CLEARPEAKS BI LAB

DATA MINING & BUSINESS INTELLIGENCE

June, 2016

Author: Javier

CLEARPEAKS

TABLE **OF CONTENTS**

DEFINITION
OF DATAMINING

The term “data mining” refers to one of the processes involved in the task of extracting knowledge from a database, also known as KDD (Knowledge Discovery in databases). However, by extension, data mining is referred to as the KDD global process because of its commercial appeal.

Understanding data mining as a KDD sub-process, we could define the term as the process of extracting underlying knowledge from a large volume of data.

Data mining is a recent development directly linked to the scientific fields of mathematics (mainly statistics), computer science and artificial intelligence. Data mining can be supported by different Business Intelligence systems, from which we can obtain several advantages.

DATA MINING
OBJECTIVES/CHALLENGES

One could broadly say that data mining has both a main challenge and a main objective, neither of them trivial:

I. THE CHALLENGE: WORKING WITH LARGE DATASETS

Data mining faces several problems with a large volume of data, such as eliminating noise, dealing with missing data, intractability, volatility, and so on. To solve this problem, we can use different DWH solutions available on the market, such as Amazon redshift or Azure.

Large datasets usually have a very large number of dimensions.

II. THE OBJECTIVE: GETTING KNOWLEDGE FROM DATA

Our aim is to extract knowledge from data that will be useful, novel, understandable and previously unknown.

Data mining seeks to obtain patterns and models from data from information systems, after being adapted for possible treatment through tasks and techniques from the following scientific and technical disciplines:

1. **Databases**

   Managers’ database systems are of fundamental importance in that they support data. A scalable and efficient system is essential to implement algorithms with adequate efficacy; likewise, systems for online analytical processing (OLAP) are traditionally used in Business Intelligence systems.

   Multidimensional Databases are a source of data from which knowledge can be extracted, and they are implemented commercially in many companies. Datawarehousing cloud services, and columnar databases also deserve special mention; there are several solutions on the market, such as Azure, Redshift, or HP Vertica.

2. **Statistics**

   One of the pillars of data mining is statistics, offering us the mathematical basis to find patterns and trends within data through tools such as variance analysis, multivariate analysis, regression, etc. It is noteworthy that classical statistics is an insufficient discipline when dealing with large volumes of data, but it is a key part of data mining. Basic statistics can be found in many sources on the internet; classical tools/languages to carry out such analyses are R, SPSS.

3. **Machine Learning**

   Framed in the discipline of artificial intelligence, machine learning is a field of computer science that allows us to take full advantage of IT tools, using techniques such as artificial neural networks and genetic algorithms. There are several tools on the market to perform machine learning tasks; some of them already have pre-built models for both predictive and descriptive learning. This is a complex topic, and beginners may want to take a look at weka.

4. **Parallel Computing**

   This allows us to develop concurrent and distributed algorithms to minimize the maximum computation and/or execution time. Big data's time has come, and parallel computing can be done with Hadoop.

KDD
PROCESS PHASES

The KKD process consists of several phases described briefly below:

1. **Integration and Data Collection ing**

   The process of building centralized data repositories, standardized as far as possible, leading to the creation of a data warehouse.

2. **Selection, Cleansing and Transformation**

   This is the most time-consuming process in the extraction of knowledge; note that it is important to have enough data for quality analysis, The ETL process also plays a vital role as it creates different datasets suitable for further processing and representation. Good vendors for ETL are Talend, Pentaho data integrator, and Informatica powercenter; you can take a look at the most relevant ETL vendors in the Gartner Magic Quadrant.

3. **Data Mining**

   The datamining process seen as a KDD sub-process stands for the creation of models and pattern extraction from the data, thus producing knowledge; it is based on descriptive and predictive tasks.

4. **Evaluation and Interpretation**

   Knowledge created models should be easily interpretable and adequate for visualization, interpretation and validation; tableau is a great option.

5. **Dissemination and Use**

   The application of knowledge to the area for which we are performing the analysis, for example, expert systems, recommender systems and support systems for decision-making.

APPLICATIONS
FOR DATA MINING

There are a vast number of tasks that can be solved using different datamining processes and techniques, but let us remember that the goal is to get new knowledge from the data.

I. DATA MINING AND BUSINESS INTELLIGENCE

In the world of big business, data mining can play a very important role in decision-making and advance analytics.

We could focus on reducing corporate costs and creating a system of recommendations to aid corporate decision-making.

For example, consider a multinational company with offices in different countries, dedicated to the sale and distribution of computer equipment.

Consider that the data available to the company is transactional accounting information systems and billing systems with other information related to human resources formatted in tables and cells, and a system with information on logistics through mobile apps.

Billing data could be sent in the form of flat text files to a central server in XML for data mobility, and Excel book sheets in the case of data coming from the HR department.

Using an ETL tool, we could populate a centralized repository with all the normalized information (this is one of the first steps to perform before getting any knowledge from the data).

Problems could occur primarily in the data sent by the mobile devices due to human error. HR data could not be standardized and appear in different formats, with typo errors, null values, missing information, etc. This may lead to difficulties finding relations in the data and join the data sources.
These sorts of problems are usually referred to as “data quality issues”, and they can be solved using different tools, or by applying some sort of filter/corrections to the databased on statistics, e.g.: finding term frequencies and replacing misspelt terms with the right ones. This step is also very important, as we are in the still in the dataset preparation phase.

At this point we now have a traditional BI system in our company, and a data warehouse to get the information from.

From the data warehouse a set of minable views can be extracted to continue the process of knowledge extraction. These minable views will be the source for a set of datamining algorithms, like decision trees, association rules or neural networks. In order to test the quality of the algorithms, other test datasets can be extracted from the data warehouse. Analyzing the outcome of the datamining processes allows us to see patterns present in the data that were unknown, and we can make informed decisions in our company; for example, our company data scientists have realized, after analyzing the joined data, that hiring people with specific profiles leads to a higher increase of company revenue.

We could consider tandem Business Intelligence-Data mining as a further step towards the traditional BI System (no mathematical model is normally created to make predictions, orno descriptive analysis of the data is made automatically). An OLAP system is the perfect starting point for the datamining process, as the information is already accessible from a single point.

II. EXTRACTING KNOWLEDGE FROM UNSTRUCTURED DATA

We could also have data from different suppliers in the form of catalogues in HTML pages, with the difficulties this implies, such as creating a usable data structure from web pages; these problems usually happen when we have to deal with unstructured data.

To troubleshoot imbalance amounts, for example, we could compare the information submitted by vendors against company accounts and generate reconciling items. With missing data, we could create fictitious values. Through data mining, we seek to get knowledge from the various integrated datasets; we could, for example:

- Discover the underlying patterns in the data
- Create a model that describes the data using unsupervised learning.
- Create a model that estimates predictions of the cost of distribution using supervised learning techniques.
- Minimize distribution costs and streamline procurement.
- Improve storage and offer deals (like discounts when buying products together) that meet customers' needs.

Other examples of unstructured data are videos and pictures. We could, for example, train a model with a set of pictures to automatically recognize and label the objects that appear in a photo taken by our camera.

III. OPTIMIZATION OF PROCESSES

Another interesting topic related to data mining, used in the KDD process, is the optimization of processes. For example, a distribution company may want to optimize the time used by their salesmen when visiting their customers by car: there are interesting approaches such as metaheuristic algorithms, which are optimization techniques that include a random component, so their outcome can vary.

Data mining is a science also related to Information Retrieval; in this scenario it provides different techniques to index the unstructured data to create different representations of the documents suitable for different types of algorithms, which is the case of IDF, IDTF, which are vector representations of the documents. Have you ever wondered how a web search works? Take a look at the document representation for the page rank algorithms.

In conclusion, data mining is the way to extract unknown knowledge from data. It works well in tandem with Business intelligence systems (OLAP), as we can use a data warehouse to easily access the datasets, as the information is already integrated and clean. So, if we use a data warehouse from our BI system, we can skip phases I and II (information integration and cleansing), some of the most time-consuming tasks.

For phase II visualizations we can take advantage of our current BI system, representing our data with tools such as tableau, and distribute this knowledge across our company.

IV. CONCLUSIONS

Data mining discovers previously unknown patterns and trends in data automatically, helping us to understand the data and learn from it.

Data mining can be applied to different kinds of business and Structured/Unstructured data, and can create predictive and/or descriptive models using different techniques.

A Business Intelligence system, or an OLAP system, Is a great starting point for the datamining process. Data mining can be used for process optimization too.

ABOUT
CLEARPEAKS

We are a specialist Business Intelligence consulting firm, headquartered in Barcelona, with offices in Dubai & Abu Dhabi and sales presence in London. We provide BI services to customers in 15 vertical industries and across several countries in Europe, Middle East and Africa. We are Oracle & Tableau partners.

Our Business Intelligence solutions place valuable information to help our customers make important business decisions around sales operations, marketing, finance, supply chain management or resource management.

We differentiate ourselves through in-depth Business Intelligence knowledge and a dedicated approach to customer satisfaction. Our pragmatic style, coupled with our customer loyalty and quality focus has been the key to our success over the years.

See more in:

CLEARPEAKS
www.clearpeaks.com
info@clearpeaks.com

Barcelona (Head Office)
Travessera de Gràcia 56,
6º 1ª - 08006

Abu Dhabi
Office No. 1954, 19th floor
Al Ghaith Tower, Hamdan Street AYA Business Center

Dubai
Office No. 101, 1st floor,
Building A, Dubai Outsource Zone P.O. Box 500069