Experimental software modeling of knowledge acquisition processes for automated knowledge bases construction in dynamic integrated expert systems

G.V. Rybina[1], A.A. Slinkov[1], D.R. Buyanov[1]

[1]National Research Nuclear University «MEPhI», Kashirskoe shossr 31, Moscow, Russia, 115409

Abstract. This paper analyzes the results of the experimental research of automated knowledge base construction for dynamic intelligent systems, in particular dynamic integrated expert systems on the basis of the so-called original combined method of knowledge acquisition with temporal extensions. The focus of this work is on some aspects of the application and development of technologies of knowledge acquisition from various sources (experts, NL-texts, data bases) in order to create new applied intellectual technologies that can be used, for example, in the field of healthcare (personalized medicine, "smart" hospital, etc.).

1. Introduction

The modern world is characterized by a significant acceleration of the introduction of methods and technologies of artificial intelligence (AI) in various sectors of the economy, production, social environment, etc., so that, according to experts' estimates, the growth of the world economy by 2024 will be at least 1 trillion U.S. dollars. Priority development directions and use of AI technologies, which are defined by the Decree of the President of the Russian Federation of 10.10.2019 (№490), contribute to the creation of favorable conditions for a significant improvement of the quality of services, including the field of healthcare (diagnosis and forecasting of risks of disease, selection of the optimal dosage of drugs and means of rehabilitation, preventive examinations, reducing the threat of pandemics, etc.).

An important place among the priority directions is given to the development of software that uses such basic AI technologies as: technologies of knowledge acquisition from various sources and intellectual analysis of Big Data; technologies of forecasting and decision support; technologies of planning and multi-agent control of targeted behavior in unstructured environments; technologies of processing natural languages (NL), etc.

Nowadays, based on these basic technologies, it is necessary to develop applied AI technologies that will be used in various fields of life. The focus of this work is on some aspects of the application and development of technologies of knowledge acquisition from various sources (experts, NL-texts, data bases) in order to create new applied intellectual technologies that can be used, for example, in the field of healthcare (personalized medicine, "smart" hospital, etc.).

2. General description of basic technologies of knowledge acquisition from various sources

As shown in [1,2], the problems of obtaining knowledge from sources of knowledge of various typologies (experts, NL-texts, DB), as well as the issues of creating effective technologies of automated knowledge acquisition are still in the focus of modern intellectual systems developers' attention, in particular, the most popular integrated expert systems (IES) with scalable architecture and extensible functionality [1].

The analysis conducted in [3-5] has shown that the most acute problem of knowledge acquisition arises when solving complex practical problems in such areas as energy, space, ecology, etc., as well as in the social field, for example, in the field of healthcare, where significant amounts of data have been accumulated, for which various Big Data technologies are used today [6], and when building knowledge bases (KBs) in intelligent systems such as the IES, several experts or groups of experts are required, which significantly increases the cost and time parameters of system development in the absence of automated means to support the process of obtaining knowledge from the expert / experts, who are the main knowledge source.

Currently, the typology of knowledge sources is no longer limited only to experts because significant amounts of expert knowledge are accumulated in the NL-texts and in the information accumulated in the databases (DB) of modern business information systems. Problems of obtaining (revealing) knowledge from NL-texts are related to the rapidly advancing technology Text Mining [7,8], and various methods and algorithms of automated knowledge acquisition from DB are included in the technology Data Mining / Deep Data Mining, Knowledge Discovery in Databases (KDD), etc. [9-11]. Success of Text Mining technology is connected with different aspects of application of textual methods of obtaining knowledge from NL-texts, which have received the greatest development in three types of modern web-oriented NL-systems - information retrieval, information extraction and understanding of NL-text (Text / Message Understanding) [2,5].

Each of the above-mentioned technologies has emerged and developed independently and today such autonomy and distribution does not allow to carry out effective monitoring of such information resources as KB, DB, and in recent years ontologies possessed by intellectual systems, in particular, IES [2]. However, at present, except for the works [2-5,12,13], there is practically no research in the field of creation of tools and technologies of distributed knowledge acquisition from various sources.

Experience of practical use of a number of applied IESs, including the most complex dynamic IESs [2,3,5], etc, developed on the basis of problem-oriented methodology (author G.V. Rybina) [1] and supporting its software - AT-TECHNOLOGY workbench, for such problem domains (PD) as express-diagnostics of blood, diagnostics of complex technical systems, design of unique mechanical engineering objects, complex environmental problems, etc., showed the effectiveness of the joint use of three sources of knowledge - experts, NL-texts and databases. For example, the analysis of experimental data obtained in the course of creation of KB of several applied IESs using the combined method of knowledge acquisition (CKAM) [1-5, 12, 13], which is an integral part of this methodology, has shown that the local use of the database as an additional source of knowledge can supplement the volume of developed BDs by 10-20%, depending on the specifics of PD.

Now on the basis of the problem-oriented methodology the intellectual program technology and the automated workplace of the knowledge engineer - the complex AT-TECHNOLOGY on which basis, including with use of three versions of means of support CKAM (local, distributed, dynamic), more than 20 applied IES for static PD are developed, and also prototypes of the most difficult dynamic IES, i.e. the IES using dynamic representations of a subject domain and solving dynamic problems [1,2] are created.

The modern stage of development of the technology of knowledge acquisition from various sources in the context of creation of a dynamic version of the CKAM and its support tools, functioning as part of the new generation of tools such as WorkBench - AT-TECHNOLOGY complex - is associated with the automation of the processes of obtaining, presenting and processing of temporal knowledge for the construction of KB in dynamic IESs. Relevance of the research is related to the fact that, despite the existence of a significant number of approaches to the representation of temporal dependencies in the context of automatic processing of NL-texts, the issues of obtaining temporal knowledge (both using manual techniques and automated methods) for constructing temporal KB in dynamic intelligent systems, in particular, in dynamic IES [5], are practically not considered.

The aim of this work is to present new results of experimental software modeling of the processes of temporal knowledge acquisition for the automated construction of KB in dynamic IES (for example, medical diagnostics).

3. Features of the combined knowledge acquisition method and means of its implementation

The basic version of CKAM [1, 2, 12] and its support facilities, created in the end of 1990s, are constantly being developed and successfully used to automate the processes of KB development in static PD, forming the core of the automated workstation of the knowledge engineer on the basis of the software platform AT-TECHNOLOGY. Today a distributed variant of computer knowledge acquisition is supported [12,13], which provides integration of three types of knowledge sources (experts, NL-texts, DB) taking into account their geographical distribution within the client-server architecture. Proceeding from the context of this work, we will focus only on those features of CKAM, which are most important from the point of view of studying the possibilities of CKAM development for the purpose of automated construction of temporal databases for dynamic IES.

The first peculiarity of CKAM is the way of organizing the process of direct obtaining of knowledge from experts by means of computer interviewing at all stages of the life cycle of the IES construction on the basis of the author's approach "orientation to the model of solving a typical problem" [1], according to which the controlling knowledge about strategies (methods) of solving specific classes of problems solved in a similar way is formed in the form of some heuristic model of a typical problem [1] (diagnostics, design, planning, etc.). Therefore, the processes of obtaining knowledge are controlled by means of sets of models for solving typical tasks, for which a number of methods and approaches have been developed and are constantly being developed, which allow to create scenarios of dialogues with experts, reflecting both the thematic structure of the dialog (i.e. the scheme of solving a typical task [1,2]) and the local structure of the dialog (steps of the dialog [1,2]), i.e. a set of specific actions and reactions between the expert and the system.

Thus, the processes of knowledge acquiring from experts and NL-texts are computer modeling, allowing in the process of dialogue with experts to build and denote all the components of the model of solving a typical problem and form both fragments of the knowledge field [1,2] (intermediate representation of structured knowledge used to verify information obtained from various sources), and the corresponding fragments of KB. To build an "action-response" scheme of partners, several implementation techniques are used, in particular, the "simulation of consultation" method, etc.

Expert interviewing processes are supported by a dialogue script interpreter, with each scenario corresponding to a specific type of task. In addition, special screen forms are provided for entering unreliable knowledge [1] (uncertainty, inaccuracy, fuzziness) and connecting the means of implementing the adaptive method of repertoire grids [1] (for example, for the implementation of procedures of differentiation of diagnoses in the case of activation of the scenario for the task of medical diagnosis). A specialized linguistic processor and a set of dynamically updated dictionaries occupy an important place in the software for supporting basic and distributed CKAM (linguistic aspects of CKAM are described in detail in [1,2,12]).

Another important feature of the CKAM is the integration of closely interrelated processes of expert computer interviewing with methods of NL-texts processing (entered both during the interview session and after the end, in the form of expert interviewing protocols), as well as methods of knowledge acquisition from the DB [1,12,13].

In order to use temporal DB [9,10] as an additional source of knowledge, the basic functionality of the means of supporting the distributed version of the CKAM has been significantly expanded by developing new algorithms of knowledge acquisition from temporal DB and means of integration of various sources of knowledge (experts, NL-texts, DB) [1,12,13]. Instead of the CART algorithm used in the local version of the basic CKAM [13], the well-known Random Forest algorithm [11] was implemented, modified to support the work with temporal DB.

The essence of the modification was the use of multivariate feature space, one of which is a time stamp. The ensemble of solution trees is constructed in accordance with the basic algorithm; however, the calculation of the value of the partition criterion has undergone changes due to the use of the multidimensional space of characters (the partition criterion will be the arithmetic mean of the calculated values of information entropy). In addition, unlike the decisive trees based on modified CART and C4.5 algorithms [1,12,13], the tree construction is performed until all the elements of the sub-sample are processed without the application of the branch cutting procedure. The solution tree algorithm is executed as many times as necessary to minimize the error of classification of objects from the test sample (classification of objects is done by voting by analogy with basic version of the Random Forest algorithm [11]).

Thus, the actual problem of the current stage of research is the further evolution of the CMKA, in order to develop methods and means of automated construction of temporal KB in dynamic IES. To date, models, methods and software for representation and processing of temporal knowledge have already been developed and tested in the creation of several prototypes of dynamic IES models [2-5,14]. Below is a description of the current results of the experimental program modeling of the temporal version of the CMKA.

4. Analysis of current results of experimental software modeling

As it has been noted above, for modeling of processes of direct knowledge acquisition from experts and NL-texts (sublanguage of business prose[1,2]) the typical problem - medical diagnostics - was used, and as PD the complex diagnostics of diseases of a mammary gland and diagnostics of traumas of a knee joint was considered. Model dialogues were conducted in the form of a "language experiment" [2,4,5] related to the search for temporal information, i.e. temporal relations both within each NL-proposal coming from the expert and/or in neighboring sentences (taking into account the current state of the local structure of the dialog) and with the search for relations indicating the time of text creation.

For these purposes, the dictionary of temporal lexemes developed on the basis of the works [15,16], a specialized linguistic processor and interviewing support tools functioning as a part of the AT-TECHNOLOGY complex were used. Scenarios and corresponding screen forms were developed and tested with the help of model dialogues.

In total, several hundred modeling sessions of interviewing were implemented with the participation of about 80 students who, according to the principle of "doctor to himself", introduced lexemes (temporary pretexts, target pretexts, causal pretexts, particles, adverbs of time, etc.) into the corresponding screen forms to build fragments of the knowledge field. On the basis of the experiments carried out, a set of modified scenarios describing the thematic and local structure of the dialogue when solving a typical problem of medical diagnostics was obtained, which made it possible to implement the elements of the "through" technology of direct acquisition and representation (in terms of an extended language of knowledge representation [2]) of fragments of temporal KB, ready to implement temporal output on the production rules [2, 14].

Thus, the use of a set of model dialogues made it possible to experimentally determine which temporal entities (markers) [5] can be detected on the basis of algorithms and software of the temporal version of the CMKA and significantly add to the current temporal lexeme vocabulary. Fig. 1 shows some results in the form of statistical data obtained as a result of "language experiments".

Another complex of experiments was carried out with the modified Random Forest algorithm, which is a part of the CMKA and is the core of knowledge acquisition from temporal DB. As an input to this algorithm, some medical temporal DB containing data in a certain format was used, and the set of medical data was exported to the database under the control of SqLite 3, and then to a separate table with the allocation of identifiers with assigned classes. Thus, a table with objects is formed, which contains their attributes at each moment of time, and a table with classes. The Random Forest algorithm builds an ensemble of trees according to this temporal DB, where each committee tree allocates the classified object to one of the classes, i.e. it votes, and wins the class for which the largest number of trees voted. A fragment of the knowledge field containing the rules of tree voting in intervals and rules in the extended language of knowledge representation is built on the trees.

The obtained fragment is suitable for further verification and integration [1,2,12,13] with fragments of the knowledge field obtained as a result of expert interviewing sessions [1,2,12,13].

Below are the results of experimental research of algorithms and software tools for the formation and verification of knowledge field elements with temporal entities. For these purposes, modal dialogues were used for medical ultrasound diagnostics and diagnostics of the knee joint. Experiment scenarios were used, including:

- addition of new events and intervals to the model dialogs, due to which the vocabulary of temporal lexemes, previously not detected in experiments carried out with the support of the algorithm for identifying temporal markers, was replenished (positive testing);
- inclusion of events and intervals in the knowledge fields without references and/or with incorrect values, to test the reaction of means of supporting the verification of the knowledge field to anomalies (negative testing)
- the use of synonymous events and intervals for subsequent experimental research of means for combining elements of the knowledge field obtained from sources of various typologies.

Figure 1 below shows examples of pie charts that display the quantitative result of these experiments.

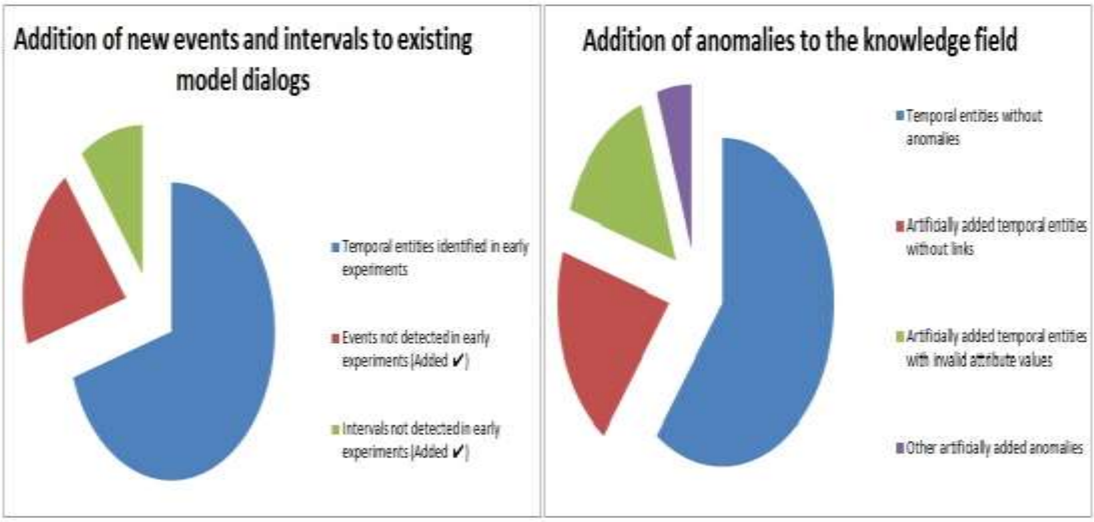

Figure 1. Pie charts with experimental results.

Now the results of experiments are presented with other sources of knowledge - temporal databases. The experimental scenario included:

- registration of the temporal database in the dynamic version of the AT-TECHNOLOGY workbench (registration means adding a file containing the temporal database to the directory where the executable file is located);
- opening the database and reading data stored in the database, namely: identifiers of objects, classes and timestamps.
- creation of files with serialization of the ensemble of trees, a description of the knowledge field in the extended language of knowledge representation, as well as a description of the knowledge field in the internal representation.

Table 1. The table contains the name of the classifying attributes with possible variants of their values.

№	Echostructure	Echogenicity	Circuit	Inclusion	Size
1	Cystic	Anechogenic	Smooth	No inclusions	1cm
2	Almost completely cystic	Hyperechoic	Unable to determine	Large comet tail artifacts	1 cm to 1.5 cm
3	Spongy	Isoechogenic	Lobed	Macrocalcinates available	From 1.5 cm
4	Mixed solid cystic	Hypoechoic	The appearance of extra-thyroid distribution	Peripheral calcification	
5	Solid				
6	Almost completely solid				

It should be noted that the Random Forest algorithm was tested on several temporal databases that have the same structure, but a different number of objects, classes, timestamps, and also with different formats of timestamps. In addition, during testing, ensembles of 1, 50, and 100 trees were built.

Consider in more detail some of the features of temporal databases used in experiments. The temporal database is represented as a set of two tables (Objects, Classes), in turn Objects (id, attrN, time_stamp), where id is the identifier column, attrN is the column with the attributes of the object, time_stamp is the column for storing the timestamp, and Classes (id, class), where id is the identifier column, class are the classes to which objects can belong. The following is a description of the model temporal database in Table 1.

Comments on table 1: Class attribute: 0/1 (biopsy required / not necessary). Number of objects: 140

Figure 2 below shows the mapping of a temporal database into elements of a knowledge field in an extended language for representing knowledge.

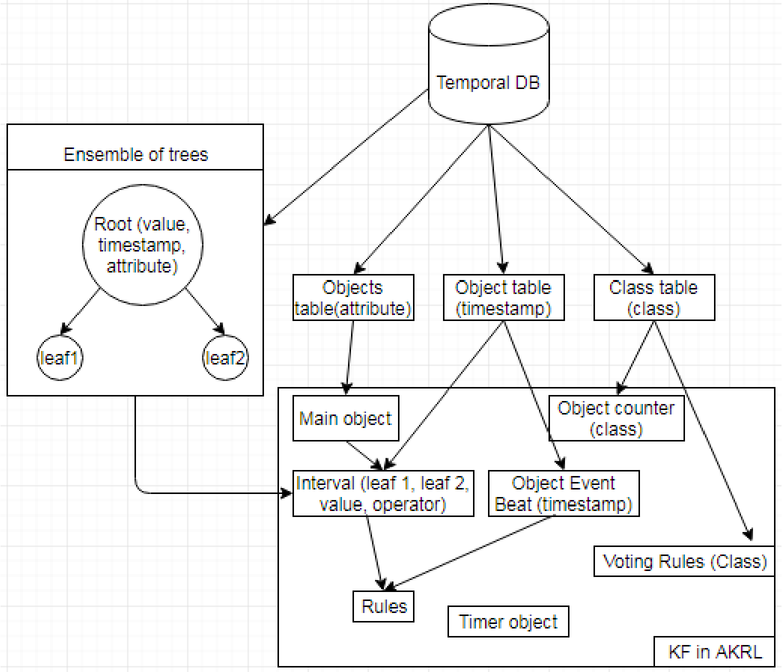

Figure 2. Mapping of the temporal database to the elements of the knowledge field in the extended language of knowledge representation**.**

5. Conclusion

The developed methods, algorithms and technologies of temporal knowledge acquisition from various sources (experts, NL-texts, DB) are especially important for medical PD, where significant volumes of temporal information are accumulated even about one patient, including all his previous conditions and diseases in a wide time range. This is a crucial task, which is necessary to improve the quality of healthcare in our country.

6. Acknowledgment

This work was supported by RFBR (project №18-01-00457).

7. References

[1] Rybina, G.V. Theory and technology of building integrated expert systems – M.: Nauchtekhlitizdat, 2008. – 482 p.

[2] Rybina, G.V. Intelligent systems from A to Z. a series of monographs in 3 books. Book 2. Intelligent dialog systems. Dynamic intelligent systems – M.: Nauchtekhlitizdat, 2015. – 160 p.

[3] Rybina, G.V. Moderne arkitekturer af dynamiske intelligente systemer: problemet med integration i moderne tendenser / / anordninger i systemet. Ledelse, Afprøvning, Diagnostik. – 2017. – № 2. – P. 1-12.

[4] Rybina, G.V. The combined method of automated temporary collection of information for the development of knowledge bases of intelligent systems / G.V. Rybina, I.D. Danyakin // Materials of the 2nd International Conference on the Development and Application of Knowledge, 2017. – P. 117-123.

[5] Rybina, G.V. Dynamic integrated expert systems: technology for automated acquisition, representation and processing of temporary knowledge // Information measuring and control systems. – 2018. – Vol. 16(7). – P. 20-31.

[6] Tsvetkova, L.A. Implementation of big Bath technologies in healthcare: assessment of technological and commercial prospects / L.A. Tsvetkova, O.V. Cherchenko / / Economics of Science. – 2016. – Vol. 2(2). – P. 139-150.

[7] Aggarwal, S.S. Mountain text data / S.S. Aggarwal, S. Zhai – Springer, 2012. – 535 p.

[8] Pusteyovsky, J. TimeML: reliable specification of events and temporary expressions in the text / J. Pusteyovsky, J. Castano, R. Ingria, R. Sauri, R. Gauzauskas, A. Setzer, G. Katz // Materials on new directions in answering the question, 2003. – P. 28-34.

[9] Kaufmann, M. Timeline index: undefined data structure for processing requests for temporary data in SAP HANA / M. Kaufmann, A. Manjili, P. Wagenas, P. Fisher, D. Kossmann, F. Fireber, N. May // SIGMIOD, 2013.

[10] Ishaq, V. Extraction of temporary formation data using sliding window technology / V. Ishaq, V. Hussein, K. Mahamud, K. Ruhana, M. Norawi // CiiT International Journal of Mining and Knowledge Engineering. – 2011. – Vol. 3(8). - P. 473-478.

[11] Tsacheva, A.A. MR is a random forest algorithm for detecting distributed action rules / A.A. Tsacheva, A. Bhagavati, P.D. Ganesan // International Journal of Data Mining & Knowladge Management Process. – 2016. – Vol. 6(5). – P. 15-30.

[12] Rybina, G.V. Pipe acquisition method, known for building bases, known for integrated expert systems/ / devices and systems. Upravlenie, Kontrol', Diagnostika. – 2011. – Vol. 8. – P. 19-41.

[13] Rybina, G.V. Processing the background of dlya automated postroeniya integrated expert system / G.V. Rybina, A.O. Dejnenko // Artificial intelligence and decision making. – 2010. – Vol. 4. – P. 55-62.

[14] Rybina, G.V. Implementation of temporal vivode in dynamic systems integrated expert systems / G.V. Rybina, A.V. Mozgachev / / Artificial intelligence and decision making. – 2014. – Vol. 1. – P. 34-45.

[15] Efimenko, I.V. Semantics of time: concepts, methods and algorithms of identification of automatic natural language processing system // Bulletin of the Moscow state regional University series " Linguistics" – M.: Publishing House of Moscow State University, 2007. – Vol. 2.

[16] Arutyunva, N.D. Logical analysis of the language: Language and time / N.D. Arutyunva, T.E. Yanko – M.: Indrik, 1997.