Intelligent Data Analysis in the EMERCOM Information System

Elena Sharafutdinova, Tatiana Avdeenko, Maxim Bakaev

Novosibirsk State Technical University, 20, Karla Marksa Ave., Novosibirsk, 630073, Russia

E-mail: bakaev@corp.nstu.ru

Abstract. The paper describes an information system development project for the Russian Ministry of Emergency Situations (MES, whose international operations body is known as EMERCOM), which was attended by the representatives of both the IT industry and the academia. Besides the general description of the system, we put forward OLAP and Data Mining-based approaches towards the intelligent analysis of the data accumulated in the database. In particular, some operational OLAP reports and an example of multi-dimensional information space based on OLAP Data Warehouse are presented. Finally, we outline Data Mining application to support decision-making regarding security inspections planning and results consideration.

1. Introduction

According to estimates by IBM and Intel, the amount of data generated and transmitted daily in the world has the order of exabytes (10^{18}), with an annual increase of about 50%. Appropriately, the analysis of large amounts of data began to develop a separate domain since the beginning of the 2000s, which is currently often called “Big Data”. Such data is not only characterized by large volumes that in some cases render impossible a long-term storage of data, but also have high rate of growth (and therefore, necessity for processing) and diversity of the structure [1]. In some cases the data can even be unstructured, causing the problem of prior structuring, i.e. bringing them to a form in that allows processing with automated data analysis methods. In general, “Big Data” can both originate online (World Wide Web, database search engines, etc.) and be created in the course of the business information systems operation. A review of the field progress and the relevant technologies is provided in [2], while [3] looks back on the development of Data Mining, which is often viewed as the most essential component for successful utilization of Big Data.

One of our previous papers was dedicated to the information system (IS) “Electronic Inspector” that we developed, launched, and supported for the Russian Ministry of Emergency Situations (MES), whose international name is EMERCOM [4]. Its State Supervision Agency (SSA) organizes and performs inspections in fire safety, civil defense, and emergency situations protection. The goal of the IS is “increasing the efficiency of management due to common information and technological structure, effective interaction between all the involved subsystems, citizens and organizations”. The three major categories of target users are general citizens, the MES inspectors, and the MES officials. The latter, who are generally managers from regional centers and headquarters, are among other things responsible for inspections scheduling and are in particular need of decision-making support based on analytical and statistical information that could be extracted from the system’s database.

The MES organizational structure is hierarchical and includes territorially distributed branches at federal, regional, subjects’, and territorial levels. Accordingly, every SSA branch installs its own database, in which its supervised objects' safety conditions and inspection results are stored. With the frequency established by the system's regulations, the data from all the levels of the hierarchy are being sent to the top federal level and are automatically joined into the integrated database (the overall architecture is presented in Fig.1).

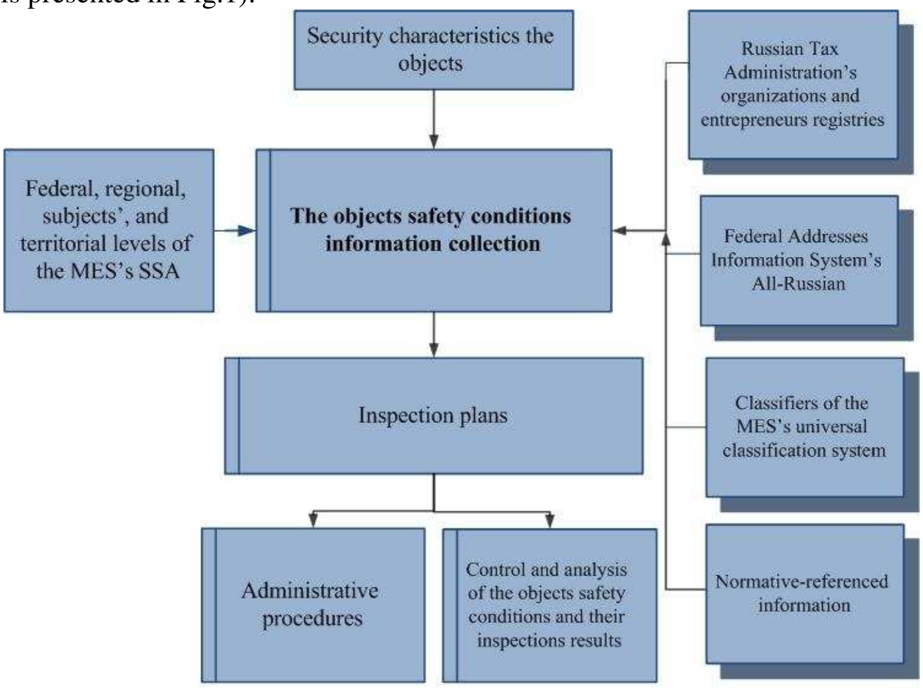

Figure 1. The logical model of the MES system's database.

For example, the "Electronic Inspector" provides the following functionality via its web interfaces:

- Visualization of supervision data in fire safety, civil defense, and anthropogenic emergencies protection for supervised objects;
- Visualization of inspection plans and results data;
- Automated and inquiry-based creation of statistical and analytical reports on safety, with various detail levels: by objects categories, by the SSA branches, by territorial subjects;
- Automated and inquiry-based creation of statistical reports on the SSA activities performance indicators, on the supervision in its authorized fields;
- Creation of statistical and analytical reports on fires and emergencies: the changes in the number of fires and the involved losses, by objects categories, by specified periods (also in comparison with another periods), by specified SSA branches or by territorial subjects.

Currently, more than 15.5 thousand inspectors work to ensure security of more than 2.3 million supervised objects (buildings, economy property, etc.). Thus, annually more than 4 million controlling actions and 7.5 million fire safety measures are performed, which, among other data, are stored in the systems database. So, the data on inspection results accumulated during the several years of the system's operation is already reflected as structured records for about 2 million supervised objects.
Consequently, pre-automation SSA data collection, processing and analysis suffered from high labour-intensiveness, latency and high sensitivity to the quality of preliminary data generalization made at the bottom levels of the organizational hierarchy. E.g., the analysis of supervision activities that was required for operational decision-making was performed quarterly and had latency of 20 days. However, a significant amount of data accumulated in the system calls for the step from purely information system functions (such as collecting, storing, and processing the data related to safety of supervised objects) towards forecasting and decision-making support for the MES officials. Indeed, modern methods of data analysis allow retrieving huge benefits from accumulated data, in particular, improving the efficiency of monitoring [5].

In our paper we propose approaches for supporting decision-making based on data accumulated on safety risks, inspection results, and emergencies. The corresponding data analysis subsystem is still in development, and the data that we have cannot be made public for all supervised objects, so the paper describes an on-going project and only preliminary results. In Section 2, we outline the methods in Data Mining and OLAP (On-Line Analytical Processing) that are used to analyze the data, while Section 3 is dedicated to their immediate application in the MES information system – building multi-dimensional information space model for the accumulated data in the OLAP-warehouse.

2. Methods

Data Mining (DM) can be defined as a process of discovering previously unknown, non-trivial, useful in practice, and interpretable knowledge in raw data, which is useful for decision-making in various human activities. Various classical and heuristic algorithms are used to search for these non-trivial patterns, including neural networks of various architectures, regression analysis, decision trees, evolutionary schemes, etc. However, in practice more important is classification not by algorithms used, but by particular problems specific features. In accordance with these features, several types of patterns discovered with Data Mining methods can be identified [6]:

1. association analysis – if several events are related;
2. sequence discovery – a chain of events linked in time;
3. classification – when indicators defining a group for an object are found;
4. clustering – unlike classification, the groups are not initially set, but the analysis system must find various homogeneous clusters of data;
5. forecasting – generally, its basis is “historical” information, stored as time-series in a database.

The development of DM is not uniform per different domains, and it is said to be problem-oriented [3]. Quite a lot of progress have been made place recently in medicine, where advancement of expert and intelligent systems remains a mainstream [7], and it’s expected that Big Medical Data classification and analysis will be able to significantly decrease the ever-growing health care expenses in developed countries. Extensive review of milestones in the field since the beginning of the millennium can be found in [8], where principal DM approaches are listed as follows: classification, regression, clustering, association and hybrid. Overall, already in this domain the desired goals seem to be agreed upon, the data collection and structuring have been carried out for a long time, and current research is to a significant degree aimed on benchmarking and perfection of various mining methods, compared per metrics of accuracy, sensitivity, specificity, etc. [7]. Such rapidly developing field as Online DM has its own particulars, and further review can be found in [9]. Lately, with the development of information technology, there emerged a popular term “information space”, which is closely linked to the concept of economic space of a region (or enterprise, or industry), as well as the term “knowledge accumulation”. Although there is currently no established definition of information space, it seems to imply the existence of several dimensions (axes) in space. The knowledge accumulation implies not just piling of knowledge, but also its structuring and formalization.

For our data analysis, we chose to use such effective technology as OLAP with Data Warehouse. OLAP technology is built on the basis of developing of commonly accepted OLTP technology (On-Line Transaction Processing), and its benefit compared to other reporting technologies is the possibility to work in interactive mode when analyzing data. In the OLAP-model of the information space, the data are organized as multi-dimensional cubes each of which represents one of the studied processes. The main components of the model are dimensions (qualitative variables) and facts (quantitative variables), which are selected from variables of the analyzed process. Such organization of information space, with possibility of data aggregation by various dimensions, provides a potent mechanism for further analysis and creation of any operative queries. Actually, OLAP can work with any OLTP-databases, but better effect may be achieved with the utilization of multidimensional Data Warehouse (DW) for implementing the subsystem of data analysis in the MES information system.

According to Bill Inmon from Prism Solutions company, “Data Warehouse is a subject-oriented, nonvolatile, integrated, time variant collection of data in support of management’s decisions” [10]. Integration is ensured by the possibility of loading into DW the information, supporting data in various formats and created in various applications – accounting systems, databases, spreadsheet editors, and other office tools working with structured data. When loading the data in different formats, they must be transformed into common representation, integral and consistent. Besides detailed data (transaction), warehouse also contains aggregated data in all necessary measurements and levels of aggregation. Non-volatile means that data must never be changed after they are put into warehouse, only adding new data is allowed. Naturally, to support the history of the data, such dimensions as time and data are introduced in the Data Warehouse architecture. The essence of multi-dimensional data representation is the following. The majority of real processes are described by many indexes, properties, attributes (many dimensions), and if collected into a two-dimensional table, this information would be too complex for visual analysis and understanding. It may even turn out to be redundant, and data extraction process from such a table turns out to be complex and confused. This problem arises because multi-dimensional data are put into a plane two-dimensional table, while OLAP cubes proved to be very effective and graphic tool for representing and operative analysis of data.

The considerable amount of information accumulated in the MES system, makes it possible to perform predictive analysis using knowledge extraction methods with Data Mining technology, implying a deeper analysis than just the operational OLAP analysis. To support and automate MES activities, all the above-mentioned types of Data Mining methods are used or planned to be used. In particular, association, classification, and sequence patterns are used in creating the algorithms of data processing as well as in implementing administrative and procedural activities. The following chapter of the paper describes the application of the described methods in the developed information system.

3. OLAP and Decision-Making Support

3.1. The Information System Rationale

In the course of the MES system development, the following objectives were identified as the major ones in multifunctional analysis of the accumulated information:

- monitoring and analysis of the safety of the supervised objects;
- inspection planning in the supervised territory;
- the execution of administrative procedures on the results of inspection;
- performance analysis to support decision-making.

In the automation of the MES SSA activities, all major types of DM patterns discovery methods are employed. So, the inspection performance algorithm considers the chain of temporary interlinked events and previously accumulated information of the previous inspections results. Briefly, it includes:

1. The inspections scheduling, which is performed with the consideration of the inspected object type; the inspection type (may include different duties and methods); the data on previous inspections – the safety, the implementation of recommendations, the existence of administrative penalties; and the independent assessment of risks involved with the object.
2. The release of the inspection order, which must be confirmed with the Prosecutor Office.
3. Carrying out of the inspection, or cancelling it if legal reasons exist for this.
4. Saving the inspection results in 3-days term. The inspection act is created which contains the discovered violations verified against the interactive normative-referenced information subsystem. It may be supplemented with recommendations for the violations correction in established terms, the administrative cases and penalties, etc.

The same principles are used e.g. in algorithms for administration the violation cases, with monitoring timeframes based on the Russian legislation requirements.

3.2. The Data Analysis Subsystem

For the data analysis subsystem, the subject-oriented Data Warehouse was designed to store the data on the results of planned and unplanned inspections for supervised objects (one inspection is one transaction in the DW). OLAP-warehouse for the data accumulated in the system can be presented as OLAP cube for the process “Inspections” containing two initial dimensions determined by qualitative variables “Inspection date” and “Inspected objects”. Facts of the cube are quantitative variables “The number of inspections”, “The number of violations” containing the detailed and aggregated results of transactions.

Multi-dimensional analysis is defined as the simultaneous analysis along multiple dimensions, and for each dimension consolidation of data can be done. Any direction of consolidation involves a series of successive levels of generalization, where each higher level corresponds to a greater degree of aggregation of data at the appropriate dimension. Thus, the dimension “Inspection date” may be determined following the direction of consolidation “Quarter” and “Year”. The dimension “Inspected objects” is consistently generalized into groups of dimensions of the system hierarchy. As a result, we have a multi-dimensional model of the process “Inspections” (Fig. 2).

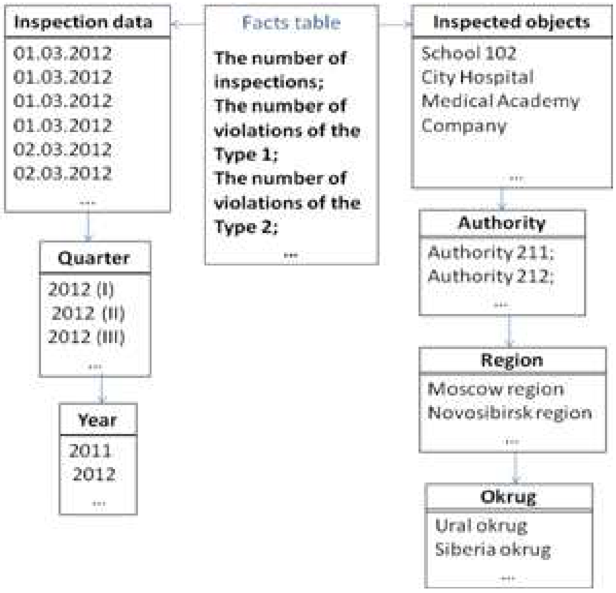

Figure 2. The “snowflake” scheme to describe the “Inspections” process cube.

After creating the warehouse structure and filling it with data accumulated in the system, they could be published in web interface for in-depth analysis. With OLAP, two-dimensional tables are extracted from the multi-dimensional cubes with aggregation using various handling operations with dimensions: rotation, aggregation and detalization, the formation of “slices”, etc. The web page “Official Records” was created in the system, which is basically an interactive query system that allows analyzing selected indicators in the context of time, considering the hierarchical structure of the inspected objects. The user can select the indicators, period and the type of analytical report for any level of the hierarchy. Additionally, the user can select a type of graphical representation, in the system’s rather robust visualization capabilities. For example, in Fig. 3 we provide a table-based comparison of the number of performed (kod_stroki = 3) and saved (kod_stroki = 4) inspections in Siberian Federal District. This analysis type is aimed on discovering discrepancies and allows to control the operation of the system. Thus the quality of MES SSA management and the final safety of supervised objects could be improved, based on the proposed data analysis methods.

4. Conclusions

While one of our previous papers [4] was dedicated to the general issues and experiences related to the development of the information system for the Russian Ministry of Emergency Situations, in our current work we put forward data analysis approaches to the significant amount of data on inspections, violations, and emergencies accumulated in the system’s database. Data Mining intellectual data analysis methodology allows extracting associations, sequences, performing classifications, clustering, and forecasting in the studied processes. We supplement it with OLAP Data Warehouse technology and “information space” concept to cover both operational and deeper analysis. Further, the data analysis subsystem, which is under on-going work in our information system development project, is presented. We outline the OLAP Data Warehouse structure, with cubes, dimensions, and facts – with the resulting interactive query system made available as the data slices in the system’s web interface.

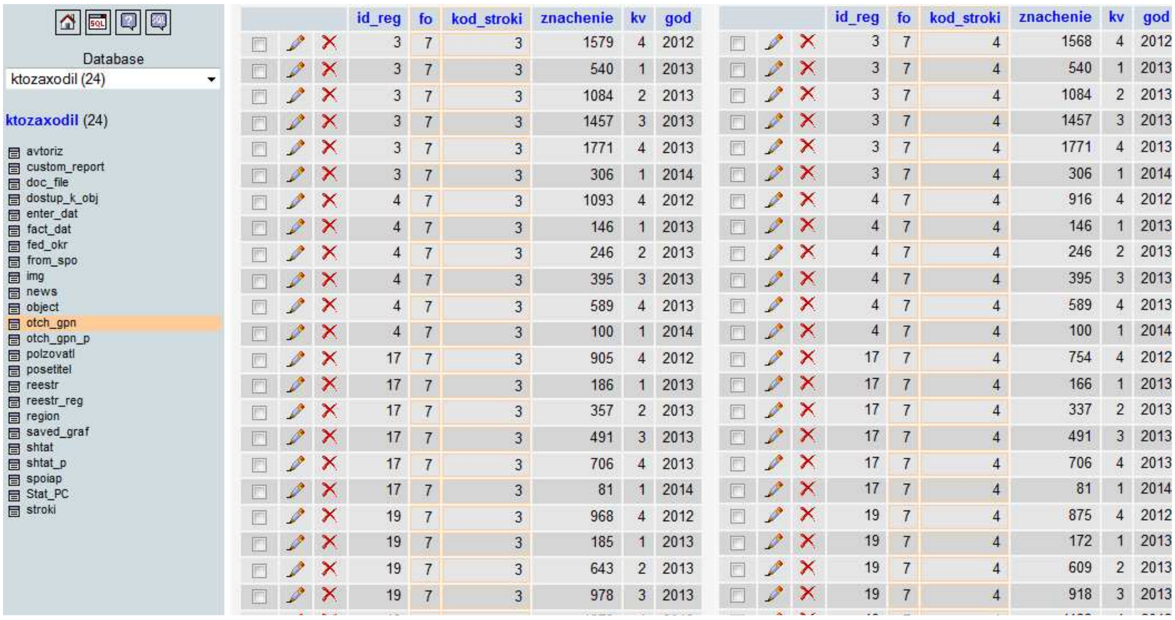

Figure 3. The example of the system's comparable "slice" report.

The research and development on the use of Big Data accumulated in our system are just starting, but they already show good potential for discovering hidden patterns in the data and supporting administrative monitoring and managerial decision-making. E.g., we plan to implement Data Mining technology using pattern recognition techniques to determine associations identifying the "black lists" of inspected objects with critical security state under responsibility of the Russian Ministry of Emergency Situations.

Acknowledgments
This work was supported by Russian Foundation for Basic Research (RFBR) according to the research project No. 16-37-60060 mol_a_dk.

5. References

[1] Beyer M A and Laney D 2012 *The Importance of 'Big Data': A Definition*. (Gartner)
[2] Chen M, Mao S and Liu Y 2014 Big data: A survey. *Mobile Networks and Applications* **19(2)** 171-209
[3] Liao S H, Chu P H and Hsiao P Y 2012 Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Systems with Applications* **39(12)** 11303-11311
[4] Sharafutdinova E, Avdeenko T and Bakaev M 2014 Design and implementation of the integrated information system for the russian ministry of emergency situations. Proc. of the 9 international forum on strategic technology (IFOST-2014), Bangladesh, 4 p
[5] Mayer-Schonberger V and Cukier K 2014 *Big Data: A Revolution That Will Transform How We Live, Work, and Think. (*Eamon Dolan/Mariner Books) 272 p
[6] Edelstein H. 1997 Data mining: exploiting the hidden trends in your data. *DB2 Online Magazine, 1997* **2 (1)** 1
[7] Seera M and Lim C P 2014 A hybrid intelligent system for medical data classification. *Expert Systems with Applications* **41(5)** 2239-2249
[8] Esfandiari N, Babavalian M R, Moghadam A M E and Tabar V K 2014 Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications* **41(9)** 4434-4463
[9] Bakaev M and Avdeenko T. 2016 Prospects and challenges in online data mining: experiences of three-year labour market monitoring project. *Lecture Notes in Computer Science, 9714: Data Mining and Big Data*, 15-23
[10] Inmon B 1992 *Building the Data Warehouse. 1st Edition* (Wiley and Sons)