Retrieval of scientific data in Esperonto

Valentina Tamma and **Michael Wooldridge** and **Ian Blacoe**
Department of Computer Science,
University of Liverpool,
Liverpool L69 7ZF, UK
email: {V.A.M.Tamma, M.J.Wooldridge, I.W.Blacoe}@csc.liv.ac.uk

Andreas Persidis
Biovista, 34 Rodopoleos Street,
Ellinikon, Athens,
16777, HELLAS
email: biovista@ath.fortnet.gr

Introduction

Esperonto is an IST project (www.esperonto.net, IST-2001-34373) whose objective is to bridge the gap between current web technologies and the Semantic Web (Fen 2003). The current web is aimed at human consumption, and is ed on markup languages such as HTML, which specify the page layout in order to render it more appealing to human users. On the other hand, the Semantic Web is intended to be used and understood by software programs, and is based on new or relatively new types of languages such as RDF(S) (Decker *et al.* 2000), DAML+OIL (DAM), and more recently OWL (OWL).

This change in the user perspective is determined by the increasing information overload which is characterising the information society (Maes 1994). More and more often, when querying search engines, we are faced by hundreds of documents which need to be at least scanned through in order to determine whether they are relevant to our needs, and search engines are struggling to ensure the quality of retrieved results. Moreover, these result do not include dynamic web content, such as the content generated by querying databases. We therefore face the challenge of taming the huge growth of the WWW by making use of novel retrieval approaches that make use of other information in additio to simple keywords, information that concerns the meaning, that is *semantics*, of the content of web pages. In fact, Semantic Web (SW) pages store both human-understandable content (the text composing the document) together with an encoding of the semantics and the structure of the digital content. The semantics and the structure of digital content is represented by means of ontologies (Studer, Benjamins, & Fensel 1998), which are explicit and machine sharable representations of the conceptualisation abstracting a phenomenon.

Ontologies are represented in one of the aforementioned SW languages; these languages are all based on XML and they provide different expressive capabilities. Software agents are able to process the digital content, and can thus offer services which make use of or retrieve this knowledge. The advantage of using software agents comes from their intrinsic characteristics of autonomy, proactiveness, and social ability which permit them to carry out complex tasks on behalf of the users (Wooldridge & Jennings 1995). This paper describes the approach we follow in the Esperonto project to the retrieval of both static and dynamic content, and we present it in the context of one of the test cases we are using to prove the validity of our approach: scientific discovery.

The Esperonto architecture

The Esperonto project (IST-2001-34373, www.esperonto.net) (Benjamins *et al.* a) started in September 2002 and explicitly aims to bridge the gap between the current web and the SW. In order to achieve this goal, Esperonto provides three main type of knowledge services that we describe below, relating them to the Esperonto architecture shown in Figure 1 (Benjamins *et al.* b).

Content availability services: One of the main challenges that research in the SW has to face is the availability of SW content. In order to get this technology up to speed it is essential to make available web pages in which the semantics of the text in natural language is defined and made explicit in a machine processable language. The amount of effort involved with

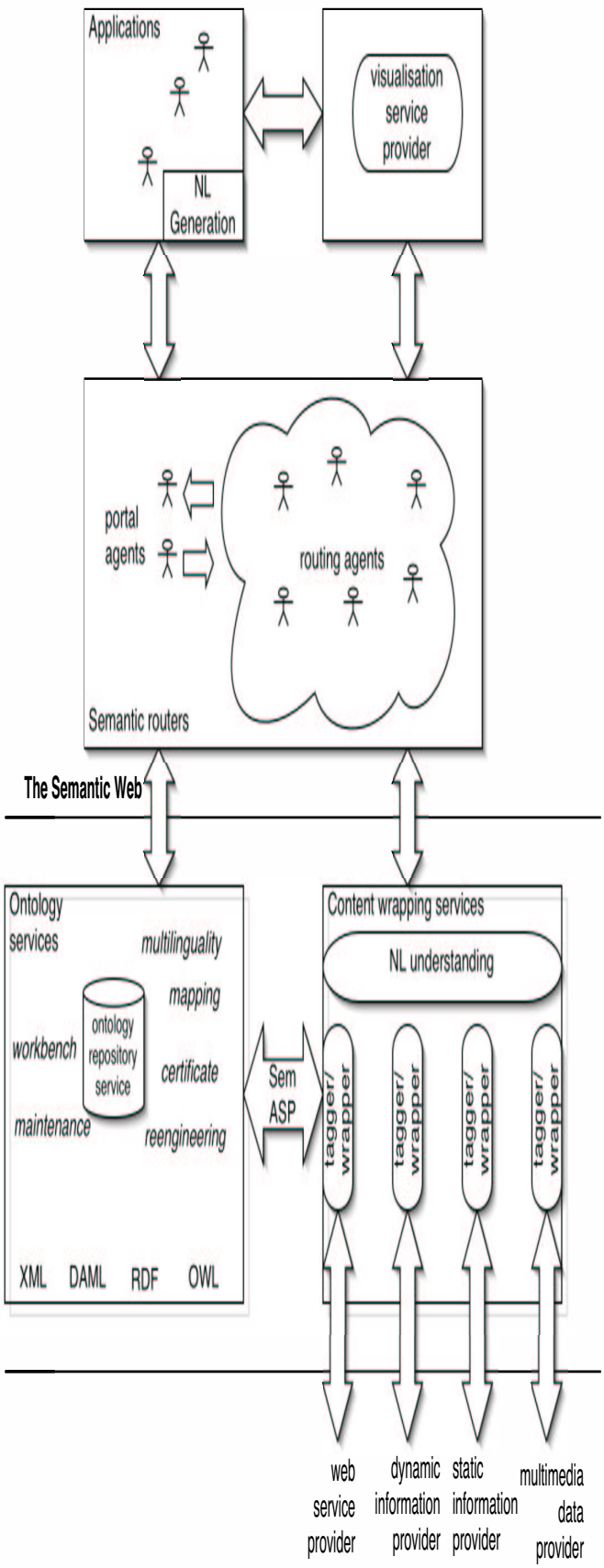

Figure 1: The general architecture of the Esperonto system

writing such pages is considerable, and some tools are being developed to reduce the burden of annotating web pages with semantic information. This effort is usually related to writing *new* web pages, whose content is annotated while the page is written. In addition, many of these tools focus only on *static* content and do not give any provision for the annotation of *dynamic* content generated from databases, which represent a considerable amount of the retrieved content. But restricting only to new web content would mean discarding the content represented in the current web: an unpractical and anti-economical choice.

Esperonto tries to solve this problem by providing annotation services to content providers in the form of tools and techniques that enable them to publish their (existing and new) content on the SW, irrespective of their native language. The content made available by the annotation services is used by software agents providing the search and retrieval services described below, and can be used to facilitate human access to the information by means of visualisation and semantic navigation tools.

Content availability services are provided by the SEMantic Annotation Service Provider (SemASP) (Benjamins *et al.* a) which has two main components: the *ontology repository service*, which provides the ontology services described below, and the *wrapper services* which perform the annotation of different types of content. The SemASP in its totality provides a suite of tools that are aimed at different types of users, and which include manual annotation and semi-automatic annotation of digital content, multilingual support in the most common European languages, and ontology services. The SemASP permits us to annotate static, and dynamic content, multimedia content and even web services.

A comparison with other annotation techniques presented in the literature can be found in Deliverable 2.1 (Esp a) downloadable from the Esperonto portal (`www.esperonto.net`).

Ontology services: Ontology services are provided by the *Ontology repository service* component of the SemASP. The services offered are those related with the management and validation of multiple related ontologies and/or multiple versions of the same ontology. They include an ontology editor suite (WebODE) (Arpírez *et al.*) which facilitates ontology construction, selection, and browsing, and translation into the most common ontology and knowledge representation languages such as, DAML+OIL, OWL, Prolog, and Jess. The suite supports consistency checking, ontology versioning, and ontology maintenance. Knowledge sharing among different user communities is also supported, by means of ontology import, alignement and mapping services.

A review on the state of the art of ontology services can be found in Deliverable 1.1 (Esp b) downloadable from

the Esperonto portal (www.esperonto.net).

Search and retrieval services: Search and retrieval services are provided by the Semantic Routers, a peer-to-peer multi-agent system that deals with user queries. This component is responsible for query decomposition, answering and aggregation. During the process of query aggregation, the retrieved web pages are also evaluated in order to avoid duplicate or incomplete data and to improve performance. The semantic routers are based on the notion of semantic indexing of the resources, where the web pages are indexed not on the basis of the keywords identifying them, but on the *concepts* corresponding to the meaning of the keywords. The semantic indices are the mechanisms that permit the routing of the queries to the competent agent on a P2P basis. A simple query is initially routed to an agent, which consults the indices to verify whether it can answer the query. If the agent cannot answer the query, or if it believes that the answer is incomplete, it routes the query to the agent with the closest interests. The decision about which agent has the closest interests is based on the evaluation of a measure of *semantic similarity* between the concepts composing the indices, which takes into account the ontological definition of the concepts in terms of degree of similarity and degree of differences (Tversky 1977; Rodríguez & Egenhofer 2002).
A literature review of the relevant efforts in the area of peer to peer systems and of information retrieval is presented in the Esperonto Deliverable 4.1 (Esp c) downloadable from the Esperonto portal (www.esperonto.net).

Scientific discovery

The technology proposed in the Esperonto architecture will be demonstrated in a prototype that aims to provide services to European citizens, exploring at the same time innovative uses of available and novel technological solutions. One of the domains we will use as a test case is *scientific discovery.*

The scientific domain offers a huge amout of information in print as well as in electronic format. Usually, this information is organised according to the structure defined by indices, hyperlinks, database schemas, etc. However, it is of crucial importance that scientific data is not only stored correctly, but also that the new information related in ways that have not been explicitly noted is stored, thus easing the process of scientific discovery. Most of this information is contained in textual sources, such as written reports, e-mail messages, journal articles, etc. Very few efforts have aimed to develop techniques that search and make discoveries in these textual resources.

Literature-based discovery methods aim to to discover new, and potentially meaningful relations between a given starting concept of interest and other concepts, by means of mining bibliographic databases such as Medline. The main principle behind literature-based discovery is that new knowledge can be obtained as the result of the combination of existing, though not connected, bibliographic information results in new knowledge. One publication may state the relationship between two phenomena A and B, while another reports on the relationship between the phenomena B and C. If no assciation has been reported between A and C, such an association can be considered new and may be of scientific interest. The crucial notion is this view is that two pieces of information are not explicitly related, but there is a hidden relation that needs to be discovered and made explicit. Useful clues to discover the relation might be one or more common aspects of the two pieces that eventually provide indirect links.

There are a number of literature-based discovery approaches (Swanson 1990; Gordon & Lindsay 1000; Weeber 2001). They mainly describe the literature-based discovery process as a two-step approach. In the first step, a hypothesis is formulated, while in the second step, this hypothesis is validated or tested by extensive bibliographical analysis. The hypothesis-generated approach is usually named as an open discovery mode, whereas the testing approach as a closed discovery mode. The main difference between the two approaches concerns textual sources that are analysed. In the open search, the literatures on a *phenomenon* and the *links* that play a role in the phenomenon are studied in order to find the *elements* which act on selected links. In the closed search, textual sources concerning the phenomenon and the related elements are studied in order to find the links.

Scientific discovery in Esperonto

Research on how to perform scientific discovery in the Esperonto project is still at an early stage, however we can make here some observations on the role that the knowledge services provided by the Esperonto architecture play in the process of scientific discovery. We have concentrated our attention on the domain of drug discovery, and in particular, of drugs related to the cure of rheumatoid arthritis (RA). The scientific discovery process will need the following components: Corpus repository, Ontology services, Semantic Annotation, Search and Retrieval.

Corpus repository: Corpus repository is the component that manages textual resources. It provides the typical functionalities of a repository, and perfoms limited consistency checking.

Ontology services: Biovista (who is the test case provider and the knowledge expert for the medical domain) designed a disease-drug-anatomical location ontology, which focussed on RA, and that described this disease in terms of the related diseases and their symptoms, the drugs that can treat it with their properties, and the anatomical part which can be affected by a disease. The ontology was originally written in Protege-2000 (Fridman Noy, Fergerson, & Musen) and later imported in WebODE by means of the import functionality provided as part of the ontology services. The ontology plays a central role in the scientific discovery process, since it permits the annotation of the corpus in the repository, but it also supports the analysis phase, by means of the explicit representation of the relationships existing between the concepts represented in the ontology. Furthermore, the semantic router component of the Esperonto architecture makes use of the ontology to build the semantic indices and compute the semantic similarity between the interests of the agents in the architecture.
The ontology services must support the creation, maintenance, and browsing of the ontology (and its instances) and must also ensure a required degree of quality.

Semantic Annotation: The SA component provides users with both manual and semi-automatic annotation services. It is able to identify the terms of interest and relate them to the ontology by means of correct relationships such as synonymy or partonomy.

Search and retrieval: This component provides the problem-solving functionalities to support literature-based scientific discovery. Two main functionalities are supported: query expansion and the provision of multiple-link explanations. The semantic router component, responsible for the search and retrieval functionalities, interacts with the users and with the ontology repository (to build its indices and to find related concepts). Query expansions are performed by merging the relevant ontology with other existing taxonomies, such as UMLS and PubMed. Multiple-link explanations is the ability to support a user query with the relevant data in one or more steps, from initial query formulation, down to the resources providing the final data.

The semantic router component is also provided with inference capabilities in order to attempt the formulation of hypotheses related to the user query. Initally, we will concentrate on candidate drugs for diseases, but we aim to generalise the method for all kind of queries. The semantic router reasons with the knowledge represented in the ontology in order to make explicit hidden links between concepts.

Conclusion

We have presented in this paper some initial experiences on scientific discovery in the context of the Esperonto project. Esperonto aims to provide knowledge services that can make it possible to bridge the gap between the current web and the semantic web. One of the applications we have considered to prove the validity of the tools we are currently building is that of literature-based scientific discovery.

Acknowledgement

The research described in this paper is funded by the IST project Esperonto. The Esperonto consortium is composed by: iSOCO (Spain), Universidad Politcnica de Madrid (Spain), University of Innsbruck (Austra), University of Saarland (Germany), University of Liverpool (UK), Fundacin Residencia de Estudiantes (Spain), CIDEM (Spain), Biovista (Greece). The authors would like to thank all members of the Esperonto consortium.

References

Arpírez, J.; Corcho, O.; Fernández-López, M.; and Gómez-Pérez., A. 2001. Webode: A scalable workbench for ontological engineering. In *Proc. of the K-CAP 2001*. ACM-Sigmod.

Benjamins, V.; Contreras, J.; Corcho, O.; and Gomez-Perez, A. 2002. Six challenges for the semantic web. In *Proc. of the KR'02 Semantic Web workshop*.

Benjamins, V.; Contreras, J.; Gomez-Perez, A.; Uszkoreit, H.; Declerck, T.; Fensel, D.; Ding, Y.; Wooldridge, M.; and Tamma, V. 2003. Esperonto application: Service provision of semantic annotation, aggregation, indexing, and routing of textual, multimedia and multilingual web content. In *Proc. of WIAMSI'03*.

The DARPA agent markup language. `http://www.daml.org/`.

de Madrid, U. P. 2003. Deliverable 1.1: State of the art on ontologies. `http://www.esperonto.net`.

Decker, S.; Melnik, S.; van Harmelen, F.; Fensel, D.; Klein, M.; Broekstra, J.; Erdmann, M.; and Horrocks., I. 2000. The semantic web: The roles of XML and RDF. *IEEE Internet Computing* 4(5):63–74.

Fensel, D.; Hendler, J.; Lieberman, H.; and (Eds.), W. W., eds. 2003. *Spinning the Semantic Web: Bringing the World Wide Web to its full potential*. MIT Press, Boston.

Fridman Noy, N.; Fergerson, R.; and Musen, M. 2000. The knowledge model of protege-2000: Combining interoperability and flexibility. In *Proceedings EKAW'00 Conference*, 17–32. Berlin: Springer Verlag.

Gordon, M., and Lindsay, R. 1000. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science* 50:574–587.

Isoco. 2003. Deliverable 2.1: State of the art on annotation. `http://www.esperonto.net`.

Maes, P. 1994. Agents that reduce work and information overload. *Communications of the ACM* 3(37):31–40.

of Liverpool, U. 2003. Deliverable 4.1: State of the art on peer to peer, and information retrieval. `http://www.esperonto.net`.

W3C web-ontology working group. `http://www.w3.org/2001/sw/WebOnt/`.

Rodríguez, M., and Egenhofer, M. 2002. Determining semantic similarity among entity classes from different ontologies. *IEEE transactions on knowledge and data engineering.*

Studer, R.; Benjamins, V.; and Fensel, D. 1998. Knowledge engineering, principles and methods. *Data and Knowledge Engineering* 25(1-2):161–197.

Swanson, D. 1990. Medical literature as a potential source of new knowledge. *Bull. Medical Libr. Assoc.* 78(1):29–37.

Tversky, A. 1977. Features of similarity. *Psychological Review* 84(4):327–372.

Weeber, M. 2001. *Literature-based Discovery in Biomedicine.* Ph.D. Dissertation, Dissertation Rijksuniversiteit Groningen.

Wooldridge, M., and Jennings, N. 1995. Intelligent agents: Theory and practice. *Knowledge engineering review* 10(2):115–152.