12-31-2002

BRIDGING THE GAP: FROM TRADITIONAL INFORMATION RETRIEVAL TO THE SEMANTIC WEB

Morgan Benton
New Jersey Institute of Technology

Eunhee Kim
New Jersey Institute of Technology

Benjamin Ngugi
New Jersey Institute of Technology

BRIDGING THE GAP: FROM TRADITIONAL INFORMATION RETRIEVAL TO THE SEMANTIC WEB

Morgan Benton, Eunhee Kim, and Benjamin K. Ngugi
College of Computing Sciences
New Jersey Institute of Technology
mcbenton@acm.org maryeun@hotmail.com bkngugi@yahoo.com

Abstract

A chief difference between the current incarnation of the World Wide Web and the new vision of the Semantic Web is the nature of information search. The Semantic Web vision reveals a radical departure from the traditional theories of Information Retrieval (IR) upon which current search engine technology is built. Semantic Web researchers are very articulate about how the pillars of the Semantic Web—semantically aware, intelligent agents, ontologies, and markup languages—will revolutionize the way that we interact with information on the web. They are less articulate about how we will get there from here. While it's true that the traditional assumptions of IR—small, static, homogeneous, centrally located, monolingual document collections—don't hold for the Web, still it is important to note the success of search engines built on IR theory. This paper calls attention to the gap between traditional IR and the more visionary Semantic Web research. We describe a preliminary roadmap bridging the two areas focusing on the concrete contributions and also calling attention to the weak points of both fields.

Keywords: Information retrieval, semantic web, ontology, markup languages, intelligent agents

Introduction

Tim Berners-Lee, the inventor of the World Wide Web, and his colleagues recently announced their vision for the future of the Web in *Scientific American* (Berners-Lee et al., 2001). They called it the "Semantic Web" because in the new Web, intelligent agent software would not only understand the natural language of web users, but also autonomously navigate that Web performing various tasks for those users based on that understanding. The articulation of this vision has generated a great deal of energy and excitement from researchers and has greatly influenced the research paths of a number of academics. However, as compelling as the vision may be, there is still a sense among some that we are flying blindly with respect to some Semantic Web research (Tuttle and McGuiness, 2001).

The problems facing the Semantic Web community are considerable. In the summer of 2001, many of the big names in Semantic Web research gathered at Stanford University for a three-day working symposium. The final reports from two of the three main tracks at the conference indicate that there is a major lack of concrete work in this field, and even worse, a lack of a concrete path for further study. In the track on interoperability, while some progress was made in terms of developing a framework for creating Semantic Web protocols, it was concluded that "crucial problems…were either missing or not dealt with" (Kashyap, 2001), indicating that researchers still haven't had time to explore fully this area and even come up with all of the important research questions. The ontology track (Tuttle and McGuiness, 2001) seemed to make even less headway and in the end gave a frustrated report with the following conclusions:

1. Researchers lack constructive outlets for their energy
2. There is no consensus on developing a process for building a common framework
3. The relative importance of developing foundational "units of meaning" is unclear
4. There is no evidence to support that Semantic Web tools are useful or practical hence calling into question whether experimentation should continue or not

5. The current standards development process is likely inadequate for ontology research
6. A Semantic Web glossary might be a useful tool (if someone were to create one)
7. "Ontology ontologies" might help in answering questions of ontology re-use
8. It should be determined who are the major stakeholders in this process

A startling realization is that even though this group of scholars recognized the lack of concrete outlets for their energy, they were not then successful at coming up with one. There seems to be a great fear of being the first to stick one's neck out and then having it chopped off. Despite this, the general conclusion of the group is that we must "start doing it" even though it was clear that no one really knew what "it" was. We feel that the field of Information Retrieval (IR) may offer some concrete avenues of research to these Semantic Web researchers.

Search engines on the web rely largely on algorithms derived from studies in the field of text-based information retrieval (IR). The pragmatic, incremental progress made in the IR field stands in antithesis to the radical leaps being attempted towards the Semantic Web. In some important ways the assumptions and evaluation methods of IR are out of date (Viles, 1996). In other ways, however, IR research continues to make solid and important contributions to our use of the Web. Semantic Web and IR researchers stand to gain a lot by being in closer conversation than our reading of the literature indicates that they are.

The goal of this paper is to integrate the research efforts of IR and Semantic Web researchers. The last section of this paper will highlight several applications that we feel represent the power of what could come from this integration. In order to understand the importance of these newer contributions we will first describe the problem that is common both to IR and the Semantic Web, then we will describe the pillars of the Semantic Web, the core ideas of traditional IR followed by an analysis of its problems. Having this background, the last section of the paper will explore the newer applications.

The Common Problem

Both IR and Semantic Web researchers are attacking similar problems, but they are coming from different directions. IR was developed before anyone had conceived of the Internet, indeed even before computer networks were common. With the invention of the Internet, the theories and tools of IR were pressed into service with great success, emerging as the core algorithms of Web search engines. The Semantic Web on the other hand is a vision of how the future might be, a reaction to the problems that exist with the Web in its current state. In this sense IR researchers are steadily moving incrementally forward from the past whereas Semantic Web researchers are working back from a revolutionary vision of what will be in the future. The common goal of both fields with respect to the Internet is enabling users of the Web to find and make use of relevant and appropriate online information in a timely fashion.

IR researchers began with very low-level structures and continue to work up, whereas Semantic Web researchers are beginning with a very high-level blueprint and are trying to work down to fill in the details. IR researchers began with the most basic structures—the bit patterns representing words in computer memory—and did bit by bit comparisons to determine if certain key words chosen by the searcher existed within a given document or document set. Their algorithms grew in power and complexity from there. Semantic Web researchers on the other hand are beginning with complex, high-level goals as expressed by humans such as "I need to make my travel arrangements for the sales conference next month," and sketching out the systems and structures that would need to be built to support the fulfillment of such goals automatically. Between the bottom-up approach of IR and the top-down approach of the Semantic Web lies a gap of what we have yet to achieve. This paper focuses on bridging that gap. First we will describe the pillars of the Semantic Web.

The Pillars of the Semantic Web

The Semantic Web seeks to address the problems of information search on the Web in several ways: by "understanding" the semantics of terms in context, by incorporating deductive and inferential rules of reasoning into relevancy determinations, and by gathering enough information about the users and their information needs to determine what is truly "relevant" in each case. Three technological pillars underlie the current vision for the Semantic Web—ontology for specifying semantics, markup languages for achieving smooth interoperability, and agents for reasoning and action (Berners-Lee et al., 2001). We will briefly describe each pillar.

Ontology

For the purposes of the Semantic Web, *ontology* refers to "a set of logical axioms designed to account for the intended meaning of a vocabulary" (Guarino 1998, p4). An ontology specifies the meanings of terms in a given vocabulary by explicitly defining the relationships between the terms. The vocabulary to which an ontology applies may be general or specific. Ontologies will naturally overlap, and two or more ontologies can be intentionally joined by relating the terms of each to a more general set of terms which applies in each vocabulary domain (Heflin 2001). Relating ontologies in this way is useful for unearthing disjoint interpretations of a given term. Maedche and Staab (2001) describe how ontologies will allow computers to understand human semantics:

> *The conceptual structures that define an underlying ontology provide the key to machine-processable data on the Semantic Web. Ontologies serve as metadata schemas, providing a controlled vocabulary of concepts, each with explicitly defined and machine-processable semantics. By defining shared and common domain theories, ontologies help people and machines communicate concisely—supporting semantics exchange, not just syntax. (p72)*

Later we will discuss some practical issues related to ontology development. Before that, however, we will describe markup languages, the vehicle for expressing ontologies in machine-readable form.

Markup Languages

The second pillar of the Semantic Web is markup language. A markup language is a tool for adding information to documents. Historically, markup has been used to denote the formatting conventions that should be applied to a text document, but with the advent of the Semantic Web, markup will be asked to provide more rich and varied types of information (Coombs et al., 1987). Ontological markup will signify the semantic interpretations that should be applied to terms within a given document. Decker et al. (2000) stipulate that to be truly useful a markup language must meet three requirements. A markup language must have:

1. Universal expressive power
2. Syntactic interoperability
3. Semantic interoperability

EXtensible Markup Language (XML) meets the first two requirements, but it is not quite adequate to capture the semantics of terms. Using XML as a base, a number of new markup languages have been developed to meet this third requirement, notably RDF, RDFS, SHOE, and DAML+OIL. Once a suitable markup language for denoting the semantics of text or other objects in a document has been developed we will be closer to a truly Semantic Web, but there is still one element that is missing.

Intelligent Agents

The third pillar of the Semantic Web will be intelligent agent software. Ontologies and markup languages will make the semantics of documents available to machines. Software agents will make use of that semantic content, actually interpreting the content of documents to perform tasks for users. For the purposes of the Semantic Web, a good agent should be (Hendler 1999):

- Communicative—able to carry on a dialogue with a user or with other agents
- Capable—able to take action, or to effect change in the world
- Autonomous—able to act without the user in control all the time
- Adaptive—able to learn from its experiences

Agents will provide the interface to the Semantic Web. Users will interact directly with agents, which will in turn sift through, sort, and even transform information on the web, and then act upon that information, or present it to the user in the most useful form. In the short run, agents will be limited to very specific areas of action or influence, and will only gradually become more able to carry out general-purpose tasks as the semantic infrastructure of the web grows. It is certain that agents will employ algorithms developed in the field of IR.

Traditional Assumptions of IR

Viles (1996) points out that:

> *...the tacit assumption in most of [the IR research] is that the collection of interest is:*
> 1. *Static (no insertions or deletions); and*
> 2. *Centrally Located (p2)*

To that we would also add three more observations: first, that traditional IR research also deals with largely homogeneous collections of documents, second that the size of the collections dealt with by IR researchers has been relatively small, and third that collections have been monolingual. Obviously none of these characteristics apply to the Web. The Web is dynamic, distributed, heterogeneous, huge, and multilingual. Yet even in the smaller, more tightly defined problem space of traditional IR, researchers have had their hands full. Since in the context of the Web, both IR and Semantic Web researchers are tackling a similar problem, if indeed the Semantic Web is the direction we're headed, examining the foci of IR research should offer insight into how to get where we're going. The following description is distilled largely from Korfhage's (1997) text on the field of information retrieval.

First it is important to understand that IR depends almost entirely on pure lexical analysis to try to identify interesting or useful information within a larger collection. Such analysis focuses on comparing the characters that make up words (and more recently words that make up phrases), in order to determine the relevancy of a given document. Lexical analysis offers little or no ability to distinguish between semantically disjoint ideas that happen to share a common lexical expression. For example, a common frustration of Web searchers is that entering a search term such as "mouse" returns a wide variety of documents, which may or may not deal with "mouse" in the way that the searcher intended—the engine may return many documents about computer hardware when in fact the user was interested in Mickey Mouse. Semantic Web researchers envision addressing this issue by creating semantically aware systems that recognize the differences between the various kinds of "mouse."

The second major characteristic of IR systems is their dependence on Boolean logic and vector-based search algorithms. Document relevancy is determined by the presence or absence of key terms specified by the information seeker. While a number of algorithms designed to systematically include or exclude certain terms—such as "stop" words, common words like "a, and, the" which occur so frequently in documents as to provide little discriminatory power—or to change the relative weight placed on each term are able to improve the performance of IR systems, the gains are not always worth the effort that is required to achieve them. An example of the attempt to weight the importance of terms in a search is known as spatial or proximity-based search. In this method, a document A is deemed more relevant than document B if, provided they both contain the target query terms, the query terms are in closer proximity to one another in document A than they are in document B. The Semantic Web would attempt to augment Boolean logic with heuristic, rule-based algorithms such as found in expert systems.

Traditional IR effectiveness measures form the third, and perhaps most important focus point of IR research. *Precision* refers to the proportion of documents retrieved by a query that are relevant to that query. *Recall* refers to the proportion of relevant documents that exist in the entire document collection that are retrieved by a given query. Most IR effectiveness measures represent some variation or combination of these two measures, but problems soon surface. For one, trying to increase one of these measures usually decreases the other; a system with perfect recall is usually very imprecise because recalling every relevant document within a set means that a high number of irrelevant documents must also be retrieved. Another core problem lies in the fact that ultimately only the individual information seeker can determine relevancy. As said before, "relevancy" to an IR system simply refers to the degree to which query terms are present or absent from a document. The Semantic Web would combine knowledge of the user with semantic awareness of the context of information search to build a notion of "relevancy" that is closer to the user's mental model.

These three elements—lexical analysis, Boolean logic, and precision/recall effectiveness measures—paint a very simplified picture of the field of IR, but it will set up the introduction to the current problems with IR that we will discuss in the next section. In each of the above cases we've briefly mentioned the way that Semantic Web researchers would address each issue, however, it must be stressed that they are still a long way from actually being able to do so. We maintain the position that more could be done to bridge the gap between these two approaches to information search.

Current Problems with Information Retrieval

Our reading of the Semantic Web literature calls attention to certain shortcomings in the IR research as applied to the Web. Although we may have alluded to some of these problems above, we will describe them more fully here. The problems we will

describe are the incongruence between the Web and typical IR document sets, the problems that users have forming appropriate queries in search engines, the problems with purely lexical analysis of text, and problems with the appropriateness of traditional IR measures of effectiveness.

Browsing the titles of the papers and tracks presented at the Text REtrieval Conference (TREC http://trec.nist.gov)—one of the major gatherings of IR researchers that has taken place annually since the early 1990's—one might conclude that the IR community has been relatively slow in re-orienting to Web-related issues. For decades IR research has been based on small, static, centralized, homogenous, and monolingual document collections (Viles 1996), whereas the Web, which emerged less than ten years ago, is huge, dynamic, distributed, heterogeneous, and multilingual. It could not, nor should not, have been expected that the entire course of IR research turn on a dime when the Web emerged in the early 1990's. However, a "very large" collection of documents in 1997 was considered to be 20 gigabytes (Hawking 1997), which has been dwarfed by the size of the Web, which in 2001 was estimated to be several terabytes (Sullivan 2001). Distributed collections (such as the Web) didn't appear at TREC until 1999, even though networked systems have been around since the 1970's. As its name indicates, the Text REtrieval Conference focuses on text documents, whereas the Web contains many other information sources including graphics, video, databases, and the more recently emerging Web services. Perhaps this slowness is what has prevented Semantic Web researchers from grabbing on to findings from IR.

On the other hand, that the latest TREC conference (2001) included a track on multilingual information retrieval is very promising. In fact there appeared to be a great deal of energy around this new research direction with ten groups submitting their findings (Gey and Oard, 2001). Indeed, this area seems to be benefiting from the excitement around the Web. In fact, XML-based markup languages as proposed by Semantic Web researchers may be one of the key tools that are used to aid in cross-lingual information retrieval.

Developing good interfaces for IR systems—i.e. search engines—is a second problem. On one hand, users' lack of expertise on a given topic—which is very likely why they search the Web in the first place—may mean they lack the specific vocabulary necessary to get good search results. On the other hand, the apparent simplicity and hidden complexity of many search engines may prevent users from developing more sophisticated search strategies. In other words the first problem is a matter of not knowing ***what*** to ask, and the second of not knowing ***how*** to ask it. To illustrate, it seems highly unlikely that a person with car trouble, but zero knowledge of auto mechanics would go to a search engine and enter a search phrase like "common carburetor problems," nor be able to specify that the engine to exclude the term "alternator." Semantic Web user agents propose to solve these problems by interpreting and understanding the user's real information need, getting them to express it in terms they understand and gradually moving towards the heart of the problem through a dialogue. A rudimentary system of this sort already exists: MELISA accepts query terms and checks them against the MeSH ontology (really not an ontology but more like a keyword hierarchy) and returns a list of more specific or related terms that the user might wish to use to make the query more precise (Abasolo and Gomez 2000). This is a promising area in which IR and Semantic Web might collaborate.[1]

Lexical analysis, as it is now used in IR, is the third problem we'd like to address. Currently, very common words—known as *stop* words—such as "a, and, the" are excluded from most Web searches because they are so common within documents as to be unhelpful for information retrieval. However, there is a strong argument that understanding the context and semantics of the stop words might restore some of their discriminatory power. The classic example "To be, or not to be" is a phrase composed entirely of stop words, all of which would be stripped from many search engines before the query is executed. Phrase-level indexing and context analysis of terms are also practices that are still at the beginning stages of development. Since it would understand the semantic importance of each word, the Semantic Web would not exclude any of the stop words and would thereby increase the effectiveness of information retrieval. Such semantic understanding would be based on ontologies; as we've seen ontology research is in a somewhat frustrated state at present. Actually, IR researchers are making some interesting findings that would give the ontology group some direction. We will discuss that in the next section.

The last problem with IR that we would like to discuss is the emphasis on and the usefulness of the traditional IR effectiveness measures: precision and recall. The problems are relatively simple. Precision is a measure of the proportion of retrieved documents (websites) that are actually relevant to the user's information search. Since a user would have to review and evaluate every website that was returned by a search (the number of which sometimes goes into the millions!!) this measure is most often

[1]It should be noted that researchers in the direct manipulation school have a number of objections to the "intelligent" agent interfaces proposed by Semantic Web researchers. An interesting debate on this topic between Ben Shneiderman and Pattie Maes (1997) brings out many interesting issues. At any rate, we don't want to portray the idea that agent-supported interfaces are the only solution that is possible, or that is currently being pursued.

not practical to calculate. Not only that, but it is of questionable usefulness, since if the user satisfies his or her information search within the first few pages returned, it is immaterial what proportion of entire set of those returned is relevant. Recall, on the other hand, is a measure of the proportion of the relevant documents in the entire collection that is returned by any given query. This measure is even more difficult (if not impossible) to calculate because it requires knowing how many relevant documents (i.e. how many relevant websites out of billions) exist in the entire document set. Again, practicality of calculation aside, such a figure is probably not of much use to the average web user. There are other issues, such as that a user might already be familiar with most of the relevant documents already—domain expertise—and therefore only be interested in new, rare, or novel documents that might not be returned by a normal query. There are a number of user-oriented effectiveness measures that have been developed by IR researchers. It is our feeling that the user-oriented nature of the Semantic Web will bring those measures into greater prominence.

Already in the preceding discussion we have made reference to some of the ways that the Semantic Web and IR can contribute to one another in the effort towards making good use of the available information on the Web. In the last sections of this paper we'll be more explicit about how these two research fields can do that.

Integrating the Semantic Web and IR

The common problem addressed by both IR and the Semantic Web is the need for Web users to sift through billions of web pages to find the information that is most useful to them regarding any given topic of interest. We contrasted the approaches represented by these two fields by saying that IR is a bottom-up, and Semantic Web a top-down approach. In this section of the paper we will illustrate how these fields complement each other and can be integrated. The particular areas we will focus on are ontology development, and query formulation.

Semantic ontologies are one of the three pillars of the Semantic Web. They are seen as the key to making the text of documents on the Web not only machine-readable, but also machine-understandable. There is little in the way of methodology that can guide researchers in the development of such ontologies. In practice, researchers have found that it is extremely laborious and time consuming to develop ontologies by hand (Guarino and Welty 2000). Words are often very imprecise in their conveyance of meaning, and so even in very small and specific domains it can be difficult to capture a vocabulary and the relationships among its terms sufficiently to support a Semantic Web type application. All of these frustrations were apparent at the Semantic Web Working Symposium in the summer of 2001 (Tuttle and McGuiness 2001). It would appear that any attempt to build an ontology that classifies not only all of the English language but all of the languages of the Web will have to make use of some sort of automation to be possible. While perhaps not a solution, IR proposes a possible path to a solution through functional ontologies.

The notion of ontologies in the Semantic Web assumes that the classification of words into an ontology will be based on their semantic meanings as interpreted by humans. While this may be a goal to strive for, in the meantime a more tractable solution would be to develop functional classifications of terms that could be discovered automatically by a computer algorithm. Such an algorithm was developed by Wu (2000). Probability Of Co-occurrence Analysis (POCA) is an algorithm that classifies terms in a document set in terms of their probability of co-occurrence in any given document and then hierarchically into concept maps. The same term may appear multiple times in the hierarchy, as it is normal for words to be related to multiple words. Wu has observed that this algorithm can be used to sort words into a sort of ontology where there is a functional relationship between terms, rather than a strictly semantic one. While in the strict sense this is not a semantic ontology, since truly semantic ontologies have in practice been so difficult to build, this may be a significant step in the right direction. The goal now will be to see if it is possible to improve information retrieval using these POCA generated ontologies. If so, the POCA concept can perhaps be extended to help create truly semantic ontologies. Not insignificant is the fact that the POCA algorithm is completely automatic, meaning that humans do not have to spend hours and days trying to fret over the finer shades of meaning in the relationships between words. Not insignificant is that a query term for document retrieval does not have to be limited to a single word. Query terms may represent phrases, which means that more precise and useful classifications of terms could be created using POCA. A next step would be to see if POCA could be used to establish useful ontologies for use in Semantic Web applications.

MELISA (MEdical LIterature Search Agent) is an example of an IR system that uses agents and ontologies to improve retrieval effectiveness (Abasolo and Gomez 2000). MELISA helps to address one of the weaknesses with IR identified in the previous section—the problem of users having trouble formulating the best queries to answer their questions. MELISA is a prototype of ontology-based agent for information retrieval in medicine. MELISA provides a framework to help users to formulate their queries based on medical category. Based on this framework, users enter the keywords related to their information need and set desired filters. Keywords are automatically checked against MeSH, a medical ontology used by MedLine, a large biomedical

bibliographic reference database. If the keywords entered by users are not valid MeSH terms, MELISA offers users valid MeSH terms that are closest to the keywords entered by users. If the keywords are valid MeSH terms, MELISA returns the list of subheadings of the keywords and lets users further specify their information need. Based on the keywords and filters set by users, MELISA formulates conceptual queries and then decomposes the conceptual queries into specific queries, which search engines can process, by combining every MeSH term associated with conceptual queries. In formulating specific queries, MELISA uses the AND operator. However, if the results turn out to be unsatisfactory after the evaluation process, a new collection of specific queries are generated by the OR operator. The core of MELISA is an IR system, but it uses ontology and agents to help users develop more effective queries. This is one of the ways that Semantic Web ideas contribute to IR.

POCA and MELISA represent just two of the areas where IR and Semantic Web concepts are able to contribute to one another. We anticipate that with further examination, many more chances to combine the bottom-up and top-down thinking styles will arise.

Conclusion

It is important to mention that we are not trying to set up an insular dichotomy here. The idea of the Semantic Web is so new that it may be hard to classify Semantic Web researchers as an entity unto themselves as they exist across many disciplines, including Information Retrieval. However, in the excitement we share with Semantic Web researchers we read a number of their papers and found that the IR concepts were not as well represented as they could have been. It is our hope that the field will become more well rounded. We also expect that other disciplines also could contribute a lot to the Semantic Web, but those contributions are for another paper.

Another issue we should bring up is that there are other visions of what the Web could be that are in competition with the Semantic Web vision. Of note, Ben Shneiderman and the direct manipulation school are concerned that over-reliance on intelligent agent software could impair our own cognitive abilities and personal sense of efficacy. Most likely the Web will continue to be what it is now, a smorgasbord of brilliance and creativity with some of everything thrown in.

Acknowledgements

We would like to acknowledge the guidance and help of Dr. Yi-fang Brook Wu in preparing this research.

References

Abasolo, J. M. and Gomez, M. MELISA, An ontology-based agent for information retrieval in medicine. *ECDL 2000 Workshop on the Semantic Web*, September 2000

Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web, *Scientific American*, May 2001. http://www.sciam.com/2001/0501issue/0501berners-lee.html

Coombs, J. H., Renear, A. H., and DeRose, S. J. Markup Systems and the Future of Scholarly Text Processing, *Communications of the ACM*, 30(11), November 1987, p933-947. http://www.acm.org/pubs/articles/journals/cacm/1987-30-11/p933-coombs/p933-coombs.pdf

Decker Stephan, Mitra Prasenjit and Melnik Sergey " Framework for the Semantic Web". Stanford University

Gey, F. C., Oard, D. W. The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries. *Proceedings of TREC-10*, Gaithersburg, Maryland, November 13-16, 2001. http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Guarino, N. Formal Ontology in Information Systems. In N. Guarino (ed.) *Formal Ontology in Information Systems. Proceedings of FOIS'98*, Trento, Italy, 6-8 June 1998. IOS Press, Amsterdam: p3-15. http://www.ladseb.pd.cnr.it/infor/Ontology/Papers/FOIS98.pdf

Guarino, N. and Welty, C. 2000. Towards a methodology for ontology-based model engineering. In Proceedings of ECOOP-2000 Workshop on Model Engineering. Cannes, France

Hawking, D., and Thistlewaite, P. Overview of TREC-6 Very Large Collection Track, *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6)*, January 1998, p93. (http://trec.nist.gov/pubs/trec6/papers/vlc_track.ps.gz)

Heflin, J. Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment. Doctoral Thesis, University of Maryland, 2001 (http://www.cse.lehigh.edu/~heflin/pubs/heflin-thesis-orig.pdf)

Hendler, J. Is There an Intelligent Agent in Your Future? *Nature*, 11 Mar. 1999

Kashyap, V. Final Report of the Integration, Interoperation and Composition Track, from *Proceedings of the International Semantic Web Working Symposium*, July 30-August 1, 2001, Stanford University, California, USA http://www.semanticweb.org/SWWS/report/

Korfhage, R. *Information Storage and Retrieval.* John Wiley & Sons Inc., 1997.

Maedche, A., and Staab, S. Ontology Learning for the Semantic Web, *IEEE Intelligent Systems,* 16(2), Mar-Apr 2001, p72-79. http://ieeexplore.ieee.org/iel5/5254/19905/00920602.pdf?isNumber=19905

Shneiderman, B., Maes, P. Direct manipulation vs. interface agents, *Interactions*, Nov.-Dec. 1997. http://doi.acm.org/10.1145/267505.267514

Sullivan, D. "Search Engine Sizes" *Searchenginewatch.com* August 15, 2001. http://searchenginewatch.com/reports/sizes.html

Tuttle, M., McGuiness, D. Report of the Ontologies and Ontology Maintenance Working Group, from *Proceedings of the International Semantic Web Working Symposium*, July 30-August 1, 2001, Stanford University, California, USA http://www.semanticweb.org/SWWS/report/

Viles, C. *Maintaining Retrieval Effectiveness in Distributed, Dynamic Information Retrieval Systems*, doctoral dissertation, University of Virginia, May 1996.

Wu, Y.-f. Automatic Concept Organization: Organizing Concepts from Text Through Probability of Co-occurrence Analysis. SIGCR2000