ONTOLOGY-BASED APPROACH TO DEVELOPMENT OF ADJUSTABLE KNOWLEDGE INTERNET PORTAL FOR SUPPORT OF RESEARCH ACTIVITIY

Yu. A. Zagorulko, O. I. Borovikova, S. V. Bulgakov, E. A. Sidorova

[1] A.P.Ershov's Institute of Informatics Systems
[2] Russian Research Institute of Artificial Intelligence
Novosibirsk, Russia
zagor@iis.nsk.su

Abstract. This paper presents an approach to the development of specialized Internet portals providing content-based access to systematized knowledge and information resources, relating to a certain branch of science. The information basis of such portals is formed by ontologies, containing the description of scientific and research activity as a whole as well as description of some branch of science, and descriptions of the Internet resources which are relevant to ontologies of the portal. The operation of such a portal is based on a multi-agent system which uses the ontologies and supports both a search for the necessary information in the information space of the portal and automatic update of the portal's content with new knowledge and information resources.

Introduction

Activitiy of individuals, groups and organizations nowadays depends more and more on the information available to them and on their ability to efficiently use this information. At present, a vast amount of information is presented in the Internet, which is already viewed by the community as a potential knowledge base. Provided with access to the ocean of information in the Internet, the user wants to receive only necessary data and documents. However, due to inefficient work of search engines what the user gets is a flood of useless information - very often hundreds and thousands of links to documents irrelevant to the matter, which hinders comprehension and choice of the needed data. The reason for this is that the modern search engines use primarily keyword search mechanisms, which are insensitive to the query semantics, and index Web resources with virtually no tools for analysis of the information presented in them.

In recent years, there have been attempts to attract thesauri and ontologies to describe the Web resource semantics. There are many examples

of tools currently being developed for semantic annotation of Web-pages and documents, when each document is linked to its semantic content. The Semantic Web initiative [1] proposed by the W3C consortium or the tools used in the earlier SHOE (Simple HTML Ontology Extensions) [2, 3] are a few to name. Using such annotations, the intelligent search agents provide more relevant responds to a user query as compared to existing engines. For example, to do this, the SHOE system supplies HTML with a set of special tags for knowledge presentation, and the Semantic Web initiative presumes supplying documents with annotations in the RDF language [4, 5]. There has been certain progress in this direction, however that does not improve the situation in general, since Web-pages annotated in such a manner are an infinitesimal drop in the sea of Web.

However, there exists another problem that cannot be solved with semantic annotation of Web resources, which lies in the fact that, to communicate with search engines, different groups of people engaged in information collection use both their own professional terminology and terms widely used in other communities with other meanings. This also leads to decrease in relevancy of information found and in search transparency, since the existing search mechanisms do not take into consideration the context in which the information exists.

Yet another problem is that search engines are capable of indexing only a small fraction of Web content. Research [6] shows that information volumes unavailable to traditional search engines (Deep Web) exceed the available fraction (Surface Web) by a factor of 400 to 550. The deep Web contains nearly 550 billion individual documents compared to the one billion of the surface Web. A large volume of that information is concealed in databases (references, databases, reports, etc.) and, while being available to their users, are inaccessible to the search engine bots, since access to these data is generally performed by means of Web-forms, and the search engines travel by links.

Dynamic resources, which already constitute a substantial part of the Web, are another impenetrable barrier to the traditional search engines. A modern website is often not a set of WebPages linked to one another, but an application, which differs from traditional desktops applications in the sense that its functional platform is the Internet. Databases containing the information necessary to the functioning of this application are, as a rule, closed for direct access from the outside; information from them appears in a certain context of the Web-application and is presented to the user as an html-page with the elements of design which, as a rule, reflect this context. Even if the search engines manage to reach such

pages, the probability is very small that the full-text page index outside the application is relevant to the meaning of that page in the application context.

One more difficulty is that, as a rule, the user submits queries in his own language, while the Internet contains resources in different national languages. Not all search engines and sites provide quality translation of the query into other languages, and many are limited to one or two languages. Hence the user cannot get a large share of information.

To solve most of the problems mentioned above, we propose to create topic-oriented knowledge portals that have to provide efficient and content-based access to Internet resources on certain topics, including both the external databases and storages directly linked to the portal. In particular, our group has worked up the concept of a specialized knowledge Internet portal which is intended for integration of resources relating to one field of knowledge into a uniform information space and for support of open and convenient access to them.

Such topic portals covering a certain field of knowledge would be useful for both scientists-researchers and students interested in getting knowledge and information allocated in the Internet. Now these people don't get that information because of access to these knowledge and resources is rather complicated, both because most of the fields of knowledge are insufficiently formalized, and because these resources are disembodied, ill-structured, distributed over various Internet sites, electronic libraries and archives, and mostly inaccessible to the traditional search engines.

The information basis of such a portal is formed by the ontologies (see [[7, 8]) describing scientific and research activity as a whole, as well as some particular branch of science, and by descriptions of Internet resources which are relevant to the ontologies.

The proposed structuring of the knowledge system, in which the domain-independent ontology of science is common for all sciences, the knowledge portal can be easily adapted to any subject domain. Thus, to construct the knowledge portal for a certain branch of science, it is enough to develop only its ontology and to connect it with domain-independent ontologies and relevant information resources.

Operation of a knowledge portal is based on a multi-agent system [9] which extensively uses the ontologies of the portal and supports both the search for the necessary information in the resources integrated in the information space of the portal and automatic update of the portal's content with new knowledge and information resources.

This paper discusses the concept of a knowledge Internet portal which should provide content-based access to the systematized knowledge and information resources of a certain scientific discipline.

1 FUNCTIONALITY OF A PORTAL

From the user's point of view, a portal is a domain-oriented Internet resource that allows him to search and view all information relating to a certain subject domain (a scientific discipline).

As an information resource, a portal has the following features:

- provides access to information on various aspects and participants of scientific activity, such as: components of the scientific discipline (subsections of the discipline, research methods, terms and concepts in use), researchers, the information about research groups, communities and organizations involved in the research process;
- allows integration of the resources on portal's subjects which are located in the Internet or in a local network;
- provides the user with advanced tools for finding the necessary information in the entire information space of the portal;
- provides the user with tools for advanced semantic-based search in the Internet;
- provides the user of the resource with information support (for example, announcements on various events and actions);
- supports a flexible user interface that takes into account the user's preferences with respect to his work with the resources and services.

2 Model of information content of a portal

The information basis of the portal is formed by the ontology of the portal and descriptions of the network resources associated with it. The ontology is understood here as a system, which consists of a set of concepts associated by relations, their definitions and assertions (axioms and rules) allowing to constrain (restrict) the meaning of concepts within some problem or subject domain.

2.1 Ontology of a portal

For a sufficiently complete and consistent representation of a field of knowledge, the ontology of a portal combines the following rather independent ontologies: ontology of science, including ontology of scientific

activity and ontology of scientific knowledge, and ontology of a subject domain that describes a certain branch of science.

Such structuring of the knowledge system in the form of ontologies that are mainly domain-independent simplifies considerably the adaptation of the portal to a certain field of scientific knowledge.

The ontology of science is based on the ontology suggested in [10] for the description of research projects. Ontology of scientific activity includes the following classes of concepts related to organization of research activity:

- *Scientist*. The concepts of this class correspond to persons involved in scientific activity: researchers, employees and members of the organizations, outstanding scientists and others.
- *Organization*. The concepts of this class describe various organizations, scientific communities and associations, institutes, research groups and others.
- *Event*. These notions include meetings, seminars, conferences, research trips and expeditions.
- *Publication*. This class serves to describe various types of publications and materials represented in printed or electronic format (such as monographs, articles, reports, proceedings of conferences, periodicals, photo and video data, etc.).
- *Activity*. This class includes the notions that describe organization of research activity (projects, programs, etc.).

Ontology of scientific knowledge contains the metanotions which specify the structures for description of considered subject domain:

- *Subdivision of science*. This class allows one to structure a science, i.e., to identify its significant parts and subparts.
- *Research method*. This class serves to describe various methods of research used in a specific scientific discipline.
- *Object of research*. The notions of this class classify the objects of research and introduce a structure for their description. For example, the object of research in the humanities can be a person, the society or a nation, as well as various objects created by a person (as a result of his activity).
- *Scientific result*. This class contains the notions such as discoveries, new laws, theories and methods of research. Usually scientific results are presented in publications.

Ontology of a subject domain describes a certain scientific discipline as a branch of science and includes the formal and informal description of its concepts and relations between them. These concepts are implementations of metanotions of the ontology of scientific knowledge.

Thus, if we consider a discipline such as archeology, then the implementations of the metanotion "subdivision of science" will be following: archeology, field archeology, etc. These concepts will be ordered in a hierarchy by the relations "generic-specific" and "whole-part". For the humanities, the methods and objects of research are very important. In particular, the methods of research in archeology are excavation and exploration, while the objects of research are monuments, tools, and other artefacts.

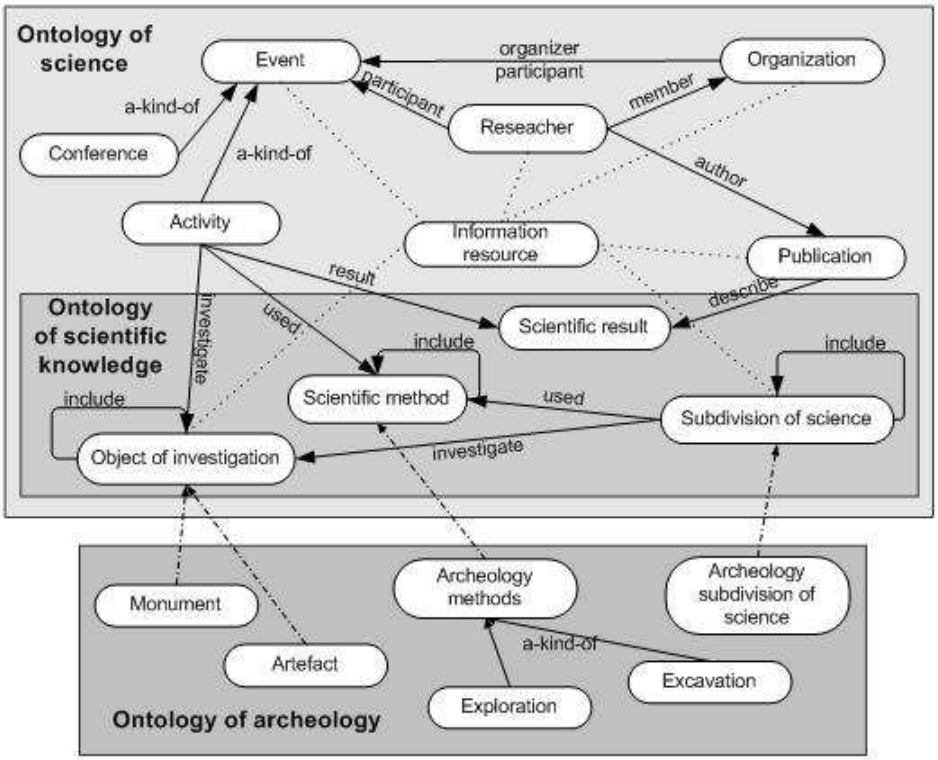

Fig. 1. A fragment of ontology of a knowledge portal for archeology.

Ontology of a subject domain is based on a thesaurus of the natural language terms that describe its significant vocabulary. Relations existing between the thesaurus terms and the concepts of ontology create prerequisites for their combined use in information search and processing.

Fig. 1 represents a fragment of the ontology of a knowledge portal for archeology. It includes the generic ontologies of science and scientific knowledge, and also a fragment of the ontology of archeology. This simplified scheme does not show all relations existing between the concepts. But as a whole the scheme represents the main concepts of the ontology of the portal and the principal relations between them. It is the basis for the construction of a complete model.

2.2 Description of information resources

The description of information resources is an important component of the information content of a portal. It includes specific attributes and relations that link the resource with the elements of the ontology.

The set of attributes and relations is based on Dublin Core [11] standard and includes the following units:

1. Title of the resource. This is the name given to the resource.
2. Subject of the resource. It specifies a topic for the content of the resource and links the resource to objects of different ontological types.
3. Resource type. It specifies the resource type (such as Internet site, database, text document) and the data format.
4. The language in which the content of the resource is described.
5. Access permissions.
6. Semantic index, i.e. a set of objects and relations representing the content of the resource in terms of ontology of a portal.
   The information space of a portal integrates structured resources (external databases), semistructured resources (HTML, XML, RDF) and unstructured resources (text documents).

3 ARCHITECTURE OF A KNOWLEDGE PORTAL

The architecture of a portal was developed on the basis of the user requirements, such as the completeness of portal content from the professional point of view and convenience and simplicity of usage.

The portal includes the groups of agents that are responsible for updating and monitoring of its content, management of user interaction,

support for integration of knowledge and search for the required information.

Let us consider the main components and modules of a knowledge portal (see fig.2).

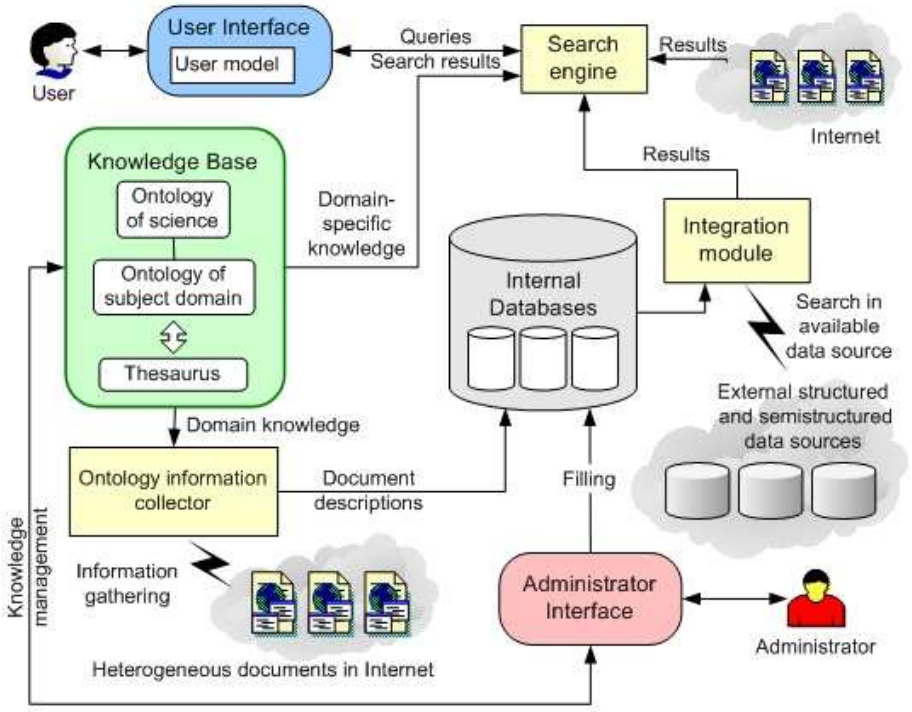

Fig. 2. The architecture of knowledge portal.

Knowledge base includes the ontology of the portal and thesaurus. The thesaurus contains terms, i.e. words and phrases of some natural language by means of which the concepts of ontology are presented in the texts and queries of a user. Relations existing between the thesaurus terms and the concepts of ontology create prerequisites for their combined use in information search and processing.

Software tools for creation and editing of ontologies and thesauri is a part of adjustment facilities of the portal.

Internal database stores all local data, including the descriptions of information resources and data sources.

The knowledge portal has both the administrator and user interfaces.

The administrator interface is used to set up and manage all subsystems as well as to maintain the internal database and knowledge base. The administrator interface provides linking to new information resources (external data sources) as well as user registration and management. For customization of the portal to a specific user or user group, the model of a user is used. The model of a user contains his subject preferences, the list of additionally connected/disabled resources, the technique for visualization of pages, etc. Note that the model of a user is updated at each logon and so it always represents his current "information portrait".

The user interface provides the users with remote access to the information space of the portal, that is, a user-friendly navigation through the knowledge base of the portal and the indexed information resources and a search for the required knowledge and data (documents). Besides, the user interface provides a user with tools for advanced semantic-based search in the Internet.

Interaction with the portal can be supported in several languages. Optional plug-in of other languages can be performed without the necessity to restructure the main system components; it is only necessary to plug in the thesaurus for the new language. This is possible due to the fact that all information interaction in the system is performed through the portal ontology, and the national language terminology is used only at the user interface level.

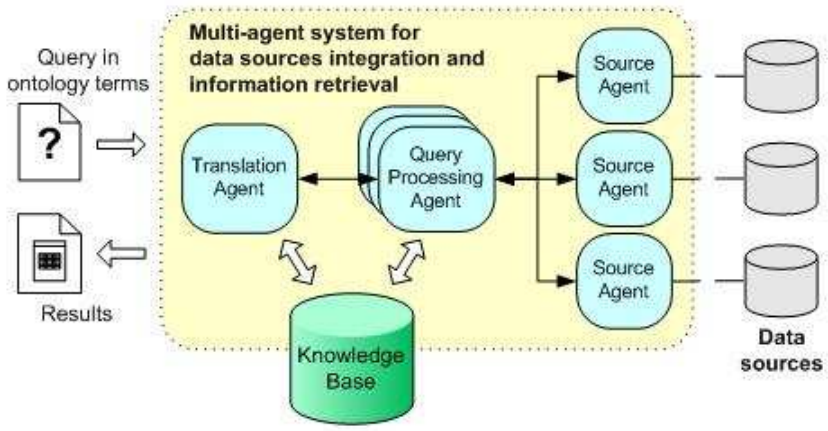

Fig. 3. Query processing.

Data source integration module is used for integration of relevant external information sources and provides a unified access to them. The specialized multi-agent system intended for semantic-based retrieval of information from a collection of distributed heterogeneous structured and semistructured data sources is used as such a module. In order that this system can process user queries formulated in terms of the subject domain of the portal, the data schema of each external data source is mapped onto the ontology of the portal. The module consists of the translation agent that translates the user query into the internal representation; the query processing agents that manage the process of search and collection of the relevant information; a group of source agents that handle connectivity issues, transform internal queries into data source queries, run them on the data sources, and transform the returned results; a set of service agents that provide access to the knowledge base of the portal and system monitoring.

The main components of a query stated in terms of the subject domain are the following:

- *Concepts* which are the terms of the ontology.

- *Constraints* that the retrieved data must satisfy. Constraints are defined as search patterns of a certain kind and/or logical expressions over the values of attributes of concepts.

Such a query is represented as a fragment of the ontology with additional constraints. Next, this query is converted into one or several queries to the internal database of the portal and/or available external structured data sources. Finally, the results of processing of these queries are merged, filtered and returned to the user (see Fig. 3).

The portal also includes a subsystem for extraction of knowledge and data from the Internet, which is called the collector of ontological information on relevant resources. The specialized information agents included in this subsystem search for semistructured and unstructured information resources that are relevant to the subject domain of the portal and perform their semantic indexing. Indexes of resources are stored in an internal database of the portal. In this manner, the information content of the portal can be automatically updated.

4 Conclusion

The paper presents an approach to increasing efficiency and convenience of access to the knowledge and information resources of a certain field of

knowledge via Internet through the development and use of specialized Internet portals.

The information basis of each of these portals is an ontology that supports integration of information resources being relevant to the subject domain of a portal into a uniform information space and provides content-based access to them. Besides, structuring of the knowledge system of the portals in the form of ontologies, when a part of them are domain-independent, makes the knowledge portal easily adjustable to a given field of scientific knowledge.

The use of the multi-agent approach and ontologies for support of the search for the required information as well as automatic update of the portal's content with new knowledge and information resources, ensures flexibility of the architecture of the portal and its scalability.

An important feature of these portals is their ability to provide both access to their own information resources and effective navigation through relevant Internet resources which were indexed during portal operation.

Let us underline the most important features of the portal that allow us to consider it as an advanced information system.

1. Using knowledge at all levels and stages of portal operation.
The information basis (model) of the knowledge portal consists of ontologies (a general ontology and subject area, and sub-area ontologies). This allows one to use the common and subject knowledge at all levels: user, conceptual, logical and physical.

2. Possibility of adjustability at all levels:

- subject domain,
- data and knowledge sources,
- the user's area of interests,
- a user as an individual.

3. Multilanguage support.
The portal can support several languages. To support a new language, it is only necessary to plug in the thesaurus for it. This is possible due to the fact that all information interaction in the portal is performed through its ontology, and the national language terminology is used only at the user interface level.

4. Simplicity of information space extension:

- stipulated by the fact that the linking of new data sources is performed through mapping the ontology into their data frameworks with the following conversion of queries expressed in the terms of

ontology into formal queries to the data sources expressed in terms of their frameworks;
- each type of data sources uses its own typical module source agent (or, if nonexistent, it is developed) which is included in the portal;
- the portal includes a subsystem which performs search for and automatic indexing of semi-structured and non-structured resources (text documents) relevant to the subject domain of the portal.

Our immediate goals are to approve the proposed concept and to develop the technology for creation of knowledge portals based on the concept. In particular, a specialized Internet portal providing content-based access to systematized knowledge and information resources relating to archeology and ethnography is currently under development. Besides, we are investigating the possibility of applying this approach to the development of knowledge Internet portals for the Artificial Intelligence field.

We plan to supplement the portal with thesauri for several languages (Russian and English). The use of ontology as a mediator between the search engines and a user (his vocabulary of terms) makes it easy to realize.

5 ACKNOWLEDGMENTS

The authors are grateful to the Russian Foundation for Basic Research (grant 04-01-00884) and the Russian Foundation for the Humanities (grant 04-01-12045) for financial support of this work.

References

1. T. Berners-Lee, J. Hendler, O. Lassila (2001), The Semantic Web, Scientific American, May 2001, pp. 28-37. http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html
2. Heflin, J. and Hendler, J. Dynamic Ontologies on the Web. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000). AAAI/MIT Press, Menlo Park, CA, 2000. pp. 443-449.
3. Heflin, J. and Hendler, J. Searching the Web with SHOE. In Artificial Intelligence for Web Search. Papers from the AAAI Workshop. WS-00-01. AAAI Press, Menlo Park, CA, 2000. pp. 35-40.
4. Brickley, D. and Guha, R. Resource Description Framework (RDF) Schema Specification, W3C (World Wide Web Consortium). 1999. At http://www.w3.org/TR/1999/ PR-rdf-schema-19990303
5. Brickley, D. and Guha, R. RDF Vocabulary Description Language 1.0: RDF Schema W3C Recommendation 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

6. Michael K. Bergman. The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing, August, 2001, Volume 7, Issue 1, (Links checked and updated August 2001) ISSN 1080-2711 http://www.press.umich.edu/jep/07-01/bergman.html
7. Gruber, T. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Workshop on Formal Ontology. 1993. March, Padova, Italy.
8. Guariano N., Giaretta P. Ontologies and Knowledge Bases. Towards a Terminological Clarification. In: N.J.I.Mars (ed.) Towards Very Large Knowledge Bases .1995. N.J.I.Mars (ed.) IOS Press, Amsterdam.
9. Wooldridge, M.Issues in Agent-Based Software Engineering. Cooperative Information Agents: First International Workshop, CIA-97 (LNAI Volume 1202), Springer-Verlag, Berlin, 1997, pp.1-18.
10. Benjamins V. R., Fensel D., et. all. "Community is Knowledge! in KA2", Proceedings of the KAW'98, Banff, Canada, 1998.
11. Dublin Core Metadata Initiative, Dublin Core Metadata Element Set, Version 1.1, http://purl.org/dc/documents/rec-dces-19990702.