A Semantic Search Engine for Learning Resources

D. Taibi [*,1], **M. Gentile**[1] , and **L. Seta**[1]

[1] Italian National Research Council - Institute for Education Technology, Via Ugo La Malfa 153, 90146 Palermo, Italy

In this paper we present an architectural overview of a search engine based on semantic web technologies to improve the search for learning resources. The use of search engines has contributed to the success of the Web. At present, many people use search engines to retrieve relevant information about a topic and students also use search engines to find learning resources.
Keyword-based searches present several problems related to the meaning of the keyword used in the search query, these limits can be overcome by applying semantic web technologies to search engines.
Semantic web meta-data can be used in e-learning fields to enrich the information content of the learning object and to develop a better search methodology. A semantic search engine can elaborate search queries semantically to find conceptual relations between documents and to retrieve learning resources in a more efficient way.

Keywords semantic web for e-learning; search engine.

1. Introduction

In the last few years learning processes have benefited from the technological evolution of the web. The diffusion of the web has permitted the introduction of new educational processes, which are more flexible for accessing the resources. For example, the use of the Internet allows universities to provide online courses or virtual laboratories to the students. Moreover, the students can not only access the learning material in a linear way, but can also carry out personal research into a subject of interest, following the links presented in a hypertext document or exploiting the enormous quantity of information available on the Internet.

The web provides an enormous amount of information which, on the one hand, offers an inexhaustible source for searching for learning material but, on the other, causes an excessive amount of noise in the search results. Currently search engines are widely used for searching, but there are a lot of unsolved problems related to their effectiveness.

According to [1] keyword-based search engines present serious problems related to the quality of the search results. It often happens that relevant pages are not indexed by a traditional search engine, in this case important information can be reached only if its specific internet address is known. Moreover, searches based on keywords are very closely related to the spelling of the word and not to its meaning, thus semantically similar queries can return different results. Then retrieving a large body of information by using a search engine is a very time-consuming task for the users who have to perform it manually. This is because the result of a search engine is a single web page, and to retrieve information it is necessary to perform several queries.

If we take a look at the structure of the web, it is composed of a huge pool of documents and links between them. Currently, web documents present human readable contents targeted at humans [2]. More and more often the web is not used only by people, but software agent communities are becoming users of the web too. All these needs have led to the development of the Semantic Web.
One of the main aims of the semantic web is to improve the existing web with a semantic layer that allows machines to understand it, or better, to enable software programs to process information efficiently.
To achieve its aims, the semantic web is based on the relationship between several layers, each of which has a specific role [3]: at the base, the XML [4] layer provides a surface syntax for structured documents, but imposes no semantic constraints on the meaning of these documents; the relative XML SCHEMA allow the structure of XML documents to be restricted and also to extend XML with datatypes. Above

[*] Corresponding author: e-mail: davide.taibi@itd.cnr.it , Phone: +39 0916809216

the XML layer, the RDF [5] layer provides a data-model for objects (resources) and relationships between them, thus providing a simple semantic; the relative RDF SCHEMA is a vocabulary for describing properties and classes of RDF resources, with a semantic for generalization-hierarchies of such properties and classes. At the top of this hierarchy is the OWL [3] layer, that adds more vocabulary for describing properties and classes: among others, relationships between classes, cardinality, equality, richer typing of properties, characteristics of properties and enumerated classes. OWL is the ontology layer that represents the formal common agreement about meaning of data. Finally, the logic layer enables intelligent reasoning with meaningful data [6].

This structure of the Semantic web is used to represent real world objects as resources linked among themselves through different kinds of relationships; in this sense the semantic web changes the existing web where documents are related by links of a single kind, into a web in which each document can have different kinds of relationships with the others [2].

Search engine technology can draw enormous benefits from structuring information according to the semantic web indications. In this way it is possible to create intelligent search engines that can return results based not only on the occurrence of keywords in a document but also on the conceptual links between documents.

In this paper we show how it is possible to improve a traditional search engine to create a semantic search engine for learning resources to support student learning activities.

2. Semantic web, search engine and e-learning

An important contact point between e-learning and the semantic web is the use of meta-data to improve the description of online resources. Meta-data are the fundamental building blocks of the Semantic Web [7], and the e-learning community recognizes their importance as demonstrated by the diffusion of learning standards for the definition of meta-data for learning objects like IEEE Learning Object Metadata, IMS, SCORM; in particular, these standards highlight the role of metadata in favouring interoperability, reusability and accessibility of learning resources.

Meta-data are used essentially to describe learning resources, providing information about who created them, where they are stored, and other pedagogical properties regarding, for example, the level of difficulty, and so on.

However, most of the previous metadata-standards lack a semantic layer [6]. According to [1], the semantic layer is necessary to support efficient retrieval of learning resources from a repository, to compose learning units from different authors, and to perform automatic management of learning resources.
In this paper we focus our attention on the first issue related to efficient information retrieval; in our opinion the technologies of the semantic web can be of great advantage in this field.

In their learning activities, students frequently use search engines to access relevant information that can be found on the Internet. The real problem is that the Internet has become a large pool of information and students using a traditional search engine run the risk of getting “lost in hyperspace”.

To improve the search activities it is possible to use a specialized search engine, that indexes pages concerning a particular domain with fewer interferences in the results than a general search engine. However, in this case, retrieving a large body of information by using a specialized search engine requires the users to perform manual operations.

Our approach proposes to combine the semantic web technologies with the use of a specialized search engine to create a semantic search engine. It allows students to retrieve relevant information from the content of web pages, as well as from the conceptual links between the concepts included in the web pages; the student can therefore submit queries that will be elaborated semantically with inferential procedures to find conceptual relations between documents.

Ontologies play a key role in achieving this goal. They were introduced into the semantic web to add the ability of performing inferential tasks. Using inferential rules it is possible to discover new facts from an existing knowledge base, in this way building conceptual links between documents is a feasible task.

If ontologies are at the base of the reasoning system it is necessary to define logical rules that permit new knowledge to be constructed.

The following sections describe our proposals for using semantic web technologies to implement a search engine for learning resources.

3. Architecture

3.1 Static and dynamic search

In this paper we suggest two possible architectures to implement a search engine which can use the Semantic Web capability. We call the first, Static Ontology Based Search (SOBS) architecture and the second, Dynamic Ontology Based Search (DOBS) architecture.

In the SOBS the search engine uses a fixed set of ontologies to organize the Web resources on semantic bases. In the DOBS we propose a more flexible engine which can process the user's query using an ontology defined by the user himself.

In both architectures we have to resolve the problem of the semantic space construction starting from the ontology. The Semantic Space is a formal representation of the Web resources by means of a terms-documents matrix, where the terms are related with the concepts expressed in the ontology. This representation permits the Web resources to have semantic annotations. We underline the use of the term "Web resources" to indicate here not only a single document but, in general, parts of documents, too, referred to as "chunks", hereafter.

The Semantic Space construction is carried out in a two step process; in the first step, the search engine works in a traditional way, using the crawler and the indexer to find documents on the Web. In the second step we designed a specific module, the Information Extractor, which by applying some techniques like the Latent Semantic Analysis (LSA), organizes the Web resources and defines the Semantic Space.

The main difference between the two architectures is in work flow: in the case of the SOBS, the Information Extractor works after the traditional search engine pipeline and the Semantic Space is constructed, using the entire space of documents (Fig. 1); in the DOBS case the Information Extractor works after the user has submitted the ontology to the search engine, and the Semantic Space is constructed considering only a subspace of the Web documents (Fig. 2), this permits faster semantic annotations and answers. We have also designed another specific module, the Reasoning Module, to infer new relationships between the chunks of documents present in the Semantic Space using the ontology. This module enables the semantic search engine to recognize new relationships between chunks, which are not immediately apparent. Finally, specific graphic interfaces have to be designed to help the user to interpret the search engine replies and to visualize the ontology based conceptual map.

In the following two sections we illustrate the two architectures with greater detail and we underline strengths and weakness of the two proposals.

3.2 SOBS architecture

The first architecture that we present in this paper has four main components, each of them was designed to be as flexible and independent as possible, in such a way that the search engine can easily be enriched with new modules. As shown in figure 1a the main components in SOBS architecture are: the document retriever, the information extractor, the inference engine, the graphical interface.

The document retriever performs the operations of a traditional search engine based on keywords. Its main components are the crawler and the indexer; they have to browse the web to find documents to index. The architecture that we propose adds an information extraction module; it analyzes the documents found by the crawler and associates semantic tags related to predefined ontologies to chunks of the documents.

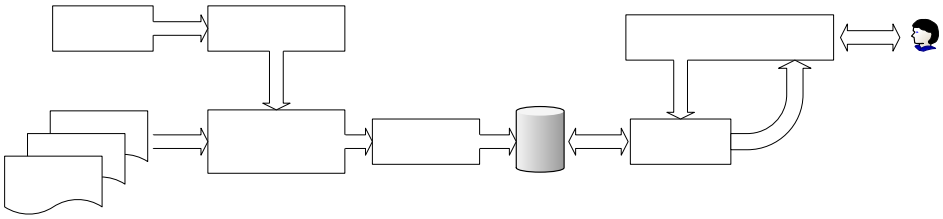

Fig. 1 *The SOBS architecture.*

This task is very important in our architecture and its execution is not easy. We are investigating the possibility of using semantic analysis techniques like LSA (Latent Semantic Analysis) in order to determine which chunk of a document belongs to a particular ontology class.

The functionalities of the information extractor module will now be explained in more detail. In a preliminary phase, for each ontology of reference, a semantic space will be constructed using a terms-documents matrix that has, as terms, the fundamentals elements of the ontology (class, subclass) and as documents, those with contents related to the ontology.

This semantic space will be used to associate a semantic concept (related to the ontology) to the chunks of the documents. The Information Extractor divides each document into paragraphs, the latent semantic analysis links each paragraph with an ontology class. The relationship between a paragraph and a class will be expressed using OWL language and stored in an OWL database (that will be used later by the reasoning module).

The reasoning module plays a fundamental role in our architecture, its function is to activate the inferential process with the aim of linking chunks of different documents which have similar semantic significance. This module is activated by a search performed by a student. It will process the information available in the OWL database to connect information which is apparently unrelated but has a similar semantic content on the basis of the ontology of reference that the student has chosen. The reasoning module works on the OWL database created by the information extractor and performs OWL queries on it. The aim of the interface module is to facilitate the use of the search engine. In the description of this module we describe a possible scenario of the SOBS architecture.

Crawler

In the user interface, the student can choose the ontology to use for its search. This choice is performed on a set of predefined ontologies regarding selected subjects. In a first test phase of the system we propose to use ontologies regarding mathematics and history of art. When the student chooses an ontology, the system will show him a conceptual map representing the ontology classes and the relationships between them. By clicking on the map the student can choose the ontology concept that he wants to study in detail.

Based on this choice, the inferential engine will be activated. Using the mechanisms described above it will return the document chunks related to the concept chosen and also those that are conceptually related to it. With regard to the visualization of results, we are considering the development of a visual presentation that shows introductory or more advanced concepts related to the student's choice.

Ontology

3.3 DOBS architecture

The above described SOBS has two principal limitations: the user can only choose an ontology from a fixed set; the OWL database is created during the recovery phase and the semantic information stored in it is not related to the user's query.

This approach is similar to the traditional way of considering the semantic search. The DOBS represents a different approach to this task. In particular, we want to explore the possibility of creating a search engine which can elaborate the user's query semantically, using ontology defined by the user, and which can organize the results in a new document where the relevant resources are indicated.

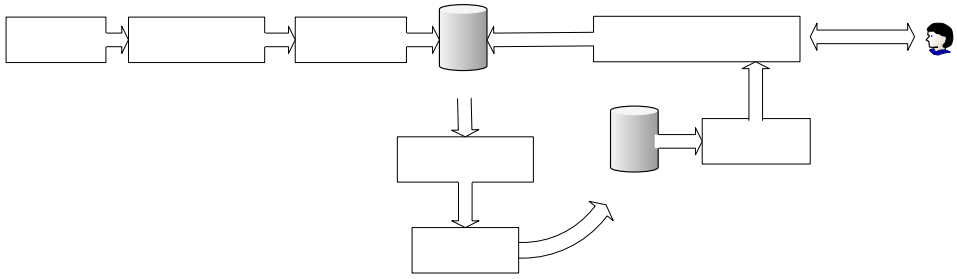

Fig. 2 *The DOBS architecture.*

The main problem to solve in this approach is the semantic space construction that must now be carried out only after the user has submitted a query and an ontology. Using the classes and subclasses present in the user's ontology, the search engine retrieves some documents and passes them to the extractor for semantic annotation. In this way the semantic space can be organized and then the user's query can be processed. Consequently, in this approach the results of the user's query must be mapped onto the semantic space, to permit semantic inferences about them, for example by means of specialized classification algorithms. To test the reliability of this architecture we have designed the following use scenario: the user can submit an ontology and a query to the semantic search engine. The ontology can be created by the user or found on the Web, and the query can be written in natural language. The DOBS elaborates this information and returns a document where links are shown to the principal resources present on the Web. Using the inferences of the ontology the reasoning module can organize the resources in the final document, highlighting the concepts involved, and the user can explore it to recover the relevant information.

The DOBS architecture is much more flexible than the SOBS one but to create a efficient engine that is based on it, fast algorithms have to be implemented to perform natural language processing and text classification. Currently a lot of techniques may be considered in order to achieve these goals and we are trying some of these out to test the reliability and the efficiency of our proposal.

4. Conclusion

In this paper we have presented two possible architectures for a semantic search engine. Our work shows how semantic web technologies can be useful in supporting students in the fundamental task of searching for learning resources.

Future work will investigate two relevant problems: at present the implementation of the reasoning module is based on inference engine technologies that do not permit the use of the expressive power of the OWL language; moreover, we will investigate the use of different approaches for the visual results representation in order to emphasize semantic relationships between information.

References

[1] G. Antoniou, F. van Harmelen. A semantic web primer, The MIT Press (2004) chap. no. 1.
[2] R. Guha, R. McCool. TAP: A semantic web Platform,. Computer Networks: The International Journal of Computer and Telecommunications Networking , **42**, 5 (2003).
[3] D. L. McGuinness and F. van Harmelen. Web ontology language. http://www.w3.org/2001/sw/WebOnt/.
[4] T. Bray, J. Paoli , C. M. Sperberg-McQueen, Extensible markup language, World Wide Web Journal, **2**, 4, (1997) p.29-66.
[5] O. Lassila and R. Swick. Resource description framework (RDF) model and syntax specification. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.
[6] L. Stojanovic, S. Staab, R. Studer. eLearning based on Semantic Web, Proceedings of WebNet 2001 - World Conference on the WWW and the Internet, Orlando, Florida, USA, 23-27 October 2001.
[7] M. Nilson, M. Palmér, A. Naeve. Semantic Web Meta-data for e-Learning – Some architectural Guidelines, Proceedings of WWW2002 - International Conference, Honolulu, Hawaii, USA, 7-11 May 2002.