Distributed Knowledge Acquisition Control with Use of the Intelligent Program Environment of the AT-TECHNOLOGY Workbench

Galina V. Rybina and Yury M. Blokhin

National Research Nuclear University MEPhI
(Moscow Engineering Physics Institute), Russia
galina@ailab.mephi.ru

Abstract. The paper discusses the problem of distributed knowledge acquisition for the construction of complete and consistent knowledge bases in integrated expert systems via the sharing of knowledge sources of different topologies (the focused in this work databases as electronic media, experts and problem-oriented texts). The work is focused on models and methods of distributed knowledge acquisition from databases as additional knowledge sources and automation of the process by using an intelligent program environment. Special typical design procedure, called "Distributed knowledge acquisition" is reviewed, which provides synchronization ofdistributed knowledge acquisition processes. This procedure uses the technological knowledge base of the intelligent planner of the AT-TECHNOLOGY workbench and special program tools.

Keywords: Distributed knowledge acquisition, task-oriented methodology, AT-TECHNOLOGY workbench, intelligent program environment, intelligent planner, typical design procedures, reusable components, integrated expert system, IES, knowledge base.

1 Introduction

The problem of knowledge acquisition has been the focus of attention for the devel-opers of current intelligent systems, among them traditional expert systems (ESs) and more complicated integrated expert systems (IESs) with scalable architecture and expandable functionalities [1]. The main subject of this work is this prime course of artificial intelligence.

The results have been widely presented in both foreign papers [2] and domestic ones [1,3,4,5]. Nevertheless, the problems of the practical use of traditional methods for knowledge acquisition and the development of automated technology of knowledge acquisition are currently topical. This is because of the severe lack of experts and custom computer systems that simulate expert skills. The research into cognitive psychology shows that the newcomer to expert path takes over 10 years, due to the long professional practice required for adaptation to successful problem solving [6].

P. Klinov and D. Mouromtsev (Eds.): KESW 2014, CCIS 468, pp. 150–159, 2014.

In solving complicated practical problems, the most critical problems of knowledge acquisition arise in medicine, power energetics, space, ecology, and so on, where the opinion of one expert is not enough.

Therefore, to construct the most complete and consistent models of problem do-mains (PDs) and to reduce the risks of expert errors, a group of experts needs to be attracted, which significantly increases the cost and time parameters of IES design [1]. Thus the urgency and the role of expert labor automation, as well as the development of custom software increases. There is also increasing role of different "acquisition shells" directed to the support of knowledge acquisition from the experts or an expert groups. This knowledge is the basic source (first type knowledge source [5]).

However, there have been few investigations into grouped knowledge acquisition from experts. Among the best known of these are papers of a theoretical and meth-odological nature [6,7] and the foreign project that describes the facilities of graphical representation of distributed knowledge [8] and papers from the French ACACIA group who are developing the KATEMES tool (Knowledge Acquisition Tool for Explainable Multi-Expert Systems) for the partial automation of engineering work based on the evidence of knowledge at the group acquisition stage [9].

On the other hand, it is not only experts who affect the topology of knowledge sources. Significant volumes of expert knowledge have been accumulated in the natural language texts (second-type knowledge sources). In the few last years, third type knowledge sources appeared. This is the knowledge from the current information systems, which are complicated technical-organizational systems with network devices, servers, DBs (DBMSs), and so on.

The problem of knowledge acquisition (detection) from second-type sources relates to the rapidly progressing Text Mining technology [10], in which the problem of automated acquisition of knowledge from DBs in artificial intelligence is related to Data Mining and the Knowledge Discovery in databases (KDD) [11]. The Text Mining technology is successful due to the application of textual methods of knowledge acquisition from natural language texts (NL-texts), which are widespread in three types of current web-oriented NL systems: Information Retrieval, Information Extraction, and Text/Message Understanding [12].

Data Mining is applied in PDs, such as scientific research (into medicine, biology, bioinformatics, and so on); the solution of business problems (banking, finances, insurance, CMR, and so on); governmental problems (protection from terrorism, searching for wanted people, and so on), and the solution of problems of web resources analysis with basic courses, such as Web Content Mining (intelligence crawlers, as well as the classification and filtration of information) and Web Usage Mining (which implies the detection of laws in the actions of a web node user or group of users); and so on.

Each of these technology have been developing independently of one another, and, now, such autonomy and distribution do not allow effective monitoring of all information resources (knowledge bases, databases, and the ontologies that were developed in recent years) belonging to the intelligence systems, especially IESs.

The investigations into the construction of tools and the development of distributed knowledge acquisition from a variety of sources of different topologies currently do not exist.

The experience of the practical use of an entire set of applied IESs developed in terms of the problem-oriented methodology (POM) and with the AT-TECHNOLOGY workbench [1,13] (including the express diagnostics of blood, diagnostics of complicated engineering systems, design of unique engineering objects, complex ecological problems, and so on) has shown the necessity of monitoring, which lies in the checking and conforming of accumulated and formalizable knowledge in corresponding knowledge bases (KBs). The AT-TECHNOLOGY workbench was developed in our laboratory, thus we use it as a program base for further researches. Comparison of AT-TECHNOLOGY with similar tools is beyond this work.

In addition to the detection of errors (defects), duplication, inconsistency and incompleteness of information in the KBs of already developed applied systems, the above mentioned problems are of great importance during the modeling of PDs and design of KBs and DBs (the control of limited integrity, consistency, agreements between the terms used in PDs, and so on). For example, to overcome the problem of the incompleteness of a developed KB (i.e., the expert does not know of and/or forgets to note some fact required for problem solving), we can invite a specific expert at times or an expert group, as well as using an independent electronic knowledge source in the form of a DB [1]. The first two ways can lead to difficulties in the modeling PDs, both due to a significant increase of labor costs and due to the "noisy" personal features of experts (misunderstandings, failures to mention, conformism, cognitive protection, self interest, the lack of semantic unification of used terms of PDs, and so on [5]). The authors of [3] also noted such factors as "cognitive self-protection", "discreteness", the lack of human knowledge, and so on.

The most neutral and independent knowledge sources are DBs. Analysis of experimental data obtained during the creation of an entire set of applied IESs showed that the local use of DBs as an additional knowledge source can fill up the volume of KBs by 1020% according to the PD assignment [1].

Thus, the development of a new automated technology of knowledge acquisition distributed by different sources occurred. We can formulate the following conceptual basis of the present work [14,15]:

1. The notion of the "distributed acquisition" of knowledge is introduced to fit the integrated information from knowledge sources with different topologies;
2. The first-type and second-type knowledge sources in the combined method of knowledge acquisition (CMKA) implemented in the framework of the POM [1] are considered to be combined, since there is a collection of well-tested technological processes in the CMKA, which allow replenishing the information from experts using information detected from problem-oriented NL-texts (this information includes the processing of protocols for interviewing experts, the acquisition of the vocabulary of a knowledge engineer/system analyst, analysis of signal lexemes in the input NL-texts, and so on);

3. The problem of integration with information obtained from a DB as a third type source for the automated construction of the most complete and consistent models of a PD.
4. Since no universal methods for solving the problem of DB completeness exist, the development and application of the technology of knowledge acquisition from DBs as additional source of knowledge are a rather new application of the Data Mining and KDD concepts for the solution of this problem.

In the present paper, the authors discuss the general characteristic of the of the combined method for knowledge acquisition and describe the typical design procedure for the application of KDD and Data Mining at different stages of its life cycle related to the automated construction of the KB of IES prototypes. The features of distributed knowledge acquisition from the databases are described in details. The implementation of typical design procedure for knowledge acquisition from databases is reviewed.

2 General Characteristics of the Combined Method for Knowledge Acquisition

According to the conceptual bases of the POM of IES construction, the most important part of this methodology is the POM of knowledge acquisition the collection of the CMKA and the technology of its use at different stages of the life cycle of an IES and web-IES construction [1]. Within the limits of basic CMKA and the media of its realization, the so-called local variant of knowledge acquisition is under consideration.

However, the "distributed" variant of knowledge acquisition, which is based on the CMKA became possible upon use of the web version of the AT-TECHNOLOGY workbench. On the one hand, this variant provides the integration of all the above mentioned types of knowledge sources and, on the other hand, it allows one to take its geographical arrangement into account, as well as to deal with the groups of remote knowledge sources in the framework of client-server architecture.

In general, the generalized model of CMKA [1,14] that takes the features of distributed knowledge acquisition into account can be written in the form:

$$MKM = \left\langle \widetilde{N}, \widetilde{S}, \widetilde{F}, K, Z \right\rangle \tag{1}$$

where $\widetilde{N} = N_{locn}$, $n = 1, \ldots, m_n$ is the set of unstructured descriptions of a PD;

$$N_{locn} = \langle IN, TN, SN, CN \rangle \tag{2}$$

where IN is the serial number of the description; TN is the type of description; SN is the source from which the description is obtained; CN is the description; $\widetilde{S} = \widetilde{S}_m$, $m = 1, \ldots, mn$ is the set of structured descriptions of a PD; F is the set of procedures for the mapping of $\widetilde{N}$ in $\widetilde{S}$; K denotes the procedures for the

conversion of the formed knowledge field (KF) into the formats of knowledge representation languages (KRL) of different tools for the ES construction (in the AT-TECHNOLOGY complex); and Z denotes the fragments of the DB in KRL formats of other tools for ES construction.

Therefore, in the course of interviewing an expert, the structuring of the information obtained in the form of a KF occurs, which is significant in the structuring of the information of the PD obtained from the expert. It provides the integrated representation and unification of basic concepts and ratios of the PD that were detected from different knowledge sources as a first step to the formalization in the concrete KRL.

According to the features of distributed knowledge acquisition, the generalized model of a KA can be presented as follows [14,15]:

$$Sm = \langle IS_m, TS_m, SS_m, O_m, R_m \rangle \tag{3}$$

where IS_m is the serial number of the structured description of PD; TS_m is the type of structured description of PD; SS_m is the source from which the description is obtained; $O_m = \{O_{mj}\}, j = 1, \dot{;} n$ is the set of objects; $Rm = \{R_{mk}\}, k = 1, \dot{;} p$ is the set of rules.

Thus, in going from the local variant of knowledge acquisition to the distributed one, the set of basic procedures of the CMKA is filled with the following procedures: acquisition of description from the distributed sources; correlation of different type acquired knowledge; refinement of the descriptions with detected inconsistencies; and grouped knowledge acquisition.

3 Applications of Distributed Knowledge Acquisition

As noted above, for the acquisition of knowledge from a DB in the framework of the CMKA, the KDD and Data Mining technologies used as an additional knowledge source to overcome DB incompleteness, because this provides the intelligent analysis of large volumes of information and the detection of the hidden laws within IESs developed in terms of the POM.

We emphasize that these terms are interpreted in the POM as follows: KDD denotes the entire process of knowledge acquisition from the DB to the representation of the obtained results, of which Data Mining is only some stage of the general process of the KDD.

According to the knowledge acquisition processes, the Data Mining concept is implemented in the CMKA in the three following ways [1]: the generation of an initial KF from a DB with further modification by an expert; the verification of a KF obtained by interviewing the expert, as well as the partial modification related to the finding of assurance coefficients for detected knowledge and the merging of the KF as a result of application of two methodologies.

One feature of KDD and Data Mining application in the framework of CMKA is the necessity of arranging access to a specific DB containing the information of the analyzed PD and its processing. Therefore, the CMKA includes many specific procedures for operation with DBs, such as[15]:

- the generation of a SQL-query to the DBMS;
- acquisition of data from the DB in accordance to the query formed by the procedure of data acquisition from the DB;
- filtration of some data subset that is then used for the construction of a set of rules (procedure of data subset filtration); and
- conversion of data to a format that can be directly used by knowledge acquisition algorithms (procedure of data conversion).

Below is the description of the procedures that are assigned for the preparation of data selection for subsequent analysis.

Based on the generation of an SQL-query the sample for subsequent application of Data Mining algorithms is formed. The knowledge engineer selects the attributes from the DB and based on this the system generates an SQL-query. Taking into account the specific character of the Data Mining algorithms used in CMKA, the knowledge engineer carries out the procedure of extracting the dependent and independent attributes (columns) in the analyzed sample. Then the processing of the unknown values of attributes occurs.

Notice that in the local variant of the CMKA, two basic algorithms for constructing the decision trees of ID3 [16] and C4.5 [17] that allow one to construct the sets of conditionaction rules in terms of the analyses of developed decision trees are used. However, the concepts of the CART algorithm [18] are preferred in going to the distributed variant of the CMKA, since this allows one to construct binary decision trees that are more convenient during visualization and rule post processing to reduce the general number of derived rules.

The procedure of data transformation converts them into a format that can be directly used by knowledge acquisition algorithms. Once the sample for the analysis has been completed, the procedures of knowledge acquisition from a DB are immediately applied, in so doing, it provides the determination of relationships in the form of if/then rules and basing on algorithm which is beyond this work.

The final procedures are the following: assessment of the model precision in terms of the textual data; estimation of the algorithm and its parameters that provide the best results in knowledge acquisition; and the conversion of obtained rules into the required format. The required format is determined by exact application where described method is used.

In going to the distributed knowledge acquisition, the emphasis is on the synchronization of the processes of knowledge acquisition from different sources by means of special typical design procedure (TDP) incorporated in POM and the technology of IES prototype construction [1]. The applied TDP uses the technological DB of an intelligent planner of the AT-TECHNOLOGY workbench and specific software for the integration of knowledge sources serving as a basis for the integration of KF fragments the obtained from different sources. The typical design procedure "Knowledge Acquisition from DB" includes the following stages: the acquisition of KF fragments in the form of if/then rules by using the CMKA (expert interviewing and knowledge acquisition from DB) and carrying out the verification of obtained KF fragments; the integration of sets of rules

due to the algorithms of comparing some KF fragments, which is based on the calculation of the adjacency coefficient [19] for each pair of rules; and verification of the united KF.

Note that the integration of rule sets is the most labor-intensive problem [20]. The automated comparison of rule sets obtained from different knowledge sources precedes this procedure. Extended decision tables (EDTs) [20] are used as the analyzed structure for the effective and rapid comparison of sets of rules in the POM. Each cell of such a table contains the data of the input and parameters of the input statement to the concrete rule, which is characterized by a headline.

Below is the detail description of the typical design procedure "Knowledge acquisition from DB" implemented in the intelligent program environment of AT-TECHNOLOGY workbench.

4 Implementation of the Typical Design Procedure "Knowledge Acquisition from Database"

Problems of development process support in developing intelligent and technological integrated expert systems (IES) with power functionality and scalable architecture are getting more and more significant and topical. For the first time these problems were reviewed in IES [1] development problem-oriented methodology and in AT-TECHNOLOGY workbench, which supports this methodology and represents knowledge engineer workstation.

The experience collected in development of multiple applied IES [1,13], tutoring IES development and usage in particular [21,22] has shown that the most of IES development problems are related to high complexity of projecting and implementation stages, and the problem domain has significant influence on organization and specificity of these lifecycle stages. The human factor also remains significant enough, because it leads to increasing of labor and development time spent. Also in the most cases the IES development technology specificity does not allow to use traditional programming methods.

Therefore a significant place in the problem-oriented methodology has been assigned to methods and instruments of intelligent program support of development processes. To all of them a common term "intelligent program environment" is applied (common methodology provisions are described in monograph [1] and another papers, for example [21]).

Today intelligent technology of integrated IES prototype development includes: development plan generation and execution with help of intelligent planner; knowledge engineer dynamical assistance based on knowledge about TDP and reusable components base intelligent program environment components; IES prototype architecture model generation; IES prototype analyzing based on knowledge about methods and models for typical problems solving; recommendation and explanation messages delivering to knowledge engineer. Intelligent planner "knows" how many and which TDPs and reusable components are registered in the workbench, and what are they designed for. Based on this knowledge, the development plan is generated.

Let us describe TDP "Knowledge Acquisition from DB" in details. Common TDP model is defined as:

$$TDP = \langle C, L, T \rangle , \tag{4}$$

where C is a set of conditions which allow TDP releasing; L execution scenario, described in internal action language; T is a set of parameters, initialized by intelligent planner when the TDP is included in IES development plan. And now, let us consider components concrete definition for TDP "Knowledge Acquisition from DB":

Component. Conditions for the TDP are defined in following way:

- a "storage" element in extended data flow diagram (EDFD) hierarchy, which represents architecture model of developing IES prototype;
- a lifecycle stage is system requirements analyzing;
- there must be at least one "unformalized operation" element in EDFD hierarchy;
- in the EDFD hierarchy a "storage" element must be connected with a "unformalized operation" element.

Component L. This TDP can be executed in two ways:

1. Initial knowledge field generation with distributed knowledge acquisition from DB algorithm without expert interviewing.
2. Knowledge field generation with distributed knowledge acquisition from DB after expert interviewing. In this variant it is necessary to execute expert interviewing task.

Component T. Context parameter P17 is set to 0 with comment that TDP "Knowledge Acquisition from DB" will be used. In the first step of TDP "Knowledge Acquisition from DB" execution, knowledge engineer selects a set of registered databases, and then forms a set of data storages with help of special program tools. These storages are analyzed with a distributed knowledge acquisition from DB algorithm [14,15]. The next step is the knowledge acquisition from DB algorithm configuring and generating with it a set of knowledge fields from each registered DB. In the third step, all knowledge fields of different types are merged. The main stag-es in this step are [15]: loading, objects merging, extended solution table and rules similarity table forming. In the stage of knowledge field fragments merging an expert sets control zones and values of float attributes coinciding. Also merging of objects, attributes types merging, and rules merging are performed. Next, a sample of rules from TDP "Knowledge Acquisition from DB" are presented.

A rule for initiating an execution of data storage forming tools:

```
<PLANRULE ID="14" Caption="Storage creation" Condition="LCStage=1
AND StorageCount(LinkToDB)>0" Parent="3" ArgType="Project"
Executor="Ware House" Action="run_warehouse" ActionType="0"
Type="1" />
```

A rule for running tools for distributed knowledge acquisition from DB:

```
<PLANRULE ID="17" Caption="Distributed knowledge acquisition from
database" Condition="LCStage=1 AND StorageCount(LinkToDB)>0 AND
AllElementCount(TDesES)>0" Parent="3" ArgType="Project"
Executor="Data Mining" Action="run_mining" ActionType="0" Type="1"/>
```

A rule for starting the tool for merging if/then rules:

```
<PLANRULE ID="18" Caption="Merge knowledge field fragmens"
Condition="LCStage=1 AND AllElementCount(TDesES)>0" Parent="3"
ArgType= "ProjectValue" Executor="Rules_src" Action="run_rules"
ActionType="0" Type="1" />
```

Samples of more complex TDPs connected with tutoring IES development are described in [22]. The difficulties of the tutoring IES development technology are caused by supporting two different work modes DesignTime, oriented to work with teachers (course/discipline ontology creating processes, different typed training im-pacts creating, etc.) and Runtime, for working with students (current student model building processes, including psychological model, etc.).

5 Conclusion

The experimental routine research of the distributed variant of the CMKA (including the collection of algorithms and procedures of knowledge processing obtained during expert interviews, as well as during the analysis of protocols of interviewing and knowledge acquisition from the DB) on several real and test DBs showed a high efficiency of the proposed approach to the solution of problems of KB incompleteness, for KBs support, and the automated updating of KBs upon the emergence of new DBs or changes of outdated ones.

In conclusion, it is necessary to point out, that today we are performing experimental research connected with intelligent support for IES prototype construction. During this experiment many weak points were already fixed, in particular connected with low performance, not sufficient technologic knowledge base content, new typical design procedures development, etc. As a result, time cost for a typical IES development prototype with AT-TECHNOLOGY workbench was reduced.

References

1. Rybina, G.V.: Theory and Technology of Construction of Integrated Expert Systems. Nauchtekhlitizdat, Moscow (2008) (in Russian)
2. Lyugger, D.F.: Artificial Intellect: Strategies and Methods for Solving Complex Problems. Williams, Moscow (2003) (in Russian)
3. Osipov, G.S.: Artificial intelligence methods. Fizmatlit Publishing House, Moscow (2011) (in Russian)

4. Chastikov, A.P., Gavrilova, T.A., Belov, D.L.: Expert System Development. The CLIPS Environment. BHV Publishing, St. Petersburg (2003) (in Russian)
5. Rybina, G.V.: Fundamentals of Intellectual System Construction. Finansy i Statistika, Moscow (2010) (in Russian)
6. Podlipskii, O.K.: Construction of Knowledge Bases by Expert Group. Komp. Issled. Modelir. 2(1) (2010) (in Russian)
7. Kobrinskii, B.A.: Extraction of Expert Knowledge: Group Varian. Novosti Iskusstv. Intell. 3 (2004) (in Russian)
8. Mendonça, D., Kelton, K., Rush, R., Wallace, W.: Acquiring and Assessing Knowledge from Multiple Experts Using Graphical Representations. Knowledge Based Systems Techniques and Applications 1, 293–326 (2000)
9. Dieng, R., Giboin, A., Tourtier, P., Corby, O.: Knowledge Acquisition for Explainable, MultiExpert, Knowledge Based Design Systems. In: European Knowledge Acquisition Workshop (1992)
10. Feldman, D., Hirsh, M.: Mining Associations in Text in the Presence of Background Knowledge. In: 2nd Internation Conference on Knowledge Discovery, pp. 343–346 (1996)
11. Finn, V.K.: About Intelligent Analysis of Data. Artificial Intelligence News 3 (2004)
12. Khoroshevskii, V.F.: Treatment of Natural-Lingual Texts: from Language Understanding Models to Knowledge Extraction Technologies. Artificial Intelligence News 6 (2002)
13. Rybina, G.V.: Problem-oriented methodology and practical applications for Integrated Expert System construction (a review of applications in static and dynamic problems). Instruments and Systems: Monitoring, Control, and Diagnostics 12 (2011)
14. Rybina, G.V.: Combined knowledge acquisition method for knowledge base construction for integrated expert systems. Instruments and Systems: Monitoring, Control, and Diagnostics 8 (2011)
15. Rybina, G.V., Deineko, A.O.: Distributed knowledge acquisition for automated integrated expert system construction. Artificial Intelligence and Decision Making 4 (2010)
16. Quinlan, J.R.: Induction of Decision Trees. Machine Learn. Journal 1, 81–106 (1986)
17. Sreerama, K., Kasif, S., Salzberg, S.: A System for Induction of Oblique Decision Trees. J. Artif. Intell. Res (JAIR) 2, 1–32 (1994)
18. Breiman, L., Fiedman, J., Olshen, R., Stone, C.J.: Classification and Regression Trees. Wadsworth Int. Group, Belmont (1984)
19. Zagoruiko, N.G.: Applied Methods for Data and Knowledge Analysis. Inst. Matem., Novosibirsk (1999) (in Russian)
20. Rybina, G.V., Deineko, A.O.: About one approach for merging IF/THEN rules acquired from different knowledge sources. Artificial Intelligence and Decision Making 4 (2011)
21. Rybina, G.V.: Intelligent tutoring systems based on integrated expert systems: Development and usage experience. Information Measuring and Control 10 (2011)
22. Rybina, G.V., Blohin, Y.M., Ivashenko, M.G.: Some aspects of the intelligent technology for tutoring construction of integrated expert systems. Instruments and Systems: Monitoring, Control and Diagnostics 4 (2013)