UDC 004.89

Kravets Yu., Kravets R.
Information Systems and Networks Department,
Lviv Polytechnic National University
S.Bandery Str., 12, Lviv, 79013, Ukraine
E-mail: rkravets@ridne.net

KNOWLEDGE-BASED DATA ANALYSIS TECHNOLOGY AND ITS APPLICATIONS

Authors consider the problem of the text documents analysis system development on the basis of knowledge-based data analysis technology and its application to the emergency situations descriptions analysis. The structure of the system and application on the example of heat supply is described in the paper.

Keywords – intelligent data analysis, systemological classification analysis, conceptual classification model, textual data warehouse, intelligent information retrieval system.

1. Introduction

Development of civilization causes development of technologies and imposes the imprint on an environment, the risk of emergency situations (ES) arising grows constantly. That is why new approaches to the preventing of ES arising through deployment of automated collection, processing and analysis of information on the basis of modern computer technologies with the advanced information-analytical support tools are needed.

Any city has utility enterprises, therefore without their permanent functioning it is difficult to present existence of whole city infrastructure and life of its habitants. A communal economy of Ukraine, in particular enterprises of heat supply, is in the difficult state enough, which is characterized small, often insufficient, financing and wearing out of equipment which remained yet from times of USSR. For today wearing out of equipment is the factor of heightened risk, because it can become the reason of breakages and spoilage of equipment and, as a result, arising of ES. In the conditions of the insufficient financing it is important to be able to estimate risks and distribute tools above all things on the removal of the most credible sources of ES arising. That is a problem, which requires the use of expert knowledge – deep knowledge and considerable experience of tasks decision in a certain problem area (PA). Not on every enterprise, especially in small towns, there are experts who perfectly know the area of ES and have a considerable experience of decision-making (DM). It causes the necessity of knowledge-based information technology deployment for providing of expert DM experience distribution for the heat supply enterprises efficient functioning.

Applying knowledge-based technology allows storing and using in the DM process the expert knowledge in the certain problem area. For efficient informational and analytical support of decision-making process in the ES area the task of intelligent system (IS) for forming of measures on warning or liquidation of ES consequences development and forming of list of most typical ES was put.

2. A review of the last researches and purpose of work

Available information enlargement leads to appearance of intelligent information processing technologies, which allow a man to fill up the knowledge about PA. Application of methods and tools of knowledge discovery in databases (KDD), knowledge management, and business intelligence allows using present information for the tasks resolving, converting the enormous volumes of information into knowledge [1, 2].

For the KDD realization the wide spectrum of methods is used. Part of methods was specially developed for the intelligent data analysis (search of association rules and sequential patterns). Other ones were developed for the different problems solving (Bayesian methods, neuron networks, case-based reasoning, decision trees and rules induction, genetic algorithms, applied statistical analysis methods etc.) [1]. It caused to a fragmentation in offered KDD approaches. Considerable part of existent methods is intended to the analysis of numeric data. In the data analysis process quantitative (numerical) patterns are used only but PA semantics and knowledge (including concepts) is not taken into account.

Existent tools is intended to ES reports registration, classification, notification about arising, liquidation plan forming and situation development prediction, task complex management support in the ES arising area. That software is dedicated to the data capture, however practically do not provide the analysis of the collected data that would allow using past experience for more efficient ES prediction, anticipation, as well as liquidation and decline of level of their consequences. The use of intelligent data analysis (IDA) tools to existent ES text descriptions for the patterns discovery in PA, which in future can be used for the tasks resolving, is describes in the paper.

The purpose of our research is to realize knowledge-based intelligent text data analysis method on the basis of systemological approach [3, 4] as an intelligent text documents warehouse analysis system that will allow using existent experience for more efficient tasks solving. Application of this method provides higher results quality, because discovered patterns take into account substantial objects properties and the resulting set of meaningful rules diminishes [5].

3. Information model of text data warehouse analysis process

The knowledge discovery process consists of such generally accepted stages: planning, data preparation, data analysis, interpretation and use of searched patterns.

At the first stage the aims of analysis are determined, a priori knowledge about PA are formed, information sources are choused. For deeper PA study and taking into account its deep semantics it is suggested to use systemological classification analysis (SCA) and to represent knowledge as conceptual classification model (CCM) of PA (multiple-aspect generic-specific classification of PA objects).

SCA is based on the use of natural classification criteria and allows to represent in classification deep knowledge about PA, substantial properties of the designed objects, to provide the classification by explanatory and prediction force. Application of natural classification schemas provides support of making optimal decisions, receipt of conclusions and recommendations, which promote efficiency and adaptiveness of the knowledge-based systems and technologies.

On the second stage classification of messages collected in a warehouse is conducted in accordance with PA CCM. For this purpose the methods of linguistic analysis and case-based reasoning are used.

On the stage of data preparation at first overhead information which is not used in the process of analysis is removed, for example, blocks of labels, HTML- tags, and then the algorithms of morphological and syntactic analysis of natural-language data are used.

The data analysis method, which takes into account PA CCM, is used for pattern search in the text messages warehouse. A method is based on the vehicle of statistical hypotheses verification. It means that it follows to understand the found patterns as statistically meaningful correlations between properties of PA objects [6].

The use of PA CCM for knowledge discovery gives next advantages. At first, patterns, which take into account substantial properties of PA objects, uncover in the analysis process. Secondly, such approach allows to get as a result the less set of meaningful rules. It is due to that first of all it is founded rules that include properties at top objects and properties hierarchy levels, i.e. at those levels, where the number of possible property values is lesser. As a result, taking into account of PA CCM in the knowledge discovery process improve the result quality and decrease costs of result processing.

4. Structure of the ES descriptions intelligent analysis system

The ES descriptions intelligent analysis system is intended to the search of existing dependences in data. The analysis of results will allow to use existent experience for anticipating and more efficient liquidation of ES consequences.

There are input data of the system:

1) the set of the short semistructured text messages by a natural language;

2) the CCM which describes the hierarchy of PA objects.

Messages have the following structure:

< DATE OF ES; TIME OF ES; PLACE OF ES; ES TYPE; text description >

The official site of the Ministry of Emergency Situations is an information source for the system.

An output data are existent dependences between information units of messages.

We will consider ES descriptions for some period of time as text documents warehouse, because it meets requirements which are put in claims to the data warehouses [2]:

- subject orientation (data are stored in accordance to areas, which it describe, instead of to applications which use it);
- time dependency (an attribute of time always is in the structure of data warehouse);
- invariance in time (getting in data warehouse information does not change already).

Based on the structure of data warehouse analysis process and necessary functionality of the system offered in a previous section, we will mark out the followings basic blocks and their functions:

- IDA subsystem – rules induction;
- PA CCM processing subsystem – creation, storing, opening, viewing, editing of CCM (nodes addition and deletion), hierarchy processing (receiving of generic and specific concepts);
- natural language processing subsystem – morphological and syntax analyzer;
- utility functions subsystem – file opening, files with ES descriptions preparation, results visualization and storing.

The structure of the system is given on fig. 1.

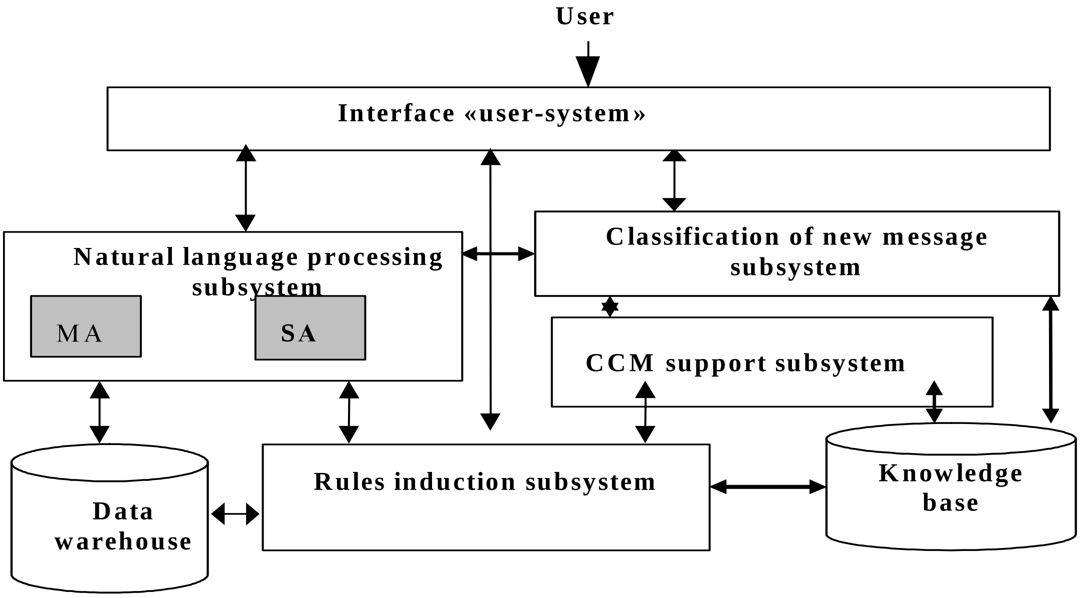
User
Interface «user-system»
Natural language processing subsystem
MA
SA
Classification of new message subsystem
CCM support subsystem
Data warehouse
Rules induction subsystem
Knowledge base

Fig. 1. Structure of the ES messages intelligent analysis system.

5. Application of the text messages intelligent analysis system in heat supply enterprise

For the decision-making support in a heat supply enterprise using text document warehouse intelligent analysis system developed on the basis of knowledge-based technology, were realized:

- knowledge base, which contains the structured information about ES in heat supplying;
- knowledge base information processing tools for DM;
- ES descriptions analysis and forming of the generalized rules tools.

Knowledge base of IS contains the fragment of CCM of ES in heat supplying, which was built by the SCA method on the basis of ES class hierarchy [5]. Most typical ES in the area of heat supplying is: heat supply violation, heating system failures, switching-off of the heating system and fuel oil leakage. The fragment of ES CCM was built.

In the DM process person who makes decision, for example, engineer on a utility enterprise, can use different aims and choice criteria. The purpose of choice in heat supplying can be, for example, choice of the most effective measures on liquidation of ES consequences, most effective measures on ES anticipating and others like that. Thus the most typical choice criteria will be: minimization of expenses or time on liquidation of ES consequences, increase of measures on liquidation of ES consequences efficiency, improving of measures on ES anticipation efficiency, determination of most typical ES for the set time period and others like that.

Use in the DM process the ES CCM, built on natural classification principles [3,5], allows to carry out an ES analysis and prediction in the DM process in heat supplying and to take into account different aims and choice criteria. In particular, at the construction of ES CCM such functional properties are taken into account in heat supplying, as *“negative influence”* (ES reason), *“violation of permanent process”* (ES consequence) and *“urgent measures are needed”* (ES development process dynamics) [5].

As there are two types of IS output data, which can be DM basis in case of difficult situations occurring – information receiving from ES CCM in the DM process and forming on the basis of separate ES texts descriptions the generalized rules – ES intelligent analysis system in heat supplying must provide implementation of the followings functions:

- ES CCM processing: creation, storing, opening, reviewing, editing of CCM (node addition and deletion), receiving of generic and specific concepts, classification of new ES descriptions;
- patterns (generalized rules) search and visualization;
- utility functions: files opening, preparation of files with ES descriptions, results storing.

ES intelligent analysis system deployed on the district utility enterprise “Zmiiv heat supply enterprise”. In the program system, accepted in exploitation, it is realized: ES CCM in the heat supplying area; an algorithm of intelligent data analysis for generalized rules forming; model and structure rules storing; a method of emergency situations development prediction and decision-making support on the basis of ES CCM and rules in it liquidation.

6. Conclusion

The ES descriptions intelligent analysis system allows to search dependencies in data, which describe existent experience of liquidation of ES consequences. The analysis of results helps in deeper PA understanding, and consequently more effective tasks solving which arise in this area. Application of systemological approach in the process of the intelligent systems development promotes quality of knowledge base, and consequently quality of task solving, because represent in a PA conceptual model and knowledge base not only superficial but also deep knowledge about PA.

Proposed conceptual classification model and program system allowed to promote decision-making quality in supernumerary situations and efficiency of preventive measures realization on the decline of emergency situations arising possibility due to the use of the structured ES description on the CCM basis, to decrease expenses on liquidation of ES consequences, and also to provide support of engineer of heat supply enterprise during determination of aims and anticipation and ES consequences liquidation measure choice criteria.

References

[1] Advances in knowledge discovery and data mining / Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (editors). – AAAI/MIT Press, 1996.

[2] Han J., Kamber M. Data Mining: Concepts and Techniques. – Morgan Kaufman Publishers, 2000.

[3] Solovyova E.A. Mathematical Modeling of Conceptual System: Method and Criteria of Natural Classification // Automatic Document and Mathematical Linguistics. Allertion Press, New York, 1991, V. 25, No. 2, P.44-56.

[4] Bondarenko M.F., Matorin S.I., Solovyova E.A. Analysis of Systemological Tools for Conceptual Modeling of Application Fields // Automatic Document and Mathematical Linguistics. Allerton Press. New York, 1997, V. 30, No.2, P. 33-45.

[5] Соловьева Е.А. Естественная классификация: системологические основания. – Харьков: ХНУРЕ, 2000, 222 с.

[6] Кравець Ю.Н. Застосування концептуальних класифікаційних моделей для підвищення якості результатів аналізу даних // АСУ і прилади автоматики, 2004, № 3, С. 41-47.