Intelligent Techniques of User-Oriented Recognition of Objects from the Web Informational Resources

Anatoly Gladun[1] and Julia Rogushina[2]
[1]International Research and Training Center of Information Technologies and Systems of National Academy of Sciences Ukraine, Kiev, Ukraine
email: glanat@yahoo.com
[2]Institute of Software Systems of National Academy of Sciences Ukraine, Kiev, Ukraine,
email: ladamandraka2010@gmail.com

***Abstract* — The idea of applying Semantic Web intelligent techniques to the recognition of information objects has resulted in a number of recent research activities and initiatives. The ontologies can be used to represent personalized, user-oriented and problem-oriented knowledge about structure of intelligent information objects in distributed Web-based applications. This article proposes the methods of ontology use for recognition of information about information objects with complex structure that are relevant to user needs.**

***Index Terms*—Semantic Web, Ontology, Recognition of Information Objects, Multimedia Resources, OWL**

I. Introduction

Applications that solve different user`s problems in the Web distributed dynamic informational space have access to a lot of multifarious inconsistent informational resources. The main problem of these applications is retrieval and recognition of structure and properties of objects that have an influence on problem solving by them. Though we need in classification and definition of these objects and in description of user and task specifics.

Informational space of the Web becomes now a major source of knowledge but retrieval of this knowledge deals not only with search for relevant data but for acquire of the facts and rules that can help to users in their problem solving. The most reused and formalised representation of knowledge in pattern recognition is based now on ontological analysis of subject domain of user's problem [1].

II. Ontology as a means of knowledge formalising

The selection of knowledge representation formalisms for the majority of usual applications is determined only by problem to be solved and different preferences of developers, but for intelligent Web-based applications this selection has its own specificity: on account of their functioning in open information space they require the constant updating of knowledge from the external environment. Therefore it is therefore advisable to use such interoperable knowledge representation as ontology which today already established recognized standards, the representation languages, instrumental tools for editing and inference, as well as available fundamental mathematical basis.

At the same time, till now, there are no accurately formulated technologies of knowledge management based on ontologies which could be directly implemented in applied systems, including the systems for recognition of certain information objects which are interesting for the user and necessary for the solving of user's problems.

The perception and recognition of information objects (IO) are the most important problems for working out of intelligent information systems which are based on knowledge [2]. Features of IO recognition in the distributed Web-based applications are their dynamism (for example, both standards and languages of the description of the metadata, and sources of data can vary) and dependence of distinguished IO structure from the specificity of solved problem.

Therefore, the creation of intelligent industrial systems based on ontologies, in environment of continuous organizational and technological changes requires methods and tools not only for ontology creation, but also for the whole complex of related problems - change management, estimations, personification, separation, mapping and integration etc.

Ontology mapping is a process of the discovery of correspondences between concepts and instances of two or more different ontologies, and the integration of ontologies (ontology merging) is a process of creation of new ontology on base of multiple source ontologies. These processes can be performed by some different ways: manually, semi-automatically or automatically.

Methods of *ontology merging and alignment* are effectively used for tasks of the Semantic Web ontology sharing and reuse.

doi:10.4304/jetwi.6.3.348-353

III. INTELLIGENT DATA OBJECTS IN AN ONTOLOGICAL ANALYSIS

An object (from the point of view of programming and, in particular, the object-oriented approach) represents some entity in virtual space which possesses a certain condition and behavior, has preset values of properties (attributes) and operations over them (methods).

Typically, during the study of objects the belonging of each object to one or several classes is allocated. This belonging in turn causes the behaviour of this object, i.e. classification is a model of this object [3].

The information object is an extension of the program object. The information object represents the certain entity comprising data in information system about any real or virtual object (subject, being, event, process etc.) that is the unique identified material or non-material entity of the real world which describes its structure, attributes, restriction of integrity and, probably, behavior.

For example, a person, a publication, an organisation, a city can be an object, and the description of some object (information about some features of this object) can be an information object. For the Semantic Web concept and the Web information space such objects of the non-material world, as ontologies, software agents, Web-services, information resources, metadata, databases, etc. also belong to the information objects.

The other aspect of IO classification is a data type viewpoint. Now the Web proposes a lot of multimedia data for different applications and with different expressiveness.

Informational Resources (IR) accessible by the Web can be classify on structures and not structured ones, textual and multimedia information, static and dynamic data etc., but content of all of that IR has some specific semantics, and this semantics deals with some subject domain (or a group of domains). There is very important for informational retrieval process to discover IR that is relevant to the domain of user interests.

The dominant approaches for document classification have some specifics hat differ them from analysis of content and link structure [4]. Methods of content analysis are oriented on text data processing and so cannot be directly used for classification of multimedia data (but they are applicable for their metadata processing).

On the other hand, analysis of link structure requires the promptly reflecting of changes in dynamic object. Link structure analysis is used for electronic libraries where related pages of IRs are connected to each other by embedded links.

Classification proposed in [5] is based of prior knowledge about the information space that can be divided on two categoriess – knowledge about structure of categories and data about users' access in this structure. This classification is oriented on prediction of access and topics. This approach can be used only for entire documents and can't consider their semantic components.

Unfortunately tags of the HTML documents are not semantic and that's why they don't capable for providing of their structure level significant for identifying and extracting information, since they are mostly used for presentation issues. Virtual document is an IR that is generated dynamically on demand and is oriented on reflecting of the new world's states. In [6]) virtual documents are used for modelling of the set of homogeneous pages using a structuring schema.

Distributed information object resolution is analised in [7]. There IO are defined through their Uniform Resource Identifiers.

The main part of the Web structures IRs (including thw metadata about all other non-stuctures IRs) are represented by means of HTML and XML formats.

The subject domain of textual IR can be define by two ways – by analysing of their textual content and by considering of metadata of these IR.

Now a lot of widespread formats are used for a storing of audio and video information, 3D-scripts and images. The indexation of multimedia resources from the Web is much more difficult in comparison with the textual information. Therefore for multimedia IR only the second way is efficient. Metadata of multimedia IR contains machine-readable information about the content of this document represented as a structured text that can be automatically processed.

The most common and perspective model of metadata is RDF (Resource Description Framework) based on XML. RDF can be processed conjointly with ontological knowledge representation. Though the ontology of the Web multimedia recourses [8] can help to user in the description of IR type for problem solving according with software for their processing.

A lot of different ontologies that describe different types of IO represented through the Web are developed. They are differs by task, size and complexity of data representation. For example, the upper ontology DOLCE-Ultralite (http://www.loa.istc.cnr.it/old/DOLCE.html) contains such patterns as kinds of realisations, relations between formal expressions, generalized expressions, schemata and IO, encodings of IO etc. (fig.1).

But this representation is too complex for the main part of the Web users. That's why we propose to use more simple IO taxonomy without specific information (fig.2).

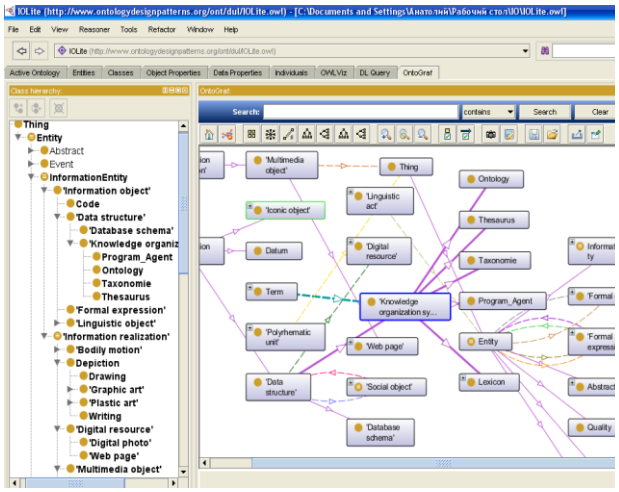

Figure. 1. DOLCE-Ultralite Ontology of the Web IO

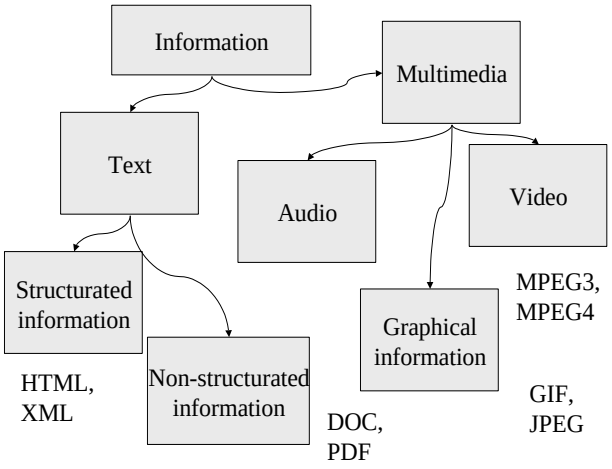

Figure. 2. Types of multimedia informational resources of the Web

We propose a new classification mechanism for different IO on base of this taxonomy.

Experimental comparing of approaches that analyse user access and collective user behaviour with other methods demonstrates better classification quality.

The methods that used for classification of IRs can be manual, automatic or semi-automatic. An examples of manual approach is Yahoo: this search engine indexes manually all it`s contents.

Another examples of manual classification are eBay and Amazon: these auction sites require sellers to classify their auctions when they upload to the sites. Manual classification is applicable for different data types and results to highly accurate results. But this approach is the inefficient if we need in processing of large number of objects with a complicated structure of categories. That's why manual classification is inapplicable for classification of the Web IRs with dynamic, heterogeneous and cumbersome data.

Automatic classification systems use domain-specific rules for classification and are able to learning by their own experience. Learning of systems is divided into supervised (classification) or unsupervised (clustering) learning [9].

Modern applications oriented on the Web are usually intelligent: they are based on knowledge about some subject domain, can process this knowledge and produce some new knowledge. For the purpose of reuse of domain knowledge in different applications interoperable forms of knowledge representation are very popular (for example, ontological approach).

That's why in the modern Web the significant part of content is intelligent and knowledge-oriented – ontologies, semantic Wikipedia items, elements of semantic markup, semantic metadata about other recourses etc (fig.3).

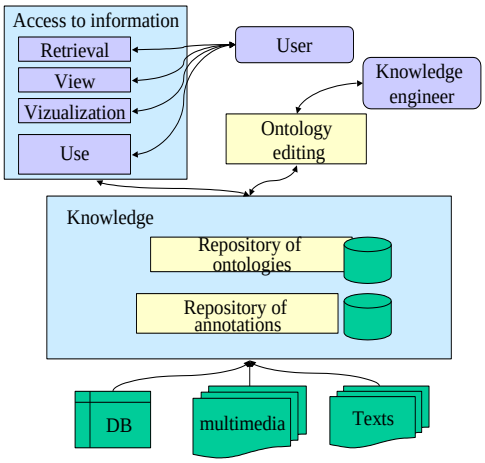

Figure 3. Ontological knowledge management

Now modern intelligent Web applications are oriented on standards, technologies and software developed by the Semantic Web project [10] of the W3C for knowledge management in the Web. One of the most important building block of knowledge in the Semantic Web are ontologies. They are based on descriptive logics and provide a shared and common understanding of a domain for communication of people and applications.

An important advantage of the Semantic Web approach for ontological analyses is an ability for users to collaboratively creation of ontologies and building of common vocabulary without any centralised control.

Ontologies of the Semantic Web consist of Semantic Web Terms (SWT). Every SWT is a building block that plays the role similar to the natural languages word. The set of SWT associates the RDF statements with formal semantics that are defined by RDF(S) with OWL statements. The social Web provides the knowledge about persons and communities that can be represented also as an ontology.

Ontological approach enables to link such information with some classes of ontology that are interested to user. This process of class recognition of heterogeneous data is a kind of pattern recognition.

Pattern recognition is an assignment of the source data to a certain class by means of allocation of the significant features that characterize these data from the total weight of non-essential data. Classification is an example of pattern recognition that tries to assign

each input entity to one of a given set of classes. However, pattern recognition is a more general problem that provides some other tasks:

- *regression* defines a real-valued output to each input;
- *parsing* assigns a sentence parse tree for describing of the syntactic structure of this natural language sentence;
- *sequence labelling* defines a class for every member of a values sequence (for example, speech tagging assigns a part of speech to word in the natural language sentence).

The classic statement of the problem of *pattern recognition*: by the set of given objects it is necessary to classify them. The set of objects consists of subsets called classes. Information about the classes, description of the whole set, information about some classified objects and information about an object, which belonging to some class is unknown but interesting to user are specified. Under the available information about classes and the object description the class is required to establish.

In modern intelligent applications this problem is usually transformed as follows - it is necessary to retrieve from the available (via Web) or derived in any other way relevant to the task informational resources (IR) the information relating to certain concepts (classes or class instances) and relations among them (their structure is reflected by domain ontology).

For example, the user needs to find a set of executors to perform a specific research project. Domain ontology allows to know that the desired object is an entity belonging to the class of "human" and having parameters such as education, experience, skills, availability of publications, degree, diplomas and work experience, etc.

But only the availability of relevant ontology allows the *problem-specific structuration* of information extracted from the IRs, because information used during the solving of another problem and extracted from the same IRs about the same objects may be absolutely different (for example, information for detection of effective ways of treatment of patients with similar symptoms differs from the information extracted from the same IRs about the same persons for finding dog owners with similar breeds).

Problem of perception, recognition and interpretation of objects is a complex task which is divided into separate subtasks [11]. Thus traditional pattern recognition, speech recognition and text recognition are only the special cases of a considerably more general problem.

Recognition of information object (IO) involves detection of information interesting to user about a particular IO in some IR content. For example, face recognition is a detection of elements that characterize the look data of IO of type "human" in the graphic IR [12].

IV. Use of Ontologies for Structurization of Recognised Data

Ontology can be considered as a basis for IO structure representation, i.e. IO is an ontology class, and various IRs are the sources for creating of class instances with interesting to user data. This approach allows to integrate information coming from different sources, and to generate the necessary to user knowledge. The task is subdivided into several sub-tasks:

- *Creation (or search) of ontology*, representing the structure of the IO (or of the set of the IO), knowledge about which is necessary to solving of the user's problem;
- *Retrieval of IRs* that explicitly or implicitly contain data about these IOs;
- *Acquisition of knowledge* about IO from IRs;
- *Representation of the retrieved knowledge* in the comprehensible and convenient to the user form.

General architecture of ontology-based recognition process links user problems with domain ontologies that are used in the recognition process of different IOs that are interested to users (fig.4). The main idea is in integration of knowledge from 3 types of ontologies: domain ontologies, user ontologies and IR ontologies.

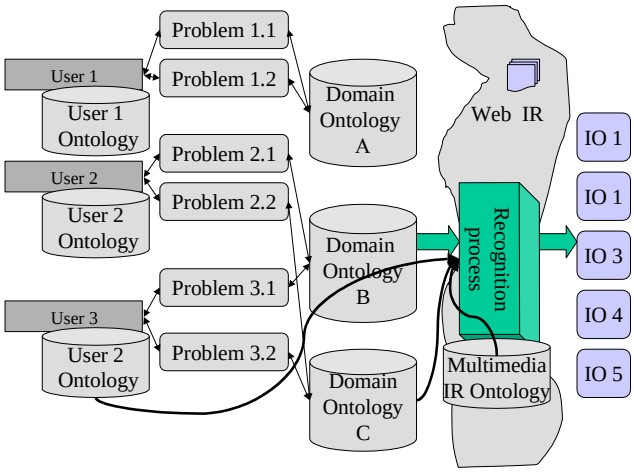

Figure 4. General architecture of ontology-based recognition process

It should be noted that in many cases this process is iterative, and during the problem resolving the information about IO has to be updated by extracting the relevant information from those IR which are available to the user (for example, via the Web or corporate network).

V. Conclusion

The main idea of this work is a design of general properties and methods of integrated use of knowledge from multimedia ontology, domain ontologies and user ontologies for semantic recognition of knowledge from different kinds of the Web informational recourses. Unfortunately, we can't compare experimentally the authors` results with other because this approach means

higher pertinence of informational retrieval that is hardly formalized by different researchers.

A lot of sites with dynamic information objects are, however, have the stable structure of their categories. Many of these sites can accumulate a large amount of additional information about their users from the process of co-operation with them. Such parameters as the structural stability and the rich personal information about users' in future can become important resources for content classification but now they don't investigated sufficiently.

This work proposes a framework for object classification into a category structure by analysing of users' traversals in the category structure. Use of ontologies as a source of information about the structuring of recognised IO is caused by properties of ontologies and their role in today's distributed intelligent applications. Furthermore, ontologies themselves can be updated by new knowledge produced as a result of the recognition process.

ACKNOWLEDGEMENTS

This work was supported in part by EU INCO-Copernicus Project 960114 – EXPERNET "A distributed Expert System for the Management of a National Network of Ukraine" and the Grant NATO NIG 971779 "National Telecommunication Networks for Scientific and Educational Institutions of Ukraine with Access to Internet – URAN".

On the other hand, this work was partially supported by Kelly Lada de Mandraka that doesn't disturb the authors along the research work and article editing.

REFERENCES

[1] D. Kirasic, D. Basch, "Ontology-Based Design Pattern Recognition Knowledge-Based Intelligent Information and Engineering Systems" // Lecture Notes in Computer Science, V. 5177, 2008. – P. 384-393.

[2] K. Kikumi Taisuoka, "Architecture of Knowledge Structures and Cognitive Diagnosis": A Statistical Pattern Recognition and Classification Approach// in Book Cognitively diagnostic assessment / University of Iowa USA, May 2003.-P.327-361).

[3] S. Dumais, H. Chen, "Hierarchical Classification of Web Content" // Proc. of SIGIR 2000.

[4] M. Balabanovic, Y. Shoham," Fab: Content-Based, Collaborative Recommendation "// Communications of the ACM, Vol. 40, No.3 199. – P. 66-72.

[5] M.Chen, A. LaPaugh, J. P. Singh,"Categorizing Information Objects from User Access Patterns". – http://www.cs.princeton.edu/%7Emaoch/publications/cikm_02.pdf.

[6] K. Pentikousis "Distributed Information Object Resolution" // Eighth International Conference on Networks (ICN), Gosier, Guadeloupe/France, 2009. – P.360-366.

[7] A.-M Vercoustre., F. Paradis "A Descriptive Language for Information Object Reuse through Virtual Documents" // 4th International Conference on Object-Oriented Information Systems (OOIS'97), 1997. – P. 299-311.

[8] J. Rogushina, A. Gladun "Semantic Search of Internet Information Resources on Base of Ontologies and Multilinguistic Thesauruses" // International Journal "Information Theories & Applications" Vol.14, 2007. – P.48-54. – http://www.foibg.com/ijita/vol14/ijita14-1-p07.pdf.

[9] A. Gladun, J. Rogushina,"Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources" // International Journal of Intelligent Systems and Applications, 2012, №1, P.11-20. – http://www.mecs-press.org/ijisa/ijisa-v4-n1/IJISA-V4-N1-2.pdf.

[10] A. Gladun, J. Rogushina ,"Management of Ontologic Knowledge for Recognition of Information Objects in the Distributed Recommending Systems" // In Proceedings 3-th Int. Conf., Knowledge Management and Bussines Intelligence», Kharkiv. Ukraine, 2013.-V.1. P. 78-82 [in Russian].

[11] J. Rogushina, A. Gladun ,"Ontology-based Competency Analyses in New Research Domains "// International Journal of Computing and Information Technology – CIT, vol.20, 2013, №4.-P.277-291. – http://hrcak.srce.hr/file/146472.

[12] A. Gladun, J. Rogushina ,"Cognitive Networks and Ontologic Knowledge for Increase of Adaptability and Quality of Service in Heterogeneous Wireless Environment"// Proc. OSTIS-2012, Minsk, Belarus, 2012. - PP.493-500 [in Russian].

Authors

Dr.Anatoly Gladun was born in Rivne, Ukraine in 1961. He received the B.Sc. and M.Sc. degrees from Technical University in Lviv, Ukraine in 1984. He holds a PhD in Department of Computer Sciences at the Electrotechnical University (Saint-Petersburg, Russia). He is Senior Researcher of Department of Intelligent Systems at the International Research and Training Centre of Information Technologies and Systems (National Academy of Sciences). He is the author of more than a 190 publications in conferences, journals and books.

His research interests include the development and application of knowledge technologies to different fields such as e-Medicine, e-Commerce, e-Learning, Retrieval Systems, Network Management. His research interests include the Intelligent Software Agents (models, architectures, methodologies of development) and their Application; Network Management and Semantic Web.

He is an Associate Professor at the Department of Computer Science (half-time) at University "Kiev-Mogyla Academy", (www.ukma.kiev.ua). Courses: "Intelligence Networks and Systems", "Computer Networks and Mobility Communications", "Developing Internet-nodes and services" for graduate students.

Dr.Julia Rogushina was born in Kyiv, Ukraine, in 1967. She received her PhD degree in Computer Science in Glushkov`s Institute of Cybernetics, Kyiv, in 1995. Now she is a senior researcher at the Institute of Software Systems, National Academy of Sciences of Ukraine.

Her research interests include the ontological analysis on base of the Semantic Web technologies, intelligent information retrieval, inductive knowledge acquisition, intelligent information systems and software agents behaviour. She has published monographs “Agent technologies”, “Knowledge management on base of ontologies for the distant learning”, “Knowledge-oriented means of the Web semantic retrieval”, several textbooks and more than 150 publications in scientific journals and conferences. Her best dog is Staffordshire terrier Kelly Lada de Mandraka.

Julia Rogushina has been involved in several national research projects, for example, “Research of intellectualisation means for multiagent information retrieval systems”.