Analysis of Automated Matching of the Semantic Wiki Resources with Elements of Domain Ontologies

Rogushina J.

Institute of Software Systems of National Academy of Sciences of Ukraine, Glushkov, 44, Kiev, Ukraine

Abstract

Intelligent information systems oriented on the Web open environment need in dynamic and interoperable ontological knowledge bases. We propose an approach for integration of ontological analysis with semantic Wiki resources: domain ontologies are used as a base of semantic markup of the Wiki pages, and this markup becomes the source for improving of these ontologies by new information

Index Terms: Ontology, semantic Wiki, information object.

1. Introduction

The transition to knowledge-based systems is one of the basic trends in the development of modern information technologies. Currently, an increasing number of applications are intelligent because they use and process knowledge in their work.

Intelligent information system (IIS) is an information system where use of domain knowledge and the methods of knowledge processing play a crucial role. Ontologies [1] are widely used in IIS for interoperable representation of knowledge – they ensure the reuse of the acquired knowledge in various applications. Ontological analysis can provide the basis for processing distributed knowledge. Today, a considerable amount of research relates to the theoretical basis of ontologies, their construction, improvement, obtaining knowledge from ontological structures, as well as with other important aspects of ontology management which differ significantly by their purposes.

The level of the system intelligence is defined both by the ability to learning and self-learning (that is, to the use of existing knowledge in new, previously unknown situations), and the domain breadth. From the point of view of ontological analysis, the intelligence of the system is determined by its ability to find and use external ontologies for various domains and the ability to replenish and improve these ontologies in accordance with the

* Corresponding author.
E-mail address:

specific tasks. Due to the complexity of knowledge obtaining the problem of ensuring their interoperability and reuse becomes important. An ontological analysis is a promising approach to solving this problem: ontologies are based on the fundamental theoretical basis of descriptive logics, and accepted standards of description, languages and software tools already exist for them. Domain ontology formalizes relations between an open information environment, IIS and its users.

The use of ontologies provides the retrieval of information from the Web at a semantic level – this task is a part of modern intellectual applications that function in an open information environment. Semantic search provides an opportunity to receive not just documents or data, but knowledge about IO that users need: this knowledge can have a rather complex structure formalized by ontologies.

IIS developers differentiate internal and external ontologies. *Internal ontologies* are created by IIS developers and display their representations about the properties and relationships of the main information objects (IO) of the processed domain. The structure of such ontology and its scope are fully determined by the goals of created IIS and by the functions that it should perform. The internal ontology remains unchanged in the process of IIS functioning and can only be supplemented with information about the instances of classes. *External ontologies* are imported by IIS from external information resources (IR) in the process of the system operation and allow dynamically update information about the corresponding domain. Such ontologies can be retrieved in repositories or formed by analysis of available IRs. There is important that external ontologies can vary considerably during IIS operations.

If IIS is oriented to work with the Web environment, then the main sources of external ontologies are the Web IRs. Such their properties as heterogeneity and dynamism cause a few problems related to the use of external ontologies. The problem of comparing external ontologies formed at different moments of IIS work arises: if the differences between ontologies exceed some measure then they have been re-processed.

2. Knowledge Management and Ontological Analysis

Modern IIS use and generate large in volume and structurally complex sets of knowledge, so the problem of their effective managing arises.

The term "knowledge management" was first proposed in 1986 by the American scientist Carl Wiig, by analogy to such concepts as "data management" and "information management". *Knowledge management* is a set of processes associated with the effective creation, preservation, dissemination and use of knowledge to achieve the user goals [2].

The main problems of knowledge management in the Web are related to:

- integration of knowledge obtained from different sources (for example, the integration of ontologies of intersecting domains constructed from different sets of IRs);
- discovery of knowledge contradictions in content of different IRs and evaluation of their reliability;
- acquisition of new knowledge from existing ones and their representation in user-understandable form;
- retrieval of knowledge required by user for solving of some specific problems;
- automated creation of metadata that correctly displays the content of IR (both textual and multimedia) on content level, and an effective search in such metadescriptions.

Every IIS can support a certain subset of knowledge management functions. For Web-oriented IIS, this task is complicated by the need to manage distributed and heterogeneous knowledge. As opposed to traditional IIS, these systems need in obtaining of information from external IRs. Some of them can be accessed directly, but in most cases, external information retrieval systems (IPS) are used as intermediaries. But IPS allows only to provide the set of relevant IRs from which it is possible to obtain information, but don't solve the problem of acquisition of the necessary knowledge from these IRs.

Knowledge management in the Web has its own specificity, which is determined by both the large volumes of processed data and the dynamism, heterogeneity and decentralization of data. One of the most popular

directions of such research is the Semantic Web project, proposed by the WWW author Tim Bernes-Lee. The main idea of this project is to transform the Web into the global knowledge base and provide standards, languages and software for processing information in these resources at the semantic level.

Applications oriented on the Semantic Web provide knowledge management for this open environment. According to the "Semantic Web Challenge", Semantic Web applications integrate, connect and logically process the information required by the user. Such applications are oriented on the open world and should meet the following minimum requirements:

1. Data values play a key role in their functioning: data values should be represented using formal definitions and be processed in a non-trivial way to obtain useful information, and processing of semantic information should play a central role in achieving results that can not be obtained with the help of alternative technologies.
2. Application use heterogeneous (syntactically, structurally and semantically) IRs with different owners (i.e. there is no possibility to control its change) that contain real-world data (and not be toy examples).
3. The retrieval is carried out in the real information space of the Web

As a rule, these applications use in some way such Semantic Web standards as RDF, RDF Schema or OWL, but this is not a prerequisite.

Analysis of modern research in knowledge management shows that, despite significant achievements in the field of Data Mining and Text Mining, the extraction of knowledge from unstructured natural language IRs remains an extremely time-consuming and completely non-automated task. Much more effective is the extraction of knowledge from the IRs with semantic markup that can be matched with structures of the formalized knowledge representation. At the same time, to achieve practically useful results, we must focus on those means of semantic markup that became quite widespread today.

IRs represented with the help of semantic Wiki by Semantic MediaWiki [3] satisfy all these requirements. This platform is widely used today, the IRs based on it are often updated and rapidly growing. The use of such sources is because Wiki-resources are dynamically updated by the entire community of users, they have a clearly defined and simple structure and provide information processing at the semantic level. The analysis of such IRs allows to form and replenish the ontologies of various domains. The built-in tools of Semantic MediaWiki provide building of OWL ontologies but these tools are imperfect and not flexible enough for development of high-quality interoperable domain ontology. In addition, there remain open problems of comparing different ontologies, improved based on different states of Wiki-resources and different means of their markup (it is important, that for the same Wiki-resource it is possible to apply a semantic markup based of different domain ontologies).

Therefore, it is necessary to develop more complete methods of forming and replenishing of domain ontologies based on Semantic MediaWiki resources.

3. Formulation of the Problem

To provide IIS an ability to import, modify, and update ontological knowledge extracted from external semantically marked Wiki resources, it is necessary to develop methods and tools to solve the following tasks:

- to construct a one-to-one correspondence between the basic elements of the domain ontology (classes and subclasses, instances, properties of objects and data, their values) and elements of the Semantic MediaWiki (categories, semantic properties and their meaning, links between pages);
- to develop a method for matching of ontologies that improve the initial domain ontology by knowledge from various Wiki resources that are semantically marked out by the elements of this initial ontology.

At the first stage, it is necessary to build (or select) the initial ontological model of domain which will be

used for semantic markup. Then we should develop a method for its transformation into the construction of Semantic MediaWiki (categories and semantic properties).

At the next stage, it is necessary to develop the method of refinement and improvement of the initial domain ontology based on the analysis of semantically marked Wiki resources. This method should convert Semantic MediaWiki constructs into ontology elements. Then it is necessary to analyze the refined ontology to estimate its properties and user's conception about domain. An iterative repetition of these actions should ensure the formation of an adequate domain ontology that can be used by various AIS as a knowledge base.

4. Components of the Semantic MediaWiki

Semantic MediaWiki (SMW) allows users to add semantic annotations to Wiki pages by means of additional markup elements (categories, semantic properties, and queries) that turns MediaWiki into semantic resource. SMW uses semantic properties (for data creating) and semantic queries (for data use). Users mark Wiki pages with categories and properties so that information becomes available for queries.

User adds [[Category: Category Name]] to its code to assign Wiki page to some category. Category pages allow to set the hierarchy of categories.

Formal model of Wiki resource $W = \left\langle P = P_{user} \cup P_{categ} \cup P_{spec}, L = \{\text{"}link\text{"}\} \right\rangle$ consists of the following elements:

- the set of Wiki pages $P = P_{user} \cup P_{categ} \cup P_{template} \cup P_{spec}$ where P_{user} is a set of user papers, P_{categ} is a set of pages that define categories, $P_{template}$ is a set of pages that define templates, P_{spec} is a set of other special pages;
- $L = \{\text{"}link\text{"}\}$ is a one-element set that defines the link from one Wiki page to another.

Formal model of the semantically enriched Wiki resource is more complex and includes more elements [4]:
$$W_s = \left\langle P = P_{user} \cup P_{categ} \cup P_{template} \cup P_{sem\_prop} \cup P_{spec}, L = \{\text{"}link\text{"}\} \cup L_{sem\_prop} = \{l_i\} \right\rangle \quad \text{where}$$
P_{sem_prop} is a set pages that define semantic properties of Wiki pages, some of the semantic properties reference on another Wiki page: $P_{sem\_prop\_page} \subseteq P_{sem\_prop}$, and other link page with values of selected type of data (data type is defined into these pages).

Semantic properties are used for binding data to Wiki pages. Each property has type, name and value, in addition, it corresponds to a separate Wiki page in the special namespace that allows to specify the type of the property, determine its position in the property hierarchy and document its use. Properties, as opposed to MediaWiki categories, have not only name, but also value. The semantic property is inserted into the Wiki page by [[Property name: Property value]]. Semantic properties are used to determine the semantics of links to other pages. For example, if content of the page "Ukraine" includes [[capital: Kyiv]] then it means that the concepts "Ukraine" and "Kyiv" are related by the relation "capital". The names of semantic properties can be used in templates and requests of Wiki pages. In addition, Semantic MediaWiki uses the links to other Wiki pages.

5. Basic Components of Ontologies

Ontological analysis makes it possible to transform the description of a certain concept of the external world into a set of terms and rules for suitable for machine processing use [5]. Ontology is an explicit specification of conceptualization. We can consider ontology as a special type of knowledge base (KB) with semantic information about certain domain. The components of ontologies depend on the representation paradigm and approach but practically all ontological models contain certain concepts (terms, classes), properties of concepts

(attributes, roles), the relationship between concepts (dependencies, functions) and constraints of their use defined by axioms. In the general case, the formal ontology model O is an ordered triple O = <T, R, F>, where T is the set of domain concepts; R is the set of relations between concepts; F - set of interpretation functions of concepts and relations. This formal model can be specified in accordance with the goal and scope of the ontology

Now IISs usually represent ontologies by different dialects of the OWL (Web Ontology Language). OWL ontology is a sequence of axioms and facts, as well as references to other ontologies [6].

Fundamental concepts of a certain domain should correspond to classes of ontology. Each individual in the OWL ontology belongs to the class *owl: Thing*, and each user-created class is automatically a subclass of *owl: Thing*. Specific root classes are defined by simple declaration of a named class. OWL also allows you to set an empty class: *owl: Nothing*. Definitions can be extensible and distributed.

The fundamental taxonomic constructor for classes is *rdfs: subClassOf*. It relates particular class with the general one. If X is a subclass of Y, then each individual of X is also an individual of Y. The relation *rdfs: subClassOf* is transitive. If X is a subclass of Y and Y is a subclass of Z, then X is a subclass of Z.

The class definition consists of two parts: a title (or a reference to it) and a list of constraints. Each expression that is directly placed in the class definition, specifies the properties of the representatives of this class. Individuals of the class belong to the intersection of these restrictions. To define an individual, it's enough to declare it as a member of a class.

Properties provide an opportunity to approve general facts about classes and individuals. There are two types of properties: - data properties – relations between class individual and RDF literals or data types defined by XML Schema; - object properties – relations between individuals of classes.

There are many ways to limit these relations -- by specification of their domain and range or as a specialization (sub-property) of an already existing property. More complex constraints are also possible. Properties, as well as classes, can be organized hierarchically. The formal semantics of OWL contain a description of how to obtain logical consequences, that is, to obtain facts that are not directly represented in ontology but are conditioned by its semantics.

The semantics of ontological languages are usually revealed through the theory of models. It defines an interpretation function that projects each ontology element into the interpretation area D.

The analysis of the various means of representation of ontologies and their formal models that offer Semantic Web technology shows that they differ significantly by their expressive capabilities and their complexity: RDF Schemas offers the simplest level for representation of ontologies, and OWL Full is the most complex one. The choice of a means of representation depends on the specifics of the problem for which this ontology is developed.

According to goals of this research work we propose to use a formal model of ontology $O = \left\langle X = X_{cl} \cup X_{ind}, R = r_0 \cup \{r_i\} \cup \{p_j\}, F, T \right\rangle$ that consists of the following elements:

- $X = X_{cl} \cup X_{ind}$ is a set of ontology concepts where X_{cl} is a set of classes, X_{ind} is a set of instances of classes such that $\forall a \in X_{ind} \exists A \in X_{cl}, a \in A$;

- $R = r_{ier\_cl} \cup \{r_i\} \cup \{p_j\}$ is the set of relations between elements of ontology where r_{ier_cl} is the hierarchical relation that can be established between classes of ontology and properties of classes and is characterized by such properties as antisymmetry and transitivity, $r_{ier\_cl} : X_{cl} \rightarrow X_{cl}$; $\{r_i\}$ is a set of object properties that establish the relationship between individuals of classes: $r_i(a, a \in X_{ind}) = b, b \in X_{ind}$, $r_i : X_{ind} \rightarrow X_{ind}$; $\{p_i\}$ is a set of data properties that establish relationships between instances of classes and values: $p_i(a, a \in X_{ind}) = t, t \in T$, $p_i : X_{ind} \rightarrow Const$, and hierarchical relations can also exist within the sets of object properties and relationship properties, and r_{ier_obj},

$$r_{ier\_obj} : \{ r_i \} \to \{ r_i \} \text{ и } r_{ier\_data} , \; r_{ier\_data} : \{ p_j \} \to \{ p_j \} ;$$

- *F* is a set of characteristics of ontology classes, instances of classes and their properties that can be used for inference (for example, equivalence, difference);
- *T* is a set of data types (for example, string or integer).

The choice of this model of ontology is determined by the following reasons. First, it most fully corresponds to the intuitive notion of ontologies built into the user interface of the Protégé ontology editor. Secondly, this model is easily integrated with the thesaurus model of the task which is used for semantic retrieval of the Web IRs. Thirdly, this model allows to compare the ontological representation of knowledge about domain with the semantic constructions of Semantic MediaWiki: for some elements, one-to-one correspondence can be formed, and the comparison of others requires additional transformations but can also be at least partially automated.

Let us consider these correspondences in more detail. Thereto we need to analyze the basic semantic elements of Semantic MediaWiki. It should be noted that although Semantic MediaWiki provides automatic construction of ontology by semantically marked Wiki pages but in current versions this process considers too few semantic parameters, and its results are of little use for further implementation.

6. Use of Ontological Analysis in Semantic Markup

Semantic processing of natural language information in based on domain knowledge that can be represented by domain ontologies. They provide connections between the fragments of the natural language (NL) text and domain concepts (for example, classes or instances of the ontology). For these purposes the thesauri and linguistic ontologies are widely used.

Today, the most common approaches to the presentation of linguistic knowledge [7] are traditional information retrieval thesauri, WordNet thesaurus and specialized (linguistic, lexical) formal ontologies. *Thesaurus* is a dictionary of the basic concepts of the language expressed by separate words or phrases, with certain semantic links between them. As a rule, thesauri are used for the semantic processing of natural language IRs, for example, for semantic search or for semantic markup. The lexical ontologies closely associated with the thesauri are ontological KB that contain knowledge about conceptual system and lexico-terminological composition of domain.

Formally, the semantic markup of an NL text X can be defined thus:

Text X, $X = \langle x_1, \ldots, x_n \rangle$ is a finite sequence of characters $x_i \in A, i = \overline{1, n}$ that belong to the finite set A, $\forall x_i \in A$. In this case, some symbols are delimiters (symbols that separate individual words of the text one from one) and belong to the set B, $B \subseteq A$. Examples of delimiters are space, period, and comma.

Domain of the semantic markup is characterized by the completed set of terms T, $T = \{t_1, \ldots t_m\}$. These terms can be used as semantic markup tags. Set T can be formed from various sources and depends on semantic markup language such as SGML, RDF Schema, microformat tags, Dublin Core elements, set of terms of ontology.

NL text fragment $\langle x_p, \ldots, x_{p+q} \rangle, 1 \le p \le n, 1 \le p + q \le n, q \ge 1$ can be associated with a term $t_i \in T$. The beginning of fragment is marked by the corresponding tag, and the end of the fragment – by the pair end tag. Markup is called *semantic* if semantics of the elements of set T is described.

The expressive capabilities of semantic markup depend in a great measure on the set of markup tags, their structure, types of relationships between these tags and the possibilities of logical inference on these structures. Ontology has the greatest expressiveness and universality as a source of tags for semantic markup: use of ontology terms as markup tags allows to establish the actual connection between the text fragments and the concepts domain that is displayed in this ontology, that is, to determine the semantic structure of NL text.

Machine processing of semantic markup can be supported by specialized *lexical ontology* (LO). LO is an easy ontology that contains two main classes: "domain concept" and "word form". instances of these classes are linked by relation "corresponds". Each instance of the class "domain concept" is associated with at least one instance of the class "wordform". The fundamental difference between LO and usual linguistic ontologies is that LO is built for some domain ontology and includes only those instances of concepts that are associated with the concepts of the domain ontology. This significantly reduces the amount of LO the time of the text processing.

7. Matching of Ontologies and Semantic Wiki Resources

The problem of mapping ontologies and semantic Wiki-resources occurs in several cases. First, under the creating of semantic Wiki-resources it is necessary to form a set of categories and semantic properties. But the built-in tools of Semantic MediaWiki allow neither visualization of this information nor checking of it's integrity and consistency. Therefore, it is advisable in the first place to build relevant domain ontology and then to use it as a basis for semantic markup. Secondly, semantically marked Wiki-resources are much more dynamic than ontologies – collaboration of the wide range of participants in their improvement and updating becomes very effective but such participants more often can work with Wiki resources and are not acquainted with Protégé or other ontological tools.

Therefore, we try to define correspondences between main elements of ontology and Wiki resource. An analysis shows that only some correspondences are one-to-one and can be detected automatically. Some others require additional clarifications and definition from the user.

Table 1. Correspondences between Main Elements of Ontology and Wiki Resource

Semantic MediaWiki	Ontology	Matching from Wiki to ontology	Matching from ontology to Wiki
Category	Class	One-to-one $P_{categ} \to X_{cl}$	Many-to-one $X_{cl} \to P_{categ} \cup P_{template}$
Hierarchy of categories	Hierarchy of classes	Many-to-one	One-to-one
Wiki page	Class instance	Many-to-one $P_{user} \to X_{ind}$	One-to-one $X_{ind} \to P_{user}$
Link to another Wiki page	The association's object relation "refers to"	One-to-one $L = \{ "link" \} \to R$	One-to-one $R \to L = \{ "link" \}$
Semantic property of the "page" type	An object property	One-to-one $P_{sem\_prop\_page} \to \{ r_i \}$	One-to-one $\{ r_i \} \to P_{sem\_prop\_page}$
The semantic property of any other type	Data property	One-to-one $P_{sem\_prop} \to \{ p_i \}$	One-to-one $\{ p_i \} \to P_{sem\_prop}$
Template	Class	One-to-one $P_{template} \to X_{cl}$	Many-to-one $X_{cl} \to P_{categ} \cup P_{template}$

The same Wiki the page can be categorized by the several categories that are matched with instances of different classes of ontology. It means that these instances of classes of ontology are matched with different fragments of analyzed Wiki page. User can choose which categories are involved in building the ontology.

Ontological classes can be matched in Wiki either with categories or with templates. It depends from the user conception of domain.

Representation of values of semantic properties can be matched with instances of lexical ontology class "wordform" as one-to-one if values of corresponding semantic properties match with instances of lexical ontology class "lexem". For replenishment of LO with wordforms, information from the construction of

Semantic MediaWiki is used: [[semantic property: term | wordform]]

Above analysis shows that if domain ontology is already built in the OWL language then it is easy enough to use it in Semantic MediaWiki. But the reverse process can not be fully automated. Moreover, with the automatic generation of an ontology by the Semantic MediaWiki, the information contained in the OWL ontology about the characteristics of classes and properties that have no Wiki equivalents (the equivalence of classes and properties, their disjointness, their range of meaning and definition). At the same time, some part of the Semantic MediaWiki content can not be directly transformed into ontological knowledge representation. For example, the fact that the pages use the same template, suggests that information pages of the same type are described on these pages, but to display this in an ontology, you need to create a specific class and associate it with the page element. But it is impossible then to understand on such an ontology what to create in the Wiki – a template or category: the choice depends on the user, because high-usage IOs with a specific and formalized structure are represented by templates, and all others – by categories.

8. Practical Use

The proposed approach to knowledge management was tested during the development of an intelligent information and cognitive support system for the functioning of the National Qualifications Framework

An ontology based on the analysis of natural language descriptions of national and European frameworks [8] was constructed. This ontology displayed the reference model of the qualification framework. Ontology formalizes semantic properties and relations of IOs associated with learning outcomes and establishes their hierarchy [9]. User semantically marks the content of Wiki pages by the elements of this ontology: defines the category (or set of categories) of any document and identifies content elements.

Use of the ontological model of the qualification framework provides the automated processing of learning outcomes and integrating the knowledge required by users.

9. Summary

Ontologies help to analyze and reuse knowledge in the domain area. A lot of software systems oriented on Semantic Web use ontologies as a base of domain knowledge for semantic markup of documents (natural language texts, semi-structures and structured texts, multimedia context etc.) by ontological concepts and relations. Now a lot of information is represented by semantically enriched Wiki resources, and many others can be easy transformed to this form. Therefor the task of automated matching between ontological structures and Wiki pages is very actual. We think that use of proposed approach to various subject domains provides more complex and adequate forms of matching and will be supported by formal methods of ontological analysis.

References

[1]   Gruber, T.R. The role of common ontology in achieving sharable, reusable knowledge bases.   – http://www.cin.ufpe.br/~mtcfa/files/ 10.1.1.35.1743.pdf.
[2]   Warren, P. (2006) Knowledge Management and the Semantic Web: From Scenario to Technology. *IEEE Intelligent Systems*. 21, 1, 53–59.
[3]   Krotzsch   M.,   Vrandecic   D.,   Volkel   M.:   Semantic   MediaWiki.   – http://c.unik.no/images/6/6d/SemanticMW.pdf.
[4]   Rogushina, J. (2016) Semantic Wiki resources and their use for the construction of personalised ontologies. *Proc. of the 10th International Conference of Programming UkrPROG'2016*, 196-203. – http://ceur-ws.org/Vol-1631/188-195.pdf
[5]   Uschold, M., & Grüninger M. (1996) Ontologies: Principles, Methods and Applications. *Knowledge*

Engineering Review, 11(2), 93–155.

[6]   OWL   Web   Ontology   Language.   Overview.   W3C   Recommendation: W3C   (2009).   –
      http://www.w3.org/TR/owl-features/.

[7]   Pedersen, T., Patwardhan, S., Michelizzi, J. (2004) WordNet: Similarity - measuring the relatedness of
      concepts. *Proc. Of the 19th National Conference on Artificial Intelligence (AAAI-04)*, 1024-1025.

[8]   Lundqvist, K.O., Baker, K.D., Williams, S.A. (2008) An ontological approach to competency
      management. – http://www.eife-l.org/publications/proceedings/ilf07/ Contribution110.doc.pdf.

[9]   Rogushina J., Pryima S. (2017) Use of ontologies and the Semantic Web for qualifications framework
      transparency. *Eastern-European Journal of Enterprise Technologies,* 1/2 (85), 25-30.

Authors' Profiles

Dr.Julia Rogushina is a currently working as a senior researcher at the Institute of Software Systems, National Academy of Sciences of Ukraine. She took her PhD degree in computer sciences in 1994 in the Institute of Cybernetics, Kiev, Ukraine. Her scientific interests deal with knowledge management, ontological analysis and semantic search, theory of software agents behavior, inductive knowledge acquisition, ontological analysis, the Semantic Web technologies. She is an author of "Agent Technologies", "Semantic Technologies: Principles and Practices", "Semantic Search for Web".