The necessity for information space mapping for information retrieval on the semantic web

Gregory B. Newby
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina
USA

Abstract

The Semantic Web offers exciting possibilities for information retrieval (IR). In IR, we would like systems that go beyond simply matching words in documents and queries, and instead match based on topic, data type, relations among data, and many other qualities. The Semantic Web, through fuzzy matching of information spaces from different sources, will provide for much more specific information seeking than current Web-based search engines or other IR systems. In order to succeed, however, it is necessary to map between the differing schema, metadata standards, namespaces and so forth used by documents on the Semantic Web. This information space mapping may be accomplished by a simple match or table lookup when document sets come from similar or otherwise well-defined domains. When the match is less precise, sets of rules or algorithms may be employed to map between information spaces. When schema or metadata are inconsistent, though, we are left with a similar data environment as the modern Web, and must rely on the context of the documents themselves to determine the mapping between information spaces.

Introduction

Information retrieval (IR) is about matching human information needs to data objects. Rich structural markup of documents will enable a next generation of IR systems to better meet the information needs of people because the markup will help to disambiguate the language used for the documents (Newby, 2000). Metadata will further enhance the capabilities of IR systems by providing high quality descriptions of content.

This work suggests that the Semantic Web, by supporting metadata and enhanced relations among data from disparate sources, will facilitate information retrieval. There are several ways the Semantic Web will improve IR, and at least one will require advanced techniques for the envisioned benefits to be realized. Below, these potential improvements the Semantic Web offers IR will be examined, and proposals for addressing the difficult components will be introduced.

The framework for this analysis is the concept of information space. Information space is defined as the set of concepts and relations among held by an information system. This is as compared to cognitive space, which is the set of concepts and relations among them held by a human (see Newby, 2001 for further detail). For IR, it has been argued that a long-term goal is for IR systems to act as extensions to human memory (also known as, exosomatic memory systems; introduced in Brookes, 1980). Were such systems to exist, finding information would be the same as remembering information already known. For this long-term goal to be realized, information spaces of IR systems would need to be tightly coupled to the cognitive spaces of their individual users.

The Semantic Web will help us take steps towards the goal of IR systems' information spaces matching their human

users' cognitive spaces. These steps will be accomplished by adding further specificity to the description of data items on the Web, and providing a deeper context for their understanding, as discussed below. The particular influence of the Semantic Web on IR that will be explored here is the merging of or translating between information spaces.

Before considering the merging of information spaces, let us consider human communication. From a simplified point of view, we can consider the act of human communication as getting a message, which has a particular cognitive understanding for the sending human, to a receiver. If communication is successful, not just the utterances of the message will be sent, but also the meaning. That is, for human communication to be successful, there must be a mapping between the sender's and receiver's cognitive spaces in which the shared message exists. This approach to human communication mirrors everyday experience, in which people must lay a context and achieve mutual understanding (that is, sharing or mapping between cognitive spaces) in order to exchange messages. In a classroom environment, for example, the instructor might spend hours or even weeks to lay out the necessary background required to understand a particular message. Without that background, which created the shared cognitive space for the students and instructor, the message would have little or no meaning or, worse, be misinterpreted.

The pertinence to IR systems' information spaces is that, as for human communication, an utterance or document cannot be effectively understood or exchanged without some shared information space. This challenge is exacerbated by the Semantic Web, in which markup and metadata will allow for shorter and more precise documents or document fragments, perhaps without the larger context of full text.

Consider these document fragments:

- The annual cost is $470 per participant.
- Subject = farm animals.
- We hold these truths to be self-evident, that all men are created equal.

The Semantic Web will allow such fragments to greatly increase the utility of IR systems by adding specificity to statements of information need (that is, human queries). Queries might include:

- What were last year's expenditures for our staff retreat?
- How shall we base our pricing strategy for membership in our trade group this year?
- Find documents about genetically engineered organisms
- What are some concerns of animal rights groups?
- Early examples of misogyny in the US
- The use of an adjective when an adverb is indicated

In these examples, queries are envisioned as matching documents or document fragments based on the topic or concept of the query and document, in addition to the words themselves. This is a critical distinction between the popular Web search engines of today and Semantic Web-enabled IR systems of the future. In the future, we would like to gain a match between the information need and the actual content of a document, rather than simply the words of a query and the words of a document.

While word matching is not to be discarded — after all, it has been demonstrably functional for many years, in many IR systems — we would like to move towards exosomatic memory. Exosomatic memory clearly requires concept matching and the merger of human cognitive space with IR systems' information spaces. In the following sections, the steps that the Semantic Web will enable towards enhancements in IR are explored.

In the next section, the role of the Semantic Web for providing specificity for IR is addressed. The following three sections address IR techniques enabled by the Semantic Web: direct mapping techniques, rule-based techniques, and context-derived techniques. A concluding section lays out the research programme for improving IR via the Semantic Web, and identifies remaining challenges to be addressed.

Throughout, the emphasis will be on the matching of human information needs with documents or document fragments in text environments. The text considered is semi-structured, such as that found in human-generated documents of all types: journal articles, news, literature, etc. In such semi-structured documents, we can identify components (such as headings, tabular data, references, etc.), but there is considerable latitude in the actual wording and formatting used in them. Information filtering, sub-document identification and similar approaches to

information retrieval are possible beneficiaries of the approaches described here. There will be little attention paid to highly structured data (databases), numeric data and non-text, although we anticipate the conceptual framework used here will apply to other data types as well.

The Semantic Web: specificity for information retrieval

The Semantic Web provides specific interpretations for data items contained within documents. This will help greatly with resolving some of the ambiguity in human language. Whereas in typical IR, systems use statistical information about patterns of term occurrence in documents to determine the best matches for a query, the Semantic Web will help to narrow the sphere of candidate document matches considerably.

Consider, as an example, a human using a modern-day Web search engine to seek information about pet care. She might type a few words as a query: "dog care and feeding."

Typical information retrieval systems (which include contemporary Web search engines) proceed to find the best matching documents by looking at the term frequency of each of the query terms (likely dropping "and" as being irrelevant). From all the documents the IR system knows about, the possible candidate set would be the Boolean intersection (or union, if an "OR" query is specified) of documents containing the query terms. In practice, there are various short cuts so that not every possible document needs to be examined, but conceptually the set of possible documents to present to the user is limited by the query from ALL documents to only those that contain all of the query terms.

For a large and diverse data set as the Web, the candidate set for most queries of a few words will be very large. To rank the "best" documents matching the query, numeric term weights are assigned to the query terms, and to their roles in each of the candidate documents. Other factors, including the length of each document and global statistics (such as the average length of a document in the collection) are considered. Many further factors may be included: the authority of a Web page, the number of incoming and outgoing links to the page, the presence of query terms in headings or meta tags, and so forth. Each of these weights must be combined somehow to yield a numeric score for each candidate document. The top-ranked candidates are then presented to the information seeker.

The main thing missing from this IR scenario, which is played out millions of times each day on Web search engines, commercial IR systems and elsewhere, is understanding of the query. Based on the query above, different Web search engines yield a variety of matches in the list of highly-ranked "hits:"

- Information for mushers on the Iditarod trail
- Pet store supplies
- Taking care of your new puppy
- Care for particular breeds (Pyrenees, pug and others)
- Holistic pet care
- Dog walking services

There are also documents that are clearly not good matches, but were retrieved nevertheless:

- Care and feeding of the press
- What to do if you find a wild dog
- Orphaned baby bunnies
- Dog interaction with iguanas

(Query tested in May 2002 against Google, Lycos, Altavista, Yahoo! and Excite search engines, all of the above were from the top-10 hits). The point here is not to illustrate imperfections in Web search engines. Rather, it is to consider what the Semantic Web could do to improve results.

If our information seeker was interested particularly in seeking a document containing vet's advice on what brand of food to purchase, the Semantic Web would help. A modified query might have content such as this:

Topic: dog food
Document type: advice
Document author-type: veterinarian

Although an experienced searcher might be able to add some of this additional specificity to a query for modern search engines, this would not change fundamental type of matching. Contemporary search engines offer no facility for metadata concepts enabled by RDF in the Semantic Web. Let us consider a simple case that would yield far superior results, and then turn to the challenges of the Semantic Web for IR.

In the simplest case, the Semantic Web will let IR systems generate a candidate set of documents from all those known based on exact and unambiguous criteria. The exact mechanisms for describing documents semantically for the Web are still under development, but the intention is clear: to develop "languages for expressing information in a machine processable form" (Berners-Lee, 1998). The languages are based on schema, document type definitions (DTDs) and the like, and the resource description framework (RDF, see http://www.w3.org/RDF/), provides the ontology used for metadata communication.

The modified query, above, would eliminate all documents from the candidate set that were not authored by vets, or did not contain advice, or were not about dog food. The resulting set of candidate documents would be ranked using standard techniques above. Alternatively, the information seeker might be presented with other available sorting mechanisms such as last update, location of the author, and so forth.

The Semantic Web will also provide for segmenting of the document collection, as well as for inheritance of attributes. Thus, one algebra for the query might be to consider each possible document in an IR systems collection, as described above. Another might first generate a sub-set of all pages that are part of a document tree authored by a veterinarian or veterinary agency, and then seek for the desired topic within that set.

This example demonstrates the potential power of the Semantic Web for IR. By providing for further specification of an information need, using the ontology and data description language(s) the Semantic Web provides, a much narrower and more appropriate set of candidate documents may be generated.

So, what's the problem? Will the Semantic Web be a panacea for IR? Not necessarily. There are at least two critical challenges to be considered:

1. Human language is highly varied. Thus, we will be left with the need for sophisticated IR techniques to determine the documents that are the best match for an information need;
2. There will be a large number of different implementations of schema/DTDs, ontologies, and vocabularies used on the Semantic Web to described documents. Thus, determining how to generate the list of candidates from a specific query, such as that given above, is itself an IR problem requiring mapping between information spaces.

The rest of this work focuses on the second point. In order to reap the benefit the Semantic Web offers IR, it is necessary to map between the information spaces of different sets of documents (held by different information systems — Web servers, database servers, etc.) in order to identify which documents actually meet a set of information need criteria.

If one document author utilizes "Document author-type" while another utilizes "Personal author," can our IR system identify that they each refer to a criterion we are interested in for our query (and, in fact, do they)? If one author inserts "dog food" in the "Topic" meta tag, and another author inserts "canine nutrition," how will an IR system determine whether to include these documents in the candidate set?

In the sections below, three approaches to addressing the issue of ambiguity in the implementation of Semantic Web metadata description and structure for IR are discussed. With direct mapping, we use a dictionary or look-up table to determine which elements match exactly some other elements. With rule-based mapping, we follow a set of procedures, possibly involving data transformation, to identify comparable data. Finally, when we have no such dictionaries or rules, we must rely on the context of the document itself to determine matching or transformation processes.

Direct mapping of information spaces

The full benefit of the Semantic Web for IR will be available when the data organization and metadata standards enable direct mapping between information spaces for different data sources. For example, if datasets containing

bibliographic records employ an "author" field to refer to the author's name for US-based data, but "contributor" for UK-based data, it is a simple matter to perform a query that matches both.

Such a query might be expressed or interpreted as:

> *If (source.location = '.uk') query(contributor="joyce, james")*
> *else query(author="joyce, james")*

On the Semantic Web, it is reasonable to anticipate considerable efforts will be made to enable such mapping. Indeed, this scenario provides the closest match, in an IR scenario, to the goals expressed by the Semantic Web working groups and publications. (We note that, for this example, the Dublin Core elements, http://purl.org/metadata/dublin_core_elements, nicely solve the problem by specifying a common label, "Creator." Yet we envision many situations when similar data will not use the same metadata elements.)

Metadata schema registries will enable people to publish their data using their own schema (or one derived from another) and to specify how it relates to the data organization of other schema. This is important, and will be quite beneficial for IR. See Heery & Wagner (2002) for a discussion of metadata schema registries. For any pair of schema, a lookup table can provide the equivalent data type, tag, entity class, etc. to map between the two schema. Within a registry, then, any possible pair that map onto each other (even if the match applies to only part of the schema) may be quickly looked up by the IR system.

The problem remains, though, for schema that are not specified in registries or, worse, that use comparable tags but for different data types or items. The following sections offer alternative approaches for when a simple direct mapping between information spaces is unavailable.

Rule-based mapping of information spaces

On the modern Web, and anticipated for the Semantic Web, are many documents which are comparable to other documents in structure and content yet have important differences. For this section, consider the example of a directory of addresses, membership, or customers. One Web site might list personnel in an academic department along with telephone numbers and research areas. Another Web site might list some of its premier customers along with the products they purchased and their area of business. Yet another might list accounts receivable.

In each of these three datasets, a record might exist of the form:

```
John Young || 222-3456 || information retrieval systems and services
```

On the modern Web, each data entry might occur in an HTML table and for IR purposes we would be hard-pressed to determine whether "John Young" were a person, a business, an employee or even a product. Similarly, 222-3456 could be anything from a telephone number to an invoice number.

The Semantic Web (as implemented through RDF, XML and other markup schemes more sophisticated than HTML) offers the ability to be specific. For the first column of entries, a label "Name" or "Personal Name" or "Faculty Name" could make the data unambiguous and enable the sort of IR process envisioned in the previous section. Labels such as "Telephone" for the second and "Research interests" for the third would help greatly.

For our IR systems on the Semantic Web, inheritance will play a critical role. Imagine the first column is labeled (or each datum is labeled), "Name," e.g., "Name=John Young." Alone, this is nearly as bad as having no descriptive label at all! Is it the name of a company, or a customer, or departmental personnel, or something else? If inheritance is available, however, other clues in the schema might help. Inheritance would be helpful for direct mapping (above) and context-derived mapping (below), although those sections describe techniques that would work in the absence of inheritance.

Abbreviated RDF syntax for a document containing the sample data above might follow this general form:

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.org/dc/elements/1.1/"

     xmlns:ex="http://yoyodynecorp.com/elements/people">
```

```
    <rdf:Description rdf:about="http://yoyodynecorp.com/">
    <dc:Creator rdf:resource="http://yoyodynecorp.com/Personnel"/>
    <ex:employee rdf:parseType="Resource">

    <ex:fullName>John Young</ex:fullName>
    <ex:phone>222-3456</ex:phone>
    <ex:interestArea>
       <rdf:Bag>
        <rdf:li rdf:resource="http://yoyodynecorp/areas/ir" />

        <rdf:li rdf:resource="http://yoyodynecorp.com/areas/irserv"/>
       </rdf:Bag>
    </ex:interestArea>
  </rdf:Description>
</rdf:RDF>
```

Does this help? In itself, it helps: we now see that 222-3456 is a phone number, and might glean other revelations. But combined with the data referenced by the document (in this case, the "ex" XML namespace linking to the company Web site), we envision that significantly more data about the document contents will be available through inheritance.

For IR, we would require a list of rules — not a simple mapping between elements — to determine which documents should be candidates for an information need. To harvest a competitor's customer list, for example, a query would specify names of individuals or organizations that list the competitor as a supplier. To build the list of candidate documents:

> *Select a document for candidacy if*
>   *It contains a list of products*
>   *AND The list includes our company*

From our example in the previous section, we could use a similar strategy to identify that a document contains names (of some sort), then use keyword matching on the "interestArea" tags' content to identify faculty with interests in a particular area such as the semantic Web:

> *Select a document for candidacy if*
>   *It contains a list of interest areas*
>   *AND terms "semantic" and "web" occur in the interest area*
>   *AND interest areas are associated with a name*

For each document under consideration, our IR system would need to examine the document plus documents from which its properties are drawn to determine whether it should be a candidate. For a well-known domain, with an available set of metadata standards and schema, pattern matching rules will allow this to happen. For less well-known domains, the information seeker might need to author his own rules. In both cases, the Semantic Web enables this type of rule-based approach, while the modern Web does not.

Context-derived mapping of information spaces

As stated above, the challenge of IR is largely due to the ambiguity of human language. On the Semantic Web, the ability of document authors to choose or create a data organization scheme they believe appropriate creates the same sort of ambiguity. One author might use "Name" while another employs "PersonalName" and a third utilizes "fullName." If the author draws from existing schema, or registers her schema centrally, we may be able to map to other schema or form rules for understanding how her data are related to other data.

In this section, we consider IR approaches that might function in the absence of such additional sources for understanding the content in a document or document set. These approaches should be particular useful during the coming years as the Semantic Web's many standards continue to emerge.

Our general goal, in this situation, is to learn about the structure and content types of documents from the documents themselves. This is not unlike the need in IR to use statistics about term frequency within and across documents to understand relations among terms and documents. In fact, this is the basis of the approaches we suggest: to derive understanding of the information space and its relation to other information spaces through the context of the documents themselves.

Given our personnel directory sample above, the single record was ambiguous. Add a few more records, however, and things become clearer:

```
John Young   || 222-3456 || information retrieval systems and services
Harold Ramus || 222-9077 || HCI
Sue Zimmer   || 232-1002 || management of information agencies
```

While the numbers remain ambiguous, the first column is starting to look very much like human names, rather than companies or accounts. This human impression would be supported by a system with wider contextual data — for example, a phone book that lists the same personal names. The third column might closely match faculty interests in another academic department for which we have more complete metadata (and perhaps with clear labeling of phone numbers). While the structure of such a document might be helpful in eliminating possible interpretations of the content (i.e., this doesn't appear to be a recipe dataset), structure alone doesn't tell us whether the first column deals with individuals, organizations, or some other entity.

Rather, the context — especially with the benefit of other documents from other sources — provides a sound notion of what the data are about. This is the "fuzzy matching" that Tim Berners-Lee suggested will be necessary for traversing the Semantic Web.

IR techniques for fuzzy matching typically involve less reliance on individual query terms, instead they expand the query to incorporate synonyms and, for XML-formatted data, they could look at the general document structure (i.e., tables versus lists; headings versus paragraphs). To illustrate, we will take the information space (IS) retrieval method used by the author (Newby, 1998).

The IS retrieval technique is closely related to latent semantic indexing (LSI, see Deerwester, *et al.*, 1990). The premise is that terms that tend to co-occur similarly with other terms are related; terms with less similar co-occurrence patterns are less strongly related. For example, the terms pig, swine, boar and hog might be anticipated to co-occur with similar patterns across a dataset dealing with barnyard animals.

As described extensively in Newby (2001), a term-by-term co-occurrence matrix is built for an entire dataset. This matrix is sparse, because the majority of terms do not occur together in any document. From the co-occurrence matrix, which contains raw co-occurrence scores, we build a correlation matrix. In the correlation matrix, the pattern of co-occurrence scores for each pair of terms is identified and normalized (the matrix is therefore dense, although many values are close to zero). The eigensystem for the correlation matrix is determined, which produces approximate (fuzzy) relations among terms when we only keep a relatively small number of the resulting eigenvectors. A further step for IS, not found in LSI, is to re-insert the terms based on their original co-occurrence scores scaled by the eigenvalues. The steps described here are also known as Principal Components Analysis (PCA).

The benefit of LSI and IS over other many approaches to IR is that all terms (and by extension, all documents containing those terms) have a measured relation to all other terms in the dataset. This is in contrast to most other approaches, which instead treat terms are unrelated to all other terms. For general purpose Web retrieval, LSI and IS can be too costly in computational resources and storage for the resulting large coordinate systems — and, furthermore, might give the casual information seeker a surprise if top-ranked documents do not contain all query terms.

For the purpose described here, though, LSI and IS offer considerable promise. For a given document of unclear structure and with an unknown schema, this fuzzy matching technique could find similar documents. Similarity, in this case, would be based on the set of documents that went into the IS computation, optimized and trained to make distinctions useful to a particular information seeker or in a particular information domain.

The author of this work is investigating application of the IS technique to general-purpose document filtering and retrieval on the Web. The prospect is to have several prototype documents, and then retrieve newly found documents that are similar in topic to the prototype. With the IS technique, the formation of Boolean candidate sets is bypassed in favor of a more exhaustive fuzzy match.

On the Semantic Web, context-derived information space matching will be a necessity for success until the day when all schema are well-known and all documents are standards-compliant. Even when that day arrives (if it ever does), there will be ambiguity in the data contained in these documents. As Berners-Lee (1998) stated, "toleration of

inconsistency can only be done by fuzzy systems."

In searching literature related to this required fuzzy matching for the Semantic Web, scarce few solutions to this challenge were found, and none which seemed to be a good solution for IR. LSI, IS and related techniques offer reasonable promise for mapping between information spaces based on the context of those spaces.

Conclusion

We desire IR systems that go beyond plain word matches. We want to match on the structure, data type and organization of documents as well. The Semantic Web offers these possibilities, and more. When fully realized, the Semantic Web will make many IR tasks similar to database tasks: the equivalent of table lookups for well-defined and clearly typed data.

While the Semantic Web is under development, however, IR will require additional techniques in order to match information needs to documents or sub-documents. Rule-based techniques will enable mapping between the information spaces of different documents or document sets based on knowledge about those documents or inherited from other documents. Such rules may be codified by IR system designers for common information needs, or they might be generated algorithmically. In some cases, rules may be specified by the information seeker himself.

Context-derived techniques for mapping between information spaces are a likely need for the Web of today and the Semantic Web of the near future. When the data themselves are ambiguous, or the schema or metadata are ambiguous or poorly typed, or when a query is expansive (such as a set of sample documents to be matched against) rather than narrow (a few terms), context-derived techniques will help.

As rephrased here, the whole point of the Semantic Web is to facilitate human communication as augmented by networked machines. In order for the contents of information spaces held by these machines (Web servers, database servers, etc.) to be usefully matched to human information needs, IR systems are needed to easily map between the information spaces. This mapping, on the Semantic Web, will consist of identifying elements, data types, topics and so forth from one set of documents that have desired features in common with another set.

The Semantic Web offers unprecedented power for information retrieval, yet at a cost. In order to realize the promise of added structure and understanding of document content, IR systems will be required to map between disparate information spaces. In the modern Web retrieval environment of simple queries and billions of documents, casual information seekers are seldom aware of their inability to express complex queries specifying, for example, particular types of data, particular relations among data elements, and particular relations to other documents. The Semantic Web will enable this complexity, and more.

Without techniques for mapping between information spaces, we are left with poor alternatives. Either our IR systems will only consider those documents for which our understanding is complete (thus eliminating many possibly useful documents), or our systems will forego the benefits of the Semantic Web and rely instead on simple word matching. Neither is attractive, so information space mapping techniques would appear to be a necessary area for inquiry in order to facilitate IR on the Semantic Web.

References

- Berners-Lee, T. (1998). "What the Semantic Web can represent." Cambridge, MA: W3C. Available at: http://www.w3.org/DesignIssues/RDFnot.html. [Site visited 9th July 2002]
- Berners-Lee, T. (1999). "Semantic Web road map." Cambridge, MA: W3C. Available at: http://www.w3.org/DesignIssues/Semantic.html. [Site visited 9th July 2002]
- Brookes, B.C. (1980). "The foundations of information science. Part III. Quantitative aspects: objective maps and subjective landscapes." *Journal of Information Science* **2**, 269-275.
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). "Indexing by latent semantic analysis." *Journal of the American Society for Information Science*. **41**(6): 391-407
- Heery, R. & Wagner, H.. (2002). "A metadata registry for the Semantic Web." *DLIB Magazine* **8**(5). Available at: http://www.dlib.org/dlib/may02/wagner/05wagner.html [Site visited 9th July 2002]
- Newby, G.B. (1998). Context-based statistical sub-spaces. *Text REtrieval Conference (TREC-6) Proceedings*, (pp. 735-746.) Gaithersburg, MD: National Institute of Science and Technology. November 19-21, 1997.

- Newby, G.B. (2000). "Will XML and information retrieval make society transparent?. XML 2000 Conference Proceedings (on CD-ROM). Alexandria, Virginia: Graphic Communications Association. Washington, DC: December 5-8.
- Newby, G.B. (2001). "Cognitive space and information space." *Journal of the American Society for Information Science and Technology* **52**(12), 1026-1048.

How to cite this paper:

Check for citations, using Google Scholar

Contents   5 1 2 2
         Web Counter   **Home**