METHODS AND TOOLS OF KNOWLEDGE MANAGEMENT AT THE SEMANTIC WEB ENVIROMENT

Julia Rogushina

***Annotation**: the main problems of ontological knowledge management for Web are analyzed, e.g. the problem of integration of knowledge from different sources, knowledge acquisition and knowledge retrieval for specific task. Methods of automated generation of metadata described the semantics of informational resources and for personalized search on base of thesauri and ontologies of user subject domain are proposed. These methods are realized in design of informational retrieval system MAIPS where the retrieval procedure is personified on multiagent paradigm and ontological analysis. MAIPS uses technologies and standards of Semantic Web and Web 2.0 (e.g. OWL – for interoperable ontology and thesauri representation, RDF – for metadata representation of informational resources, tag clouds – for visualization of search thesauri, social services – for user interaction), some set-theoretic operations on thesauri and creation of thesauri by natural language texts are realized. Text readability criteria are used for retrieval of information pertinent to personal informational needs of user.*

***Key words**: Semantic Web, knowledge management, ontology.*

***ACM Classification Keywords**: I.2.4 Knowledge Representation Formalisms and Methods*

Introduction

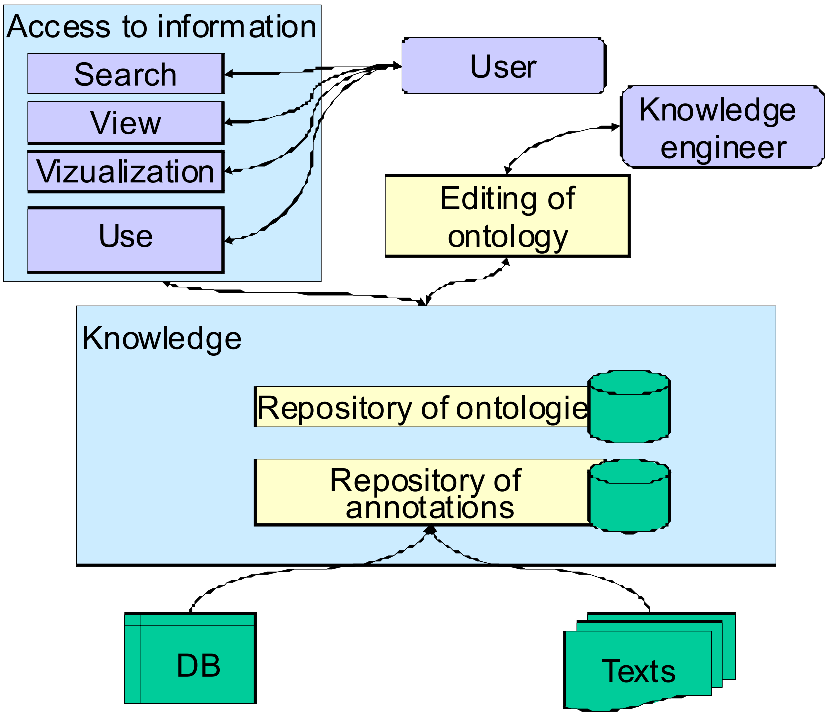
Now a lot of Web applications are intelligent and use knowledge about some subject domain or produce some new knowledge. In such applications knowledge is represented in interoperable form and can be reusable. For such representation ontological approach is widely used because ontologies have a fundamental theoretical foundation (descriptive logic).

Ontologies typically provide some general vocabularies that describe different domains of user interest or specialization of informational resource and define the meanings of terms used in the vocabulary. The ontology representation contains data and conceptual models, for example, sets of terms, classifications or theories.

Problems of knowledge management for Web

Main problems of knowledge management for Web deal with (pic.1):

- Integration of knowledge from different informational resources (e.g. integration of ontologies built on base of different texts from one subject domain);
- Search of inconsistency of knowledge acquired from content of different informational resources and rating of their adequacy and security;
- Knowledge acquisition from accessible information and it's representation at form understandable to user;
- Search of knowledge that user needs for solution of some specific tasks;
- Automation of metadata creation and improvement that correctly describes the content of informational resources (textual or multimedia) on semantic level, and efficient search of such metadata.



Pic.1. Main elements of ontological knowledge management

A lot of other examples of similar exists bat all of them come to the following ones:

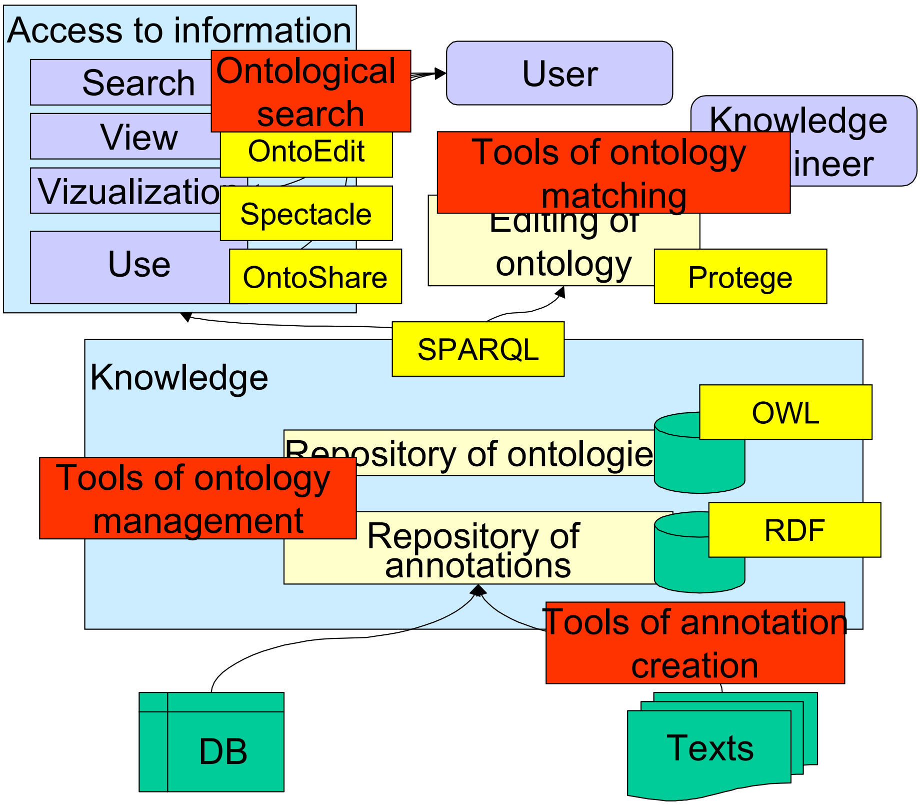
1. Selection of means for knowledge representation (sufficiently powerful to satisfy the different requirements of users but available to rapid processing and understandable for human): Now for these goals ontologies are widely used but the problem deals with selection of ontology representation language version (OWL 1.0 versus OWL.2.0, OWL Lite, OWL DL, OWL Full, RDF, RDF Schema etc.) [1]. Domain ontology is the certain part of knowledge that describe important concepts and relations that can be used for solution of problems at this domain.
2. Methods of acquisition of new knowledge on base of some informational resources (for example, creation of metadescriptions of informational resources, inductive and traductive inference): new knowledge can be acquired from implicit, uncertain, contradictory textual representations but large capacity of such information necessitates some automated methods of their processing. Availability of RDF – language for metadata representation – is a necessary but not sufficient condition for it. For example, automated creation of metadata that describes the natural language document on semantic level requires to use: 1) methods of linguistic analysis; 2) knowledge of subject domain (e.g. domain ontology); 3) application-dependent methods of inductive, deductive or traductive inference oriented on processing of specific structures of knowledge (e.g. RDF triplets).
3. Methods of matching of different informational objects on semantic level (e.g. integration of two ontologies or detection of differences between them, matching of informational query and informational resource relevant to this query, discovery of subject domain of informational resource by analysis of it’s content): these problems are not trivial and don’t reduce to traditional search because they have to analyze rules and knowledge of subject domain and their formal

representations by special matching algorithms. The matching operations deal with following challenges: large-scale evaluation, performance of ontology-matching techniques, discovering missing background knowledge, uncertainty in ontology matching, matcher selection and self-configuration, user involvement into the process of matching, explanation to user of matching results, collaborative ontology matching, alignment management and reasoning with alignments [2].

4. Quality rating of new knowledge (veracity, consistency, actuality, completeness). It needs to develop the different models of knowledge representation, to use the appropriate mathematical apparatus (e.g. first-order sentence theory, descriptive logics) and to evaluate the quality of ontologies concerning to real world and informal knowledge about real world.

At the present stage of IT in the majority of cases Web applications use standards and technologies of knowledge management developed by Semantic Web project. Knowledge management in Semantic Web environment needs in creation of adequate tools for retrieval, acquisition, store and use of knowledge subject to such properties of up-to-date Web as dynamics, heterogeneity, very large capacity and orientation on semantics.

The main component of Semantic Web conception is an ontology use that allows to formalize knowledge about subject domain. Ontology in contrast to XML Schema is a knowledge representation not a message format. Different instrumental tools provide following possibilities: ontology creation and their linking with different informational resources, checking of ontology's consistency, refinement of ontology and executed of inference operations on ontologies (pic.2).



Pic.2. Knowledge management architecture on base of Semantic Web

Ontologies as a knowledge representation means

Analysis of publications shows that ontologies are the adequate and effective means for knowledge modeling about different subject domains, informational resources and other objects. Different authors represent various formal models of ontology but all these models include [3]:

- the set of concepts that can be subdivided into the set of classes and the set of individuals;
- the set of relations between concepts where the some subclasses of relations («class-subclass», hierarchical, synonymy etc.) and functions (as special relation where the n-th element of relation is uniquely defines by other n-1 elements) can be separated;
- axioms and interpretation functions of concepts and relations.

Formal model of ontology is a triple O=<X, R, F>, where X is a set of concepts, R – a set of relations between concepts from X and F – interpretation functions for concepts from X and relations from R. This is a general model, and in practice more precise models are used. For example, in [4] ontology is defined as a structure that includes identifiers of concepts, identifiers of relations, identifiers of attributes, data types and hierarchies of concepts and relations. In [5] ontology is defined as a tuple that, in addition to sets of classes, individuals, relations and data types, contains a set of values and some special relations (specialization, exception, creation of individual and assignment).

Existing technologies of the Semantic Web propose various means of ontology representation that differ one from others by their expressiveness and their complexity: RDF Schema is the simplest representation and OWL Full is the most powerful. Decision of ontology representation depends of problem specifics.

Languages for ontology representation can be viewed as syntacsic variants of Description Logic (DL). The fundamental modeling concept of a DL is the *axiom* - a logical statement relating roles and/or concepts. There are many varieties of Description Logic and there is an informal naming convention, roughly describing the operators allowed: F – Functional properties; E – Full existential qualification (Existential restrictions that have fillers other than owl: Thing); U – Concept union; C – Complex concept negation; S – An abbreviation for ALC with transitive roles; H – Role hierarchy (subproperties - rdfs:subPropertyOf); R – Limited complex role inclusion axioms; reflexivity and irreflexivity; role disjointness; O – Nominals. (Enumerated classes of object value restrictions - owl:oneOf, owl:hasValue); I – Inverse properties; N – Cardinality restrictions (owl:cardinality, owl:maxCardinality); Q – Qualified cardinality restrictions (available in OWL 2.0, cardinality restrictions that have fillers other than owl:Thing); (D) – Use of datatype properties, data values or data types. The prototypical DL Attributive Concept Language with Complements (ALS) is a simply AL with complement of any concept allowed, not just atomic concepts. The description logic SHIQ is the logic ALC plus extended cardinality restrictions, and transitive and inverse roles. The naming conventions aren't purely systematic so that the logic ALCNIO might be referred to as ALCNIO and abbreviations are made where possible ALS used instead of the equivalent ALUE. The design of OWL is based on the family of DL. The Protégé ontology editor supports SHOIN(D). OWL 2.0 provides the expressiveness of SHOIQ(D), OWL-DL is based on SHOIN (D), and for OWL-Lite it is SHIF(D).

Semantic search as an important component of Web knowledge management

We think that one of the most important tasks in knowledge management for Web deals with semantic informational search – in a lot of intelligent Web application informational retrieval is a part of a system or is called as an external service. Semantic search is a superstructure on traditional retrieval procedure where (i.e. more efficient satisfaction of user's informational needs) processing of knowledge (about user, his/her personal informational needs and interests; about informational resources accessible for retrieval mechanism) is used for the purpose of increasing of search pertinence [6]. The result of semantic search can be not only the concrete Web document or fragment of such document but some more complex informational object:

1. interesting to user information acquired from accessible informational resource (textual or multimedia) where this information contains implicitly;
2. a list of informational resources with some semantic annotations deal with user's query and user's personal preferences;
3. integration of knowledge contained in different informational resources;
4. informational object of specific for subject domain class (corresponding to some concept of domain ontology) – for example, organization, geographical object, human or scientific article;
5. composition of classified informational objects (e.g. human with some characteristics that work in organization of specific type and live in some concrete city).

On base of analysis of current state of work in sphere of informational content representation and methods of programming for Semantic Web we can mark out some main problems that we have to solve in process of design of intelligent Web application realized the semantic search procedure (i.e. the questions that have not now some universal standardized methods of solving and for which open software products are not realized):

- automated creation of meta-descriptions of informational resources that reflect not only formal characteristics of documents but their semantics that deals with some subject domain;
- generation of semantic markup of natural language documents by ontological concepts;
- automated creation and enhancement of ontology (at initial stage and for existing ontologies) on base of informational resources processing and by use of expert knowledge, particularly:

- formation of thesaurus of natural language informational resource;
- formation of initial ontology of subject domain by the set of natural language documents selected by user;
- enhancement of ontology of subject domain by the set of natural language documents;
- acquisition of ontological information from meta-descriptions of informational resources;
- use of inductive inference for discovery of relations between the ontological concepts;

- operations on ontologies (the most necessary operations are consistency valuation of ontology, matching of pair of ontologies and integration of terminological base of different ontologies);
- semantic search that take into consideration ontological knowledge about subject domain, user and task that user try to solve.

It is not easy for user to formalize the query for semantic search that reflects his/her informational need (as a user we consider either human or agent – software entity with some goals and intentions) because this formalization has to reflect:

1) the description of problem that needs some information for it's solving;

2) what information user has before this query;

3) what level of complexity and form of knowledge representation user can understand;

4) how to acquire the necessary knowledge from accessible documents.

Semantic sears has some important differences from traditional one realized in usual information retrieval systems (IRS) that operate at Web environment:

	Traditional IRC	Semantic IRC
Query	The set of keywords	Informational need deal with some subject domain and problem
Information for search personification	History of user queries	Models of user and his informational needs
Search results	Document with keywords	Knowledge acquires from relevant to query documents that describe some interesting for user object (document, human, organization etc.)
Source of information about accessible IR	Index DB of IRS	Index DB of IRS and their metadata
Description of subject domain interesting for user	-	Domain ontology

Linguistic methods in creation of ontologies of natural language informational resources

The algorithm of semantic markup of natural language texts is proposed. This algorithm factors into morphological and syntactical properties of natural language and knowledge about subject domain. As a result of this algorithm we receive the text where some fragments are linked with concepts and relations of domain ontology. The other result of this step is a set of rules that provide links between ontological entities and word forms of natural language. The input of first stage are: O_0 – initial ontology of subject domain containing the most obvious for user concepts and relations; T_0 – the set of natural language texts that describe interesting for user domain (texts from glossaries, manuals, textbooks, Wikipedia articles, other well structured definitions of domain terminology). O_0 and T_0 can be empty.

On the next step the rules of markup are used for new texts. If in one sentence two or more fragments are marked up by ontological concepts but no fragments are marked by ontological relations then we can add (if necessary) new relation to domain ontology. If in one sentence one fragment is marked up by ontological concepts and other one – by ontological relations then we can add (if necessary) new concept to domain ontology.

This algorithm can mark up not only classes bat individuals as well. In natural language the equivalents of individuals are named entities (names, titles etc.).

The results of such semantic markup can be used for development and improvement of ontologies together with linguistic approach. If some text paragraph contains two fragments linked with ontological concepts and a fragment linked with ontological relation and if linguistic analysis of the sentence shows that in this sentence these fragments are associated but the domain ontology don't contains such relation of these concepts then the ontology can be enriched by this relation. Ontology is enriched by the new concept: if one fragment of the text paragraph is linked with some other concept and other fragment – with some ontological relation and linguistic analysis helps to search the fragment that is semanialally associated with these fragments than user can determ a new ontological concept associated with this frafment.

In prosess of linguistical analysis we propose to create and devlope a lexical ontology of domain that contains information about natural text fragments that are associates with concepts and relations of domain onntology. This ontology is created in process of semanic markap of domain texts and then enreached in dialog with user durin the analysis of other texts.

Use of thesauri in semantic search

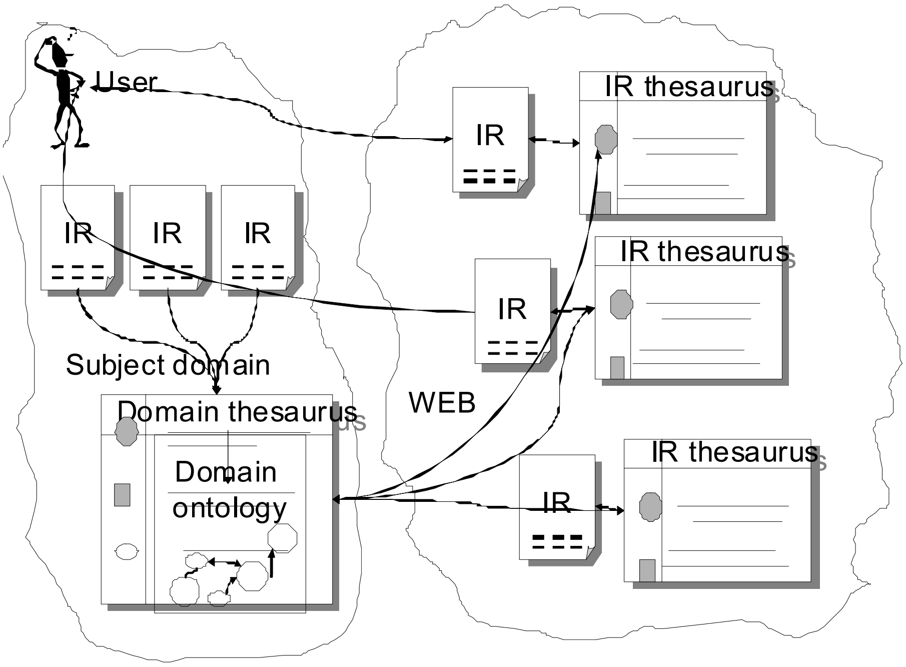
Thesaurus is a special case of ontology T=<X, R, ∅>. In some situations we can match in processof semantic search the thesaurus of domain that is interesting for user with thesauri of avaluable informational resources. The replacement of the ontologies by the thesauri reduces the problem of their generation because we can create thesaurus of natural language document much easier (by lexical analysis) then ontology.

The thesaurus of domain is created as a union of sets that represent thesauri of natural language documents selected by user to describe the sphere of his/her interests. Then user can refine this thesaurus according to IDEF5 methodology for development of ontological models (www.idef.com/IDEF5.html).

If we have an ontology of some domain or informational resource than we can reduce it into the thesaurus. In some situations for retrieval procedure we can take into account only the set X (concepts) and then matching of ontologies can be reduced to comparison of these sets (there is not a deep semantic analysis but this procedure can help to reject informational resources without corresponding terms.

For modeling of ontological relations mereological apparatus can be used. Mereology as a formal theory about parts marks out seven types of relation «the part of», for example, component-object, part-mass, material-object. This classification helps in refining of ontologies if user in process of adding of new relation to ontology explicitly states the mereological type of this relation.

For analysis of the lot of IR an algorithm of thesauri building is proposed: term vocabulary is building by the general list of document words, and then words from user list are thrown away from that vocabulary. User list can contain stop-words for soma subject domain or natural language. If IR has some metadata describing it's semantics (for example, in RDF) then words for vocabulary can be acquired from this metadata. Then this vocabulary is matched with user thesaurus. User thesaurus can be built by extraction of concept names from domain ontology in OWL, as a union of vocabularies of IRs selected by user, manually by user or by combination of these methods (pic.3).



Pic.3. Informational retrieval on base of thesauri

Realization of semantic search in IRS MAIPS

The results of described above research work were used in realization of semantic search system MAIPS. This IRC is oriented on users that have stable informational interests into the Web and needs in regular acquisition of corresponding information. At this system ontologies and thesauri are used for formalized definition of subject domain that is interesting for user, and inductive inference methods provide acquisition of additional information about users by analysis of their permanent query history (e.g., preferences in informational sources, language and size of the text). In addition, the search is personified with a help of individual indexes of natural language text readability that provides the most understandable and valuable information to user.

MAIPS integrates ontological representation of knowledge, multi-agent paradigm and Semantic Web technologies for the purpose of semantic search. The main features of MAIPS:

- use of OWL language for domain ontologies and thesauri interoperable representation;
- realization of set-theoretic operations on thesauri;
- automated thesauri generation by natural language documents;
- use of Web 2.0 technologies (tag clouds – for search thesauri visualization, social services – for user cooperation;
- original sequencing algorithms for searched IRs with account of ontological concepts;
- use of natural language texts readability criteria for informational retrieval with account of personalized user needs;

- original inductive inference methods for generalization of MAIPS operation experience;
- use of multiagent paradigm for modeling of intelligent IRS behavior on base of BDI architecture;
- use of intelligent Semantic Web services paradigm for interoperable description of MAIPS functions.

Inductive inference in MAIPS

IID3M Algorithm. A significant drawback of the well-known algorithm of inductive generalization ID3 [7] consist in the fact that it builds a classification rule only for the two classes. The IID3M algorithm [8] generalizes ID3 to an arbitrary number of classes and takes into account the level of accessibility of attribute values. This algorithm also detects the situation attributes that which carry the most information about the result and thus help in constructing of the smallest decision tree. At the each step the algorithm searches an attribute Ai

$$C(A_m) = \sum_i \sum_j \frac{C(A_m = a_{mi}, R = R_j)}{T(A_m)} = \\ = \max_s C(A_s) = \max_s \sum_i \sum_j \frac{C(A_s = a_{si}, R = R_j)}{T(A_m)} \quad (1)$$

where C (X, Y) - the amount of information $C(X,Y) = \sum_i \sum_j p(X = x, Y = y) * \log p(X = x, Y = y),$ where p (X = x, Y = y) is a probability of combined occurrence of the events X = x and Y = y, and T (Am) is the cost of obtaining the value of Am.

The time for classification of the object by classification rule built IID3M upon the average is not exceed the classification of the object in any other classification rule built on the learning sample. This follows from (1).

MID3 Algorithm. Attribute selection criterion (1) usually gives a good result but the decision tree branching at every step for all possible attribute values causes a number of problems: the specialized rules are built and the number of examples in the nodes is reduced. Separation of the attribute values into two subsets increases the computational complexity by the choice of these subsets. In this regard, we propose an algorithm MID3 - pseudo-binary generalization of IID3M that avoid complex calculations bat allows to remedy these deficiencies. Instead of branching for each value of attribute chosen by (1) it can branch some individual attribute value and other values in the form of a common branch and at each node of the decision tree attribute is a conditionally binary and accepts only two values – "X" and "not X".

For the same attribute these X may be different at different nodes of decision tree. The choice of attribute values that is allocated to the separate branch is doing on base of information entropy measure (2). We choose the value of an attribute that carries the most information about the result:

$$a_{ki} : \sum_p C\left(A_k = a_{kp}, R = r_j\right) = \max_m \sum_j C\left(A_k = a_{ki}, R = r_j\right) \quad (2)$$

Procesing of incomplete data. IID3M and MID3 algorithms are designed for processing of complete data during the consultation. But often it is necessary to classify objects where a full investigation is impossible (because of the complexity, cost and other reasons). Data are incomplete (Maybe-data) if their values is currently unknown but although they can be determed later. On base of these data it is not always possible to unambiguously classify the object but we can select a subset of classes that object can belong on various methods of completions of incomplete data. We propose a method for constructing of such subsets - a method of yellow-green branches [8].

The most adequate way of formalizing and processing of incomplete data is proposed by Codd method "Null Values" [9] according to which data is incomplete if the property value for this object is currently unknown, although the property is inherent to the object and can be determed later. This unknown value can be defined by special constant, and any occurrence of such value may be substituted by the concrete value from the set of acceptable ones. The work with unknown values requires a special three-valued logic with the epistemic truth values (T-yes, F-no, W-maybe) and the corresponding truth table for all logical operations. The application of this logic to incomplete data sorts them into two classes: True-data that values are always accessible, and Maybe-data that values can be not available.

The following technology for inductive generalization of incomplete data is proposed:

Step 1: all n attributes are sortes by two classes according to a priori knowledge about their incompleteness: m attributes whose values are always available in the process of consultation, m≤n, and k attributes whose values during the consultation can be unknown,n = k + m; then from the training set matrix X' obtained from the matrix X by reordering the columns so that the first m columns of X' is formed by True-data;

Step 2: matrix X' is divided into a set of matrixs – matrix A containing m columns and matrixs B [h] containing k columns that the matrix A consists of such rows that for any row of the matrix A there a row of the matrix X ' exists where substring of the matrix A is a substring containing the first m attributes, and there is non another row of A where the first m values of which coincide with the values of this row, and each of the matrixes B[i], $0 < i \le h$, consists of such lines that for any row of the matrix B[i] there a row of X' exists that is a substring of it and the first m values of it a substring of the i-th row of the matrix A.

Step 3*:* decision trees building by the inductive inference algorithms for each of the obtained matrixes wher another meaning - "unknown" (the attribute value is missing, can not be obtained, not known precisely, and so on - the data type Maybe) – is added to the list of possible values for each attribute of B[i] matrix. This value during the consultation is interpreted in a special way and is not considered in decision tree constructing because the situation with an attribute value "unknown" is possible only in the consultation process.

Such indictive methods can be used for Semantic Web knowledge management in two ways: 1) for ontological knowledge acqisition from natural language documents (where the rows of learning sample are the occurrences of ontological concepts into some text and the results are the correlations of text with some domain); 2) for ontology enhancement by new relations and concepts. In MAIPS inductive inference is used also for acquisition of personal preferences of users (by generalization of system experience) and for clusterization of users with similar informational needs.

Summary

A method of use of user subject domain ontologies and thesauri is proposed to increase the pertinence of semantic informational retrieval as an important component of knowledge management. An algorithm of the automated acquicition of ontological knowledge from subject domain natural texts is developed. These methods are realised in intelligent IRS MAIPS oriented on users with permanent informational needs. MAIPS allows to personify the informational retrieval by inductive generailasation of search expirience and by taking into account of personal readability of informational resources.

All proposed technologies can be used for task of competence identification of scientific researchers or learning cources [10] as a part of research planning. This task is an example of a problem that needs an integrated use of different methods of Web knowledge management because knowledge about potential researchers and subject domain has to be acquised from the available Web resources: structured descriptions of individuals (e.g., FOAF) and institutions (organizational ontologies) and their posibilities (for example, in form of Web services) with account of their confidence level (with the help of Web 2.0 technologies and social networks) and from natural language and multimedia documents (and metadata that describe their content) that fix the results of researsh work (articles, monographis, reports, presentations etc.) and then methods of srmantic matchmaking have to be applied to founded information.

Bibliography

1. OWL-S: Semantic Markup for Web Services. – http://www.daml.org/services/owl-s/1.0/owl-s.html.
2. Shvaiko P., Euzenat J. Ten challenges for ontology matching // Proc. of The 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), 2008.
3. Cimiano P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. – Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006. – 347 p.
4. Gladun A., Rogushina J., Martínez-Béjar R., García-Sanchez F., Valencia-García R. Integration of Financial Domain Knowledge on Base of Semantic Web Technologies // Information Models of Knowledge // Edited by K.Markov, V.Velychko, O.Voloshin. – IT H E A, Kiev-Sofia, 2010. – P.106-112.
5. Euzenat J., Shvaiko P. Ontology matching. – Springer-Verlag Berlin Heidelberg, 2007. – 332 p.
6. Gladun A., Rogushina J. Use of Semantic Web technologies in design of informational retrieval systems // in Book “Building and Environment”, 2009 Nova Scientific Publishing, New-York, USA.-P.89-103 .
7. Quinlan J.R. Discovery Rules from a Large Collection of Examples: A Case Study // Expert Systems in the Microelectronic Age -Edinburgh: Edinburgh University Press, 1979, pp. 52-64.
8. Processing of Incomplete Data in Example-Learning Systems / N.I.Galagan, J.V.Rogushina, E.I.Nechvalenko, E.N.Borovaya // Proc. of EWAIC, Sept. 7-9. - Moscow, 1993, - P. 301-305
9. Codd E.F. Extending the Datebase Relational Model to Capture More Meaning // ACM Transactions on Datebase Systems. - 1979. - Vol.4, N.4. - P.397-434.
10. Gladun A., Rogushina J., Garcia-Sanchez F., Martinez-Bejar R., Fernandez-Breis J.T. An application of intelligent techniques and Semanic Web technologies in e-learning environments // Expert Systems with Applications, An International Journal, 2009, V.36. – P.1922-1931.

Information about author

Rogushina Julia – candidate of physico-mathematical sciences, senior staff scientist, associate professor, Institute of Software Systems of National Academy of Sciences of Ukraine, Kyiv-187, 03680, Academician Glushkov avenue, 40, email: ladamandraka2010@gmail.com