Exploiting the Small-Worlds of the Semantic Web to Connect Heterogeneous, Local Ontologies

Henry M. Kim & Markus Biehl

Schulich School of Business, York University, 4700 Keele St., Toronto, Ontario Canada M3J 1P3; (416) 736-2100 x77952 [phone]; (416) 736-5687 [fax]

hkim@schulich.yorku.ca | mbiehl@schulich.yorku.ca

Abstract

The WWW makes data widely accessible; the semantic Web makes data widely interpretable, ensuring that data can be shared as intended by their creator. However, how can a semantic Web software agent find the right intrepretation (ontology definition)? In this paper, a parallel is drawn between this semantic Web search problem and how people are able to find strangers using a surprisingly short chain of acquaintances—a result from the "six degrees of separation" experiment. The experiment relied on shared understanding of the phrase, "someone you know on a first name basis" to define an acquaintance relationship. Web searching relies on standardized use of the hyperlink relationship. Hyperlinks are constituted from universally accepted meta-data: Anchor and bookmark HTML markups. Say that heterogeneous local ontologies are all marked-up using standard meta-data. Then, the meta-data and some universally accepted semantics constitute a shared ontology, which can be used to bridge local ontologies, much as highly connected people who belonged to many cliques (small-worlds) were used disproportionately often in the search for strangers. This paper outlines the framework for approaching the semantic Web search problem using meta-data based shared ontologies inspired from small-worlds theory of sociology. This approach is exciting for universal or large-scale data integration because it 1) enables data sharing over the semantic Web without *post hoc* modifications to local ontologies, and 2) uses meta-data, which in many situations are already commonly available and implemented in XML

Keywords: semantic Web, meta-data, ontologies, small-world networks

Pre-print version of article published in *Information Technology and Management*, Vol. 6, pp. 89-96, 2005.
Posted with permission at http://www.yorku.ca/hmkim under http://www.yorku.ca/hmkim/files/ITM03.pdf

Printed article can be found at
http://springerlink.metapress.com/(cy04j545ghasapvppbsfo055)/app/home/contribution.asp?referrer=parent&backto=issue,5,7;journal,4,19;linkingpublicationresults,1:101768,1

1. Introduction

Humans can interpret that if the term ‘jaguar’ appears along with words like ‘warranty’ and ‘driver’ on a Web page, the term refers to a car and not a feline. In the context of Ford Motors’ intranet, automated processing of the data instance ‘jaguar’ will not likely be problematic. Standardized terminology and business practices can be brought to bear for intra-enterprise data integration. However, large-scale or even universal inter-enterprise data integration is difficult when automation is desired, but standards do not exist. How can subtle interpretations to differentiate between a car and a feline be automatically made if pre-programmed routines or software agents, not humans, are executing large-scale or universal Web services?

Tim Berners-Lee, the oft-acknowledged inventor of the WWW, states that machine interpretation required for widespread adoption of Web services will be possible with the realization of the *semantic Web* (Berners-Lee *et al.* 2001).

> Computers will find the meaning of semantic data by following hyperlinks to definitions of key terms and rules for reasoning about them logically. The resulting infrastructure will spur the development of automated Web services such as highly functional agents.

In this vision, meanings that computers can find and reason about are represented using *ontologies*, data models that “consists of a representational vocabulary with precise definitions of the meanings of the terms of this vocabulary plus a set of formal axioms that constrain interpretation and well-formed use of these terms” (Campbell and Shapiro 1995).

There is likely much to be researched about the semantic Web because it is so nascent. In predicting how ontologies for the semantic Web will evolve, Kim (2002a) posits that research emphasis should be placed on “developing de-centralized and adaptive ontologies, which have value in of themselves, but whose full potential will only be realized if they are used in combination with other ontologies for data sharing.” The open question stated is how such ontologies would be developed and organized. It is this question that is addressed here.

There are different levels for knowledge sharing—i.e. people (knowledge agents) share knowledge represented as content (knowledge conceptualizations) in Web pages (knowledge models). For the Web, the content is words in HTML; for the semantic Web, it is represented as ontology expressions. A data instance can be shared unambiguously if the right ontology expression to interpret it can be found. This search and find is a lot like trying to connect two random nodes in a network with limited local, but not global, information, since one universal, omniscient ontology is unlike to materialize, but rather local, heterogeneous ontologies exist. Knowledge sharing levels are discussed in Section 2.

Social and Web page networks exhibit small-world properties. These properties can be exploited to effectively find a chain that connects two random nodes in such networks. Strategies for finding a connecting path between a data instance and an ontology definition or axiom that appropriately applies to it as a chain of relationships can similarly exploit the expected small-world properties of the semantic Web. Small-world networks are discussed in Section 3.

Relationships can be defined and constrained from widely shared, hence globally known, knowledge model meta-data, such as anchor and bookmark HTML markups for Web pages. Such definitions and axioms comprise a meta-data base shared ontology of some knowledge model type. When used for the semantic Web, these shared ontologies—which are constructed from meta-data common to all local ontologies of certain type and do not represent much about their content—form a mediating layer to find a semantic relationship between one local ontology’s term to another’s. They then enable correctly applying one local ontology’s semantics to interpret a data instance populated using another local ontology without *post hoc* modifications to these local ontologies. These meta-data based shared ontologies are discussed in Section 4.

Then, strategies for applying these shared ontologies to search for and find relationships for the semantic Web are stated in Section 5. Finally, in Section 6, concluding remarks and future works are stated.

2. Knowledge Sharing Levels: Conceptualizations, Models, and Agents

A Web page is sought and read by humans for the knowledge that can be formulated from concepts represented in its contents. Furthermore, these concepts are publicly sharable via the WWW. As far as software agents are concerned, concepts are represented in and shared using ontologies—another definition is *shared*, explicit *specification* of a *conceptualization* (Gruber 1993). Shared conceptualizations can be as informal and implicit as cultural norms expressed in conversations, informally and explicitly represented in documented standard operating procedures, and formally and explicitly represented in data or knowledge bases. Semantic Web ontologies must be of the last kind, for they must be formal, and obviously explicit, insofar as they must be machine-understandable (Ding *et al.* 2002). Formal representations are expressed in restrictive syntax and semantics such that a given expression has one interpretation, and so machines can algorithmically infer that interpretation. This also means that other machines (software agents) can infer the same interpretation, i.e. share, as long as they can process the syntax and semantics.

Software agent understandable conceptualizations represented in a formal language and collected in ontologies for machines to share on the semantic Web are akin to human understandable conceptualizations represented in informal, natural language and collected in Web pages for humans to share on the WWW. Varying from the "knowledge level" (Newell 1982), three abstract levels for knowledge sharing can be considered: An *agent level* of humans or software agents; a *model level* of Web pages and ontologies; and a *conceptualization level* of words expressed in a natural language like English, or terms, definitions, and axioms expressed in a formal language like OIL (Ontology Inference Layer) (Fensel *et al.* 2001). Knowledge agents create and share knowledge models; models are used to create and share knowledge conceptualizations.

Here is a practical search problem for the semantic Web and Web services. A software agent aware of some local ontology and its underlying conceptualizations has to process a command, "buy a jaguar," to execute a Web service. To do this, the agent must find the right definition of 'jaguar,' but doesn't know where to find it. That is, it must relate an isolated data instance, 'jaguar,' to a term in an unknown ontology for which an appropriate definition exists. Many semantic Web formalisms—e.g. DAML+OIL (Bechhofer *et al.* 2001)—define and constrain an ontology term by its semantic and Boolean relationships to other terms. An ontology is represented as a semantic network (Brachman 1979), from which one definition or axiom constitutes a sub-net. Ontology conceptualizations can be considered sub-nets in a semantic network. It is believed that the semantic Web will be comprised of numerous, locally consistent but globally heterogeneous ontologies; no central ontology aware of these local ontologies is likely to exist (Kim 2002a).

Then locally, an ontology is a meaningful semantic network; globally, one semantic network may be meaningless to another, though some conceptualizations in one ontology may be semantically related to some in a disparate ontology. This is similar to social networks. People in a community generally know each other. Geographically dispersed communities are generally not aware of each other, though people in some communities know people in other far away communities. Milgram (1967) showed that though people form clusters of small-worlds they are nevertheless effective at getting a letter to someone unknown far away by iteratively sending the message on to acquaintances. Can the characteristics of social networks that yielded this result be of use for the semantic Web searching problem? This is explored in the next section.

3. Small-World Networks

A small-world network is comprised of groups of highly clustered small-worlds, which are not heavily connected to each other. Yet, a connection between two random nodes in the network can be achieved with surprisingly few intermediate nodes because there are a few special, "bridge" nodes that belong to many small-worlds (Watts and Strogatz 1998). Societies of film actors (Reynolds 2002), power grids in the Western US, and neuron transmission networks also have been posited as small-world networks (Watts 1999).

In Milgram's experiment, subjects in Nebraska were each given a letter to be received by one target person in Boston, and told to mail it to the person if they knew his address or send it to someone, whom they knew on a first-name basis, who could eventually get the letter to the target. All subsequent subjects who received the letter were given the same instructions. On average, six people handled the letter, which was eventually received by the target. This is the root of the "six degrees of separation" phenomenon. Just 11% (3 people) of penultimate letter recipients were responsible for sending 48% of the letters directly to the target (Travers and Milgram 1969). Subjects perceived that these three knew many from different social circles, and hence sent the letters to the three. Subjects also sent letters to those whom they believed closest to the target. Geographic proximity was primarily used to evaluate closeness. Occupational proximity—it was known that the recipient was a stockbroker—was also considered. Not surprisingly then, the penultimate recipients all lived in the Boston area and two were stockbrokers.

This experiment outlines the two natural person search strategies in social networks. First is to send the letter to someone with many relationships to others, regardless of the nature of the relationship between the sender and recipient, and recipient and others. The second is to send the letter to someone based on the nature of the relationship of the recipient to the target—e.g. the recipient lives closer or has a similar job to the target. The first is a non-semantic search; only the *structure* of the network characterized by relationships is exploited. The second is a semantic search; the specific *nature* of these relationships is exploited. Can both types of strategies be employed for the semantic Web search problem? First, though, how relationships are discerned must be explored. This is done in the next section.

4. Shared, Meta-Data Based Ontologies

Search strategies entail exploiting known relationships. Minimally, there must be some meanings globally accepted throughout the network to indicate that some relationship exists. Milgram's experiment relied upon a reasonably standardized interpretation of the phrase "someone you know on a first name basis." For the Web, HTML provides such a standardized language. A hyperlink relationship between two Web pages is constituted from an anchor in the domain page to the bookmark in the range page. Generally then, meta-data, such as anchor and bookmark HTML markups, are used to relate knowledge models as long as meta-data for all models in a given network are represented in a standardized way so that they can be interpreted commonly.

Meta-data comprise a common vocabulary. An ontology that formally defines and constrains proper use of the vocabulary ensures that meta-data instances are interpretable by machines, and hence sharable by software agents. Meta-data are sharable because commitment is made to only represent what is common about knowledge models, not the conceptualizations they collect, i.e. their contents. For instance, the basic terminology of a meta-data based ontology of academic articles comprise of relationships like 'has author' and 'article cites'; some terms formally defined are 'author' entity and 'writes with' relationship; and an axiom is "an article cannot cite itself." Beyond a relationship like 'has keyword,' there is no commitment to represent article contents in this ontology. So, it cannot be used to *precisely* relate similar or equivalent concepts in different articles. However, it *may* be used for drawing *imprecise* relationships. That is, generally, meta-data based, sharable ontologies of knowledge models may be used to approximate relationships at the conceptualization level. For the semantic Web, this means that shared ontologies representing meta-data about local ontologies, not what is expressed in them, may be used in search strategies to match 'jaguar' with an appropriate definition. How this may work is explored in the next section.

5. Towards Semantic Web Search Strategies

Using the academic article ontology, a 'collaborates with' relationship can be precisely defined between, say, two co-authors. This is a very strong proxy for their having met before and a better-than-random indication that they will co-author again. Generalizing then, meta-data based sharable ontologies of knowledge models can be used to infer precise, and imprecise, but better than random, relationships between knowledge agents, and between agents and models.

Also, 'hyperlink to' relationship relates Web pages, and 'collects'" relates a Web page to its content. Actually, "hyperlink to" is a proxy; it indicates that there is a relationship between the content of two Web pages. However, the nature of the link—whether it denotes a 'part-of,' 'has-additional-info,' etc.—is not specified. Also, the domain and range values of the hyperlink relationship are unbounded—any phrase can be anchored and bookmarked—and hence cannot generally be exploited to discern the semantics of the link.

Yet, search engines effectively use this proxy. The Teoma™ search engine (2002), for instance, manipulates clustering of inter-linked Web pages as a proxy for clustering of related content. Its basis is that the most relevant pages containing search keywords are found in same-subject Web communities in which these keywords are most often found, not in random isolated pages (Davison *et al*. 1999). Existence of these communities (Kleinberg 1999) follows from the observation that the WWW is a small-world network (Adamic and Huberman 2001). Moreover, users apply implicit understanding shared between them and Yahoo!™ taxonomists to limit their interpretation of a hyperlink in Yahoo!™'s subject trees to one of a handful of taxonomical relationships such as 'has-topic' and 'part-of.' Generalizing then, meta-data based sharable ontologies of knowledge models can be used to infer precise, and imprecise, but better than random, relationships between knowledge conceptualizations, and between conceptualizations and models.

Now, practical semantic Web search strategies can be outlined based on explorations of knowledge sharing levels, small-world networks, and meta-data based shared ontologies. Searching over the global span of local ontologies, which are semantically heterogeneous with respect to each other and represented as semantic networks in a language like DAML+OIL, is possible without a global ontology consistent with all such networks. This is possible because sharable ontologies based on local ontologies' meta-data, which are likely represented using XML, can be used to infer precise, and imprecise, but better than random, relationships between terms in disparate, local semantic networks. The small-world properties of the semantic Web can be exploited to estimate the likelihood that a data instance is somehow related to the small-world within which a given local ontology belongs. The higher the likelihood, the more likely some relationship between that instance and a term in that ontology's semantic network exists. Also, some semantics of local ontologies can be manually defined and included in the shared ontologies, but only necessarily for those local ontologies that are the most connected. This is practical since by virtue of the "small-worldness" of the semantic Web, a software agent traversing shared ontologies' relationships is disproportionately likely to search meta-data of these "popular," local ontologies. The agent can quickly and more precisely discern whether a term it is trying to define can be related to a term in the ontologies' semantic network without having to fully "understand" that network.

6. Concluding Remarks and Future Work

In this paper, a realistic approach to enable universal, inter-enterprise data integration using semantic Web ontologies is presented. It is realistic because it does not require major modifications to existing heterogeneous local ontologies, and assumes existence of global meta-data standards. Shared ontologies of local ontologies' meta-data are used as a mediating layer to match a term in one local ontology to definitions and axioms in another ontology required for appropriately interpreting and sharing an instance of that term. The basis of this exciting approach is the following: This matching is similar to people finding, and inasmuch as small-world properties of social networks enable effective searching based on personalized knowledge with limited shared cues, software agents can effectively search based on their local ontologies with shared ontologies of limited meta-data.

There are two exciting projects under way to research semantic Web search strategies further. Taking advantage of the availability of citation indexes on the Web, meta-data from over 60,000 article including 200,000 citation links have been automatically parsed into a knowledge base and cleaned. The aim is to develop an ontology of academic articles and test out search strategies to answer questions about themes and topics in the articles. Currently, a proof-of-concept search engine is being developed, wherein the abstracts of only the most central, "popular" articles (1-2% of all the articles) are cached. When the contents of the abstract of one or more of these central articles is matched to search conditions, the

vocabulary idiosyncratic to the matching articles becomes the shared ontology for a subset or the entire set of 60,000 articles. Most simply, the shared ontology may be the vector of keywords for matching central articles. All other articles can be ranked based upon similarity with this ontology. A non-semantic approach also conceptualized draws small-worlds around the matching central articles along citation and co-authorship relationships, and discern key articles, authors, institutions, and keywords that appear in these small-worlds. The inference is that these are much more relevant than random to fulfill the intent of the search. Also, starting from documents in which Shakespearian plays are represented in XML, an ontology for discerning familiar and conversational relationships has been constructed (Kim 2002b). The aim is to test out search strategies to answer questions about conversational topics even though they are not explicitly represented as meta-data.

7. References

1. Lada A. Adamic, and Bernardo A. Huberman, "The Web's Hidden Order", *Communications of the ACM*, Vol. 44, No. 9, pp. 55-9, 2001.
2. Tim Berners-Lee, James Hendler, and Ora Lassila, "The Semantic Web", *Scientific American*, May 2001.
3. S. Bechhofer, C. Goble, and I. Horrocks, "DAML+OIL Is not Enough", In: *First Semantic Web Working Symposium (SWWS-01)*, July 30-August 1, Stanford, CA, 2001.
4. Ronald J. Brachman, "On the epistemological status of semantic networks", *Associative Networks: Representation and Use of Knowledge by Computers*, N. V. Findler (ed.). Academic Press: NY, pp.3-50, 1979.
5. A. E. Campbell and S. C. Shapiro, "Ontological Mediation: An Overview", In: *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, AAAI Press, Menlo Park, CA, 1995.
6. Brian D. Davison, Apostolos Gerasoulis, Konstantinos Kleisouris, Yingfang Lu, Hyun-ju Seo, Wei wang, Baohua Wu, "DiscoWeb: Applying Link Analysis to Web Search", *Eighth International World Wide Web Conference*, Toronto, Ontario, May 11-14, 1999.
7. Y. Ding, D. Fensel, M. Klein, B. Omelayenko, "The Semantic Web: Yet Another Hip?", *Data and Knowledge Engineering*, Vol. 41, No. 3, pp. 205-27, 2002.
8. D. Fensel, F. van Harmelen, I. Horrocks, D. McGuiness, P. F. Patel-Schneider, "OIL: An Ontology Infrastructure for the Semantic Web", *IEEE Intelligent Systems*, Vol. 16, No. 2, pp. 38-45, 2001.
9. Thomas R. Gruber, "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, Vol. 5 No.2, pp.199-220, 1993.
10. Henry M. Kim (a), "Predicting How the Semantic Web Will Evolve", *Communications of the ACM*, Vol. 45, No. 2, pp. 48-54, 2002.
11. Henry M. Kim (b), "XML-hoo!: A Prototype Application for Intelligent Query of XML Documents using Domain-Specific Ontologies", 35th Hawaii International Conference on Systems Science (HICSS), January 4-7, 2002.
12. Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", *Journal of the ACM*, Vol. 46, No. 5, pp. 604-32, 1999.
13. S. Milgram, "The small world problem", *Psychology Today*, Vol. 61, No. 1, pp. 60-7, 1967.
14. Allan Newell, "The Knowledge Level", *Artificial Intelligence*, Vol. 18, No. 1, pp. 87-127, 1982.
15. Patrick J. Reynolds, "The Oracle of Bacon at Virginia", http://www.cs.virginia.edu/oracle, Last Accessed: July 20, 2002.
16. Jeffrey Travers, and Stanley Milgram, "An Experimental Study of the Small World Problem", *Sociometry*, Vol. 32, No. 4, pp. 425-43, 1969.
17. Duncan J. Watts, Small Worlds: The Dynamics of Networks Between Order and Randomness, Princeton Press: NJ, 1999.
18. D. J. Watts, and S. H. Strogatz, "Collective Dynamics of 'Small-World' Networks", *Nature*, No. 393, pp. 440-2, 1998.
19. Teoma, "Teoma™ – search with authority", http://www.teoma.com, Last Accessed: August 10, 2002.