Knowledge Enhanced Searching on the Web

Afraz Jaffri

Dependable Systems and Software Engineering Group
School of Electronics and Computer Science
University of Southampton
a.o.jaffri@ecs.soton.ac.uk

Abstract. The move towards a semantic web has been in progress for many years and more recently there have been applications that make use of semantic web technology. One of the features that made the Web so easy to use is the ability to search web pages in a matter of seconds through the use of search engines. Now that the use of OWL and RDF as a knowledge representation format is increasing, the possibility appears to improve the quality of searching by using the semantic web to enhance the 'ordinary' Web. This paper outlines an architecture for using distributed knowledge bases to assist and improve searching on the web.

1. Introduction

The semantic web promises a new generation of World Wide Web infrastructure that will make it possible for machines to 'understand' the data on the web instead of merely presenting it. The increased adoption of RDF and OWL as knowledge representation formats are enabling the production of semantic web systems that can manage, manipulate and display data in novel ways [1, 2]. However, there are also those who believe that the semantic web is merely a dream that will never be fulfilled [3]. In order to encourage the increase of semantic web technologies there have been suggestions that a 'killer app' may be needed to convince those that are still unsure about the benefits that semantic web technologies can bring.

The search engine is an example of a potential 'killer app' that has been responsible for increased usage of the current web. However, despite the improving quality of modern search engines, statistics show that only 17% of people find exactly the information they were looking for [4]. Furthermore, a study has shown that the recall of some search engines can be as low as 18% [5]. There is therefore a need to improve the quality of search results and user experience. The semantic web provides an opportunity to achieve such a goal. The use of RDF and OWL as knowledge representation formats can provide structured content to describe a given domain or set of domains. Using this knowledge, it should be possible to add a sense of 'understanding' to a search engine when searching for results whose knowledge has been partly or fully described in a knowledge representation format.

In the past, such a proposal may not have been viable due to the lack of ontologies and RDF resources available on the web. However, at the present time there are estimated to be more than 5 million RDF or OWL documents available on the web [6].

Even if most of those documents contain knowledge about a limited set of concepts, RDF data from sources such as DBpedia (http://dbpedia.org/docs) and Wordnet (wordnet.princeton.edu) provide a suitable base from which to begin exploring the enhancement that can be made to ordinary web searches. More importantly, there will be a number of knowledge bases that will be used to store RDF instance data and OWL ontologies that have the ability of being queried using the SPARQL query language. Therefore, there are a number of components that need to fit together in order to achieve semantically enhanced querying that will be presented in Section 4.

2. Related Work

There have been many projects under the general heading of 'Semantic Search' that work towards different goals and objectives. The Swoogle [7] search engine attempts to index all semantic web documents (SWD) on the web. The query a user makes is usually in order to find an ontology that they can use that contains descriptions about their query item. This is a purely semantic web service, i.e. it deals only with SWD and not any other type of document available on the web. There are a few ontologies that have a high rank because they are imported by other ontologies and therefore are returned frequently back to the user.

The SemSearch search engine integrates ontologies and RDF data to provide a search facility for a departmental university website [2]. Queries are semi-structured and require users to input a subject keyword as well as free text. This system is also a closed world system i.e. it does not interact with the web and it cannot make use of knowledge from other repositories.

The project that shares the same aims and objectives as those stated in this paper; to improve the quality of search on the web, is the TAP project [8]. TAP is both a semantic web application by itself and also has the ability to interact with the WWW. TAP does not try to model concepts or definitions, but instead concentrates on modelling real world entities such as movies, athletes, musicians, places, people etc. The only limitation with the system is that it does not bring in knowledge contained in other repositories. This means that when a user searches for things that are not inside the TAP knowledge base, very little useful information is returned.

The latest project that is being developed by the Linking Open Data (linkeddata.org) project is DBpedia. This consists of a large knowledge base containing structured information gathered from Wikipedia (http://en.wikipedia.org). The idea behind the project is that people will expose their data and interlink it with the RDF from DBpedia. Such a large scale repository provides an ideal starting point for implementing a semantic search engine and such a system has already been produced (http://dbpedia.org/search). However, DBpedia is by no means an exhaustive reference for every concept or entity. The project is also in the initial stages and not everyone can be expected to provide data that has been linked to their own. Nonetheless the SPARQL endpoint provided by DBpedia provides access to a rich set of knowledge that is being used in our system.

A gap exists in the domain of semantic searching that has not been covered by existing systems. The system proposed in this paper aims to be an open world semantic

web application. The system will use heterogeneous knowledge bases that are accessed through the SPARQL query language which means that any repository that has a SPARQL access point can be used in the system. The system will not be limited to a particular domain or a particular web site. Through the use of external links embedded in RDF, searches made on the entire web can be semantically enhanced as will be described in Section 4.

3. Consistent Reference Service

As described in Section 2 the DBpedia project is an effort to try and capture the information from structured data provided by Wikipedia. The RDF data they have gathered has also been linked with data from other sources through the use of owl:sameAs predicates. Whilst the community developing these projects advocates such linkage between disparate data sources, this may not always be practical. You may not be sure of which URI refers to the same entity as your own URI and how many such URI's exist on the semantic web. Our proposal is to use a CRS (Consistent Reference Service) to manage the referential integrity of semantic web resources. Each site or endpoint that provides access to RDF data maintains knowledge about 'bundles' of resources that it considers to be identical. This service has been implemented in the ReSIST project [9]. Thus, authors of papers who have different URI's from DBLP, Citeseer and their own site are bundled together with one URI chosen as the canonical representation. The CRS is being adapted for use in the system proposed in this abstract by being able to identify the similarity between concepts other than 'Person'.

4. Proposed Architecture

The system architecture comprises of three main components as shown in Figure 1. A user enters their keywords (KW) in a normal Google style search box without the need for using special syntax or constructs. These keywords are then fed to the knowledge manager who passes them on to the knowledge mediator.

The knowledge mediator has access to an arbitrary number of knowledge bases and a CRS that are accessed over HTTP using SPARQL. The mediator then queries the CRS for concepts that match or are similar to the given keywords. The CRS returns a 'bundle' of resources that have been found to be the same. The knowledge Mediator then issues a DESCRIBE query on each URI to find the properties and literals for each resource. The RDF returned is then passed on to the Knowledge Manager who looks for links such as 'foaf:page', 'rdfs:seeAlso' and 'dbpedia:reference' so that a list of web pages associated with the query can be returned. The knowledge manager then displays the knowledge and web links to the user according to the type of concepts returned. For example, the distributed knowledge bases contain definitions, articles, links, publications and other assorted information; these results are returned to the user as views relating to one of the concepts. The user can then select whichever view they choose to explore the results further. Formally:
Let *Concept* be the set of concepts contained in all knowledge bases.

Let *URI* be the set of URI's that represent each concept.
Let *Bundle* be the set of bundles whose elements are URI's representing a concept.
Let *V* be the set of property values of a URI.
We have a function BundleOf that returns one bundle for a given concept:

$$BundleOf : Concept \rightarrow Bundle$$

We have a function Describe that gives the property values for a given URI:

$$Describe : URI \rightarrow \mathrm{P}(V)$$

The result of a query for a concept from the Knowledge Mediator is:

$$\forall a \in Concept, \mathrm{Re}\,sult(a) = \bigcup_{i \in BundleOf(a)} Describe(i)$$

The results that are returned to the user are both document links and knowledge in the form of RDF statements that were returned from the knowledge mediator when the original query was sent. This enables the user to understand how the results were achieved. An investigation into whether the knowledge alone is enough to satisfy a user's query is planned as part of the research.

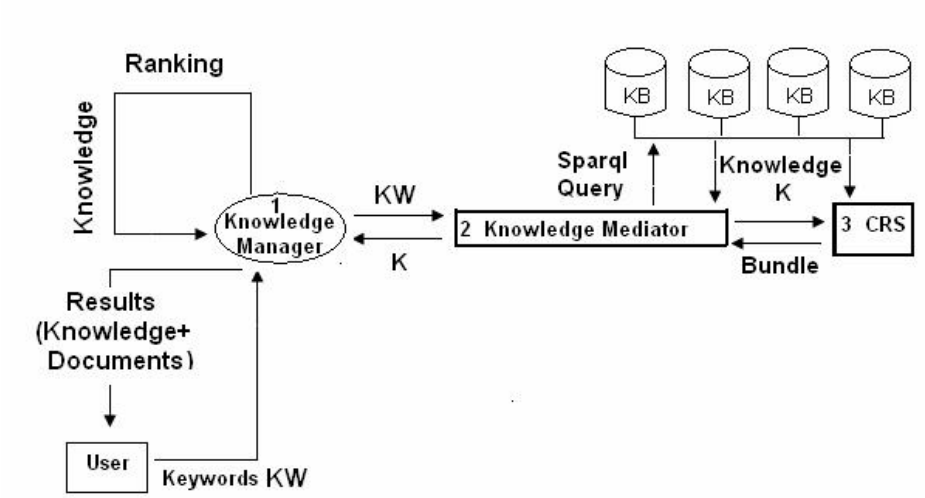

Fig. 1. The system architecture diagram shows the three main components numbered 1, (*Knowledge Manager*), 2, (*Knowledge Mediator*) and 3, (*CRS*).

5. Research Methodology and Future Work

The initial research that led to the architecture described in Section 4 being formed has been successfully completed. A proof of concept system is now being implemented that will take the form of a Google Maps mesh-up of Wikipedia(http://labs.systemone.at/wikipedia3), DBpedia, CIA Factbook

(https://www.cia.gov/cia/publications/factbook/index.html) and Geonames (www.geonames.org) data. These four sources contain an extensive set of information about all the countries of the world. The four sources each have their own SPARQL endpoint that can be queried from the web.

The interface for the system takes the form of a Google Map where the user can select any country. The system then performs queries over the RDF and returns information about the country from the different sources. The results are presented so that different types of information are separated on screen. The types of information vary from country to country and include geographical, social and political information as well as the people associated with a country and the events that have taken place in a country. Each type can be explored in more detail so that links to web sites can also be seen.

The CRS for the system identifies URI's from each knowledge base that refer to the same country. The DESCRIBE query issued to each URI then performs similarity matching on the properties to filter out duplicate entries. The Knowledge Manager then looks at each property to determine which type of information is being referred to. The results are then presented to the user under the Google Map.

Once the prototype system has been fully implemented the research will then focus on broadening out the searches for queries on any kind of entity, not just countries. The evaluation for the current system will be performed by users who will provide feedback so that any features can be included in a future version. The system will also be compared with Google and also with results obtained from each knowledge base individually to assess the improvements that can be made by distributed querying.

References

1. Shadbolt, N., Glaser, H., Harith, A., Carr, L., Chapman, S., Ciravegna, F., Dingli, A., Gibbins, N., Harris, S., & Schraefel, M. C., 2004. CS AKTiveSpace: Building a Semantic Web Application.*The Semantic Web: Research and Applications, First European Web Symposium, (ESWS 2004),* Springer Verlag, 417-432.
2. Lei, Y., Uren, V.S. & Motta, E.. (2006) SemSearch: a search engine for the semantic web. *Proceedings EKAW 2006*, pp.238-245,Podebrady, Czech Republic
3. Shirky, C., 2001. *The Semantic Web, Syllogism, and Worldview* [online], http://www.shirky.com/writings/semantic_syllogism.html [15 Feb, 2007]
4. Fallows, D., 2005 Search Engine Users, PEW Internet and American Life Project [online], http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf [16 Feb 2007]
5. Shafi, S.M., & Rather, R.A. 2005. Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology. Webology, 2 (2), [online], http://www.webology.ir/2005/v2n2/a12.html [20 Jan 2006]
6. Ding, L., 2006. Enhancing Semantic Web Data Access, PhD Thesis, University of Maryland [online], http://ebiquity.umbc.edu/paper/html/id/317/Enhancing-Semantic-Web-Data-Access
7. Finin, T,Ding, L.,Pan, R.,Joshi, A.,Kolari, P.,Java, A. & Peng, Y. Swoogle: Searching for knowledge on the Semantic Web. In AAAI 05 (intelligent systems demo), July 2005.
8. Guha, R. & McCool R., 2003. TAP: A Semantic Web Platform. Computer Networks, 42 (5), 557-577
9. Millard, I., Jaffri, A., Glaser, H. & Rodriguez, B. (2006) Using a Semantic MediaWiki to Interact with a Knowledge Based Infrustructure (Poster). *Proceedings EKAW 2006*, Podebrady, Czech Republic