TOWARDS AN INFORMATION ASSESSMENT FRAMEWORK FOR USE WITH THE SEMANTIC WEB

Heidi J. C. Ellis
Department of Engineering & Science, Rensselaer at Hartford, 275 Windsor St., Hartford, CT 06120 USA

Gregory W. Hislop
College of Information Science & Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104 USA

Todd Kowalczyk
Personal Lines Internet Consumer Application Support, The Hartford, 8 Farm Springs Road, Farmington, CT 06032 USA

Keywords: Semantic Web, Trust, Information quality

Abstract: The extension of the existing Web with meaningful information to form the Semantic Web holds great potential for allowing applications to carry out much more sophisticated tasks than supported by the current Web. As part of carrying out these tasks, Semantic Web applications must access and integrate information from a variety of sources including databases, services, programs, sensors, personal devices, etc. The ability of Semantic Web applications to dynamically assess this information with respect to its trustworthiness and quality is a key contribution to the successful completion of tasks. In addition, increasing interest in the research areas of security and information assurance highlight the need for an information assessment framework for the Semantic Web that incorporates aspects of trust and information quality as both of these aspects are necessary components to information security and electronic commerce. This paper presents a research agenda for the development of an information assessment framework encompassing information quality and trust management or trust agency for the Semantic Web.

1 INTRODUCTION

Automated decision making on the World Wide Web is difficult due to the fact that very little of the information contained on the Web is understandable to the machines that are responsible for processing, categorizing and querying this information. Tim Berners-Lee's vision of the Semantic Web (Berners-Lee, Hendler, & Lasilla, 2001) augments the existing human-readable Web with structured information that will allow computer applications to easily access and process this information. The Semantic Web vision is for applications to integrate a variety of existing Web applications and to bring together databases, services, programs, sensors, personal devices, etc. (Hendler, Berners-Lee, & Miller, 2002). One goal of this expansion of Web information is to support a wider range of applications that can perform more complex tasks than are supported by the current Web. A second goal of the Semantic Web is to allow computers and humans to more easily cooperate in solving complex tasks as computers obtain information from a wider range of sources and more efficiently process this information. The accomplishment of these complex tasks requires Semantic Web applications to make informed decisions. In order for these Semantic Web applications to make correct decisions, the information upon which the decisions are made must be of high quality, or at least be of known quality, and the application must have a reasonable level of trust in the information. However information quality is always a problem in large collections of information and on the Web, quality is certainly uneven and generally difficult to determine. Quality deficits reduce the value of information/knowledge obtained from the Web because such deficiencies increase the risk of getting bad information.

J. C. Ellis H., W. Hislop G. and Kowalczyk T. (2004).
TOWARDS AN INFORMATION ASSESSMENT FRAMEWORK FOR USE WITH THE SEMANTIC WEB.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 447-450
DOI: 10.5220/0002654904470450

Semantic Web applications operate in a dynamic, heterogeneous environment where information sources and Web Services are constantly being added and removed from the Web. There is no central authority that can testify to the reliability of information sources or the trustworthiness of individual services. Applications must be able to dynamically locate services and information sources of interest and make decisions about the information quality and trustworthiness of an entity even if the application does not have prior experience with the entity. A comprehensive model that incorporates aspects of trust and information quality would provide a foundation for accurate decision-making by these applications.

Research in the area of data and information quality (Parssian, Sarkar, & Jacob, 1999; Pipino, Lee, & Wang, 2002; and Kahn, Strong, & Wang, 2002) has identified key criteria by which information quality can be measured, and several researchers have investigated the application of a set of information quality criteria to static Web sites (Katerattanakul & Siau, 1999; Zhu & Gauch, 2000; and Zhang, Keeling, & Pavur, 2000). However, none of this work has addressed the complex and dynamic environment of the Semantic Web. In addition, when dealing with a diverse and ever-changing environment, the trust aspect of information must also be considered. While the quality models include some coverage of trust in components such as authority and reputation, a more comprehensive consideration of trust would be valuable.

The topic of managing trust on the Semantic Web is an emerging area of research interest. An initial survey of models of trust on the Web (Finin & Joshi, 2002; Richardson, Agrawal, & Domingos, 2003; Ramchurn, Jennings, Sierra, & Godo, 2003; and Golbeck, Parsia, & Hendler, 2003) has resulted in a set of trust criteria that includes reputation, endorsement, personal experience, history, temporality, context, credentials, policies, transitivity, disposition, institutional backing, and annotations. Maximillien and Singh (2001; 2002) describe a model that uses reputation as a mechanism for determining the trustability of a Web Service. By exchanging reputation information and endorsing others, agents are able to dynamically locate trustworthy services even though potential interaction partners have not had any prior experience with one another. Gil and Ratnakar (2002) investigate the trustability of Web content based on the source and describe an interactive tool, TRELLIS, which aids individuals in the assessment of the trustworthiness of a particular source. Richardson, Agrawal, and Domingos (2003) present an approach based on a web of trust where entities exchange respective trust assessments of other entities with one another. And in a work that combines some aspects of information quality and trust, McGuiness and Pinheiro da Silva (2003) introduce the inference web, a framework for assessing the quality of source data through proofs.

While the area of trust on the Semantic Web has been receiving increasing attention, many of the existing efforts have focused on the security aspects of trust (e.g., certificates, authentication, etc.) and have not addressed trust with respect to information quality. In addition, McBride (2002) and Sandberg & Ellis (2002) both identify the pressing need for practical applications that are tolerant of error to be developed in the near future in order to ensure the widespread adoption of the Semantic Web. A comprehensive information assessment framework that incorporates information quality assessment and trustability assessment provides the foundation required for these applications to accurately select between alternatives.

2 RESEARCH QUESTIONS

The proposed research described in this paper is expected to result in a comprehensive information assessment framework that embodies trust and information quality. To provide a structure in which to discuss the research effort, we define the following research questions.

In order to ensure that the assessment framework contains the necessary aspects of both information quality and trust, the critical aspects of each must be determined and clearly defined. Therefore, the initial and foundational research question is:

R1: What information attributes could support an information assessment framework about quality and trustworthiness of information sources?

This general framework must then be implemented and evaluated within the context of the Semantic Web. Therefore, the second research question is:

R2: Does this information assessment framework support usefully accurate determinations of quality and trustworthiness of Web information sources?

These research questions provide an outline for the direction of the proposed research.

2.1 Significance

Developing an information assessment framework that includes aspects of information quality and trust is essential for the widespread adoption of the Semantic Web. In particular, accurate information

assessment is a crucial capability for commercial and military Semantic Web applications that make critical decisions based on this information. In the dynamic, distributed environment of the Semantic Web, the decision-making process is difficult due to incomplete information, conflicting information, and insufficient information. Therefore, the ability to evaluate the trustworthiness and reliability of information is critical. The proposed research into the information assessment framework will aid in the decision-making process, particularly with information of questionable quality or questionable trustworthiness.

A second impact of this effort will be the ability to assess the behavior of portions of the Semantic Web. For instance, the information assessment framework could be applied to evaluate the same information within the same application at various time intervals and the differences in trust and information quality factors examined. An examination of these results would lead to observations about the impact of changes in information or Web sources on trust and information quality.

3 METHODS

The overarching goal of the work discussed here is **to construct a comprehensive information assessment framework that embodies trust and information quality.** In order to achieve this goal, we will build on existing work on a Semantic Web application (Blough & Ellis, 2003). This work produced a prototype tool, DANDI (Data Analyzer / Data Integrator) that provides an extensible test bed for working with Semantic Web information. DANDI uses a proxy server to observe Web browsing activity, and passes the browser stream to a data analyzer. DANDI has a flexible structure, created with reuse and evolution in mind.

In the initial phases of the work, we will develop preliminary attribute models of information quality and trust. Where possible, measures for these attributes for the Semantic Web will be defined. The data analyzer component of the DANDI test bed will be extended to support the integrated trust and information quality models and to apply them to information assessment. The prototype produced in the initial phases of the research will be used to demonstrate the assessment framework's utility in Semantic Web applications for several domains.

3.1 Research Steps

The achievement of the overarching goal will be accomplished in compatibility with the existing W3C standards and activities (http://www.w3.org/2001/sw/) and involves the following steps:

Step 1: Select a comprehensive set of attributes of trust. In order to ensure the completeness of the information assessment framework with respect to attributes of trust, we will first conduct a broad investigation of prior work in approaches to trust as well as researching the emerging approaches to trust management on the Semantic Web. As a result of this examination of approaches to trust, we will select a comprehensive set of attributes of trust.

Step 2: Select a comprehensive set of attributes of information quality. The motivation for selecting a complete set of information quality attributes is similar to that for the set of trust attributes; these attributes are necessary to ensure that all aspects of information quality relevant to the Semantic Web environment are included in the assessment framework. The result of this study of information quality characteristics will be the selection of a comprehensive set of information quality attributes.

Step 3: Define and revise the trust and information quality attributes. For each member in the initial set of trust and information quality attributes resulting from steps 1 and 2, we will identify a measurement approach, in the form of a formula, function, method, etc., or a heuristic. Attributes that do not appear to have reasonable representation on the Semantic Web will be discarded. The result will be trust and information quality models that consist of attributes that can be applied in the Semantic Web context.

Step 4: Extend the DANDI prototype to include the information assessment framework. The data analyzer in DANDI will be extended to support the information assessment framework produced in Step 3. The data analyzer will passively gather items of interest from the information stream between the browser and server and will evaluate this information using the information assessment framework to determine trust and information quality appraisals for the information.

Step 5: Application of the information assessment framework. In order to test the framework, DANDI will then be applied to various domains that involve information from a range of sources with varying quality and trustworthiness values and the results will be calculated. The application will first be tested using information

with known trust and information quality values. Then the application will be tested using existing Semantic Web pages that have been utilized for other trust management applications (where possible). The results from our application will be compared to other published results.

Since a key aspect of Semantic Web applications is the ability to make decisions based on reliable assessments of information quality and trust in a diverse and dynamic environment that is undergoing continual change, the assessment framework will consider the following aspects: 1) the dynamic nature of the Semantic Web environment; 2) compatibility with existing trust determination approaches used on the Semantic Web; and 3) flexibility of support for both agent-based applications as well as Semantic Web applications with different architectures.

4 CONCLUSION

In order for the Semantic Web to be commercially and widely adopted, Semantic Web applications must have the ability to make accurate decisions based on information with known trust and quality values from a variety of sources. This paper reports on a research agenda for the development of an information assessment framework that incorporates trust and information quality to support such decision-making capabilities. Successful completion of the proposed research and framework will provide the ability to accurately assess the information quality and trustworthiness of information used by Semantic Web applications. These assessments will aid in the decision-making process, particularly with information of questionable quality or questionable trustworthiness.

REFERENCES

Berners-Lee, T., J. Hendler, & O. Lasilla, May, 2001. The Semantic Web, *Scientific American*.

Blough, R.T., & H.J.C. Ellis, 2003. A passive real-time semantic Web framework for information gathering, *Rensselaer at Hartford Technical Report, RH-DOES-TR 03-02.*

Finin, T., & A. Joshi, 2002. Agents, trust, and information access on the Semantic Web, *SIGMOD Record, (31)*4.

Gil, Y., & V. Ratnakar, 2002. Trusting Information sources one citizen at a time, In *Proceedings of the First International Semantic Web Conference (ISWC)*, Florida, USA.

Golbeck, J., B. Parsia, & J. Hendler, 2003. Trust networks on the Semantic Web, In *Proceedings of Cooperative Intelligent Agents 2003*, Helsinki, Finland.

Hendler, J., T. Berners-Lee, & E. Miller, 2002. Integrating applications on the Semantic Web, *Journal of the Institute of Electrical Engineers of Japan, (122)*10.

Kahn, B.K., D.M. Strong, & R.Y. Wang, 2002. Information quality benchmarks: Product and service performance, *Communications of the ACM, (45)*4.

Katerattanakul, P., & K. Siau, 1999. Measuring information quality of Web sites: development of an instrument, In *Proceedings of the 20th International Conference on Information Systems*, North Carolina, USA.

Maximillien, M., & M. Singh, 2001. Reputation and endorsement for Web Services, *ACM SIGecom Exchanges*, 3.1.

Maximillien, M., & M. Singh, 2002. Conceptual Model of Web Service Reputation, *ACM SIGMOD Record (31)*4.

McBride, B., 2002. Four steps towards the widespread adoption of a Semantic Web, In *Proceedings of the 1st International Semantic Web Conference*, Italy.

McGuiness, D, & P. Pinheiro da Silva, 2003. Infrastructure for Web Explanations, In *Proceedings of the 2nd International Semantic Web Conference*, Florida, USA.

Parssian, A., S. Sarkar, & V.S. Jacob, 1999. Assessing data quality for information products, In *Proceedings of the 20th International Conference on Information Systems*, North Carolina, USA.

Pipino, L.L., Y.W. Lee, & R.Y. Wang, 2002. Data quality assessment, *Communications of the ACM, (45)*4.

Ramchurn, S.D., N.R. Jennings, C. Sierra, & L. Godo, 2003. A computational model for multi-agent interactions based on confidence and reputation, In *Proceedings of the 6th International Workshop of Deception, Fraud and Trust in Agent Societies*, Melbourne, Australia.

Richardson, M., R. Agrawal, & P. Domingos, 2003. Trust management for the Semantic Web, In *Proceedings of the 2nd International Semantic Web Conference*, Florida, USA.

Sandberg, M., & H.J.C. Ellis, 2002. Major integration and scalability roadblocks to the large-scale deployment of a semantic search engine, In *Proceedings of the 2002 International Conference on Information and Knowledge Engineering*, Reno, NV.

Zhang, X., K.B. Keeling, & R.J. Pavur, 2000. Information quality of commercial Web site home pages: an explorative analysis, In *Proceedings of the 21st International Conference on Information Systems*, Brisbane, Australia.

Zhu, X., & S. Gauch, 2000. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, Athens, Greece.