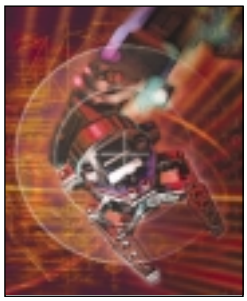

Trusted Autonomy

Michael N. Huhns • *University of South Carolina* • *huhns@sc.edu*
Duncan A. Buell • *University of South Carolina* • *buell@cse.sc.edu*

Last issue, this column described how an agent basis for software could lead to robustness. In this issue, we will describe how agents are the right building blocks for constructing trustworthy systems. As we'll show, these two thrusts — robust software and trusted autonomy — represent the future for agent technology and software engineering.

Reaching the Next Generation

While visions for the Web's future abound, most are based on the tenets that the Web will be ubiquitous, will provide services as well as content, and will continue to be dynamic, growing substantially in the short term. Furthermore, most agree that the Web will continue to have no central authority, with its components remaining autonomous, and will support cooperative peer-to-peer interactions as well as client-server interactions (see Figure 1).

Most visions also recognize the following problems with the Web, as it currently exists:

- information on the Web is not organized;
- information on the Web might be inconsistent, incomprehensible, and inaccurate;
- Web searches typically earn low values for precision and recall;
- Web services are rigid, procedural, and strictly client-server; and
- state information is purely local.

If there were a central authority with a global ontology to which all Web components adhered, and if the Web's components were static, and if the identity of the components were fixed, and if there were a small fixed number of component types, these problems would disappear. But then the Web would no longer be the vibrant, useful place on which the global economy and modern society are based. Researchers must work to overcome the problems we've listed, realizing the visions in a way that will foster the Web's growth without compromising its utility. We believe that the keys to the next-generation semantic Web[1] are cooperative services, systemic trust, and semantic understanding, coupled with a declarative agent-based infrastructure.

While the Web's size and dynamism present problems, the Web provides a means for solving its own problems. For example, there might be an overload of information for a given topic, with much of it redundant and some inaccurate, but a system can use negotiation and voting techniques to eliminate information that is not consistent or agreed upon. Or there might be many potential service providers — some of which are not trustworthy — competing for many potential clients, but a system can use a Web-based "reputation network" to assess credibility. Finally, different sites might use many different ontologies, but this multiplicity of ontologies can yield a global, dynamically formed, consensus ontology.[2]

Trusted Autonomy

Autonomy is a characteristic of agents — and of many envisioned Internet-based applications. Among agents, we generally refer to social autonomy, where an agent is aware of its colleagues and is sociable, but nevertheless exercises its independence in certain circumstances, such as by refusing a request when it might harm the agent's interests. Autonomy is in natural tension with coordination or with higher-level notions, such as commitments. To be coordinated with other agents or keep its commitments, an agent must relinquish some of its autonomy, but an agent that is sociable and responsible can still be autonomous. It would attempt to coordinate with others where appropriate and keep its commitments as much as possible, but it would exercise its autonomy in entering into those commitments in the first place.

Information systems are becoming increasing-

 1089-7801/02/$17.00

ly autonomous. Automated planners are scheduling military combat missions, for example, and the US Air Force routinely uses fire-and-forget armaments. NASA missions might be away from human contact and control for several years. Unmanned aerial vehicles are already being outfitted with weapons to defend themselves and attack targets, and soldiers will soon share battlefields with robotic tanks and artillery. One day, global supply chains will even control the complex movement of goods from raw materials to customers without human intervention.

As such systems and missions become more complicated and of longer duration, the software systems controlling them will necessarily be large and complex. No one can anticipate all the situations the systems will face, so the systems cannot be fully tested. We will basically have to trust them — and we need a principled basis for that trust.

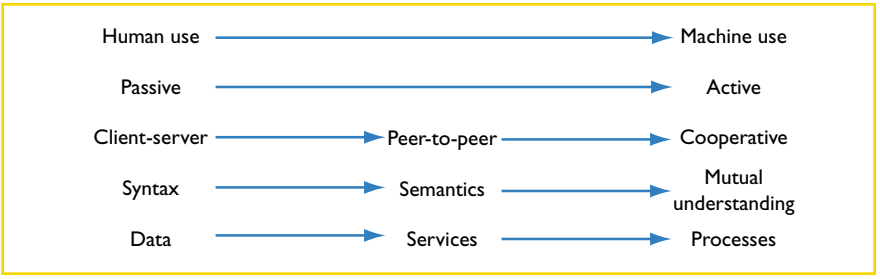

Figure 1. Trends in the Web's use and character, showing that it will become increasingly more suitable for use by automated processes.

Systemic Trust

Both agents and Internet-based applications are dependent on and driven by trust. Fundamentally, the information that agents and applications re-trieve must be accurate, or characterized accurately, and the information they contribute must be used appropriately. This requires mechanisms for tracking sources' reliability and reputation, as well as for specifying constraints on usage and ensuring that dependencies are preserved and maintained. The result is that information items have credibility and domains of utility. What are some of the ingredients of trust?

- *Understanding.* You are more likely to trust something if you understand it. Unfortunately, as systems become more complex, they are harder to understand. So, we need to describe, program, and use them at higher levels of abstraction where the complexity is hidden. At a high level, systems become team members and behave like agents and, where there are a lot of them, assume societal roles.
- *Interaction management.* Trust goes beyond security in that it is about managing interactions at the application level. For example, where security is about authenticating another party and authorizing actions, trust is about the given party acting in your best interest and choosing the right actions from among those that are authorized.
- *Philosophy and societal conventions.* We suggest that endowing agents with explicit ethics and a philosophy that we understand and endorse can lead to systems that we trust. For example, Western philosophy teaches that there is value in each human life; so if a robotic tank embodied such a philosophy, we could be confident it would not roll over its own troops in its zeal to carry out an assignment.

Trust can be established through agent-based components, an architecture that embeds explicit philosophical principles into the agents, and means of organizing agents that are akin to human social systems. Human organizations and societies have evolved ways of maintaining social order by adopting ethics, norms, and conventions. Complex agent-based systems can be constrained by sets of agent societal laws similar to Asimov's laws.[2,3]

Research on current agent-based applications has so far demonstrated that

- agents can glue together independently developed legacy systems,
- control of a system can be distributed among autonomous agents and still maintain global coherence, and
- coherence and capability improve greatly when systems (represented by agents) cooperate.

Crucial technical issues remain unresolved, however, and their resolution is important for the next generation of Internet applications:

- whether adopting human social concepts can enable agents to achieve the same flexibility and robustness exhibited by some human societies,
- whether having an explicit philosophy can enable agents and the complex systems they form to handle emergencies and unanticipated problems or circumstances,
- whether philosophical agents are more effective members of robotic-agent-person teams, and
- whether we can raise the abstraction level at which we program and use complex systems to sufficiently simplify their deployment.

Such research will evaluate the feasibility of systems controlled by a philosophical agent society and create a road map for developing, implementing, and deploying future Internet systems. Using philosophical agents in future information systems should support longer, more complex applications than possible under the current model of human control and should lead to

less costly applications because the amount of support they require can be minimized. Furthermore, communication time lag will be eliminated as a significant factor, providing the ability to react to and take advantage of serendipitous events, and making applications significantly more robust.

Reputation and Credibility

Systems will be trustworthy when the information they use and provide is credible, which requires that the information's sources be reputable. Our research team at the University of South Carolina is investigating how to assess and maintain information credibility in a large-scale environment of autonomous information sources. This is a key component in an overall research objective aimed at increasing information's security, while ensuring its availability. We have begun to analyze the role of reputations and how to propagate and compute them efficiently. At its heart, our approach relies on large-scale systems of computational agents that interact to maintain up-to-date reliability assessments of information sources in specific domains.

Society is rapidly approaching a world in which computing is ubiquitous. Sensors, wearable computing devices, and portable computing and communications devices will let us rapidly generate an enormous database of individual data items pertaining to any large organization's human and equipment resources. To enhance decision-making processes, both strategic and tactical situations will require that we can assess the reputation of the system's information.

The Credibility Problem

Our suggested approach draws on solutions to several problems that resemble the credibility problem. For a theoretical underpinning, we first draw on the extensive research on the rumor problem and on Bayesian network solutions to that problem. Next, we include methods used in Web search engines for estimating the quality of a particular Web site's information. Finally, because credibility estimation relies on interconnectivity and interrelatedness of information that is subject to corruption or repudiation, we introduce methods used successfully to compare phylogenetic trees in the presence of mutation and change in an underlying genetic structure.[5]

At its core, the credibility problem reduces to a problem of quantifying parameters on the nodes of a graph. Each node represents an item of information, and directed arcs link one item to any other item whose credibility depends on the first. The credibility itself is a numerical quantity or quantities attached to the node. Some Web-search engines, most notably Google,[6] have adopted a basic technique for determining data's reputation that involves considering the number of Web links pointing into the site in question. Under the presumption, perhaps, that 40 million Web sites are not likely to be wrong, the more links into a site, the more reputable the data. Thus, when Google ranks the sites returned from a search, it gives a higher ranking to a site to which a large number of other sites point.[7]

Determining Reputation

We can take a similar approach with regard to reputation, with some changes. Data items that are mutually verified by independent sources are more likely to be reputable than those that track back to a single source, regardless of how many times the item has been retransmitted by someone downstream from the original source. This is similar to the rumor problem, which has been attacked successfully by Bayesian techniques that not only manage uncertainty, but also are sufficiently robust that a single garble or corruption will not produce a catastrophic ripple effect. On the other hand, some individual sources can be more credible than others, and items generated from within a secured environment are more likely to be trustworthy than are items generated in the field from sensors or communications devices that could have been physically compromised since their initial placement.

Viewed this way, the problem of determining information's reputation relates strongly to the problems of building "come-from" trees and inferring a given node's credibility from the weighted credibility of nodes from which the node in question derives information. This is a classic problem in computing, but often a difficult one to solve given the need for substantial computing resources and an everything-against-everything computation. Even after a computation is performed to produce a starting point, a problem persists in how to add information to the credibility trees incrementally rather than being forced to do the computation over and over again.

The problem as stated also resembles the problem of phylogenetic comparisons in computational biology. Of great interest to computational biology is the determination of a path by which one genetic sequence can transform into another through mutation and natural selection. These transformations can be represented as trees, and one computational problem in phylogenetics is to quantify the similarity of one tree against another. Weightings exist on the arcs of the trees because some changes are more likely than others based on the underlying biology. And, just as with the information reputation problem, the computational result of the problem as modeled might produce wrong or irrelevant answers because irrelevant similarities can exist among phylogenetic trees.

An Information-Theory Approach

Finally, from an information-theoretic standpoint, trustworthiness and robustness in an information system arise from redundancy (for example, parity bits in a data word). Ubiquitous information sources, if organized appropriately, can provide the redundancy needed to detect, correct, and

compensate for errors, whether arising from inexact sensors, algorithmic mistakes, or disinformation. A further complexity exists in the problem in that "trust" and "mistrust" of information are not necessarily symmetric; trust is often conferred only by the combined use of several indicators, but revoked based on a single indicator, because a minimax principle obtains and a conservative approach to minimizing loss must be adopted. In other situations, the possibility of any success must outweigh the possible costs. We expect our investigations to yield fundamental advances in understanding the limits of large-scale distributed reliability assessments.

The result will be an improved understanding of the credibility of information and the reputation of information sources. This will have the practical benefits of improving information utility for Internet applications, where typical searches yield huge numbers of documents with unknown credibilities, and for information-intensive military scenarios, where uncertainties and disinformation might be widespread.

Conclusion

As Web uses (and thus Web interactions) become more complex, it will be increasingly difficult for one server to provide a total solution and increasingly difficult for one client to integrate solutions from many servers. Web services currently involve a single client accessing a single server, but soon applications will demand federated servers with multiple clients sharing results. Cooperative peer-to-peer solutions will need managing, and it appears that an agent basis is what is needed. Agents can balance cooperation with self-interest, and they also have a property of persistence, which is necessary for establishing trust.

Moreover, agents typically interact via the exchange of declarative messages. The current models for Web services are based on the exchange of procedures via Corba, .Net, or remote method invocation (RMI), but the history of computing has shown that declarative approaches are ultimately favored. Just as Structured Query Language (SQL) supplanted information management service (IMS) procedures, so will agent communication languages, such as the FIPA Agent Communication Language (ACL), supplant the Web Services Definition Language (WSDL).

Significant research is needed to enable autonomous services to be federated on demand and with acceptable delay. Coupled with research advances in semantic reconciliation and the assessments of trust and credibility, the result will be a more efficient and more useful Web in a ubiquitous computing environment.

Acknowledgments

The National Science Foundation supported this work under grant number IIS-0083362.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, May 2001.
2. I. Asimov, *I, Robot*, Gnome Press, New York, 1950.
3. I. Asimov, *Foundation and Empire*, Gnome Press, New York, 1952.
4. L.M. Stephens and M.N. Huhns, "Consensus Ontologies: Reconciling the Semantics of Web Pages and Agents," *IEEE Internet Computing*, vol. 5, no. 5, Sept.-Oct. 2001, pp. 92-95.
5. D. Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge Univ. Press, Cambridge, UK, 1997.
6. N. Manocha, D.J. Cook, and L.B. Holder, "Structural Web Search Using a Graph-Based Discovery System," *ACM Intelligence*, vol. 12, no. 1, Spring 2001, pp. 20-29.
7. J.M. Pierre, "Practical Issues for Automated Categorization of Web Sites," *Electronic Proc. ECDL 2000 Workshop on the Semantic Web*; http://www.ics.forth.gr/proj/isst/SemWeb/program.html.

Michael N. Huhns is a professor of computer science and engineering at the University of South Carolina, where he also directs the Center for Information Technology.

Duncan A. Buell is the chair of the Department of Computer Science and Engineering at the University of South Carolina, where he also is conducting research in high-performance computing and information security.

IEEE Internet Computing

Editorial: *IEEE Internet Computing* targets the technical and scientific Internet user communities as well as designers and developers of Internet-based applications and enabling technologies. Instructions to authors are at computer.org/internet/edguide.htm. Articles are peer reviewed for technical merit and copy edited for clarity, style, and space. Unless otherwise stated, bylined articles and departments, as well as product and service descriptions, reflect the author's or firm's opinion; inclusion in this publication does not necessarily constitute endorsement by the IEEE or the IEEE Computer Society.

Circulation: *IEEE Internet Computing* (ISSN 1089-7801) is published bimonthly by the IEEE Computer Society. IEEE headquarters: 3 Park Avenue, 17th Floor, New York, NY 10016-5997. IEEE Computer Society headquarters: 1730 Massachusetts Ave., Washington, DC 20036-1903. IEEE Computer Society Publications Office: 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720; (714) 821-8380; fax (714) 821-4010. Subscription rates: IEEE Computer Society members get the lowest rates and choice of media option — US$37/30/48 for print/electronic/combination. For information on other prices or to order, go to http://computer.org/subscribe. Back issues: $10 for members, $20 for nonmembers. Also available on microfiche.

Postmaster: Send undelivered copies and address changes to *IEEE Internet Computing*, IEEE Service Center, 445 Hoes Ln., Piscataway, NJ 08855-1331. Periodicals postage paid at New York, N.Y., and at additional mailing offices. Canadian GST #125634188. Canada Post International Publications Mail Product (Canadian Distribution) Sales Agreement #1008870. Printed in USA.