

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,
  2) cleansing of data,
  3) Business Understanding

- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
  1) selection of data,