Future Academy®
Offering excellence in Social and Behavioural Sciences

The European Proceedings of
Social & Behavioural Sciences
EpSBS

eISSN: 2357-1330

WELLSO 2015 - II International Scientific Symposium on Lifelong Wellbeing in the World

Issues of Application Design for Social Studies

S. Romanchukov[a*], E. Berestneva[a]

* Corresponding author: berestneva_l @mail.ru

[a]*National Research Tomsk Polytechnic University, Institute of Cybernetics, Tomsk, Russia (634050, Tomsk, Lenin Avenue, 30), e-mail: tpu@tpu.ru*

Abstract

http://dx.doi.org/10.15405/epsbs.2016.02.12

The given paper presents the relevant problem of software design for social welfare researches oriented towards the life satisfaction and other psychology studies which are important for social development. The aim of this paper is to show most important issues which can complicate application of algorithms and software for social studies e.g. testing and interviewing results analysis. It summarizes the elements of the Data Mining concepts which are vital for processing big amount of complex data characterized with weak mathematical formalization (as results of sociological studies). This article also includes description of requirements for software systems meant to be implemented in social studies and names existing software decisions in order to show their limitations. This allows us to conclude that this software implementations can be expanded and coordinated by means of new information system, created especially for solving this problem. And also includes short description of «MultiTest» web-portal designed to solve a number of problems associated with the social studies.

Keywords: Welfare; everyday practices; quality of living; social indicator; social welfare model; data mining, social researches, analytics subsystem.

1. Introduction

In modern society, the human factor, the level and quality of life plays an increasingly important role. It creates the need for different kinds of social and psychological research. Amounts of data obtained during the execution of these works require special software to process them (Zharkova O.S., Berestneva O.G., Moiseebko A.V., Marukhina O.V., 2013).

Each specific type of sociological study due to the nature of the goal, put forward tasks, whereby there are three main types of sociological research: exploratory, descriptive and analytical, all kinds of researches implies a certain sequence of steps:

1) forming a research program;
2) collection of primary information;
3) digital processing of the received information;
4) analysis of the processed information, formulation of conclusions and recommendations.

Obviously the use of network information resources and analytic software speeds up the process of any research. After the presentation of the results of surveys and tests in varying scales it becomes possible to start their mathematical processing using a variety of statistical methods. However, they have their limitations as regards, first of all, the methods of parametric statistics, which are based on the assumption that the random vector variables forms a distribution. If this assumption is not supported, you should use non-parametric methods of mathematical statistics and DataMining - search the raw data previously unknown, nontrivial, practically useful and accessible interpretation of knowledge (Barseghyan A.A., Kupriyanov M.S., Frost I., , Tess M.D., Elizarov S. I., 2009).

There are five groups of methods DataMining: association, consistency, classification, clustering and prediction. In the case of Social Studies at the first row out of the problem of clustering and classification, as using the classification of identified features that characterize a group to which belongs to a particular object, and using clustering isolated groups of data that were not set in advance, creating new products (Mehmed Kantardzic, 2011).

2. Problem statement

As the initial information during the study we are collecting details of testing results and surveys, including:

1) Personal data of respondent (if the test was not anonymous);
2) Information about the connection (IP-address, region, use the browser, etc.);
3) Discussed questions;
4) Answers;
5) Test results;
6) Test duration.
7) Whatever...

In addition to individual results database tables must store information about the structure of the tests and so on. OLAP-processing of this datasets is ineffective this is just cause for implementing Data Mining procedures.

The most important Data Mining methods for solving tasks of social studies are clustering and classification.

Cluster analysis - a multidimensional statistical procedure, collecting data, containing information about the sample objects, and then organizing objects into relatively homogeneous groups (clusters).

The main purpose of cluster analysis is partition of objects and attributes into homogeneous groups (clusters). A significant advantage of cluster analysis is that it allows you to partition and grouping objects on a set of attributes.

As a quality criterion of clustering somehow includes a number of informal requirements (Mandel I.D., Black L.M., 1988):

http://dx.doi.org/10.15405/epsbs.2016.02.12
eISSN: 2357-1330 / Corresponding Author: Romanchukov S., Email: inoy@vtomske.ru
Selection and peer-review under responsibility of the Organizing Committee of the conference

1) the relationship of objects within a group

2) objects of different groups should be far from each other;

3) the distribution of objects in groups, ceteris paribus should be uniform.

The main objective of cluster analysis is allocating n-dimensional subsets of homogeneous data, such that objects within a group are close to each other and moving away from the objects of the other groups.

Traditionally, the classification is divided into hierarchical and non-hierarchical (sometimes called structural). The basis of hierarchical algorithms is clustering closest and then consistently and increasingly distant from each other elements. Most of these algorithms is based on a matrix of similarity (distance), and each element is considered first as a separate cluster.

Overall a hierarchical clustering algorithm can be represented as three repeated operations to measures distances between objects (clusters):

1) find the shortest distance between the objects (clusters);
2) combine them into a single cluster;
3) calculate the distance from the cluster obtained before to any other.

There are a number of methods of cluster analysis based on the Euclidean distance measure, including the method of Ward, nearest neighbor, a distant neighbor method and the median method. In practice, as a result of testing various algorithms researchers was formed by a series of recommendations for their application:

1) Various algorithms for cluster analysis give different partition;
2) The results of the classification are inversely related to the dimension of the space and a straight line from the target number of clusters;
3) The most resistant to noisy data algorithms k-means, Ward metod, the least - the median method and the nearest-neighbor;
4) In a small number of features and a large number of classes are best shown themselves Ward method, group of medium and long-neighbor method, bad - k-means and method of the nearest neighbor;
5) With a large number of features and a small number of classes the most effective method of Ward and k-means least - median, centroid and nearest neighbor;
6) In the wide range of conditions cases work well enough distant neighbor method and the method of Ward, bad - the median and nearest neighbor;

Given the sensitivity to noise and is capable of recovering data structure is the best algorithm Ward worst - nearest neighbor method (Mandel I.D., Black L.M., 1988).

3. Existing software solutions

There is a variety of widely used software for implementing statistical methods, which can be divided into three categories:

1) Universal statistical packages (STATGRAPHICS, Statistica, etc.).
2) Professional statistical packages (SPSS, SAS / IDS, BMDP).
3) Specialized statistical packages (EQS, heuristic).

It should be noted that the opportunities for statistical treatment of data provided by the tools require large computer processing resources, and moreover some of them supports a separate operating system or a significant problem when working with a network supporting only work over a local network or interfering with the large number of restrictions on the size and structure of the processed data, the value of exchange data with a server, etc. Studies related to the processing of personal data, also face legal restrictions that exclude the use of foreign servers (Marukhina O.V., Mokina E.E., Berestneva E.V., 2015).

The other obstacle to the development of these programs is the time to be spent on training. Researches in social studies cad hardly handle with this kind of software. So because of lack of knowledge the power of statistical packages is often not used in a proper way.

An important drawback is the high price of these software products and their excess capacity. Most of functions, provided by typical statistical package are unclaimed for social or medical studies.

Also there are some legal issues with using databases and servers located abroad (e.g. art. 18, art. 22 of the Federal Law of the Russian Federation "On Personal Data" dated July 27, 2006 N 152-FL after changes in accordance with the Federal Law N 142-FL dated June 4, 2014)

All this brings us back to the need to develop software solutions tailored to the needs of real-life research groups, together with the participants.

4. System requirements

Analytical module is being developed as part of an extensive network system, which imposes a number of restrictions on it.

1) The product should be based on web-oriented languages and interact with other modules fitting into common "framework" of the system.

2) It is necessary to have the ability to handle data in a wide range of formats, reading files of different types.

3) Equally important is the opportunity to work with the most common databases (MySQL, Oracle, etc.).

4) The developed module should be available through a network of Internet, it required testing for the most common browsers.

5) Service should allow separation of users into different categories and implement the separation of access rights, ensuring the security of data storage and protection against unauthorized access.

6) In terms of localization of storage should be noted that it is desirable location of servers and storage in the territory of the Russian Federation in accordance with Russian law.

7) Preference is given to free software with open source.

The basic idea of the development of the whole system is to improve her mobility, compared to traditional applications used in the art that is essential for solving tasks, in a situation where study covers the whole region.

The three main features of the developed system should be distinguished from its analogues: the possibility of obtaining the results of surveys and tests on the network as soon as possible, the

http://dx.doi.org/10.15405/epsbs.2016.02.12
eISSN: 2357-1330 / Corresponding Author: Romanchukov S., Email: inoy@vtomske.ru
Selection and peer-review under responsibility of the Organizing Committee of the conference

possibility of processing results in the mode of on-line, and, most importantly, the freedom of the researcher, not tied to a particular machine. Enough to work a browser and Internet access, the computation is done on the server and the user does not have to download all the necessary data and install cumbersome client on his car goes only necessary "working set" requested results (Bobrova M.V., Berestneva E.V., 2015).

5. MultiTest web-portal

Scientific-educational laboratory of information technologies in health care and social studies developed «MultiTest» portal.

A significant amount of information has been already accumulated in its database and holds information on the many years of research, using MySQL database. It allows to develop a module to interact with MySQL databases, performing data export for processing and maintaining a results database.

Universal data format XML is used to describe testing methods and results, which allows, for example, to create new tests include it into the portal by means of third-party programs, including the usual text editors. The same format is carried out and storing in a database information about the test results (Berestneva O.G., Fisochenko O.N., Moiseenko A.V., Shcherbakov D.O., 2013).

Analytical module will be located on the same server with the existing components of the portal, which is a HTTP-server Apache on the FreeBSD platform and implemented using php, html and AJAX, similar to already placed components of the system to maintain uniform standards with them.

In addition the service should allow the separation of different categories of users, supporting various functions and carry out the separation of different categories of access rights, ensuring the security of data storage and protection against unauthorized access.

In simplified form the scheme such system, on an example of multipurpose portal «MultiTest» is shown in Figure 1.

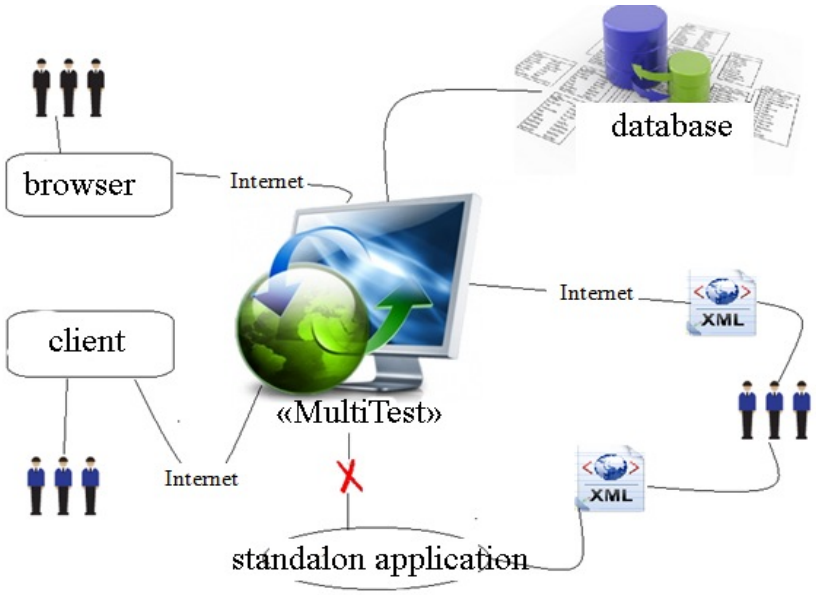

Fig. 1. Scheme of multipurpose web-portal "MultiTest"

Analytical module of this software system is still in work-in-progress stage (realization of selected statistical methods). Implementation of Data Mining techniques is possible in different ways. This problem can be solved by using various programming languages, but best results are expected with R language. This is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls and surveys of data miners are showing 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. R functionality has been made accessible from several 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.

Server base for R applications has been already set up and now is in set-up process.

6. Conclusion

Proposed 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 packages. 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.

This approach allows us to consider component being developed as an interface between the human operator, the database that contains multiple test results and applications, which performs analysis of the incoming data. Hopefully it will allow our partners and associates to operate most data sets gathered during their studies in more convenient way than now.

Application design is in progress, but we are confident that we will achieve the desired result in a reasonable time.

Acknowledgements

The research was conducted in Tomsk Polytechnic University was supported by RHSF, research project # 15-03-00366: Sociocultural factors of the new industrial modernization in the regions (based on statistical research in the Tomsk region) and by Ministry of Education and Science of the Russian Federation in line with accomplishing the research scientific work on “Evaluation and improvement of social, economic and emotional welfare of elderly people”, contract № 14.Z50.31.0029.

References

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

Bobrova M.V., Berestneva E.V. (2015) Information processing system for social and medical data by data mining, *Proceedings of the VII International Student Science Conference "Student’s Scientific Forum"*

Computer data processing method from: http://denisvolkov.com/wp-content/uploads/2011/03/KMOD-0.pdf (access date: 23.12. 2014)

Barseghyan A.A., Kupriyanov M.S., Frost I., , Tess M.D., Elizarov S. I. (2009) Gregory Pyatetskii-Shapiro, DataMining and information overload *The preface to the book: Data Analysis and Process* W ed. Revised. and add. SPb .: BHV-Petersburg, p.13.

Mandel I.D., Black L.M. (1988) Experimental comparison of algorithms for cluster *analysis Automation and Remote Control,* No 11.

Marukhina O.V., Mokina E.E., Berestneva E.V. (2015) Application of Data Mining methods for identifying hidden regularities in the task of medical data analyzis *Fundamental research*. 4, 107-113

Mehmed Kantardzic (2011) Data Mining: Concepts, Models, Methods and Algorithms– New Jersey, P. 249-253.

Statistical packages review from: http://www.sciencefiles.ru/section/46/ (access date: 27.11. 2014)

Statgraphics Developer's Website. Terms of medium access Statgraphics Online from: http://statgraphics.com/statgraphics_online.htm (access date: 29.12. 2014)

Types of regularities revealed by Data Mining from: http://fsecrets.ru/2010/10/типы_закономерностей_выявляемых_мет/ (access date: 25.12. 2014)

Virtual complex "Political Science" from: http://read.virmk.ru/s/SANZ_SOC/g-014.htm (access date: 20.12. 2014).

Zharkova O.S., Berestneva O.G., Moiseebko A.V., Marukhina O.V. (2013) Psychological computer testing based on multitest portal, *World Applied Science Journal*. 24, 24, 220-224.