SUMMARY OF ALGORITHMS AND SOFTWARE FOR PROCESSING RESULTS OF SOCIAL RESEARCH

S.V. Romanchukov
(Tomsk, Tomsk Polytechnic University)
inoy@vtomske.ru

Annotation. This article is devoted to algorithms and software for analytics module of multifunctional web portal for social researches. It consists of «MultiTest» web-portal short description and also summary of the elements of the Data Mining concept, used in the current system and programming language, used for their realization.

Keywords. Data Mining, statistical analysis, social researches, web-portal, tests, R lang.

Introduction. Modern universities are interested in a wide range of social research aimed at understanding of their own staff and applicants, such as career guidance tests for possible entrants, tasks of quality measurement of learning material, teachers qualifications, common level of student's knowledge and their opinion about univercity. And one of the main means to carry out all this testing we can use network resources, whether multipurpose or specialized. In addition to the collecting the data such resources provides possibility of processing gathered information about the participants and their results in order to create or adjust strategies for the future policy of the university or one's part. There are some problems in developing such analysis module in this kind of system:

1. Storing data in a form less dependent on the platform.
2. Possibility of statistical processing of the stored information.
3. Possibility solutions weakly formalized goals supported by the available data.
4. Possibility of data mining.

For data storing we can use DBMS like MySQL or Oracle. The flexibility of this DBSM allows us to realize many types of tables in variety of supported data formats. For description of tests and procedures we can use universal format XML, which allows, for example, to create new tests for the portal, with help of side-party programs, including even the usual text editors [1].

But traditional relational DBs store information about the test results themselves; they are not suitable for analysis of the gathered information. The very nature of social research complicates the formal statement of the problem by the query language of DBMS. Traditional means of the mathematical statistics are not suitable for such tasks. Methods of mathematical statistics are useful first of all while checking formulated hypotheses and exploratory analysis which is the foundation of online analytical processing (online analytical processing, OLAP)[2]. That is why data mining techniques take on special significance.

Data Mining. Data Mining is a discovery process in the raw data for previously unknown, non-trivial, practically useful and accessible interpretation of knowledge required for decision-making in various spheres of human activity [3]. A feature of Data Mining is, as already noted, the non-triviality of wanted templates. Desired patterns should reflect the unobvious, unexpected regularity in the data that make up the so-called hidden (deep) knowledge.

There are five common types of patterns, which may be obtained by Data Mining: association, consistency, classification, clustering, and forecasting. In this case the most important tasks are data classification and clustering.

Using the classification allows us to identify features that characterize a group that includes a particular object. This is done through an analysis of already classified objects and formulates a set of rules. Clustering differs from classification in that the groups themselves are not set in advance. Using Clustering Data Mining tools independently identify different homogeneous groups of data [4].

Implementation of Data Mining techniques is possible in different ways. This problem can be solved by using various programming languages, but we tried it with R language. It is a software environment and programming language specialized for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R's popularity has increased substantially in recent years.

R is free, open-source product and also can be used as a part of server software for web-applications. It can be used from scripting languages such as PHP, Python, Perl, Ruby, F# and Julia. R with PL/R extension can be used alongside, or instead of the PL/pgSQL scripting language in the PostgreSQL and Greenplum DBMS. All this turns R language in ideal instrument for this task.

Design of analysis module for the web portal facilitates the existence of a sufficient number of software products that implement data mining algorithms, from stand-alone applications (such WizWhy and WizRule) to embedded function libraries or modules within the mathematical packag-es. There are a lot of Internet application performed in web-oriented programming languages e.g. PHP, JavaScript, C++ and associated with the MySQL and Oracle DBMS.In simplified form the scheme of such system, on an example of multipurpose portal «MultiTest» is shown in Fig. 1.

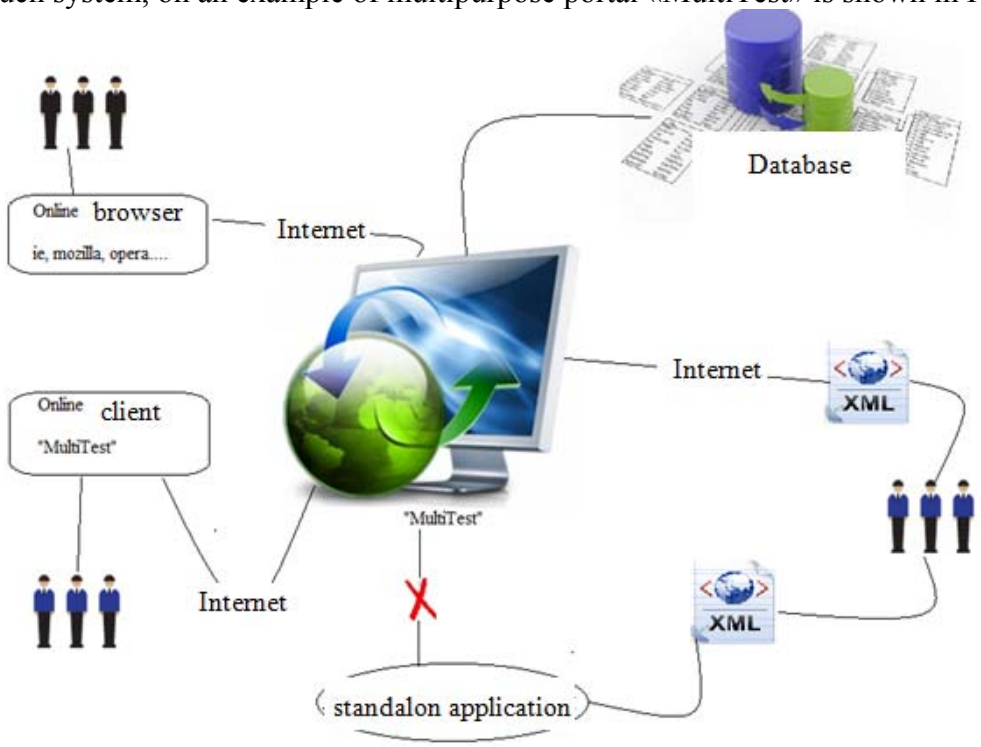

Fig. 1. Scheme of multipurpose web-portal «MultiTest»

The study is carried out on the basis of Tomsk Polytechnic University with financial support from the Ministry of Education and Science of the Russian Federation in the framework of the sci-entific research works in the direction of «Evaluation and improvement of the social, economic and emotional well-being of older people», contract № 14.Z50.31.0029.

References

1. Berestneva O.G., Fisochenko O.N., Moiseenko A.V., Shcherbakov D.O., Career-oriented decision support system development for the students of the National Research Tomsk Polytechnic University / / Internet Journal of Science of Science. – 2013 – №. 4.

2. Data Mining – intellectual data analysis [electronic resource] / / VA Duke, St. Petersburg Institute for Informatics and Automation of RAS

3. URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html # 1. What is Data Mining?

4. Gregory Pyatetskii-Shapiro, Data Mining and information overload / / preface to the book: Data Analysis and Process / A. Barseghyan, M.S. Kupriyanov, I. Frost, M.D. Tess, S. I. Elizarov. 3 ed. rev. and add. St. Petersburg.: BHV-Petersburg, 2009. 512 p. C.13.

5. Data Mining: Concepts, Models, Methods and Algorithms/ Mehmed Kantardzic – New Jersey, 2011.– C. 249-253.