ITHEA

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

2009 Volume 3 Number 4

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE

Volume 3 / 2009, Number 4

IJ ITK is official publisher of the scientific papers of the members of the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK as well as the **subscription fees** are given on *www.ithea.org*
The camera-ready copy of the paper should be received by e-mail: info@foibg.com.
Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.3, Number 4, 2009

Edited by the **Institute of Information Theories and Applications FOI ITHEA®**, Bulgaria,
in collaboration with the **V.M.Glushkov Institute of Cybernetics of NAS, Ukraine**,
and the **Institute of Mathematics and Informatics, BAS, Bulgaria**.

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Printed in Bulgaria

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

MULTIAGENT SYSTEM FOR SMART DOCUMENT ANALYSIS

Vyachslav Lanin, Elena Mozzherina

Abstract: In the article you can find an intermediate outcomes for the complex approach to implementation of development electronic document management subsystem in the CASE-system METAS. The system meant for creating distributed information systems that allow dynamic settings for changeable environment and user requirements. The operational efficiency of electronic documents processing is suggested to increase considerable because of automatic smart document analysis. Agent- and ontology-based approaches are suggested as a solution. Ontologies are used to present the semantic and the structure of the document explicitly. Agents are used for simplification of analysis, made it scalable and upgradable. The results of smart analysis and processing of the documents taken from the heterogeneous sources can be used not only for automatic document classification and clusterization in information system in user friendly form, not only for reducing the labor intensity for subject field analysis and design of information system, but also for intellectualization the processes of reports creating based on information from the database of the system.

Keywords: ontology, agent, multiagent system, smart search, document analysis, adaptive information systems, CASE-technology.

ACM Classification Keywords: D.2 Software Engineering: D.2.2 Design Tools and Techniques – Computer-aided software engineering (CASE); H.2 Database Management: H.2.3 Languages – Report writers; H.3.3 Information Search and Retrieval – Query formulation.

Introduction

Nowadays a large number of CASE-systems, which automate the most labor-consuming stages, which connected with the business flows and interface creation, in Information System (IS) development, exist. The stage of the subject area analysis, which is not usually automated by the CASE-systems, becomes the longest and the most labor-consuming. Thus, one of the most promising trends in the CASE-system development is the automation of this process. This task becomes especially vital in the CASE-systems, which are oriented for the creation of IS with the dynamic adaptation during their usage, and where the stage of the subject area analysis lasts for all the time when the system is used. Usually during the usage of such systems the task of development is entrusted (even partly) to the users, which are experts in the subject area but not in the programming, so components for the automation analysis become the most important. In other words, if we pose the problem of the dynamic setup of IS in changeable environment, then the basis of creating the tools for its dynamic adaptation become the tools for restructuring the data in the Data Base (DB). And these tools allow making changes in the data model based on the results of the subject area analysis, on the normative reference and administrative documents, which regulate the activity in this area. Hence follows the need for support the stage of analysis in dynamically adapted systems, which is one of the most complex and labor-consuming stages in the IS development. As information source for the analysis the documents of different types can be served, because the activity of any business-system is build on the normative documents. The support of the business with the help of the IS tools requires the reflection in the data model the system of the standards, fixed in the normative-reference

and administrative documents, in the form of limitations (attributes, subject area objects' properties and relations between them) and operations, assigned on the data [1].

As a result of the analysis the system of the interconnected document will be constructed:

- the documents, that belong to the determinate directions of the business activity (to the specific concepts, to the objects of the subject area);
- the documents, that reflect the relations between these concepts (one or several documents can be connected with each concept and the connections between documents reflect the relations between the concepts);
- the documents, that contain the normative information, which also can be extracted with the help of the documents content analysis.

On the basis of the constructed system of the interconnected documents it is possible to partially automate the process of the analysis of changes in the subject area and introduction of changes in the subject area model (i.e. to implement the support of the development process and adaptation of IS). Thus the document management system becomes not only the «wrapper» above the IS and its DB, that allow to view the data processing results, that are stored in DB, in a user-friendly form, but also becomes the basis of the IS development tools, tools for data restructuration.

Description of Documents with Ontologies

For increasing the efficiency in the processing an electronic document requires the presence of metadata that describe the structure and the semantics of the data. One of the possible approaches to describe the information placed in the document is the one based on ontologies. By ontology we understand the knowledge base of the special type that can be easily read, understood, alienate from the developer and/or divided physically by its users [4].

Ontology based approach has the following advantages:

- it is understandable for people;
- the user who develops the ontology do not need a special qualification;
- one document can be described with several ontologies.

An ontology based approach was chosen as the one for solving the task described above [1]. In this approach the ontology describes not only the structure but also the content of the document. According to the suggested approach the ontology is used for describing the semantics of the data in the document and its structure. Let us take into account the specific of the tasks solvable in this article and specify the definition of the ontology. We will consider the ontology as a specification of a certain subject area, which includes the dictionary with terms (concepts) from this area and a set of connections between the terms which describe how the terms are correlated between them in this particular area.

The following base types of the relations are used for constructing the hierarchy of the ontology concepts:

- "is_a";
- "part_of";
- "synonym_of".

You should remember that the given types are the base types and do not depend on the ontology, but the user needs the possibility to add the new relations, which would take into account the specific of the subject area.

The ontology also includes two types of nodes. Nodes from the first group describe the structure of the document (for example, table, date, occupation and etc.). These nodes describe the common concepts which do not depend on the subject area. Nodes from the second group describe the specific concepts which belong to this particular document. We will call structured nodes those from the first group and semantic nodes those from the second group. In Fig. 1 structured nodes have the dark color, and semantic nodes have the light color.

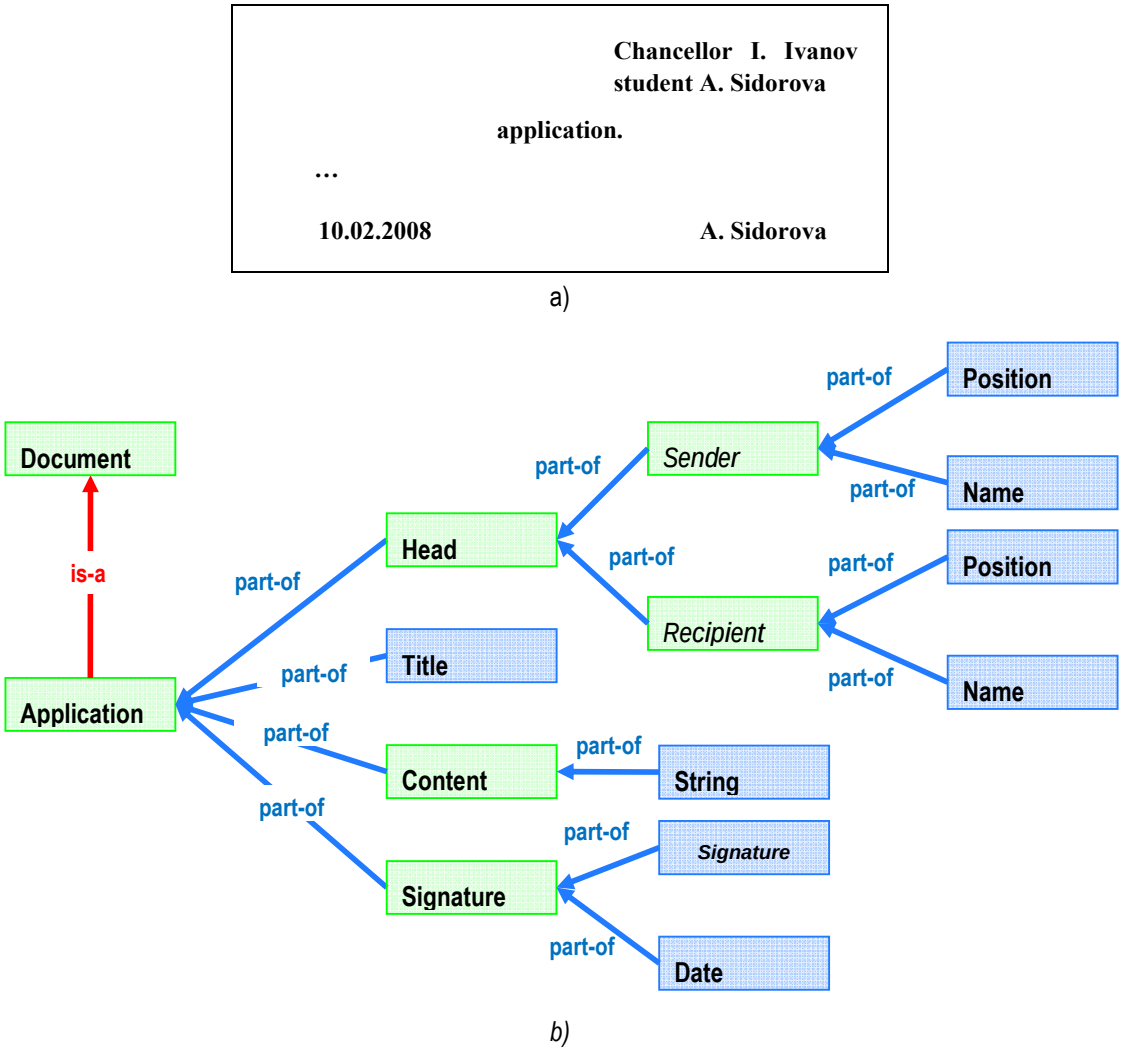


Fig. 1. Example of the document «Statement» (a) and ontology that describes the class of the documents «Statement» (b)

Actually in this context ontology is the hierarchical conceptual basis of the considered subject area. The ontology of the document is used for the document analysis, thanks to it we can obtain the required data because we know where to search and how to interpret the information.

If we present the document with the help of the ontologies then the task of comparison the ontology and existing document is reduced to the task to look for the concepts of the ontology in the document. As a consequence the system needs to answer the question: does this ontology describe the document or does not? The last question can be answered affirmatively if in the process of comparison all the ontology concepts were found in the document. It is necessary to search for the nodes which describe the structure of the document before we have a look for the nodes which contain the concepts of the document. Thus the original problem is reduced to the task to look for the common concepts in the text of the document on the basis of the formal descriptions.

In the given example (Fig. 2, b) all the nodes of the ontology are divided into two planes that is take into account under the comparison of the document (Fig. 2, a) and its ontology.

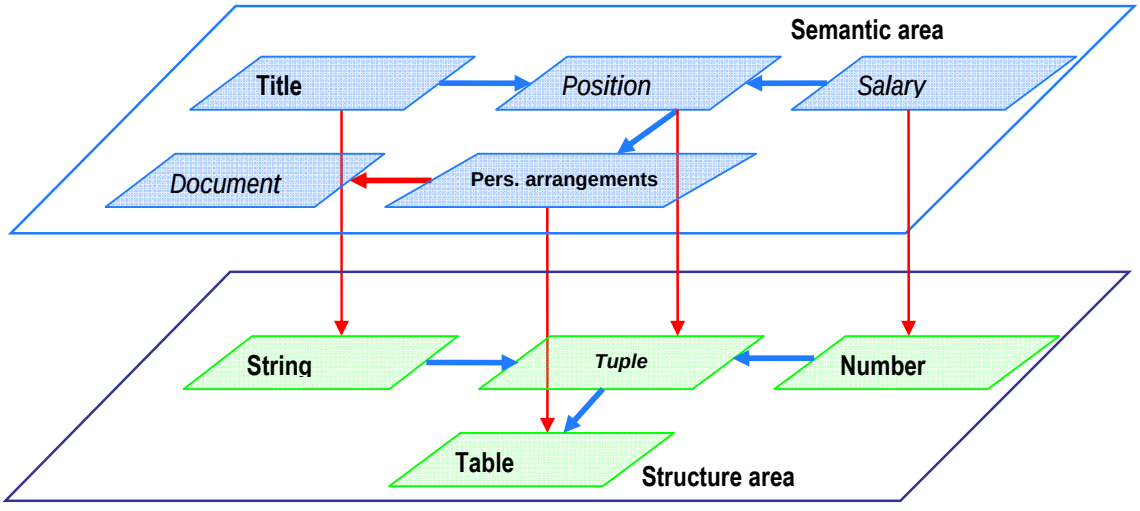
Order № 1
from 01.11.2005

Accept from 01.12.2005 personnel arrangements:

Assistant professor	**€ 6 000**
Associate professor	**€ 10 000**
Head of department	**€ 15 000**

Chancellor I. Ivanov

a)



b)

Fig. 2. Example of the document «Order» (a) and partitioning the ontology nodes of the document into two areas (b)

A repository of ontologies consists of 3 ontology levels (fig. 3). Ontologies which describe objects of specific system and take into account it features locate on the first level. Objects which are invariant to data domain locate on the second level. Objects of first level are described in terms of second level using relation «is_a» and «part_of». Objects of third level describe common conceptions and axioms which are used for describing objects of underlying levels. Third level and second level can be delivered to two part: describing of structures and describing of templates.

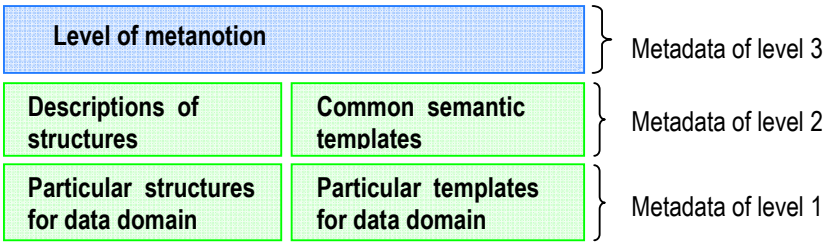


Fig. 3. Levels of metadata

Agent-Based Approach to Document Analysis

There are some requirements to the search process itself:

- the high rate of processing the large volumes of data;
- fault tolerance;
- scalability and
- adjustability to the users' need and changing conditions.

To solve the problem of extracting the common concepts base on formal descriptions we suggest an agent-based approach [2]. By the agent we understand a system that is aimed at specific goal, capable to interact with the environment and other agents to reach the goal [3]. This approach will satisfy all the requirements that were presented above to the search process, if during the system implementation all the advantages of multiagent systems will be realized.

With the use of this approach for each ontology node, that contains the common concept, the agent, which searches for this specific concept in the document, is created. The presence of the knowledge base (KB) is necessary condition to create an intelligent agent. Thus to determine the agents, that act in the system, we need to select the way of describing the KB, interaction with the environment and the way of collaboration with other agents. Tools to represent the KB are considered in the following section.

One of the most important properties of agents is a social significance or capability for interaction. As it was said earlier the agent is created for each ontology node that contains the common concept (semantic node). According to the classification of agents such agent is intentional.

The agent is aimed to solve two problems:

1. It divides into single components the entire list of the concepts patterns and starts simpler agents for search structure nodes.
2. It assembles all results obtained by the simpler agents.

Simpler agents mentioned above are reflector agents. They obtained the pattern and they aimed to find the fragments of the text that are fall under this pattern.

The communication between the agents is also an important question. All the mechanisms of communication can be divided into direct and indirect. A model of interaction called «contract network» is an example of direct communication. The mechanism of indirect communication is realized through the model «blackboard»:

- Contract network. This model assumes that all agents in the system can be divided into two classes: customers and contractors. The heart of the model is in solving the tasks by choosing for this the most appropriate agent. Customers are responsible for the distribution of tasks between the agents. Potential contractors analyze the requests provided by the customers, and if they can do it they send the request to the customer.
- Blackboard. This model is based on the model of the class board on which is presented the current state of the system where the agents are operating. All agents are continuously analyzing information on the board attempting to find a use to their possibilities. If at a certain moment the agent finds out the possibility to resolve the current task then it leaves a note on the board about the beginning of the work and it will place the results on the board after it completes the work.

Taking into account the special features of the task the combination of two models (both «contract network» and «blackboard») is implemented.

Both the architecture of the multiagent system and the process of document analysis are presented in Fig. 4.

Agent's Knowledge Base Presentation

One of the most important questions in the system is the one about presenting the KB of the agent. Nowadays presentation of the KB of the agent can be done in three different ways: with the use of ontologies, with the use of regular expressions and with the use of productions.

Presenting the knowledge of the agent with the help of the ontologies is the most expressive method which uses all the advantages of the explicit representation of the knowledge (Fig. 5). The advantage of this method is that we can use different ways to «proof» the node. For example it can be done through the simple coincidence of key phrase or request to the DB of the IS. Ontologies make it possible to describe different situations if it is impossible to find the exact coincidence. We can find more general or more define concept and so on.

The content of the analyzed document is presented in the form of special object model as the base of which was taken the object model of the document from Microsoft Word. The API functions were developed to give an access to this object model. The functions give an opportunity to operate with identical concepts when you work with documents in different formats. The API functions include functions to syntactic analysis of sentences, functions to evaluate different metrics between the concepts, functions to extract the information about the structure of the document. If additional operations are necessary to find the concept from the node they can be

described in the script with the help of the API functions mentioned above. In the script you can also use the requests to the object model of the IS.

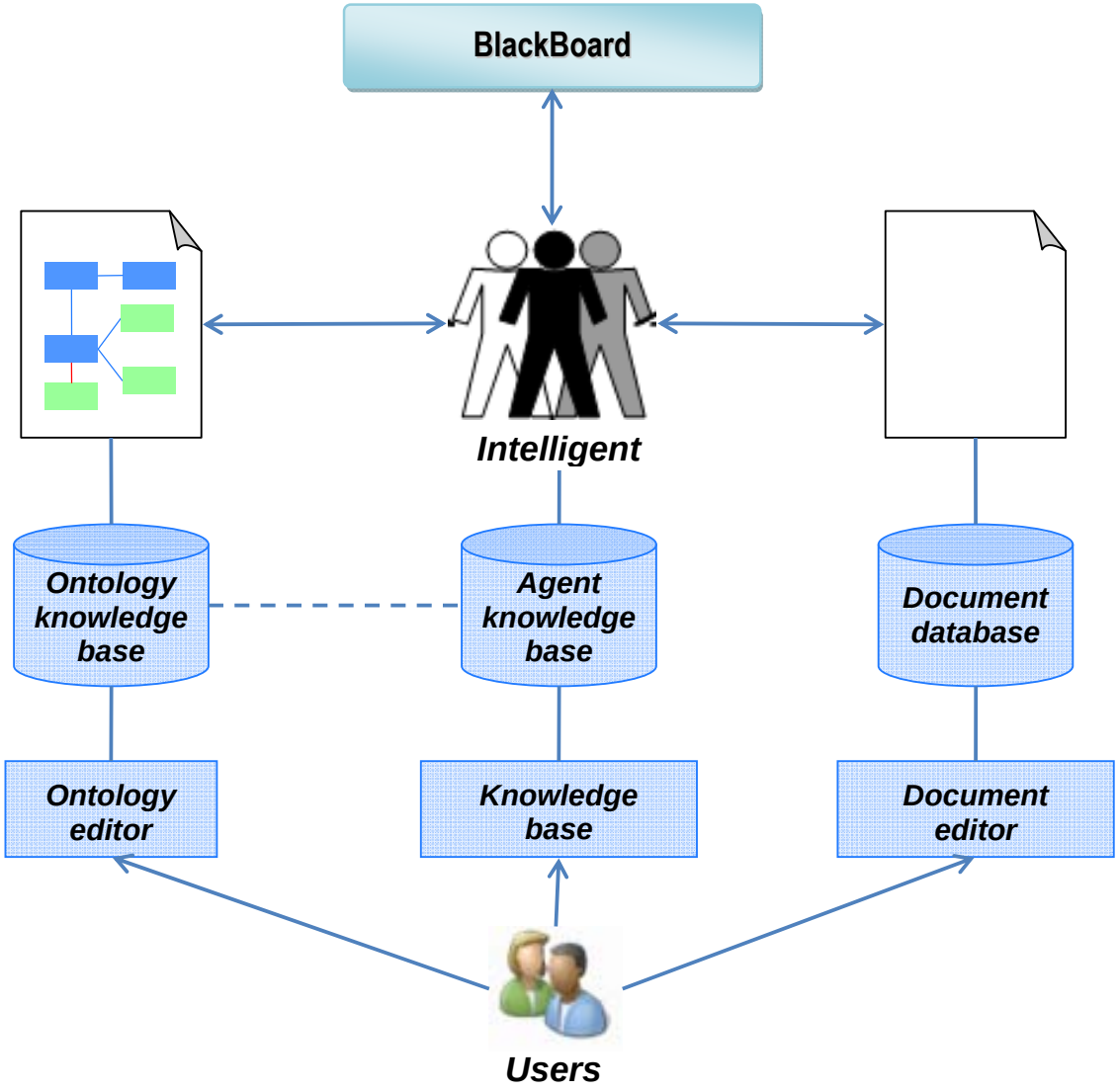


Fig. 4. Architecture of the SemanticDoc system

The second approach is the one that uses regular expressions. Expressions make it possible to consider various forms of the word and to work with large volumes of information [5]. It is necessary to remember that sometimes, especially for the users without special qualification, it is hard to construct a regular expression correctly. There is a special editor in the system to simplify the task. The editor allows working with the regular expressions in natural language. For example, the equivalent to the phrase «five-digit number» is «\d{5}» and etc. Furthermore the function of building a regular expression «upon the pattern» is also very useful. This means that on the

examples given by the user the system can build the regular expression automatically. For example, the user typed two dates as examples «1.12.08» and «15.07.2006». The system will build an expression that is correspond with both given dates: «(\d{1,2}).(\d{1,2})((\d{4})|(\d{2}))».

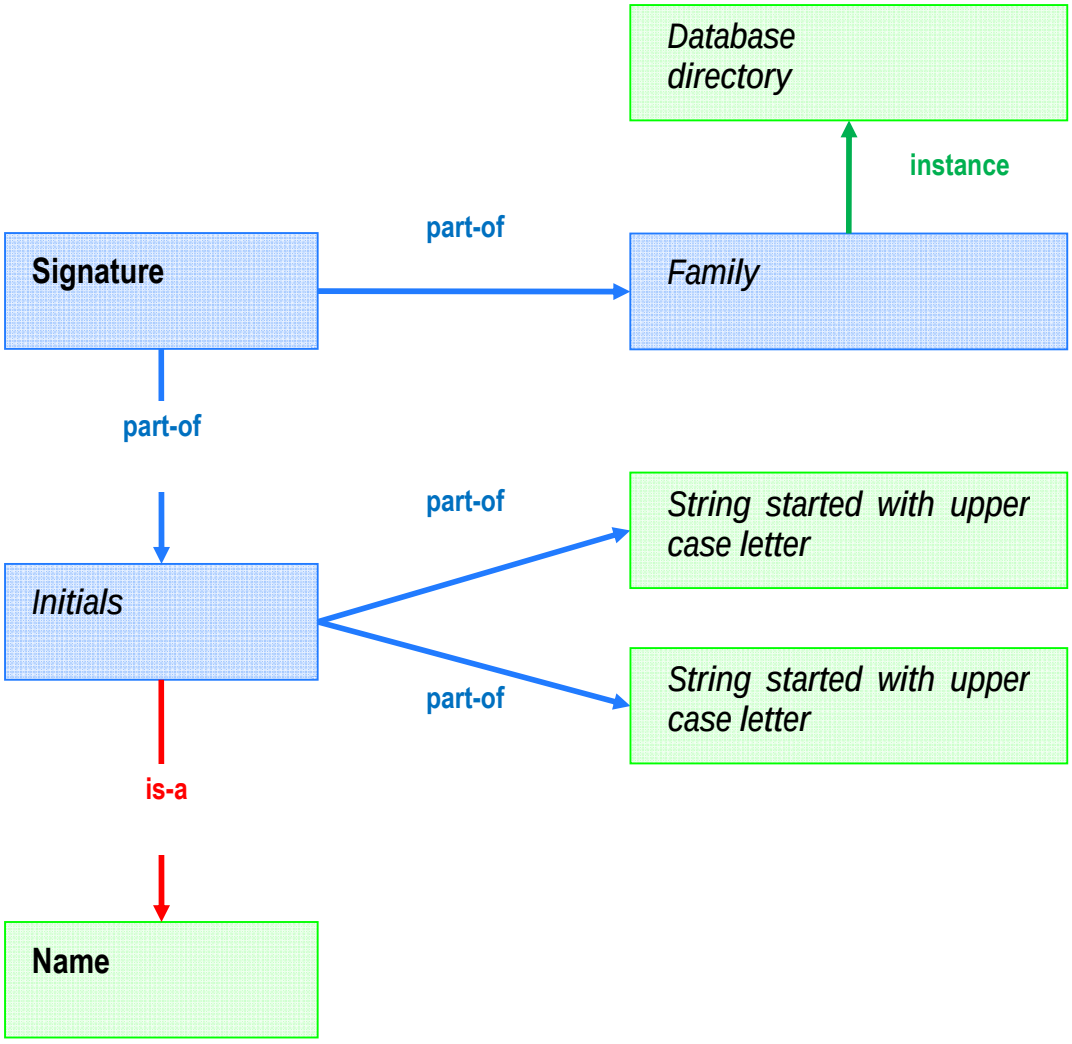


Fig. 5. Agent's knowledge base presentation with the help of the ontology

A drawback in the regular expressions is the fact that they do not take into account the location of the desired word or phrase during the search. It is possible to use both regular expressions and production rules, which are the third method of presenting the KB of the agent, to eliminate this defect.

Usually production rules are used to analyze the structure of the document. Special concepts can be used to set the conditions. For example, the rule is looking for the headings in the text can be formulated in the following way:

If (type of the paragraph differ from the type of previous and next paragraph) and (paragraph has center alignment),

then paragraph is the heading.

Conclusion

At the given moment the result of this work is the system SemanticDoc which was implemented on .NET. SemanticDoc is a multiagent system that matches the document and the ontology.

Two characteristics were brought in information retrieving to compare the quality of the results. The first is precision, the second is recall [6]. We can bring the similar characteristics for the system which matches the document and the ontology.

By precision (*P*) we will understand the number of correctly matched documents divided to the number of all matches done by the system. Under the recall (*R*) we will understand the number of correctly matched documents divided to the number of all existing matches.

Let the *N* be a number of all existing matches between the document and ontology, *M* – a number of matches that was done by the system, *A* – a number of correctly matched documents and ontologies. Then:

$$P = A/M \quad \text{and} \quad R = A/N .$$

Usually recall conflicts with precision and in practice it is impossible to reach both precision and recall.

The work on estimation this parameters was not conducted but the next stage will be evaluating the *P* and *R* after the experiments on the real documents.

The tools for document analysis can be used both for reduction the labor intensity of the users and for support the solution of subject area analysis done by the developers. In this case it is proposed the deep integration of the functional subsystem which includes not only the development tools, but also the tools for the end users. This given the possibility of designing the CASE-technology intended for the creation dynamically adjusted IS with the possibility to adapt to the changeable environment base on the feedback and smart document analysis.

Within this work the formal model of electronic document and ontology with regard to this task is developed, and base on it the existing object model of IS, metadata and algorithms to manage the documents become more specific.

So, listed above tasks can be solved by described approach:

- semantic indexing of documents;
- intelligent search;
- intelligent classification of documents;
- information extraction from not structured documents ;
- support of analyst work.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (Project 08-07-90006-Бел_а) and by the Russian Humanitarian Scientific Fund (Project 09-02-00373В/И).

Bibliography

[1] Ланин В.В. Интеллектуальное управление документами как основа технологии создания адаптируемых информационных систем // Труды международной научно-технической конференций «Интеллектуальные системы» (AIS'07). Т. 2 / М.: Физматлит, 2007. С. 334-339.

[2] Тарасов В.Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. М.: Эдиториал, УРСС, 2002.

[3] Рассел С. Искусственный интеллект: современный подход. М.: Издательский дом «Вильямс», 2006.

[4] Хорошевский В.Ф., Гаврилова Т.А. Базы знаний интеллектуальных систем. СПб.: Питер, 2001.

[5] Фридл Дж. Регулярные выражения. СПб.: Питер, 2003.

[6] Weal M.J., Kim S., Lewis P.H., Millard D.E., Sinclair P.A.S., De Roure D.C., Nigel R. Ontologies as facilitators for repurposing web documents / Shadbolt. Southampton, 2007.

Author's Information

Vyachslav Lanin – Perm State University, Postgraduate student of Computer Science Department; 15, Bukirev st., Perm, Russia; e-mail: lanin@psu.ru.

Elena Mozzherina – Perm State University, Candidate for a master's degree of Computer Science Department; 15, Bukirev st., Perm, Russia; e-mail: mozzherina@gmail.com.