УДК 004.4

DATA MINING AS A PART OF THE INFORMATION TECHNOLOGIES MARKET OF THE HUMAN RESOURCES

Gil M. V., Fonotov A. M.
Donetsk National Technical University, Donetsk
Department of Computer Engineering
E-mail: mary_kukla@mail.ru

Annotation

Gil M. V., Fonotov A. M. Data mining as a part of the information technologies market of the human resources.*We study the selection of profiles of candidates for interviews with the employer in this article. Data mining can be a solution for data analysis, faced by many companies in the solution of this problem.*

Statement of the objectives of the study. *In this paper, the task is the selection of profiles of candidates for interviews with employers.*

Solution of the problem and the results of research.

Data mining as a part of the information technologies market of the human resources. Data mining software is one of a number of analytical tools for analyzing data now. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified [1].

Data mining (the analysis step of the knowledge discovery in databases process, [4] or KDD), a relatively young and interdisciplinary field of computer science [6] is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems [5].

Highlights [2]:

1) Provides both theoretical and practical coverage of all data mining topics.
2) Includes extensive number of integrated examples and figures.
3) Offers instructor resources including solutions for exercises and complete set of lecture slides.
4) Assumes only a modest statistics or mathematics background, and no database knowledge is needed.
5) Topics covered include; predictive modeling, association analysis, clustering, anomaly detection, visualization [2].

Depending on the specifics of the data can exist the following classes of systems, Knowledge Discovery and Data Mining [3]: Data Mining, Text Mining, Image Mining, Video Mining, Audio Mining, Web Mining, Multimedia Mining, Spatial Mining, Temporal Mining, Streams Mining, Social Networks Mining (http://www.asonam.org), etc.

The main goal: We must choose the required data of candidates, that would they will pass an oral interview with the employer for the different sectors of activity of enterprises based on the analysis of certain parameters - the length of service, age, occupation, sex, education, foreign languages, etc.

This problem relates to the problems of evaluation and debriefing.

Input information:

1. List of candidates;
2. Information from the questionnaires of candidates

Output information:

1. List of candidates applying for the vacant post, with all the necessary characteristics and results of preliminary testing;
2. Reports with additional information about the candidates generated and sorted by criteria.

The selection of profiles of candidates for interview consists of the following steps:

1. Formation of Data from the questionnaires candidates filled by hand, in electronic form;
2. Search the database of candidates for an interview with the employer criteria;
3. Testing of applicants and entering the results into the database;
4. Creating a list of candidates, and statistical reporting.

The problem of "Filter whether the candidate for an interview to the employer?" With Data Mining methods will be solved as follows.

1. The database agencies, excluding the current candidates divide into two classes:

1) The last interview with the employer;

2) Do not pass the interview with the employer.

We excluded from the two classes of candidates, who have the type of activity does not correspond to the desired. Activities in system are denoted as a three-level directory with the following hierarchy levels: group, subgroup, and appearance.

3) Based on the analysis of class "did not pass the interview with the employer" is defined (as common) and put into a separate table of "failure" list of attribute values of the candidates according to their personal data entered into the system. For each applicant determined the following list of attributes:

1. Name (Not analyzed);
2. Your passport, TIN (not analyzed);
3. Address;
4. Date of birth;
5. Total length of service work;
6. The floor;
7. Specialization and qualification;
8. Positions held and duties of the last three places of work;
9. Reasons for dismissal;
10. Availability of work experience in managerial positions;
11. Experience with the computer;
12. Language skills, etc.

Each attribute in the application the applicant has a "measure of importance" - a number from 1 to 20. Sort the resulting table in descending order, "the importance of the criteria."

4) When you receive information about each new candidate for an interview we determine the main "features" potential "objector", by a comparative analysis of the attribute values of the questionnaire with the appropriate values of the table, "Failure," taking into account "the importance of the criteria." The maximum error in the calculations assumes a 10%.

5) Based on calculations made by the candidate define a class.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining).

This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and used in further analysis or for example in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Computer science conferences on data mining include:

1. CIKM – ACM Conference on Information and Knowledge Management
2. DMIN – International Conference on Data Mining
3. DMKD – Research Issues on Data Mining and Knowledge Discovery
4. ECDM – European Conference on Data Mining
5. ECML-PKDD – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
6. EDM – International Conference on Educational Data Mining
7. ICDM – IEEE International Conference on Data Mining
8. KDD – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
9. MLDM – Machine Learning and Data Mining in Pattern Recognition
10. PAKDD – The annual Pacific-Asia Conference on Knowledge Discovery and Data Mining
11. PAW – Predictive Analytics World
12. SDM – SIAM International Conference on Data Mining (SIAM)
13. SSTD – Symposium on Spatial and Temporal Databases
14. Data mining topics are present on most data management / database conferences.

The knowledge discovery in databases (KDD) process is commonly defined with the stages Selection Preprocessing Transformation Data Mining Interpretation/Evaluation. It exists however in many variations of this theme such as the CRoss Industry Standard Process for Data Mining (CRISP-DM) which defines six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment or a simplified process such as Pre-processing, Data mining, and Results validation. Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable timeframe. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate datasets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data. Data mining involves six common classes of tasks:

1. Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors and require further investigation.
2. Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
3. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
4. Classification – is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam.
5. Regression – Attempts to find a function which models the data with the least error.
6. Summarization – providing a more compact representation of the data set, including visualization and report generation.

The final step of knowledge discovery from data is to verify the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called overfitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish spam from legitimate emails would be trained on a training set of sample emails. Once trained, the learned patterns would be applied to the test set of emails on which it had not been trained. The accuracy of these patterns can then be measured from how many emails they correctly classify. A number of statistical methods may be used to evaluate the algorithm such as ROC curves.

If the learned patterns do not meet the desired standards, then it is necessary to reevaluate and change the pre-processing and data mining. If the learned patterns do meet the desired standards then the final step is to interpret the learned patterns and turn them into knowledge.

Conclusions

The algorithms used in Data Mining, require a large number of calculations on large amounts of data. Previously, it was a deterrent to widespread practical application of Data Mining, but now the growth performance of modern multiprocessor servers took the acuteness of the problem. Now, within a reasonable time can be a qualitative analysis of hundreds of thousands or millions of records.

Publications

1. Data Mining. Technology Note prepared for Management 274A. Anderson Graduate. —Bill Palace, Spring, 1996
2. Introduction to Data Mining. Pang-Ning Tan, Michael Steinbach, Vipin Kumar Addison-Wesley, 2005. ISBN: 0321321367.
3. Feldman R., Sanger J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge.: Cambridge University Press, 2006.
4. Cabena, Peter, Pablo Hadjnian, Rolf Stadler, Jaap Verhees and Alessandro Zanasi (1997). Discovering Data Mining: From Concept to Implementation. Prentice Hall, ISBN 0-13-743980-6.
5. Feldman, Ronen and James Sanger. The Text Mining Handbook. Cambridge University Press, ISBN 978-0-521-83657-9.
6. Guo, Yike and Robert Grossman, editors (1999). High Performance Data Mining: Scaling Algorithms, Applications and Systems. Kluwer Academic Publishers.