SEMANTIC SEARCH OF INTERNET INFORMATION RESOURCES ON BASE OF ONTOLOGIES AND MULTILINGUISTIC THESAURUSES

Anatoly Gladun, Julia Rogushina

Abstract: the approaches to the analysis of various information resources pertinent to user requirements at a semantic level are determined by the thesauruses of the appropriate subject domains. The algorithms of formation and normalization of the multilinguistic thesaurus, and also methods of their comparison are given.

Key words: an information resource, ontology, thesaurus, informational retrieval.

ACM Classification Keywords: I.2.7 Natural Language Processing, I.2.4 Knowledge Representation Formalisms and Methods (F.4.1).

Introduction

During last years the Internet becomes one of the main means of the information publication. It is dynamical distributed environment, and the information resources (IR), presented in it, are heterogeneous. The effective retrieval of Internet IR by expand of network amount and complexity becomes more and more difficult and laborious. Thus critical is not the search time but selection of IR that satisfy to real information needs of users.

The quality estimation of information retrieval systems (IRS) is a complex question [1]. The problem concerns with parameters of IRS estimation. A lot of existing techniques analyze such IRS parameters as relevance, completeness, accuracy and their various combinations. Relevance is a thematic correspondence of the information, received as a result of search, to request. The completeness of search is a ratio of the correctly found documents amount to the total relevant documents known to IRS. Accuracy of search is a ratio of correctly found documents amount to the total amount of the documents given by IRS in reply to request.

However it is necessary to take into account, that the formal request to IRS is the user attempt to formalize his/her information need that, unfortunately, not always really reflects this need. It results in degradation of Internet use. Therefore more important such parameter of IRS quality estimation as pertinence – a ratio of amount of the information interesting for user to total amount of the received information. To increase the pertinence of informational retrieval IRS requires information about area of the user interests. This information applies by IRS for choose among accessible resources what are interesting to user and not only formally correspond to request. Such information should be submitted in the form suitable for automatic processing and reuse, and their formation must be automatized.

Internet Informational Resources

Among IR, potentially accessible to the Internet users, still prevails the textual information mainly in HTML and XML formats however it's share constantly decreases due to multimedia IR increase. The subject domain that is characterized by these IR can be represented by two ways: 1) analyzing textual information and 2) considering metadata of these IR.

Metadata contains machine-readable information about the document, which can be automatically processed by computer. Now the most perspective and common metadata model is RDF (Resource Description Framework) based on XML. With the help of RDF one can describe both structure of a site and connected with appropriate domain. RDF describes informational resources in oriented marked graph form - each IR can have properties,

which in turn also can be IR or their collections. Most widespread set of elements for metadata specification is Dublin Core Metadata Elements. Metadata can be built in IR or be stored and updated independently of resources.

Multimedia data. Recently Internet IR along with the textual information includes the graphic elements, video, sound etc. There is a great deal of the widespread formats for a storing of audio and video information, 3D-scripts and images. The multimedia resources are accessible for indexation much worse than textual information. If the information about multimedia IR is not submitted by their provider explicitly in any format known for indexing mean, it is a necessity to apply the complex and laborious operations (image recognition, speech recognition etc.). Now MPEG group develops a number of standards for representation of multimedia information metadata (for example, MPEG7 and MPEG21). In spite of significant differences between multimedia and textual IR, most acceptable for realization of information retrieval (taking into account time of its fulfilment and data level of index BD) is their description with the help of the same means, as textual IR: key words, file size and date of it's creation etc.

Web-services. Initially World Wide Web technology was focused on work with static hypertext documents represented in the Internet. But then sites offering to the clients not only the documents, but also service (for example, sites of e-commerce) began to occur. Many such sites use application servers, which not only return the document but can process the data entered by the user (queries, completed form etc.) and dynamically generate the documents depending on the parameters, specified by the user. Such dynamic component of the Internet grows much faster then static one and requires application of more complex information technologies. In this connection it is possible to consider a separate class of IR - Web-services.

Web-service is a set of logically connected and program-accessible through the Internet functions. There is the program identified on UR. It's interface can be determined by XML structures. Web-services are based on three basic Web-standard: SOAP (Simple Object Access Protocol) - the protocol for sending of messages by the HTTP and other Internets protocols; WSDL (Web Services Description Language) - language for the description of program interfaces of Web-services; UDDI (Universal Description, Discovery and Integration) - indexing standard of Web-services.

Statement of Problem

For effective search of the information that user needs (textual and multimedia documents, information services etc.) there is necessary to generate the model of user interests domain (for example, as ontology) and use this model when IRS fulfils the user's query.

Thesauruses and ontologies as means of domain knowledge representation

Every domain has phenomena that people allocate as conceptual or physical objects, connections and situations. With the help of various language mechanisms such phenomena contacts to the certain descriptors (for example, names, noun phrases).

For the successful solution of an informational retrieval task it is necessary to present user knowledge about domain of her/his interests in some form suitable for computer processing. The specifications of high-level domain are formed by integration of the domain structures of low-level domains. It is important to achieve an interoperability of domain knowledge representation. Ontological approach is an appropriate tool for solution of this task. Ontology is an agreement about common use of concepts that contains means of representation of subject knowledge and agreements on methods of reasons. It can be considered as the certain description of the

views on the world in some specific sphere of interests. Ontology consists of: 1) a set of the terms; 2) a set of rules of their use that limit their meanings in the context of concrete domain [2].

The ontology is knowledge base of a special kind with the semantic information about some domain. It is a set of definitions in some formal language of declarative knowledge fragment focused on joint repeated use by the various users in the applications.

Ontological commitments are the agreements aimed at coordination and consistent use of the common dictionary. The agents (human beings or software agents) that jointly use the dictionary do not feel necessity of common) knowledge base: one agent can know something that don't know the other ones, and the agent that handles the ontology is not required the answers to all questions that can be formulated with the help of the common dictionary.

Every domain with the certain subject of research has it's own terminology, original dictionary used for discussion of typical objects and processes of this domain. The library, for example, involves the dictionary relating to the books, references, bibliographies, magazines etc. Thus, pattern of domain is discovered by its dictionary - the set of words that are used in this domain. Clearly, however, that the specificity of domain is shown not only in the appropriate dictionary. Besides, it is necessary: (i) to provide strict definitions of grammar managing of combining the dictionary terms into the statements, and (ii) to clear logic connections between such statements. Only when this additional information is accessible, it is possible to understand both nature of domain objects and important relations established between them. Ontology - structured representation of this information [3].

The formal model of domain ontology O is an ordered triple $O = < X, R, F >$, where X - finite set of subject domain concepts that represents ontology O; R - finite set of the relations between concepts of the given subject domain; F - finite set of interpretation functions of given on concepts and relations of ontology O.

Until recently term "thesaurus" was used as a synonym of ontology, however now in IT with the help of the thesauruses frequently describe domain lexicon in a semantic projection, and ontology apply to modelling semantics and pragmatists in a projection to representation language [4]. The models either of ontologies or of thesauruses include as the basic concept the terms and connections between these terms.

The term "thesaurus" for the first time was used still in XIII century by B.Datiny as the name of the encyclopaedia. In translation from Greek "thesaurus" means treasure, riches. The thesaurus is the complete systematized data set about some field of knowledge allowing the human or the computer to orient in it.

The thesaurus is a dictionary where the descriptors of the certain field of knowledge with ordering of their hierarchical and correlative relations are represented. The descriptors are given in alphabetic order but they are grouped semantically; the search is carried out from concept to a word. Collection of the domain terms with indication of the semantic relations between them is a domain thesaurus. The thesaurus can be considered as a special case of ontology. The thesaurus is a pair $Th = < T, R >$, where T - finite set of the terms; and R - finite set of the relations between these terms.

The multilingual thesaurus is a coordinated set of the monolingual thesauruses containing equivalent descriptors on languages-components necessary and sufficient for interlingual exchange, and including means for the indication of their equivalence. At an recognizing of equivalence of descriptors of the various monolingual versions it is necessary to distinguish on different languages-components the following degrees of equivalence of the terms: 1) complete; 2) incomplete; 3) partial; 4) absence of the equivalent term. Incomplete equivalents are the terms, for which the volumes of concepts, expressed by them, are crossed. Partial equivalents are the terms, for which volume of concept expressed by one equivalent, is included into volume of concept expressed by other equivalent. One way to the recognizing of equivalence to a various degree bases on appropriate domain ontology

use: every word from the monolingual thesauruses refers to one of the ontology terms that helps to make connection between words of the various thesauruses. If some words from thesaurus refer to one ontology term then they are equivalent. If some words refer to ontological terms being a subclass one another then these words are in relation of incomplete equivalence.

Use of thesauruses for IR retrieval

For taking into account semantics of area of user interests in process of retrieval of IR satisfying his/her informational need o it is necessary (fig. 1):

1. to generate the domain thesaurus corresponding to information needs of the user (by analysis of IR that this user considers relevant to this domain [5];
2. to construct the thesaurus for every IR known to IRS (simple dictionary without stop-words);
3. to compare the thesauruses of IR relevant to user query to IRS with the domain thesaurus and to find those ones that contain the maximum number of words in intersection.

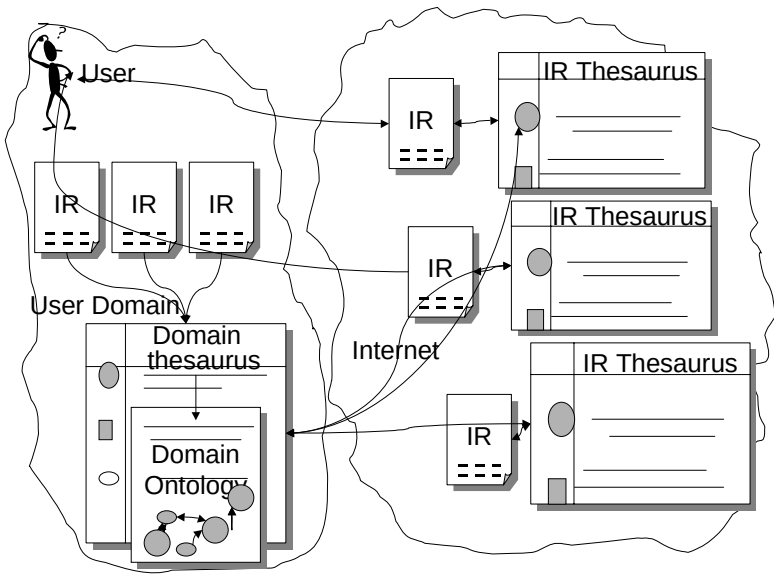

Fig.1. Informational retrieval on base of thesauruses

At thesaurus construction it is necessary to use ontologies of the appropriate areas (with higher level in comparison with user domain to normalize the multilingual thesauruses). Normalization procedure is similar to stemming and provides for integrated processing of words in different morphologic forms and multilingual representations. Normalised thesaurus contains relation between equivalent terms in different languages. As every thesaurus is constructed from the user point of view (which is reflected in user domain ontology), therefore it's forming is the user task.

Constructing of domain thesaurus

At first user should independently select the set of IR that he/she considers relevant to domain of his/her interests. Every IR is described by not empty set of the textual documents connected with this IR - text of content, metadescriptions, results of indexing etc. The domain thesaurus is formed as a result of the automated analysis

of these documents (the user actions are reduced to constructing of semantic bunches - by linking of each word of the formed thesaurus with some term of domain ontology. Algorithm of domain thesaurus construction consists from the following steps:

1. Formation of initial set of the textual documents relevant to domain. At the input of algorithm the set A of the textual documents describing chosen IR comes (each of documents from A can have the coefficient of importance and the coefficient of IR relevance *И*Р that allows defining differently weight of words from these documents for the IR description).

2. Creation of domain information space. For every document from A $a_i \in A, i = \overline{1,n}$ the IR thesaurus $T(a_i)$ - dictionary that contains all words occurred in the document a_i - is constructed. The IR thesaurus is formed as union of the thesauruses a_i: $T_{IR} = \bigcup_{i=1}^{n} T(a_i)$, and domain thesaurus - as association of the IR thesauruses.

3. Clearing of the thesauruses. User should specify dictionary for every $a_i \in A, i = \overline{1,n}$ containing a stop-words (for example, prepositions and conjunctions of language of the document are stop-words for it but prepositions and conjunctions of other language used as examples do not concern to them) $s_j, s_j \in Voc$. It is necessary to remove words contained in $s_j, s_j \in Voc$ from the thesauruses. Then all service information is rejected (for hypertext, for example, there are marking tags). The cleared thesauruses $T`(a_i), \forall p \in T(a_i) \Rightarrow p \in T`(a_i) \vee p \in s_j, T`(a_i) \cap s_j =$ thus are formed. The cleared thesaurus *И*Р is under construction as association of the cleared thesauruses $a_i$: $T_{ИР} = \bigcup_{i=1}^{n} T(a_i)$ $T`_{ИР} = \bigcup_{i=1}^{n} T`_{ИР}(a_i)$, and cleared domain thesaurus - as association of the IR thesauruses.

4. Linking of thesaurus with domain ontology. To integrate processing of words with equivalent semantics (for example, synonyms, translations of the term on different languages, various kinds of a spelling) the domain thesaurus is associated with some domain ontology (the user can form it himself, use ready ontology or it's modification).

Each word from the thesaurus it is necessary to link with one of the ontological terms. If the relation is lacking the word is considered as a stop-word or marking element (for example, HTML tag) and should be rejected. $\forall p \in T`(a_i) \exists t = Term(p,O) \in T_O$. The group of the IR thesaurus words terms connected with one ontological term named *the semantic bunch* $R_j, j = \overline{1,n}$ is considered as a single unit. $\forall p \in T`_{ИР} \exists R_j = \{r : r \in T`_{ИР}, Term(p,O) = Term(r,O)\}$. It allows to integrate processing of semantics of the documents written on various languages and, thus, to ensure the multilinguistic analysis of the Internet IR.

5. Extension of ontology. If the IR thesaurus contains words that can't be linked with ontological terms but user considers that these words are significant than it is necessary to add the appropriate terms to domain ontology, specify their connection with other terms of ontology and return to step 4.

6. Construction of the normalized domain thesaurus , i.e. association of all terms of domain ontology that are connected with words from the normalized IR thesaurus (fig. 2):

The normalized thesaurus is a projection of set of the IR thesaurus words on set of the domain ontology terms. $L_{ИР} = \{t : p \in T`(a_i), i = \overline{1,n}, t = Term(p,O) \in T_O\}$, and normalized domain thesaurus is a union of the normalized IR thesauruses (fig. 3).

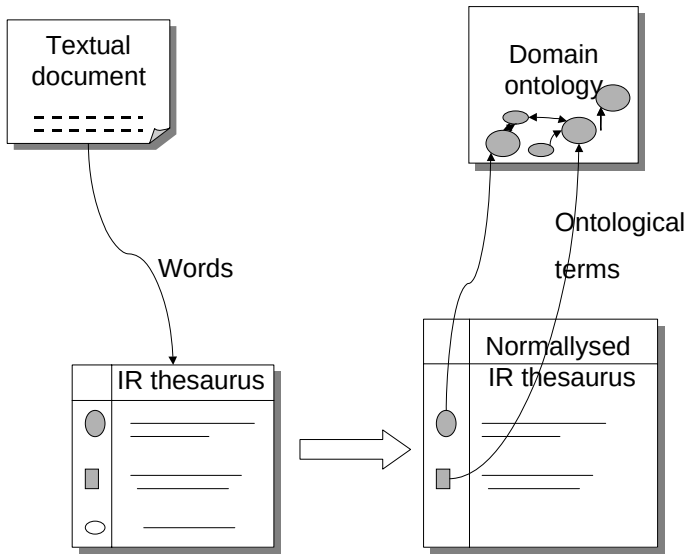

Fig.2. Building of normalized IR thesaurus

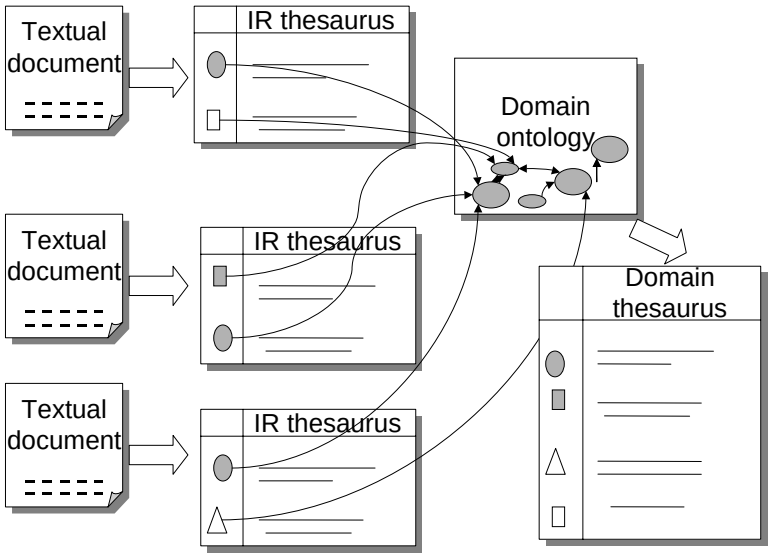

Fig.3. Building of domain thesaurus

Building of IR thesaurus

The thesaurus of IR found by IRS as a result of the user query execution is simple a dictionary that does not contain the relations between words (discovery of such connections from the text is rather difficult and in this case is not justified).

The algorithm of the IR thesaurus building consists of the following steps:

1. Formation of the initial IR set U, $U = \{IR_j, j = \overline{1,m}\}$.
2. Formation of the IR thesauruses from U. For each IR a thesaurus is formed and cleared.
3. Construction of the normalized IR thesauruses: for normalization the semantically bunches generated by the user during formation of the domain thesaurus are used .

Algorithm of domain and IR thesaurusus comparison

The normalized IR thesauruses L_{IR} and domain thesaurus L_{domain} are the subsets of the domain ontology terms O chosen by the user: $L_{IR} \subseteq Term(O)$, $L_{domain} \subseteq Term(O)$. If IR description contains more words linked with terms of domain interest for user (that is reflected in the normalized domain thesaurus) then it is possible to suppose that this IR can satisfy informational needs of the user with higher probability than other IR relevant to same formal query. Thus, it is necessary to find IR q satisfied the conditions $f(q, L_{domain}) = \max f(L_{IR}, L_{domain})$ where the function f is defined as number of elements in crossing of sets L_{IR} and L_{domain}: $f(A,B) = |A \cap B|$. If the various terms of the normalized thesauruses have for the user different importance it is possible to use the appropriate weight coefficients w_j that take into account their importance. In that case the criterion function is $f(A,B) = \sum_{j=1}^{z} y(t_j)$, where the function y is determined for all terms of domain ontology and $y(t_j) = \begin{cases} 0, t_j \notin A \vee t_j \notin B \\ w_j, t_j \in A \wedge t_j \in B \end{cases}$.

Conclusion

The proposed approach to use of domain ontology for creation and normalization of the IR thesauruses allows fulfilling informational retrieval at a semantic level abstracting from language of the IR description. The application of thesaurus measure of the information allows to offer to the user only understandable to him/her items of information that provides pertinence of information retrieval.

Bibliography

1. S. Bechofer and C. Goble. Thesaurus construction through knowledge representation. Data & Knowledge Engineering, 37:25–45, 2001.
2. Gruber T.R. A translation approach to portable ontologies // Knowledge Acquisition, N 5 (2), 1993. – P.199-220..
3. IDEF5 Method Report. Knowledge Based Systems, Inc.1408 University Drive East College Station, Texas 77840, 1994. – 175 pp.
4. Takeda H., Takaai M., & Nishida T. Collaborative development and Use of Ontologies for Design // Proceedings of the Tenth International IFIP WG 5.2/5.3 Conference PROLAMAT 98, September 9 – 10 – 11, 12, Trento, Italy, 1998.
5. Gladun A., Rogushina J., Shtonda V. Ontological Approach to Domain Knowledge Representation for Informational Retrieval in Multiagent Systems // International Jornal "Information Theories and Applications", V.13, N.4, 2006. – P.354-362.

Authors' Information

Anatoly Gladun – PhD, since 1997 works as Senior Researcher in International Research and Training Centre of Information Technologies and Systems, National Academy of Sciences and Ministry of Education of Ukraine, 44 Glushkov Pr., Kiev, 03680, Ukraine, e-mail: glanat@yahoo.com

Julia Rogushina – PhD, since 1997 works as Senior Researcher in Institute of Software Systems, National Academy of Sciences of Ukraine, 44 Glushkov Pr., Kiev, 03680, Ukraine, e-mail: jjj_@cybergal.com