

Using the Semantic Web for Web Searches

Norman Noronha and Mário J. Silva
Faculdade de Ciências
Universidade de Lisboa
{normann,mjs}@di.fc.ul.pt

Abstract

ReQuest is a semantic search system for specialized domains. It aims to offer context based searches by integrating semantic web technology, such as ontologies and resource description files. ReQuest was built to evaluate and compare the relevance of semantic searches and regular searches used in current information retrieval systems with a user survey.

Keywords:

Semantic Web, RDF, Ontologies, Web Search, User Evaluation.

1 INTRODUCTION

The World Wide Web is an enormous source of information. The time necessary to find relevant information depends on the usefulness of automated machine tools, such as search engines, and the amount of human effort on browsing different types of information. Existing search tools do not consider the semantics behind information sources, leaving this for the user to decide. These restrictions require the user to filter information, even though some of these tasks could be automated.

The Semantic Web is a proposal for a Web of machine understandable data for the purpose of easily automating user and computer tasks [BLHL01, W3C03]. The Semantic Web would extensively use metadata for this purpose. Metadata is data that describes attributes and properties about other data, such as Web pages. Metadata for a Web page can be as simple as a set of data that includes information about its author, when it was published and where is it available. Dublin Core is an example of a Web metadata standard [Ini04]. By providing more metadata, the author helps improve the quality of information and makes it more accessible. The Semantic Web aims to provide a structure that will enable machine processes to maximize the utility of data and metadata structures.

To understand the semantics of information, we need a data structure that defines the meaning of the available data. In the Semantic Web, Ontologies are used to define how to understand the data and metadata of the Web [Fen00]. Ontologies are shared vocabularies which provide a context for machine processes to understand. As vocabularies, ontolo-

gies can help users formulate their information needs. By associating a context, users can narrow their search as well as browse an ontology to search in. It should be easier for a user to express his information need by the meaning of what he wants rather than an exact word or expression in a set of keywords.

The purpose of this work was to evaluate the usefulness of an intelligent system structure that uses the Semantic Web to improve searches on the Web. It was hearted on the validation of the hypothesis that *Searches based on ontologies improve user satisfaction and reduce effort by eliminating irrelevant results.*

The aforementioned hypothesis was validated through a series of experiments comparing user behavior and results obtained with ReQuest, an ontology-based search engine prototype against traditional search engines (see Figure 1).

The ReQuest prototype was configured to operate on the Portuguese news domain, providing enhanced search capabilities for a few Portuguese Web newspapers. The experiments involved the evaluation of a set of queries that operate on the restricted subject domain of journalism and news publications.

This paper details the development and construction of the ReQuest prototype, and its role in comparing and evaluating semantic search against current search engine technology. The next section discusses the design and architecture of the ReQuest prototype. The following section presents its implementation for the news and journalism domain, and statistics and results collected from its evaluation. Finally, we outline the conclusions and point to further work suggested by this research.

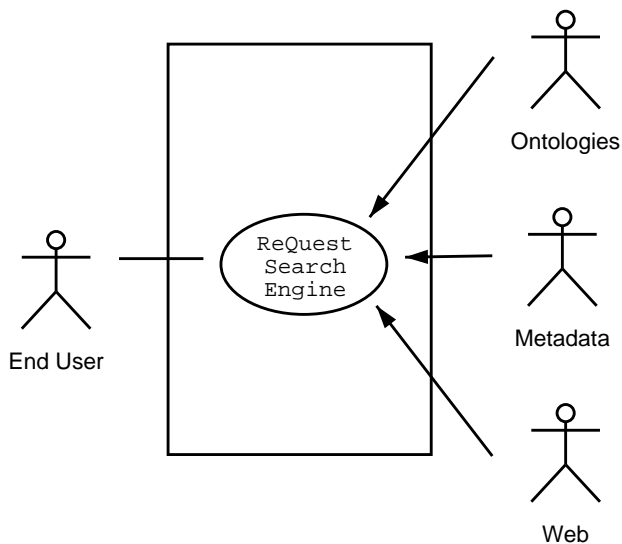


Figure 1: Simple ReQuest Architecture. The search for information could be optimized with tools that consider data semantics. This would reduce the cost of browsing as well as eliminate unnecessary results from the final analysis by the user. The ReQuest search engine works as an intermediate between the end user and the information contained in Web documents. ReQuest uses metadata and ontologies to add semantics to the data.

2 REQUEST'S ARCHITECTURE

ReQuest uses ontologies to provide context-oriented queries over RDF Documents to extract information [BLBMS01]. Ontology enabled searches increase precision of results by restricting to the domains intended by the user and help users in formulating queries by providing domain vocabularies.

The work described in this paper is detailed in Noronha's Master Dissertation [Nor04]. Figure 2 shows the main use cases of ReQuest as an information search tool that provides users with the ability to query data semantically, as long as ontologies and metadata (RDF) information sources are properly configured. The range of searches that ReQuest can perform depends on the number of ontologies and data files associated. Users and administrators are the actors that interact with ReQuest. Users perform global or domain searches and can use equivalences to help search for information. This section describes only two of the top level use cases of the ReQuest search system, Global Search and Domain Search.

The Global Search use case is invoked when the user is unsure of his exact information need. This search will, in general, produce lower quality results than the more restricted Domain Search. The Global Search use case provides the following searches:

Search in all contexts with keywords - This is the most simple to use search in ReQuest, since it only requires the user to input his query terms without identifying any

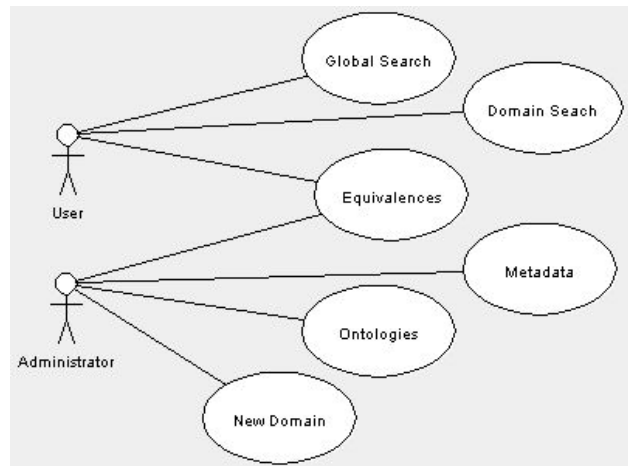


Figure 2: Use Cases for the ReQuest Search System. In ReQuest, the user has the ability to search domains for context sensitive information with Domain Search, or to invoke a Global Search on the System. The user can also use Equivalences to search between similar properties. An administrator must set up ReQuest by configuring the origin of the required input files, such as the ontology and metadata files, as well as define equivalences and create new domains within ReQuest.

context.

Display few results from multiple contexts - ReQuest replies with a few results displayed from different contexts to help the user browse the right context and find the necessary result. For example, if we wanted to search for the phone number for Dr McDonalds but wanted to avoid any fast food phone numbers, we could search for McDonalds and then select the context for Person or Doctor instead of the context of Restaurants or Fast Food Outlets.

Search in specific context with keyword - After discovering the right search context, the user has the ability to execute a domain search with the same query by following a link placed after the last result.

Domain searches are extremely useful for a user when he has a vague idea of the context where results are expected. The Domain Search use case provides the following functions.

Browse search tree with ontologies - When a user searches for information in a specific domain, he starts browsing existing ontologies and selects which context to examine.

Select ontology and choose concept to search - After selecting an ontology, ReQuest presents to the user all the concepts contained within the ontology.

Submit search term for at least one property - Each concept contains multiple properties and the user will have to enter some search term in at least one property. For

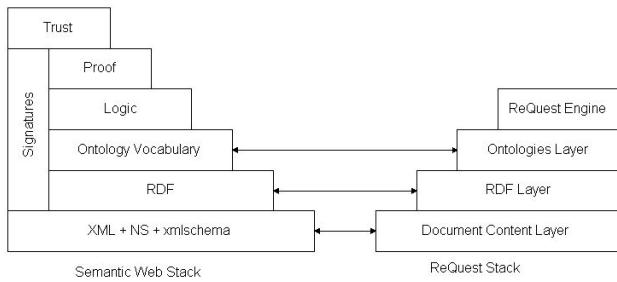


Figure 3: Mapping of Semantic Web and Request Layers. Each layer in the Request architecture can be directly mapped to a corresponding layer in the Semantic Web stack.

example, to search news articles about the city of Porto not related to the sports club, the user would have to browse the ontology tree in ReQuest for *news documents* and search within the property *subject* for local city news. This would eliminate from the search results any sports news.

2.1 Components

The main components of the ReQuest Architecture are:

database - contains ReQuestDB and holds all the data extracted from the metadata files stored in tables constructed from information processed from ontologies. ReQuestDB also contains configuration data for each domain, including equivalences.

ontology - responsible for retrieval and processing of Ontologies.

metadata - retrieves periodically metadata, processes and stores data in ReQuestDB.

the user interface - Web interface with servlets for search queries and administrative tasks.

2.2 ReQuest Architecture and the Semantic Web

Each layer in ReQuest's architecture can be directly mapped into a layer of the Semantic Web (see Figure 3). It is important to recognize that ReQuest is built directly above the ontology layer of the Semantic Web. Thus, ReQuest does not involve any construction of rules or other elements that are part of the logic, proof and trust layers of the Semantic Web.

In ReQuest, the Document Content layer references Web documents. ReQuest does not harvest the contents of Web documents into its repository. Instead, it gathers metadata files that represent the documents. The Ontology Subsystem in ReQuest is located at the same level on the Ontology Layer in the Semantic Web. In Figure 3, ReQuest appears



Figure 4: Global and Domain Search in ReQuest (top). The index page of ReQuest contains the domain search tree to browse on the left hand side and the global search query box on the right. At the bottom, a screen shot of ReQuest's domain search within the class Person. The user should search in the appropriate properties.

to be at the same level as the logic layer. However that does not mean that ReQuest performs any kind of logic inference. We only intend to show that this Semantic Web application operates above the Ontology layer.

3 USER EVALUATION

To validate the hypothesis of this work, we configured ReQuest for searching a specific domain, News. Ontologies were created or imported for the news domain, which consists of newspapers publishers, news articles in Web editions and the journalists that work at the newspaper publisher and write the news. Metadata for newspapers was created manually and, for Web edition's metadata, Rich Site Summary (RSS) documents from Linxs.pt [Lin04] were used.

Users can search for information with Global Search or Domain Search features (Figure 4). This section presents the experiments and results of testing and evaluating the prototype within the news domain.

To validate the prototype, we conducted a survey with five (5) volunteer users. They were introduced to the system through a FAQ (Frequently Asked Questions) documenting

ReQuest and *ReQuest for News*. A series of information queries were presented to each user, to measure difficulties and satisfaction quotients. The evaluation had nine (9) information queries about newspaper publishers, workers and Web articles. Each query produces at least one valid result. After execution of each query both in ReQuest and a conventional search engine, each user was asked to fill a questionnaire that evaluated the results obtained. Finally, after finishing all the queries, each user had to respond to one final questionnaire that gathered general information about the test.

3.1 Results

Detailed results for each of the nine test queries are presented below. The first three queries were intended as simple exercises in using ReQuest. They were necessary for the user to experiment different searches and features intended to provide the information from various processes. In the next three queries, the information is found in a different context than that where the search begins. Finally, the last three queries are complex scenarios to determine how users will react when it is necessary to use various features of ReQuest searches. All users chose Google as the conventional search engine to test against ReQuest. We will refer to Google by name as our conventional search engine during the tests. Information for all queries could be found on the Web by using either ReQuest or Google.

The user's responses to the queries will now be detailed:

Query 1 [Q1] Use ReQuest to find the post office address for the publisher *Publico*.

With query Q1, only one user experienced difficulties. The same user subsequently noted that his difficulty arose from automatically formulating the search in the form of keywords. All but the same user easily found relevant information in less than 2 minutes, both from ReQuest and Google. Users found the information by testing both domain and global search features in ReQuest successfully. One user mentioned that ReQuest's case sensitivity was a minor setback. Users were evenly divided between both systems about ease of use and which system reduced the task effort.

Query 2 [Q2] Use ReQuest to find five (5) workers in the *Cultura* department.

The results from user's evaluation of query Q2 show that almost all users found very relevant information with ease in ReQuest. Also, 80% referred that semantic links were helpful in this search process. Only one user managed to use Google with ease for this query and only another user was satisfied with the number of relevant results discovered. In this query, ReQuest was unanimously considered easier to use.

Query 3 [Q3] Use ReQuest to find three (3) articles about the Euro 2004.

In query Q3, most users satisfied their information need with ReQuest and Google. One user was unable to find

the information with ReQuest, while another user did not find it with Google. For this query, almost all users used global search in ReQuest. Almost all users felt that ReQuest helped reduce the effort of analyzing new results and trying new searches.

Query 4 [Q4] Use ReQuest to find the email of the editor of the *Sociedade* department.

One user did not succeed with ReQuest, but found the same information easily with Google. All other users managed to complete the query task without complications in ReQuest, even though only one of them found semantic links to be helpful. This majority had an average or difficult experience using Google to resolve this query. Overall, ReQuest was superior for 80% of users than Google for Q4.

Query 5 [Q5] Use ReQuest to find two other articles of the journalist that wrote the article entitled *mp violou código do processo penal*

Query Q5 tests user interactions with queries that include different contexts. Results are very similar to those of query Q4. One user had considerable difficulty in using ReQuest but could find all the information with Google. All but one user had extensive experience with search interfaces and were successful with both systems. The one user that was unsuccessful complained that ReQuest produced an excess amount of information to examine.

Query 6 [Q6] Use ReQuest to find the name of the editor that supervises the journalist *João Pedro Henriques*.

One user was unable to find satisfactory information. Other users found relevant results with ReQuest, while only two users found relevant results with Google. 60% of users used ReQuest's domain search and found semantic links helpful. Most users had difficulties in searching for the information requested in query Q6.

Query 7 [Q7] Use ReQuest to find the name of the editor of the section that is responsible for the article <http://jornal.publico.pt/publico/2004/01/07/Desporto/D41.html>.

Query Q7 has results from only four users (one user skipped this query). Two users found exactly what they wanted with ReQuest, but did not find the same information with Google. However, the two other users found the information with ReQuest and Google with one user showing some difficulty with ReQuest. Semantic links proved helpful for this query to three out of the four users.

Query 8 [Q8] Use Request to find how many journalists wrote articles about Iraq.

Only one user experienced difficulty with ReQuest. Another user was not satisfied with the results obtained with ReQuest. All users had a tough time searching for information with Google and none was satisfied with

the results found. Overall query Q8 was more successful with ReQuest than with Google, as it was easier to use, reduced effort and provided better results.

Query 9 [Q9] Use ReQuest to how many distinct articles were published by *Publico* about *Futebol* between the 5th of January, 2004 and the 7th of January 2004.

Finally, with query Q9 we observed that two users did not find the information with either ReQuest or Google. The three other users found the information with domain search but without the use of semantic links. All users who found the query difficult with ReQuest also found no satisfactory results with Google. Two users suggested that restricting document searches by time or date of document would be useful.

Each user completed a second questionnaire at the end of the tests. The intent of this questionnaire was to gather information about user knowledge, test experience and observations after using ReQuest. As a result, ReQuest's users had at least adequate knowledge of using Google but were not completely satisfied. Users also gave a positive response to the utility of semantic links. The following are the main observations and suggestions received from participating users after testing ReQuest:

Changes to Existing Features One user suggested that, in the domain search interface, if a property has restricted values, these should be available as the only options. Others suggested that ReQuest can be improved by a better designed interface. A couple of users asked for case insensitive search queries. Ranking search results by some algorithm such as TF-IDF would present the most useful results first. One user noted that ReQuest presents too much information and would prefer a reduced version of results to see result sets with less results.

New Features Most users asked for more search options as new features, such as searching within results to help search inside a large subset of results or searching with multiple properties in different contexts, that is, querying ReQuest with domain search with properties from different contexts instead of only properties from within the same context. This option would be simple to implement given ReQuest's architecture, but it would require a new design for the search interface.

Domain Search Domain search was preferred over global search by users that grasped how to use ReQuest's interface. Users that did not interact positively with ReQuest used global search more often. Domain search is more intuitive, easier to use and produces better results than the advanced search option in conventional search engines. However, the search interface of the existing prototype still needs much improvement.

4 CONCLUSIONS

A number of questions was included in the user evaluation questionnaires to check how the initial hypothesis holds. The questions evaluated the following aspects:

Information Need Satisfied Users whose information needs were met responded positively to questions specific to the relevance of results achieved with ReQuest and the conventional search engine. In the 9 test queries, 3 out of 5 users achieved greater success in finding relevant information with ReQuest approximately 7 times. Another user found ReQuest marginally better than Google leaving only one user more satisfied with Google's results than ReQuest's. Google's results were considered better in only one query (Q3).

Less Irrelevant Results The goal of some of the questions was to determine the amount of irrelevant results produced by each system for the user to analyze. Excess irrelevant information has to be considered when studying user effort to find information. Since Google has a much larger collection to analyze than ReQuest, only results from the first page were compared to evaluate the quantity of irrelevant results produced. 80% of users found fewer or no irrelevant results with ReQuest than Google. One user found ReQuest and Google to be comparable for one-third of all results, while producing the same quantity of irrelevant results for the remaining one-third. Comparing questions individually, Google was superior in Q3 and Q9, but equal in Q4 and Q8. ReQuest showed less relevant results in the other 5 queries. ReQuest was more precise than Google for 48.9% of all questions, while Google was more precise for 24.4%. This leads us to believe that ReQuest performed better than Google in producing fewer irrelevant results.

Reduced Effort If users responded that semantic links were helpful and domain search was useful in reducing effort, then we considered that ReQuest succeeded in reducing effort. A majority of users were successfully aided by ReQuest, approximately 7 times, while only 20% managed to resolve more than half of the queries with less effort with Google. While some users were almost evenly divided between Google and ReQuest, others did not produce a single query where Google required less effort than ReQuest. Therefore, the survey shows that ReQuest managed to reduce the effort necessary for finding the required information.

In the context of the experiments explained in the previous section and carried out in this work, **Searches based on ontologies improved user satisfaction and reduced effort by eliminating irrelevant results.** We must however stress that these results are not statistically significant. The hypothesis is valid only for the user population that tested the prototype responding to the questions answered in each test survey in the news domain. This work shows that there is interest in examining the quality of semantic searches in

a larger context. Overall, We also found that offering users the ability to select the context to search for information is a better method for expressing the information need than only through a set of keywords.

The ReQuest prototype was built using RDF and ontologies from the Semantic Web. The idea behind the Semantic Web looks reasonable because it works on top of the existing Web. The Semantic Web does not require the extinction of the World Wide Web, rather it builds on it to offer greater features. Currently, the success of building a Semantic Web of quality will lie with the tools available to produce metadata, ontologies and other relevant data objects. If a regular Web site content provider can see the benefit of using the enhanced features of the Semantic Web and create metadata, ontologies and rules with the same ease, then we will see a growth in the Semantic Web. The lack of simple tools and the discussion about document formats demonstrates that the Semantic Web is only in its early days. Currently, Web sites that are regularly updated such as news, journals and blogs, automatically produce RSS documents of their Web site content for user applications that keep users informed. Besides syndicating news, in an era of Web Television, user applications could be informed of new episodes of shows to watch or download. RSS can also be coupled with peer to peer network applications such as Bit Torrent to create a new distribution channel [Coh04]. We are now beginning to see the applications of producing metadata for Web sites. With the development of newer tools for ontologies, we will be able to use the metadata in more efficient ways. Semantic Web technology has a long way to go in becoming a reality, but as long as it continues to produce easy to use tools with practical benefits at each stage it might just reach there.

4.1 Future Work

A big advantage of semantic searches is that they can provide greater quality searches in multiple languages. Differ-

ent languages can be searched by a common ontology and properly configured equivalences. In the case of the prototype *ReQuest for News*, RSS documents from different language news sites can be searched concurrently. This would be a more efficient way to fight the separation of the Web in large blocks of data divided by languages.

REFERENCES

- [BLBMS01] BERNERS-LEE T., BRICKLEY D., MILLER E., SWICK R. R.: Frequently asked questions about RDF. <http://www.w3.org/RDF/FAQ>, July 2001.
- [BLHL01] BERNERS-LEE T., HENDLER J., LASSILA O.: The semantic web. *Scientific American* (May 2001).
- [Coh04] COHEN B.: Bit torrent: Protocol specification. <http://bitconjurer.org/BitTorrent/protocol.html>, March 2004.
- [Fen00] FENSEL D.: *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. ISBN 3-540-41602-1. Springer-Verlag, 200.
- [Ini04] INITIATIVE D. C. M.: Dublin core metadata initiative. <http://purl.org/DC>, March 2004.
- [Lin04] LINXS.PT: Linxs.pt. <http://linxs.pt>, March 2004.
- [Nor04] NORONHA N.: *ReQuest: Validating Semantic Searches*. Master's thesis, Faculdade de Ciências, Universidade de Lisboa, 2004.
- [W3C03] W3C: W3C semantic web. www.w3.org/2001/sw/, July 2003.