Popova M.S.

PROVIDING A MULTILINGUAL ACCESS TO AN INFORMATION RESOURCE OF THE SPECIFIED SUBJECT DOMAIN AND ITS CATEGORIZED KNOWLEDGE

This paper describes an approach to provide a substantive multilingual access to a knowledge portal integrating knowledge, as well as to information resources which are relative to the specified subject domain based on ontologies. The requirements for a conceptual structure of ontologies are laid down. The analysis was performed in relation to directions of integration of semantic retrieval systems with ontologies for the description of a search subject domain, as well as the requirements set by retrieval systems for the standards of ontological metadata and forms of their representation.
Practical realization of these requirements shall provide users with the most applicable tools to solve the urgent tasks of multilingual access to information resources of the specified subject domain, and to complete the task of definition of different types of information objects, identified by a retrieval system.

Keywords: *information systems, ontologies, knowledge portal, thesaurus, multilingual thesaurus, information resource.*

With a continuous growth of information volume referring to different subject areas, it is necessary to provide efficient information support of a scientific and production process. And in most current information systems (IS) (for instance, a corporate information system) and information resources (for instance, a web-catalogue or a portal) the data are normally represented in form of text documents.

The most natural form of data submission still means the representation in form of several interrelated facts. Such representation of information and efficiency of its retrieval can be provided only by IS's, which use not only basic knowledge about the world, but also the knowledge about subject domains (SD), covered by the system. Today this knowledge is represented by means of ontologies [1].

In this paper we have defined the requirements for a conceptual structure of ontologies when providing a multilingual access to information resources of a subject domain. We have also analyzed the directions of integration of semantic retrieval systems with ontologies for the description of a search subject domain and the requirements set by retrieval systems for the standards of ontological metadata and forms of their representation.

Ontologies are the interoperable representation of knowledge with already created generally accepted standards, representation languages, tools of editing and logical deduction, as well as fundamental mathematical base. However, there are up to now no clearly defined methods to control the knowledge based on ontologies, which could be directly realized in application systems, particularly – the methods to recognize different information objects (IO), being of interest for a user and necessary to solve the current problems. Perception and recognition of IOs are the most important tasks at the development of intelligent information systems based on knowledge [1, 2].

IO is a certain essence containing the information about any real or virtual object (subject, creature, event, process, etc.) which is a uniquely identified material or non-material essence of the real world, describing its structure, attributes, constraint and, probably behavior.

Access to IO can be based on the ontology of the respective SD. Every IO corresponds to a certain ontology class and it is a sample of the certain class, as well as of the structure specified by this class. There could be different links between IOs, which semantics is defined by the relations specified by the respective ontology classes. These relations could also be used for retrieval. A substantive multilingual access to the classified knowledge and information resources of the specified subject area is ensured due to advanced means of search and navigation [3].

A problem of perception, recognition and interpretation of objects in information technologies is a complex problem which could be split into separate sub-tasks [4]. A traditional recognition of images, recognition of speech and recognition of text are just private cases of a more generic problem. Recognition means detection in any information resource (IR) of the information about any IO a user is interested in. In case of the ontological approach, a recognition task can be reduced to a task of classification, when IR's and their fragments are connected with different classes and samples of classes of the ontology a SD user is interested in. It should be noted that such classification depends on the ontology of a SD.

In modern intelligent applications a task of recognition is usually formed as follows: it is necessary to find:

- IR's relevant to a task specified by a user;
- information referred to several classes the ontology of a SD or to their samples;
- structures reflected by means of the ontology of a SD and which are essential from the point of view of the problem set in front of the user.

Today for interoperable representation of knowledge people more often apply an ontological approach which ensures a repeated and joint usage of the accumulated knowledge.

Different sources offer different formal modes of ontologies representations. All of them contain:

- a variety of terms (notions, concepts), which could be split into a variety of classes and variety of samples;
- a variety of relations between notions with clear selection of the relations "class-subclass", hierarchy (taxonomic) relations and relations of synonymy (resemblance), as well as the function which is a special case of relations for which the n-th element of the relation is uniquely defined by $n - 1$ of previous elements;
- axioms and functions of interpretation of notions and relations.

Formally the ontology O is represented by a triple $O = <X, R, F>$, where X is a variety of concepts, R is a variety of relations between concepts, F is a function of interpretation of concepts from a variety X and the relations from R [1]. This model is general, whereas in practice more definite specific models are used, in particular, those related to the standards and languages of ontologies representation Web Ontology Language (OWL, ontological language for information networks). For OWL representation and processing there is a theoretical basis in form of set of DL logics, ensuring a proof of logical deduction based on ontologies, and different means of logical deduction help to conclude based on structured data (OWL and RDF). Having analyzed expressive power of different tools of representation of ontologies and formal models of ontologies, we can assert that the existing technologies offer means of description of ontologies which differ by capabilities and complexity: RDF Schemas are the simplest level for the representation of ontologies, and OWL Full is the most complicated one. A choice of mean to represent ontology depends on of a problem's characteristics for which it is being developed.

Ontology could be considered as a basis for representation of an information object's structure, described by an ontology class, and different IR's – as the sources to create the samples of this class. Such approach helps with the integration of information coming from different sources, and with the formation of knowledge required by a user. And the task here is divided into several subtasks:

- formation (or search) of the ontology reflecting the structure of an information object (or a variety of objects), the knowledge of which are necessary for a user to solve the problem he faces;
- retrieval of IR's clearly or unclearly containing the information about these SD's;
- elicitation of knowledge about SDs from IR;
- representation of the elicited knowledge in the form which is comprehensible and convenient for a user.

Ontology can be used as the specification of information system, as it defined the knowledge required to carry out the tasks of the system being developed. Joint and multiple usages of ontologies referred to different SD's and applications may significantly improve the architecture of intelligent information systems.

Retrieval of information is the process of recognition, in the amount of available information, of those objects that satisfy certain (clearly or unclearly specified) conditions. The result of such retrieval can be ontology, a separate document, a document fragment or the information about any objects of the type set by a user (a geographic point, a human, an organization, a product, etc). Information retrieval is currently a greatly developing field of science, which popularity is caused by an exponential growth of information amount. Its aim is to help a user to meet his information needs.

Retrieval of information consists in 4 stages:

- definition of an information need of a user and construction of a retrieval query, reflecting this need;
- definition of a cluster of available IR's and their characteristics;
- elicitation of information from the recognized IR's;
- provision of a user with the results of retrieval in the form which can help to satisfy his information need.

Existing information retrieval systems (IRS) have a number of serious disadvantages. Even those IRS that use

knowledge and ontologies, do not always properly analyze the context to solve the problems of homonymy and synonymy in a natural language (NL). Besides, the elicitation of information from the determined sources is realized insufficiently effectively, as well as the comparison of information with the information object a user is interested in, and integration of information from various IR's, that requires from a user to carry out this routine work by himself. It is connected with a fact that these operations require the knowledge of a retrieval SD's, represented in an interoperable form, and available for a retrieval system for a re-use.

On some extent solving of the problems of automated creation and replenishment of ontologies can be made by means of a word sense tagging. Based on semantically marked texts one could automate the creation of ontologies with the terms corresponding to the tags of such word sense tagging, elicit information about the links between terms from the links between the text fragments marked accordingly. But the problem is that ontologies created like this use only those tags that reflect information interests of a certain association in whole, and not a specific user solving a specific task.

Semantic search is a sort of an automated information retrieval with consideration of semantic aspects of a user's query, available information resources (IR), among which a search is performed, and context of the query [1]. At a semantic search its subject can be not a certain IR, or its fragment, but and information object of a specific class.

Systems of semantic search are often reduced to retrieval systems capable of the processing of NL queries, or to the systems processing metadata about resources. A semantic search is however a wider notion. As a rule it is a semantic analysis of natural language constituent part of the objects and of a user's query. For semantic analysis one could apply a content analysis, a method of semantic cases, association analysis, a method of subject classification based on a model of structural representation of a text, semantic differential, latent semantic analysis, etc.

A key point of a semantic search is that not only formal parameters of the considered objects are being analyzed, but also their semantics. Efficiency of search can be significantly improved by means of intelligent analysis of the objects, for which an agent and ontological approaches are applied. Ontologies can be applied for the description of semantics of content of a certain document and its structure, and for the description of the objects information of which is necessary for a user.

As we go to a semantic search there occurs a problem to recognize different IO's. Recognition of IO's implies the detection in any of IR's of the information about an IO, a user is interested in. Recognition of IO's can be considered as a special case of recognition of images, which is defined as an attribution of initial data to certain classes by a separation of essential features, that specify this data within a total mass of data.

Recognition of images requires a classification of objects from a specified variety by available descriptions of objects and classes [1]. A standard task of image recognition is that for a set of objects M, described by a set of features and represented in form of combination of disjoint subsets (classes) $M = \bigcup_{j=1}^{P} M_j$, such that $\forall j \neq k$, $M_j \cap M_k = \varnothing$, and for the K-part of the objects from M, $\forall K$, $\in K$, $i = 1, n$, $K \subseteq M$ we know the class they refer to, it is required to define the class of these objects by a set of values of features of the objects from $L = M \setminus K$.

The creation of intelligent systems for retrieval and recognition of different information objects in a specified subject domain requires a development of methods and tools not only to elicit and process new knowledge but also re-use the knowledge received earlier. One of possible solutions of this task is a usage of ontologies ensuring the storage, search, estimation and safe application of ontologies, as well as change management, control of personification, separation, presentation and integration. And there occurs a problem of formation of the requirements to meta-descriptions of ontologies and development of a single standard realizing these requirements.

Consequently it becomes clear that it is necessary to study which particular information about ontologies help to ensure their usage, for example, for the tasks related to a semantic search and recognition of different types of information objects, corresponding to the classes of ontologies, as well as which functional capabilities of ontologies can be used herein.

Apparently, for a semantic search not only methods and algorithms of processing of IR semantics are required, but also formally represented knowledge about the search SD's, in particular, the ontologies of the respective SD. In general one could choose one of three possible sources of such ontologies – to create a new ontology, to modify the already available one, or to find a one created earlier and satisfying a user's needs. For the creation of ontologies one could use a manual construction of ontology by a SD specialist, automated processing of metadata about IR's, acquiring of ontological knowledge from natural language texts, application of an inductive inference. Modification of ontologies can be carried out manually or in an automated manner by means of logical operations with the existing ontologies (crossing, combinations, difference, etc.).

A user needs either to create ontology reflecting his information interests in a specific SD by himself, to use it later at the searching of IR, which is rather complicated, or to re-use ontologies, created by other researchers and covering the field of his interests, or without any changes, or extending and modifying it. But for this a user needs the tools for a search of the ontologies which are connected with a required SD and which have a required degree of complexity and detail.

Replenishment of SD ontology can be made by means of a multilingual linguistic analysis of the texts defined by a user in accordance with his information needs. Currently a number of methods and tools is developed for an automated construction of ontologies and thesaurus by full-text IRs.

There are means of comparison of queries and resources, oriented towards the provision of a multilingual access to information resources and towards a semantic search of a specified subject domain, which could be used for a search of other types of resources, as well as means of comparison of ontologies.

Therefore, there is a task to support several languages in a subject domain. This task can be solved by extension of available ontologies with required linguistic knowledge. However, introduction of additional essences and relations to an ontology shall make it rather awkward and opaque, it will hider its development and maintaining. That is why it is logical to extend a system of knowledge by another component which is a multilingual thesaurus [2, 3, 4], whose introduction can make a portal adjusted to the "understanding" of multilingual resources, to the support of navigation within its information space and perception of queries in different languages.

Usage of thesaurus and ontology has been frequently discussed in research papers. Basically in theory applications of dyads "thesaurus-ontology" are considered in a task of processing of texts and in information retrieval [4, 5]. The approach described in this paper helps to provide a joint usage of ontology and thesaurus not only to solve these tasks, but also to integrate this knowledge and resources relevant to the specified SD's in a single information space and provision of efficient navigations within this space by means of use of a natural language by a user. In this context we propose to use all relations in ontology, which are necessary for representation of knowledge about this SD, with the construction of pure linguistic relations (synonymy, equivalence, etc.) for thesaurus.

The creation of a subject domain ensures a substantive access to the knowledge and information resources of the specific topics; it helps to solve the tasks related to the processing of text resources, representation of their content in form of interrelated facts, organization of retrieval of the required information and its visualization in different languages. Solution of these tasks lies in the usage of knowledge about the structure and terminology of this SD, about the structure and typology of resources, as well as knowledge about the properties of languages, the texts if these resources are represented in.

Technology of creation of such SD proposes to organize its systems of knowledge based on the integration of multilingual thesauruses and ontology of the related SDs.

We shall describe a knowledge system (*KS*) of the portal as a quadruple of the form $KS = (Os, Th, ICs, Ir)$, where *O's* is the ontology of knowledge portals; *Th* are multilingual thesauruses of subject and problem domains of knowledge portals, *IC's* are information replenishments of knowledge portals, which are built on the basis of the structure specified in the ontology *O's*; *Ir* is an information resource integrated into an information space of the knowledge portal.

Representation of ontology requires a formalism that ensures flexible means for the description of the notions of its problem and subject domains and all various semantic links that occur between them. Important requirements to it lie in a possibility to organize the SD's notions into a hierarchy "general-private" and in a support of inheritance of the properties by these hierarchies. This formalism shall also be ensured as a possibility to set restrictions towards the values of SD objects properties and descriptions of relation semantics in form of axioms.

To provide with formalism which satisfies the described above, the following meta-ontology is proposed $O = (C, R, T, D, A, F, Ax)$, where $C = \{C_j, ..., C_{,,}\}$ are finite non-vacuous sets of classes which describe the data of subject or problem domains; $R = \{R_1, ..., R_m\}$, $R_1 \subseteq C * C$, $R = \{R_T, R_P\} \cup R_A$ are finite non-vacuous sets of binary relations specified on the classes (notions); R_T are antisymmetric, transitive, non-reflexive binary relations of inheritance, which specify a partial order on the sets of notions C; R_P are binary transitive relations of inclusions ("whole-part"); R_A are finite sets of associative relations множества; $T = \{t_1 ..., t_k\}$ are the finite non-vacuous sets of standard types; $D = \{d_1 ..., d_k\}$ are the sets of domains $d_1 = \{s_1 ..., s_k\}$, where s are the values of a standard type "string"; $TD = T \cup D$ are the generalized types of data including the sets of standard types and sets of domains; $A = A_C \cup A_R = \{a_1 ..., a_w\}$ are the finite sets of s=attributes which describe the notions properties C ($A_C \subseteq C * TD$) and relations R_A ($A_R \subseteq R_A * TD$); F are the sets of restrictions for the values of notions attributes and relations, i.e. predicates of the type $p_i = (e_1, ..., e_m)$, where e_k is a name of attribute ($e_k \in A$), or a constant ($e_k \in td_j$ where $td_j \in TD$); Ax are the sets of axioms that define semantics of classes and relations of ontology. The relations of the inclusion "whole-part" R_p have the transitivity properties, due to which the transitive closures could be performed in IRS.

A set of associative relations R_A is defined by a user. These relations help to organize an IRS, as well as the navigation within the portals' content. An important feature of the relations R_A is that they have their own attributes specialized in a link between arguments.

There are IRS's oriented on a search within structured data (particularly, represented in the formats OWL and RDF). The system notates documents in the formats N-Triples, RDF/XML and N3 (RDF). Here there is a processing of the documents entirely composed by means of these languages, as well as subject domains including the elements of a semantic code. However in these IRS a search as carried out usually by key words and that is why there occur the problems with the recognition of ontology interesting for a user and with a tolerable complexity.

Besides, a major part of IRs represented in the specified subject domain is not attended by RDF metadata (and if it is, then a reliance on this data remains an open point) or by any ontologies and the construction of IR ontologies can be only party automated and anyway it requires a human participation at a number of stages, being nevertheless rather long and оставаясь при этом достаточно длительным и labor-intensive process.

Over the last years ontologies have been applied in different application intelligent information resources. It should be

noted that despite a great interest of the Russian researches in various aspects of an ontological analysis, an automated construction of ontologies, their comparison, replenishment and analysis, it is exactly the questions of access, storage and creation of meta-descriptions of ontologies that still remain almost uncovered. However, little progress has been made around the world for an efficient solution of the aimed at re-use of ontologies.

The developments related to the use of ontologies are provided by new technologies in this field. In a specified SD there are many ontologies from different SD's. But due to a complexity of the ontologies' structure and due to their large amount, it is difficult for a user not only to modify and supplement them, but to find ontology by a topic and a level of complexity in general.

The ontology built this way describes subject and problem domains, as well as specifies the structures by the representation of real objects and relations between them. That is why the data is represented as the sets of interrelated IO's, corresponding to certain ontological notions with the structures specified by it. Semantics of the relations in IRS and between IOs are defined by the relations which are specified between the respective ontological notions. The sets of such IOs and their relations make an information contents.

Thesaurus provides the possibilities of interoperating in several languages, including navigation, search and processing of information resources which are represented in different languages.

Thesaurus looks as the follows $Th = (Tr, At, Rt, RTO, Axt)$, where $Tr = \{Tr_1, ..., Tr_n\}$ is a finite non-vacuous set of terms representing a notion and relation of a certain SD; from this set of terms Tr we should следует distinguish sub-sets of basic terms $Trb \subseteq Tr$, considered to be the most appropriate for the representation of the notion's name and the relation; $At = \{at_1, ..., at_w\}$ are finite sets of attributes which describe a property of terms Tr; $Rt = \{Rt_1, ..., Rt_m\}$, $Rt_i \subseteq Tr * Tr$, $Rt = R_{SBT} \cup R_{SNT} \cup R_{AT} \cup \{R_{USE}, R_{UF}, R_{LE}, R_{TO}\}$ are finite non-vacuous sets of binary relations which are specified by SW terms in accordance with the accepted standards of GOST and ISO [1]; R_{SBT} iare sets of binary relations that associate a certain term with a more generic one; R_{SNT} are sets of inverse relations to R_{SBT}; R_{AT} are finite sets of associative relations between terms; R_{UF} are binary relations, connecting more appropriate terms with the synonyms in the same language; R_{USE} is an inverse relation to R_{UF}; R_{LE} are relations of lexical equivalences between terms, which define the same notions but in different languages; R_{TO} are relations being established at the correlation between the terms of thesaurus and the notions or relations of an ontology, i.e. $R_{TO} \subseteq Trb * Eo$, where Trb are sets of the basic terms of thesaurus , $Eo = C \cup R$ are multiple notions and relations of ontologies; Axt are sets of axioms that define semantics of links between all terms.

The first step in this direction is the approaches to create collections of ontologies and the respective resources. Initial projects of acquisition of the existing ontologies offered the creation of systems of ontology control, with different functions to manage, adaptation and standardization of ontological groups. These systems are important tools for a grouping and reorganization of ontologies, for their further re-use, integration, technical maintenance, presentation and version handling.

The system of ontology control is software for the storage, organization, modification and elicitation of knowledge from ontologies, which sustains semantics for the functions of storage, organization, modification and elicitation of knowledge form a specified SD. Moreover, in different realizations it may contain and support many other functions related to the processing of ontologies. It is an analogue of a database control system (DCS) intended to work with a specific content – ontologies, with consideration of their semantics, and its major aim is to support a multilingual access to knowledge and its re-use by a human or machine.

Another problem is that ontologies which are trying to represent knowledge by a rather wide SD turn out to be too awkward for an efficient use. But the ontology's modularity is not sufficient for a re-use of ontologies, if the developers cannot efficiently find the required modules. That is why there occurs a necessity in the respective infrastructure, which could support an intelligent research of ontologies and the choice of them by end-users.

From a technical point of view practical implementations of ontologies are substantially different. There is a problem of interoperability between them, as the developers apply different methods and technologies for an integration and use of metadata. Besides, most existing ontologies do not support such functions as modularity and versioning well enough, as well as the relations between ontology and a development environment for the support of the whole life cycle of ontology.

The scientific literature describes a lot of variants of design of the systems of access to information resources and information retrieval. More often the estimations of information retrieval systems use used for this purpose [1]. For instance, it is accuracy and completeness that are mostly used for the estimation of information retrieval.

A user can turn to the ontologies created by other users – he can overview them, set a search context by them, copy their required fragments, but he has no right to modify them. IRS may ensure a search of ontologies that contain the terms introduced by a user, as well as a search of ontologies similar to the ontology chosen by a user. It helps to create groups of users with common information interests and prevent from duplication under the addressing of similar multiple queries by different users.

One of the major problems is related to the fact that to describe the information needs to a SD a user had to make significant efforts to create a respective ontology or to describe it not clearly enough using one of the offered ontologies.

When using external ontologies, this problem is not solved completely but is considerably simplified – a scope of content, as well as the availability of the standard for a description of semantics of the represented ontologies helps

to find the ontology which is sufficiently close to a user's interests. Moreover, they provide with a number of qualitative assessments of the ontologies being stored in them. That is why the integration of ontologies made possible a significant extension of the field of a retrieval system application, it made it efficient in the recognition of different types of information objects, as well as it made possible its joint application with different intelligent permissive systems with a multilingual access.

A proposed approach ensures a multilingual substantive access to knowledge and information resource of a specified subject domain based on joint uses of ontologies and thesauruses. The existing relations between terms and notions are provided by a visualization of information in different languages, and besides the preconditions of their joint use during retrieval and processing of information are made.

A relevancy of the problem of development and use of ontologies and thesauruses is confirmed by the research initiatives analyzed in the article. Whereas ontologies are a mechanism of interoperability and data exchange between information objects, the ontologies themselves are almost always created independently. There is no formulated common understanding of annotation for ontology by means of metadata, and no common ways of identification of ontology's version. Different ontologies use various technologies and methods of annotating and editing of ontologies, not limited by any standard agreements during a whole life cycle. To solve these problems it is important to achieve interoperability between ontologies through common interfaces, standard formats of metadata.

Nowadays scientific researches are being actively carried out aimed at the organization of relations of standards development for meta-descriptions of ontologies. The solution to this problem would ensure to perform a global search of ontologies and their components not only by key words, but also on the level of their semantics, which shall provide a multilingual access to re-use of represented ontologies of the specified DS. Practice of integration of the existing ontologies with semantic retrieval systems shall help to form the requirements to the new standards of ontological metadata and forms of their representation.

References

1. **Gladun A. Y.** Ontology as a tool of recognition of intelligent information objects in distributed applications and systems / A.Y. Gladun, Y.V. Rogushina // Trans. of the XVIII-th International research and practice conference "Information technologies in economics, management and business". – K.: European University, 2012. – P. 51-55.

2. **Zhuravlev Y. I.** About an algebraic approach to the solution of the tasks for recognition and classification / Y.I. Ahuravlev // Problems of cybernetics. Issue 33. – M.: Science, 1978. – P. 5-68.

3. **Velichko V.** Structuring of associations ontology for the abstracts of natural language texts / V. Velichko, V. Gladun, L. Svyatogor // International Book Series, N.2. Advanced Research in Artificial Intelligence. Supplement to the International Journal "Information Technologies & Knowledge". – 2008. – V. 2. – 153 p.

4. **Efimenko I. V.** Ontological modeling of corporate economics and economics of industries in modern Russia: Part 1. Ontological modeling: approaches, models, methods, tools, solutions / I.V. Efimenko, V.F. Khoroshevsky // Mathematical methods of solution analysis in economics, business and policy. – M.: Publ. House of Higher School of Economics, 2011. – 76 p.

5. **Guarino N.** Formal Ontology in Information Systems // Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press. – P. 3-15.

6. **Zagorulko Y. A.** Construction of portals of scientific knowledge based on ontologies // Computational technologies. – 2010. – Vol. 12. – Spec. issue 2. – P. 169-177.

7. **Dagobert S.** Multilingual thesauri and ontologies in cross-language retrieval // AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. – Stanford: Stanford University, 2007. – 306 p.