Applied informatics

Options for validity analysis and presentation of data from databases.

ZEMÁNEK, Z. - PLUSKAL, D. - ŠUBRT, Z.

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Options for validity analysis and presentation of data from databases.

1. Data mining
2. Data validation in decision making proces
3. Data Mining -Text Mining
4. Assignments

Aims of the lecture

1. Provide students with information to data mining.

2. Provide data for validation options in the decision-making process.

3. Clarify data mining - Data Mining, Text Mining.

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
UNIVERSITY OF DEFENCE
INVESTMENTS IN EDUCATION DEVELOPMENT

Data mining

- The trend is a huge increase in the number of data stored in databases.
- It is generally known that up to eighty percent of the stored data in databases worldwide is in the form of text, i.e. unstructured data. [1]

- It was only in the early 90th of the 20th century, the idea of using data primarily from computer databases, originally intended only for registration purposes, as well as the source of the automated acquisition (mining) knowledge. [2]

- The main factor for the development of a new field has been interested companies to process their data in order to obtain better information management in the company and be able to respond better and faster to the market, to be competitive. [3]

Basic terms

Information is communicated to the recipient's knowledge that makes sense and reduces the degree of uncertainty in his decision.

Data are encrypted information in the form understandable recipients.

Knowledge is structured summary of interrelated findings and experience in a particular area or for any purpose. We are fetching it particularly by practice or study.

Database (or data base), a certain ordered set of information (data) stored on the storage medium.

Knowledge discovery in databases is seen as an interdisciplinary discipline mainly because it requires consuming process share a number of disciplines.

Data validation in decision making process

- When obtaining data from different sources, as well as for statistical assessment of technological processes (such as compliance with the prescribed standards), it is important to examine the validity, i.e. the validity of the results obtained due to the available facts. The process of ensuring validity is then called validation, such as validation test.
- Qualitative or quantitative independent validation is particularly important where the phenomenon under review can not be completely separated from other influences and where the interpretation of results difficult.

Data validation in decision making process

- Data are more and more extensive. The aim to draw useful conclusions from them becomes increasingly complex:
- Challenging the decision-making processes with ICT.
- Millions of financial transactions.
- Millions of calls per day for telecommunication operators.
- …

Data validation in decision making process

- Data are more and more extensive. The aim to draw useful conclusions from them becomes increasingly complex:
- Search for hidden dependencies in the data.
- Comparing patterns of behavior.
- The prediction using segmentation methods, neural networks, etc.
- Finding opportunities, risk prediction.
- ...

What is Data Mining?

Who needs it?

Executive and management.

What implements?

Information about the individual objects and transactions.

What is the purpose?

To support management.

How to implement?

Using database systems.

Data Mining

- Data Mining - Extraction of data is sometimes understood as an analytical part of KDD (Knowledge Discovery in Databases)
- DM scans existing databases, based on the special methods to seek some new knowledge.
- Search for valuable information in large volumes of data.
- The process of identifying valid, unknown, potentially useful and easily understandable knowledge from data (E.g. susceptibility to purchase, fraud, etc.) [2]

What is Data Mining good for?

- An increasing amount of data stored in databases:
  - We continuously generate data
  - Business and banking transactions
  - Communication, biological, astronomical, system data, ...
- Database technology is getting faster and cheaper
- Database systems are able to work with more extensive data
- Nontrivial finding of hidden dependencies between data entities (E.g. susceptibility to buy, fraud, etc.) [5]

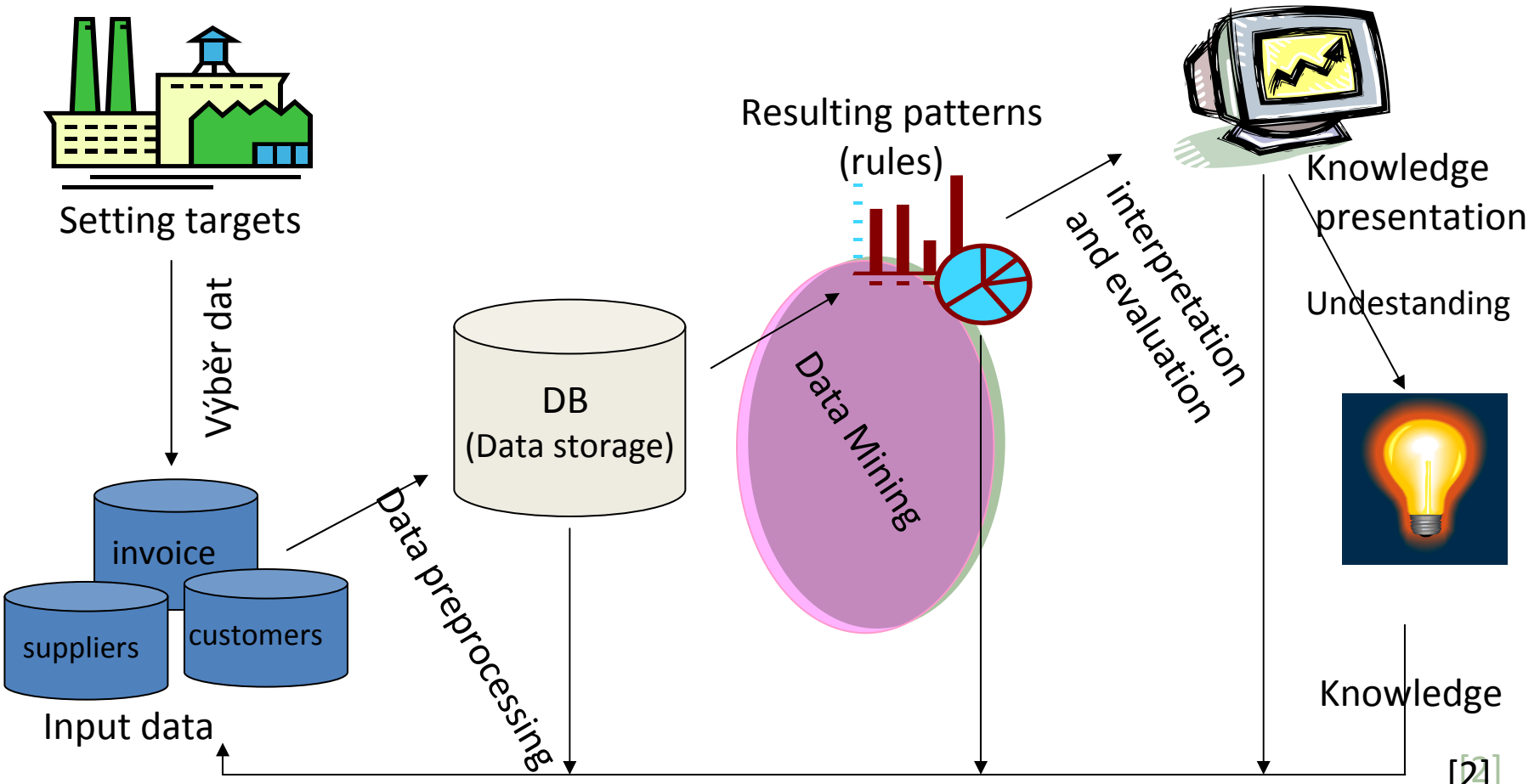
Where Data Mining is being used?

- Common applications are primarily in the areas of:
  - finance (e.g., risk estimation, the search for fraud)
  - Direct Marketing (selection of clients to reach)
  - telecommunications (client segmentation, sales programs, ...)
  - monitoring activities on the Internet in order to detect activity of fraudsters and potential terrorists)
  - Internet Sales (analysis of transitions between pages, efficiency of advertising, ...). [4]

Examples of Data Mining

- designing and monitoring the effectiveness of marketing campaigns
- designing security measures for complex industrial plants and machinery,
- Analysis and optimization of server solutions
- examining patterns of climate change by long time series of meteorological measurements,
- creating various sociological forecasts
- Planning of stock market and currency speculation. [2]

Process of data mining



[2]

Data Mining Process

- Setting goals
- What kind of knowledge we want to find?
- Over what data we process the data mining?
- Is the problem solved?
- Will the results will be useful in practice?
- In what shape and form we see the results of the acquisition of knowledge?
- Are our data appropriate for the method?

Data Mining Process

- Select the data source
- Types of databases in terms of content
  - Customer database - Information about the customer or on its activities
  - Database of transactions - information about the activities of customers (mostly anonymous)
  - Database of history of offers - the database of addressing clients in campaigns
- External data - WWW

Text Mining

- Text Mining generally belongs to a set of data mining methods - but they are typically working with numbers, or with nominal or ordinal variables (such as category names, etc.)
- Text Mining is working with unstructured text. It can be defined as the process of extracting valuable information from the text, this method may help in the actual data mining analysis.

[1]

Data Mining Methods

Extraction of the message meaning from the unstructured text

- The number and structure of words can help to identify the topic and sense. The document does not have to be a large Yearbook or a multi-page thesis, but for example a web page.
- A more interesting possibility is based on definition of specific keywords or links (term of the language).

[1]

Data mining methods

Extraction of the message meaning from the unstructured text

- We look for objects in the body (individual words or important connections) – discount loan, cystic fibrosis, Gothic monuments for example, the word trauma indicates higher indemnity, because the client was likely to be seriously injured.
- Terms are then displayed in an array of words that is created on the basis of frequency analysis (frequency of occurrence).

Data Mining methods

Automatic document sorting

- Even more interesting properties of text mining tools are the possibility to find specific or similar text records based on cluster analysis.
- Textual records are classified and categorized into clusters according to their similarity.

[1]

Data Mining methods

Automatic document sorting

- The figure shows individual text records (documents, forms, applications, etc.) that were subjected to cluster analysis.
- Records that are outside the cluster, somehow differ from most documents, and therefore the analytical department could pay attention to.

[1]

Data Mining methods

Presentation of data analysis

- The output visualization does not bring anything new, but presentation of data and analysis results can greatly facilitate the understanding and subsequent interpretation.
- The results of calculations on the data can take different forms.
- The simplest form of numeric, organized into sets, tables, etc., usually an expert for further work.
- Much more illuminating are complementary outputs the charts, under the general rules of good design.

[3]

Data Mining methods

Example: Automated text sorting

- Fraud management is an area that focuses on early detection of deception.
- Text Mining as a tool in this area is used for the purpose of internal control. It automatically reads e-mails of employees and when it detects a certain word or phrase that indicates suspicious E-mail then a relevant department pays attention to it.
- In the same way text mining tool also analyzes the electronic application, orders via internet, etc., which come from outside. Inputs are classified into meaningful clusters, for example, and can reveal a suspicious order, etc.

[1]

Data Mining methods

Conclusions

- The trend nowadays is a huge increase in the number of data stored in databases.
- Qualitative or quantitative independent validation is important where the interpretation of results is difficult.
- Acquisition (mining) of knowledge from the data is a non-trivial process of acquiring implicit, previously unknown and potentially useful and valid (valid) information from the data.
- Data-mining methods for working with numbers, or with nominal, or ordinal variables, such as category names, etc.
- Text mining is working with unstructured text, it can be defined as the process of extracting valuable information from the text. The method can help in the actual data mining analysis.

Assignments

On the Internet find relevant information for:

- Data extraction,
- Data validation in decision-making process,
- Data Mining, Text Mining.

Resources:

1. ULDRICHT, Miloš. *Text mining aneb Kladivo na nestrukturovaná data.* [online]. [cit. 2013-10-29] č.12/2011, IT SYSTEMS: Business Intelligence Dostupné z: http://www.systemonline.cz/clanky/text-mining-kladivo-na-nestrukturovana-data.htm
2. Datové sklady: *Data mining.* [online]. [cit. 2013-10-23]. Dostupné z: http://kix.fsv.cvut.cz/~vanicek/vyuka_l13/sklady.ppt#295,28,Shlukování – některé metody
3. ŠARMANOVÁ, Jana. *METODY ANALÝZY DAT - Učební text.* [online]. [cit. 2013-10-26] © 2012, Ostrava: VŠB-TU. 170 s. ISBN 978-80-248-2565-6 Dostupné z: http://www.person.vsb.cz/archivcd/FEI/MAD/
4. BERKA, Petr. *Aplikace systémů dobývání znalostí pro analýzu medicínských dat.* [online]. 24. 10. 2002 [cit. 2013-10-24]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=uvod
5. Data mining. *ORACLE* [online]. [cit. 2013-10-27]. Dostupné z: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/odm-techniques-algorithms-097163.html