OntoSearch: An Ontology Search Engine[1]

Yi Zhang, Wamberto Vasconcelos, Derek Sleeman
Department of Computing Science,
University of Aberdeen,
Aberdeen, AB24 3UE, Scotland, UK
Email: {yzhang, wvasconc, dsleeman}@ csd.abdn.ac.uk

Abstract

Reuse of knowledge bases and the semantic web are two promising areas in knowledge technologies. Given some user requirements, finding the suitable ontologies is an important task in both these areas. This paper discusses our work on OntoSearch, a kind of "ontology Google", which can help users find ontologies on the Internet. OntoSearch combines Google Web APIs with a hierarchy visualization technique. It allows the user to perform keyword searches on certain types of "ontology" files, and to visually inspect the files to check their relevance. OntoSearch system is based on Java, JSP, Jena and JBoss technologies.

1.   Introduction

Reuse of knowledge bases[2] is an important area in knowledge technologies. Determining the principal topic of an existing knowledge base (KB) is very important for the reuse of knowledge bases. Identify-Knowledge-Base (IKB) [2, 3] is a tool to identify the principal topic(s) of some particular knowledge base by matching concepts (extracted from the KB) against a reference taxonomy (extracted from a reference ontology). Finding (normally from the Internet) a relevant reference ontology for a particular KB is the key point in the use of the IKB system.

---

[1] This work is part of the Advanced Knowledge Technology (AKT) project, which is funded by EPSRC, [1]. The IKB system [2, 3](Aberdeen University) and the ExtrAKT system [4, 5, 6] (Edinburgh University), which incorporate with OntoSearch system, were built for the AKT consortium as well.

[2] Knowledge Reuse: http://www.aktors.org/technologies/reuse/

The Semantic Web[3] provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It envisions the globally interconnected network of machine-processable information, made possible by means of the sharing of semantic data models or ontologies. Locating suitable existing ontologies to capture the user-required information from the Internet is a big challenge in the current research of the Semantic Web.

Finding a suitable ontology from the Internet is a hard task. There is still no good tool to handle this problem. Google offers a powerful web search engine. However, with regard to the ontology searching, it has its own problems, such as a lack of visualization facilities. Google APIs[4] give us a chance to develop our own tool (OntoSearch) to search the relevant ontology files to meet the user requirements.

In this article, we discuss the issue of searching for relevant ontologies on the Internet and introduce our tool, OntoSearch. In section 2, we give some background to our research and list some current problems. In section 3, OntoSearch is introduced in detail. In section 4, some discussion and future work are given followed by a brief summary.

2. Background

2.1 IKB: Identify Knowledge Base

Reuse of knowledge bases is a promising area in knowledge technologies and many researchers are focusing on how to reuse existing knowledge bases for different applications [1, 2]. Such requests for reuse are often specified as a knowledge base (KB) characterisation problem:

Require knowledge base on topic T, conforming to the set of constraints C [2].

There are two key points here:

- Decide what the principal topic (T) of a given knowledge base is.
- Decide whether a KB conforms to certain constraints C.

As we noted, determining the principal topic of an existing knowledge base (KB) is an important step in the reuse of knowledge bases. **Identify-Knowledge-Base (IKB)** [2, 3] is a tool to suggest the principal topic(s) addressed by a knowledge base. It matches concepts extracted from a particular knowledge base against some reference taxonomy, where the taxonomy can be pre-stored or extracted from ontologies which are either stored on the local machine or are accessible through the WWW. The 'most specific' super-concept subsuming these extracted concepts is said to be the principal topic of the knowledge base.

---

[3] W3C Semantic Web: http://www.w3.org/2001/sw/
[4] Google Web APIs: http://www.google.com/apis/

Here we give a simple example about a taxonomy of Food. Suppose we already have the taxonomy depicted as in the next page:
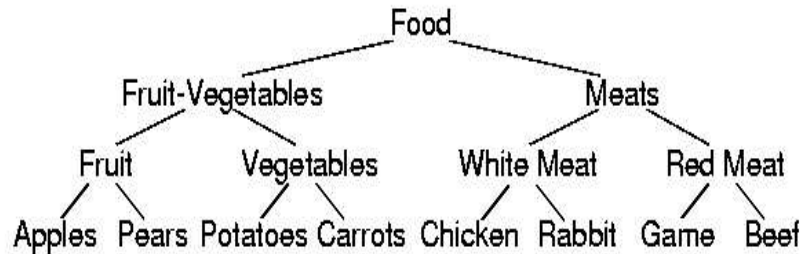

Figure 1: Taxonomy showing different kinds of food

If the concepts {Apples, Pears} are extracted and passed to the IKB system, the system would suggest that {Fruit} might be the focus of the knowledge base. Similarly, if the concepts {Apples, Potatoes, and Carrots} are extracted, {Fruit-vegetables} would be the output. If the set of concepts {Potatoes, Chicken, and Game} is provided, topic {Food} would be returned as the result.

The IKB system is implemented in Java[5]. Jena[6], a Java API, is used to manipulate RDF[7] models. The ExtrAKT system [4, 5, 6] developed at Edinburgh University is used to extract concepts from a Prolog knowledge base and then passes them to the IKB system.

There are two main inputs in the IKB system: extracted concepts from a KB and a reference taxonomy. The concepts can be extracted by the ExtrAKT system. However, choosing a suitable reference ontology is very hard. In using the IKB system, we found that there are a huge number of ontologies available online; but finding a relevant reference ontology for some particular KB is not an easy job at all. (More discussions will be given later in the next section.) However, finding a relevant reference ontology taxonomy is essential for using the IKB system.

2.2   Semantic Web

"The Semantic Web is an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation."

---- Tim Berners-Lee, James Hendler, Ora Lassila,
The Semantic Web, Scientific American, May 2001[7]

The Semantic Web [8] provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It

---

[5] Java: http://java.sun.com/
[6] Jena: http://www.hpl.hp.com/personal/bwm/rdf/jena/
[7] Resource Description Framework (RDF): http://www.w3.org/RDF/

envisions a globally interconnected network of machine-processable information, made possible by the sharing of semantic data models, which is also known as ontologies.

The Semantic Web is a collaborative effort led by the World Wide Web consortium[8] with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML[9] for syntax and URIs for naming.

There are many people working in this area to improve, extend and standardize the Semantic Web. Many documents and tools have already been developed. However, Semantic Web technologies are still in the infancy and there are many challenges in this area. One of the most important issues is to locate suitable existing ontologies to capture the user-required information from the Semantic Web. For example, if you want to publish your top ten favourite music tracks in Semantic Web, you would like to find some ontologies that represent real-world things like "artist", "track title", and "album". Otherwise, you will have to build these ontologies yourself. However, to locate suitable ontologies from the Semantic Web is currently far from easy and there is still no handy tool to help the users as we know. So, we need to build a kind of "ontology Google" tool to kick-start this process.

2.3   Google Application for ontologies

Nowadays, Google[10] is widely used to search for information on the Internet. With the powerful facilities offered by Google, we can rapidly search many resources on the web. The next question is: Can one use Google to locate an existing ontology, which conforms to the user's requirements? The answer is "Yes". As pointed out in [9], we can simply use the Google facility "filetype:" to limit the type of searching file. For example, if we search in Google for "filetype:RDFs Food", then Google will return all the RDFs files with the keywords "Food". So the user can use Google to search for existing ontologies in different formalism, such as DAML (+OIL)[11], RDFs[12], OWL[13], etc. and use (or reuse) them for their own needs.

It seems Google is a good way to help the user find suitable online Ontology resources. However, after some experiments (basically focused on finding RDFs files), we found it does not perform as expected; it is very hard to use Google to search for suitable ontology files. There are several problems:

Firstly, ontologies are not always available for a particular topic/domain. Some domains have many resources while others have very few.

---

[8] W3C: http://www.W3C.org

[9] Extensible Markup Language: http://www.w3.org/XML/

[10] Google: http://www.Google.com

[11] DAML(+OIL): http://www.daml.org/

[12] Resource Description Framework (RDF) Schema:
http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

[13] OWL: http://www.w3.org/TR/2004/REC-owl-features-20040210/

Secondly, Google returns links of relevant files, and the user will have to check if they are really relevant. This can be very time consuming because Google does not offer a quick way to browse ontology files.

Last but not least, Google searches files based on keywords supplied by the users. It does not check the real content and structure of the files. Some (usually many) irrelevant files will be returned to the user, just because they have the keywords somewhere in their files. We quite often find many RDFs files, which contain the required keywords, but on further examination of the ontology, we realised that the files do not match our needs at all; that is, they do have the required keywords, but they are not situated as required. For example, when we searched for a food ontology using the keyword concept "Food", ontologies about the Animal domain are also returned, because the file contains a statement, such as "**animal food vegetarian**". Obviously, it is not really what we want. This kind of "mistakes" can cost the user more time to find acceptable ontologies. Thus, Google's keyword searching is not good enough as an ontology search tool.

Google Web APIs are a free beta service to help programmers develop their own google-based applications. With the Google Web APIs service, software developers can query more than 4 billion web pages directly from their own computer programs. Google uses SOAP[14] and WSDL[15] standards so a developer can program in his or her favourite environment, such as Java, Perl[16], or Visual Studio .NET.[17] So, with the support of Google Web APIs, we can develop a more specific tool to search for user-required ontologies from the Semantic Web.

3. Empirical Studies

3.1 Design of OntoSearch

As mentioned in the last section, finding ontologies to satisfy user requirements is a very important issue, in both KB reuse and Semantic Web areas. There is no existing tool to solve this problem. Google does have the power, but does not seem to be specific enough to give good results.

After some experiments, we noticed that the problem arises because Google does not offer a good visualization function for the ontology files (in different formalisms, such as RDFs, etc.). As the user cannot view the ontology in an intuitive graphic format, they have to look through the ontologies as structured text files. This process takes a lot of time and cannot guarantee a good result, as the plain text of the ontology cannot show the internal structure of the ontology clearly.

After reviewing some Ontology tools, we find that showing the hierarchy (structure) of an ontology is very important to help the user to understand the nature of the ontology. Most of the tools, such as ReTAX [10], Protégé [11], OntoEdit [12, 13],

---

[14] SOAP: http://ws.apache.org/soap/
[15] WSDL: http://www.w3.org/TR/wsdl
[16] Perl: http://www.perl.org/
[17] .NET: http://www.microsoft.com/net/

OilEd [14], WebODE [15] and OntoRAMA [16, 17], offer a facility of hierarchy viewing to support the user to build and edit ontologies. A hierarchical view of ontology seems to be a good way to give the user a quick overview of the selected ontology. In this piece of work, we investigate the applicability of visualisation techniques for ontology searching on the Internet.

To answer this question, we developed a visualization tool, OntoSearch, which combines the Google search engine together with the RDFs ontology (hierarchy) visualization technology. It helps the user search for relevant (based on keywords) ontology files on the Internet and displays the files in a visually appealing way—a hierarchy tree. The hierarchical view allows users to quickly review the structures of different ontology files and select the relevant ontology files.

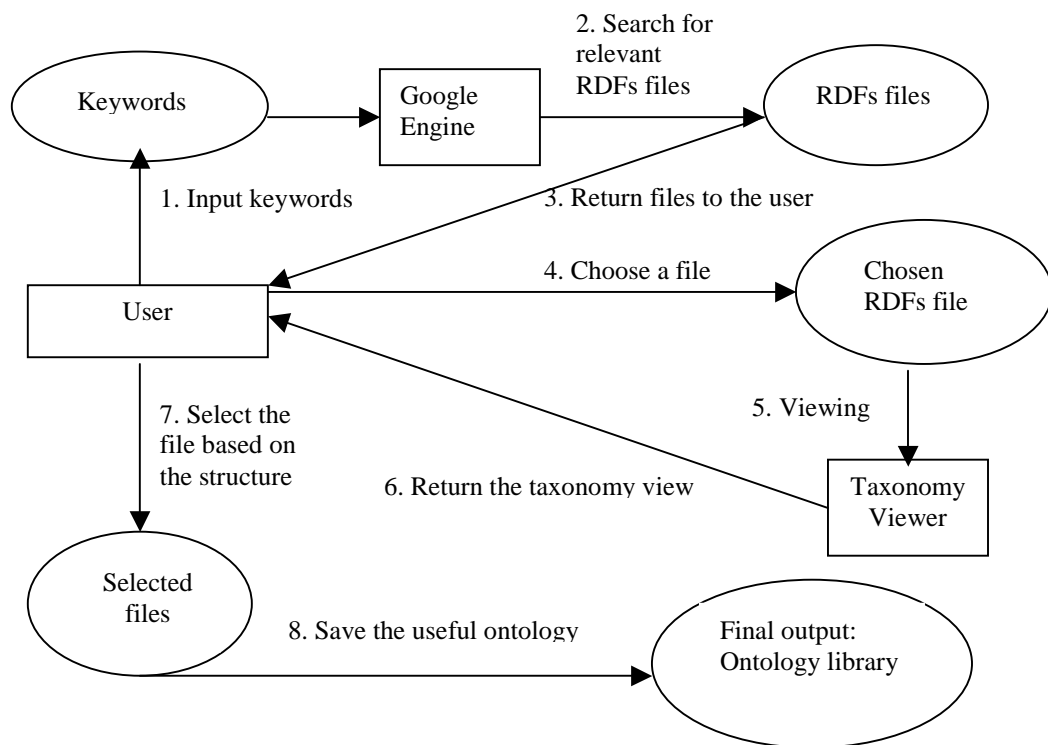We show a diagrammatic overview of OntoSearch in Figure 2:

Figure 2: Overview of OntoSearch[18]

The user inputs to OntoSearch the keywords to describe the nature of the required ontology. Then OntoSearch applies the Google engine to search for RDFs files related to the keywords and returns a list of relevant links (URLs) to the user. The user then chooses some of the returned RDFs files and displays their structure, and

---

[18] The rectangles in the figures represent processes while the ovals represent data or information.

decides which of the files are relevant. Finally, the user select the relevant RDFs files and saves them in a taxonomy library for future use.

As we now have the ontology-searching tool OntoSearch, we can link it to our other tool IKB. Figure 3 discusses links between them and demonstrates how they interoperate.
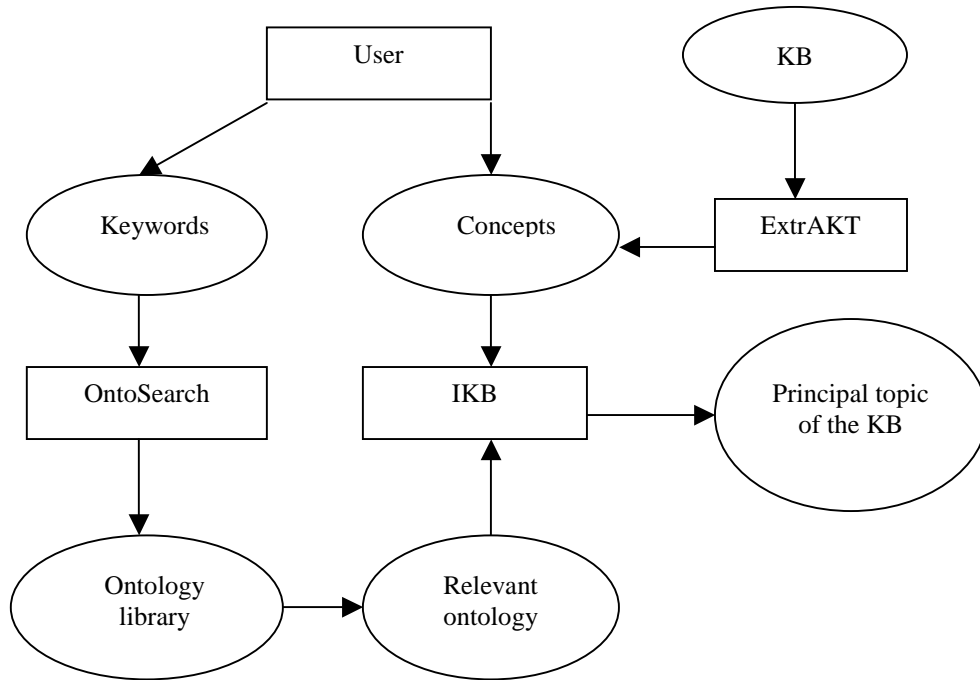
Figure 3: The relation between OntoSearch, IKB and ExtrAKT

3.2 Development of OntoSearch

The OntoSearch system is implemented in Java and JSP[19]. It is a web-based system, which can offer online service based on JBoss[20]. Jena, a Java API for manipulating RDF models, is used to read the ontology (RDFs file) into Java. Google Web APIs contribute to the Internet search engine. One JSP tag (tree tag[21]) is applied to visualize the hierarchy structure of the ontology.

The user can browse and use the OntoSearch interface using any web browser. The user inputs keywords to describe the nature of the required ontology on the

---

[19] JSP: http://java.sun.com/products/jsp/

[20] JBoss: http://www.jboss.org/index.html

[21] JSP Tree Tag (Version: 1.5): http://www.guydavis.ca/projects/oss/tags/

keyboard. Then, OntoSearch will apply the Google Web APIs to search the Internet for relevant files (the file type is restricted as RDFs now but can be changed) and return all the URLs on the screen. The user can select the files to inspect their structures in a hierarchy tree view. Thus, the user can get a general idea of the content and structure of the returned ontologies. Finally, the user can save the relevant ontology on local disk.

3.3 Demonstration of OntoSearch

Next, an example of using OntoSearch is given. Suppose the user is looking for some ontology in a Food domain. The required ontology should contain some real-world knowledge about food and related issues. The user inputs the keyword "Food" into OntoSearch. After searching, some RDFs files are returned as results, which are shown in Figure 4.
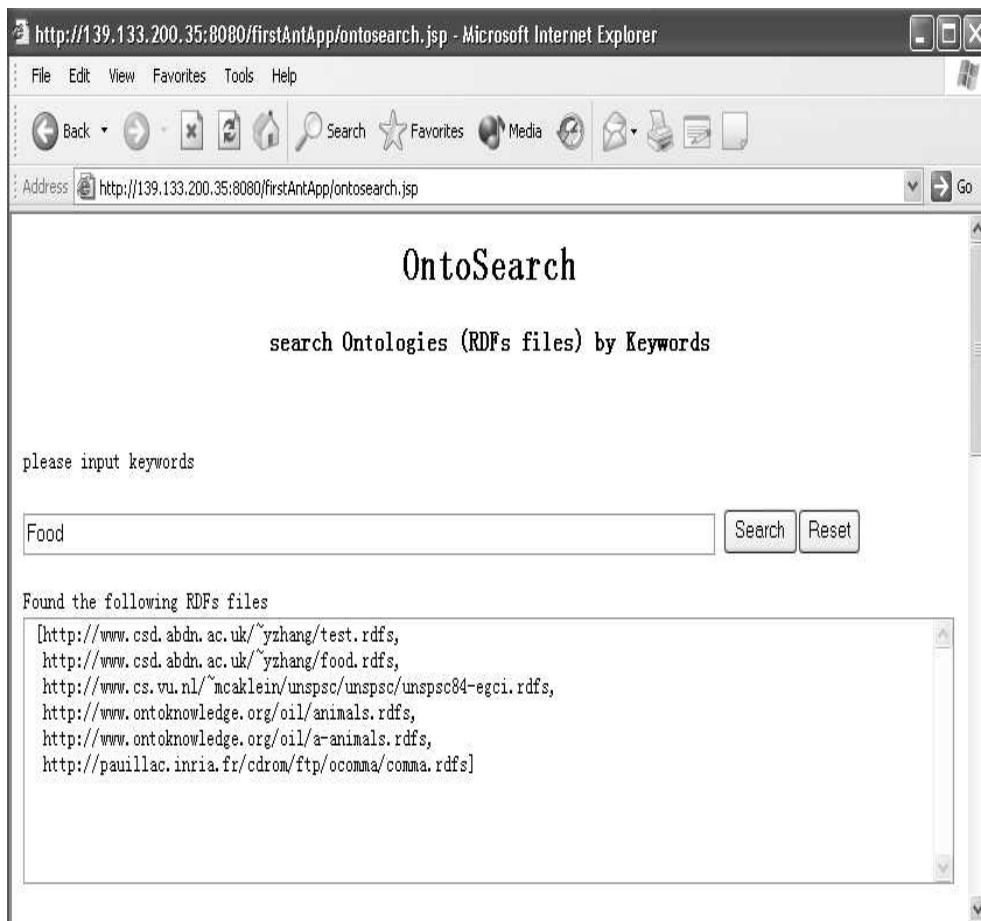

Figure 4: Search ontologies by keywords

As often many RDFs files are returned, the user then has to inspect them to check if these files are really about the Food domain. As there is one file named "Food.RDFs", the user selects that one first. The content of that RDFs file is shown as triples in Figure 5.

Figure 5: Show the triples of the example ontology

As shown in Figure 5, there is only one kind of triple in this ontology. All the triples are "subClassOf" type of triple. All the concepts in this ontology are subclasses (within several levels) of the food concept. Thus, we can think this ontology is a hierarchy of different kinds of foods. In Fact, this ontology file does match the user's needs.

Figure 6 gives the hierarchy of that ontology. Obviously, this format is much easier for the user to understand than the triple format which is shown in Figure 5.
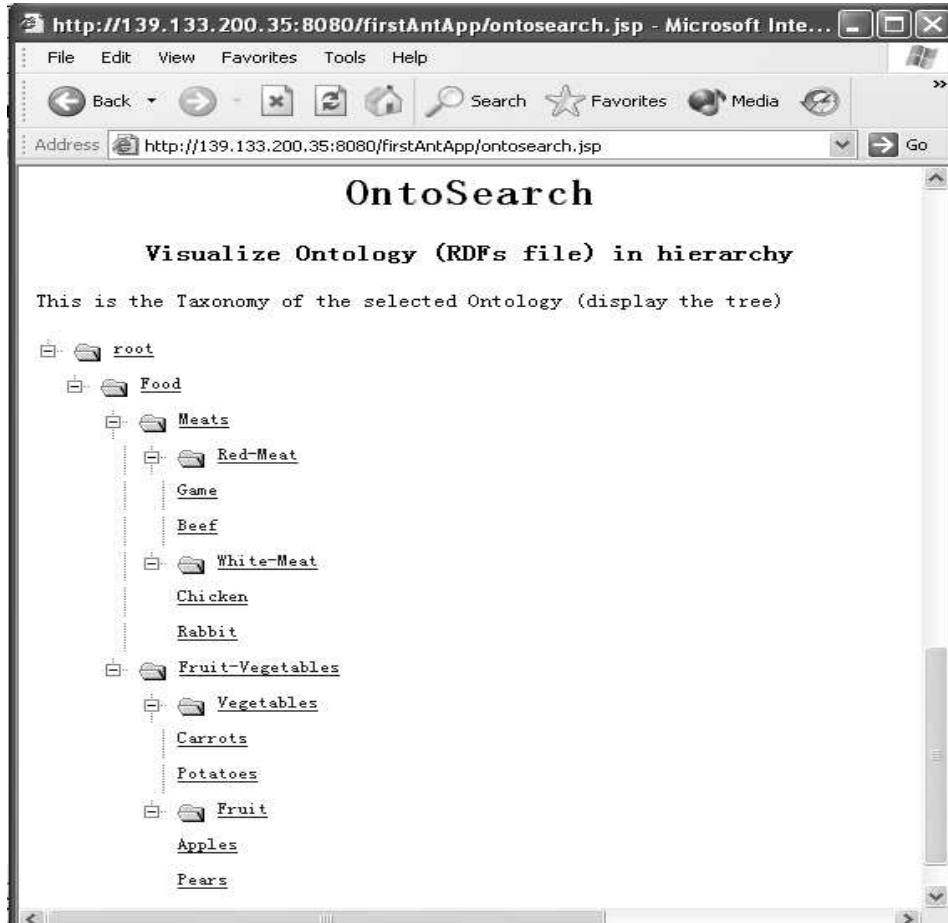
Figure 6: Hierarchy visualization of selected ontology

After viewing the hierarchy of the select ontology, the user makes the decision whether the ontology is relevant to the requirement, and then proceeds to check further returned ontologies.

4.   Summary, Discussion and Future Work

As mentioned earlier, the OntoSearch system is a useful tool which can search for ontology files from the Internet and visualize them as hierarchies. The next stage of our work will be developing an advanced mode for OntoSearch system:

The current OntoSearch system is quite simple. It can only search for one type (RDFs) of ontology file, and it only compares the user keywords with the contents of the ontology files wherever they occur. And so it matches indiscriminately the keywords both from concepts and comment fields. A future version of OntoSearch

will allow the user to choose different representational formalisms used to express ontologies, and it will allow the user to specify the type of entity (concepts, attribute or comments, etc.) to be matched.

Other future work includes:

- ## Creating a "library" of the Taxonomies

More experiments will be carried out, especially on particular domains to test our OntoSearch system. The user-acceptable ontologies will be stored in a repository for future use (eg. for use with IKB).

- ## WordNet[22] application

The synonym problem is not well addressed in the current version of OntoSearch. We are planning to incorporate WordNet in future versions so that our tool will be more effective, ie it will retrieve a large number of relevant ontologies.

Acknowledgements

References

1. Advanced Knowledge Technology (AKT project) http://www.aktors.org/akt/
2. Sleeman D, Potter S, Robertson D, and Schorlemmer W.M. Enabling Services for Distributed Environments: Ontology Extraction and Knowledge Base Characterisation, ECAI-2002 workshop, 2002
3. Sleeman D, Zhang Yi, Vasconcelos W. Characterisation of Knowledge Bases, Proceedings of AI-2003 (the twenty-third Annual International Conference of the British Computer Society's Specialist Group on Artificial Intelligence (SGAI)), 2003
4. Schorlemmer M, Potter S, and Robertson D. Automated Support for Composition of Transformational Components in Knowledge Engineering. Informatics Research Report EDI-INF-RR-0137, June, 2002

---

[22] WordNet: http://www.cogsci.princeton.edu/~wn/

5. Sleeman D, Potter S, Robertson D, and Schorlemmer W.M. Ontology Extraction for Distributed Environments, In: B.Omelayenko & M.Klein (Eds), Knowledge Transformation for the Semantic Web. pub Amsterdam: IOS press, p80-91, 2003

6. ExtrAKT system: a tool for extracting ontologies from Prolog knowledge bases. http://www.aktors.org/technologies/extrakt/

7. Berners-Lee T, Hendler J, Lassila O, The Semantic Web, Scientific American, 2001 http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2

8. Sean B. Palmer, The Semantic Web: An Introduction, 2001-09. http://infomesh.net/2001/swintro/#ontInference

9. DuCharme B, Googling for XML, February 11, 2004, http://www.xml.com/pub/a/2004/02/11/googlexml.html

10. Alberdi E & Sleeman D, ReTAX: a step in the Automation of Taxonomic Revision, Artificial Intelligence, 91, p257-279, 1997.

11. Musen M. A., Fergerson R. W., Grosso W. E., Crubezy M., Eriksson H., Noy N. F. and Tu S. W., The Evolution of Protégé: An Environment for Knowledge-Based Systems Development, International Journal of Human-Computer Interaction (in press)

12. Sure, Y., S. Staab, J. Angele. OntoEdit: Guiding Ontology Development by Methodology and Inferencing. In: R. Meersman, Z. Tari et al. (eds.). Proceedings of the Confederated International Conferences CoopIS, DOA and ODBASE (2002) Springer, LNCS 2519, 1205-1222.

13. Sure, Y., S. Staab, M. Erdmann, J. Angele, R. Studer and D. Wenke, OntoEdit: Collaborative ontology development for the semantic web, Proc. of ISWC2002, (2002) 221-235.

14. Bechhofer S., Horrocks I., Goble C., Stevens R. OilEd: a Reason-able Ontology Editor for the Semantic Web. Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence, September 19-21, Vienna. Springer-Verlag LNAI Vol. 2174, pp 396--408. 2001.

15. Corcho, O., Fernandez-Lopez M., A.Gomez-Perez and Vicente O., WebODE: An Integrated Workbench for Ontology Representation, Reasoning and Exchange, Prof. of EKAW2002, Springer LNAI 2473 (2002) 138-153.

16. Eklund P, Roberts N and Green S. OntoRama: Browsing RDF Ontologies using a Hyperbolic-style Browser http://www.kvocentral.com/kvopapers/ontorama.pdf

17. Eklund P, Cole R, and Roberts N. Retrieveing and Exploring Ontology-based Information, Handbook on Ontologies. International Handbooks on Information Systems Springer 2004, ISBN 3-540-40834-7 (2003) 405-414.