1999

Personal Ontologies

Michael N. Huhns
University of South Carolina - Columbia, huhns@sc.edu

Larry M. Stevens
University of South Carolina - Columbia, stephens@cec.sc.edu

Publication Info

Published in *IEEE Internet Computing*, Volume 3, Issue 5, 1999, pages 85-87.
http://ieeexplore.ieee.org/servlet/opac?punumber=4236

PERSONAL ONTOLOGIES

Michael N. Huhns • *University of South Carolina* • *huhns@sc.edu*
Larry M. Stephens • *University of South Carolina* • *stephens@sc.edu*

> There is nothing more basic than categorization to our thought, perception, action, and speech. [1]
>
> —George Lakoff

Larry and I seem to be drowning in information. While searching my office for an article I needed for this column, I came across two other relevant articles—neither of which I had even remembered saving.

We have shelves and file cabinets full of books, magazines, reports, and papers, and we are accumulating more every day. And because we work in a supposedly "paperless" organization, we are also accumulating online documents, e-mail messages, and data files. Our organization maintains numerous databases that we access, and we maintain pointers to many other sources across the Web.

Too much information is just as bad as no information. If you can't find it when you need it, and it gets in the way of other information you're trying to find, then you might as well not have it. Unfortunately, you never know in advance which information to keep and which to discard.

Information expires or is superseded, and there is no easy way to determine when this happens. On the Web, it is impossible to distinguish among links to pages that no longer exist, links to pages that have new or updated information, and links to pages that haven't changed.

Corporations also suffer from too much information, and it is often inaccessible, inconsistent, and incomprehensible. The corporate solution entails knowledge management techniques and data warehouses. In fact, one objective of data warehousing is to enable individuals to access corporate data for decision making. But data warehousing projects are usually massive efforts. "In our experience, data warehousing projects typically cost around $2 million and take at least two years to implement," says Mark Demarest, president of DP Applications, a data warehousing solutions vendor. "Even with that effort, only about half are successful." [2]

But the corporate answer—putting all information in a warehouse—won't work for us: our office information is too complicated. Data warehouses usually consist of data and information culled from already structured sources, such as operational databases. Unfortunately, the information in our offices (and probably yours, too) is unstructured and disorganized. Our office information is also inherently multimedia and heterogeneous. Note too that the problem's magnitude does not depend on the amount of information in gigabytes, but rather on the number of concepts involved. Corporate data sources may be large, but they don't necessarily involve many concepts. Office workers like us cannot afford to spend two years and $2 million, so what can we do?

A Personal Ontology

One promising approach is an organization scheme based on a model of our office and its information—an *ontology*—coupled with the proper tools for using it. [3]

An ontology is a computational model of some portion of the world. It is often captured in a semantic network—a graph whose nodes are concepts or individual objects and whose arcs represent relationships or associations among the concepts. Properties and attributes, constraints, functions, and rules governing concept behavior augment the semantic network.

With an ontology, we can organize keywords and database concepts by capturing the semantic relationships among the keywords or among the tables and fields in databases. The semantic relationships provide an abstract view of the information space for our offices.

Magnitude and Requirements

What we need is a simple means of constructing an ontology for our information and a simple way to access and maintain it. Besides simplicity, we need

- Support for offline as well as online information—that is, the information in our bookcases and file cabinets.
- Support for both browsing and searching. Searching is more useful when you already know what information you're seeking and the collection is small enough to be well understood. Browsing is more useful when you're not sure what information is available, which is often true when the collection of information is very large.
- Support for temporal ordering so that, for example, we can locate our most recently saved document.

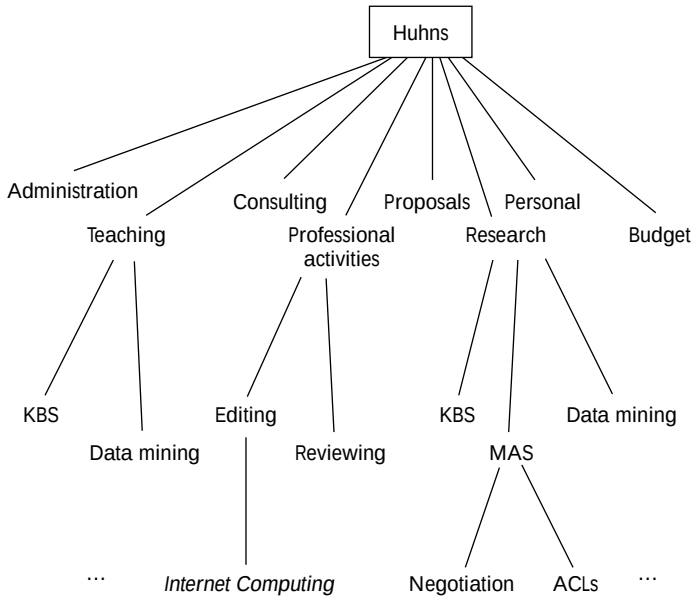

Figure 1. Part of a three-level ontology for Huhns' information space.

- A balance between coarse- and fine-grained classifications and between the depth and breadth of branching for browsing. If our information is organized into too many categories, we have trouble locating the right category. If it is organized into too few, then a given category lacks sufficient discrimination. If the categories are not grouped into higher level concepts, it is too difficult to navigate among them.

Now let's analyze the problem's magnitude and clarify the above requirements. We would like to index our information to the individual paper, file, or chapter level, and I have 16,000 such items, distributed as follows. My bookshelves contain about 600 journals with five papers each, 300 books with eight chapters each, and 60 conference proceedings with 100 papers each. I also have six file cabinet drawers, each with about 100 items. My online information consists of 2,800 files, 800 e-mail messages, 100 browser bookmarks, spreadsheets with 200 distinct headings, and personal database tables with 100 different fields.

It would be reasonable to group about 32 items in each category, resulting in about 512 base categories. If these categories were themselves grouped into higher level concepts, we could have a hierarchy three levels deep with a branching factor of 8, as shown in Figure 1. We could browse this hierarchy of concepts with just three mouse clicks. The ontology is not necessarily a strict hierarchy, as shown, but can be a directed acyclic graph.

Bootstrapping

Most of us probably maintain a directory of online files via MS Explorer, MacOS Finder, or the Unix file manager. This provides a good start for a personal ontology. Next, many journals now have their tables of contents on the Web, and we could incorporate these into our personal ontology for the journals on our bookshelves. We would need a tool—perhaps a personal information agent—to help us. This agent would also link our ontology to a directory service, such as Microsoft's Active Directory, which stores enterprise information about people, applications, and resources.

Titles of books and proceedings would have to be entered manually, but these publishers too may have the contents online, for which we would need another processing tool. After the contents are integrated into our ontology, our request for information on some topic would cause a search through the contents, ultimately yielding a pointer to the correct book on our shelves.

Finally, we would need a Web crawler (spider) to parse our online documents and produce an index of keywords. The keywords would then have to be integrated into our ontology, hopefully with the assistance of yet another tool. Such Web crawlers are available now, but the other tools need to be developed.

Once constructed, a personal ontology might be an information resource in its own right. We could borrow and modify ontologies from colleagues working in similar research areas.

Performance

Just as in classical information retrieval, the important concepts are precision and recall. A high-precision retrieval would not contain any irrelevant information, and a retrieval with high recall would contain all relevant information. These are complementary concepts—retrieving every document would yield maximum recall but poor precision, while retrieving no documents would yield maximum precision but poor recall. The goal is to maximize both concepts at the same time.

Other Available Technology

Search engines have become an essential technology for dealing with the huge amount of information available on the Web. They have also been deployed on corporate intranets and individual Web sites. They work in two steps. First, they use Web crawlers to find documents or compute the keywords that represent each document or Web page and to organize the keywords into an index. Second, when users specify the keywords they are interested in, the engines search the index for documents that match the keywords most closely and return pointers to the document, usually in an order that depends on the degree of match.[4]

A personal search engine would help with my online text documents. However, online documents have many different encodings (for example, .doc, .ppt, .xls, and .mdb), making it difficult for search engines to index their contents or for other search programs (such as grep and its variants) to find possible keyword matches.[4]

Semio Corp. solves this problem by

providing an indexing engine that extracts key concepts from sources of multiple types, relates the concepts to each other, and organizes them into a two-level taxonomy (that is, an ontology limited to a hierarchy of concepts arranged from general to specific). The source documents are linked to the nodes of the taxonomy. Semio also provides a library of upper level ontologies for general domains, such as sports, health, or gardening. The most specific concepts in the topic library are linked manually to the most general nodes in the taxonomy. The result, SemioMap, is displayed graphically for convenient browsing and retrieval of the source documents.

How Can Agents Help?

The information you need might not be in your office or on your PC, but personal information agents can help you find it. If everyone had a personal ontology and an agent to manage it, these agents could communicate about the information they have and the information they need. By providing a vocabulary and relationships among the terms in the vocabulary,[4] their ontologies would enable them to understand each other. We could then find documents in our co-workers' offices and cubicles.

Systems of the Bimonth

To try out a system that organizes documents into a network of related concepts, visit the SemioMap at Semio Corp.'s Web site.

For a description of how XML can be used to encode semantic mappings among concepts, including concepts represented in different natural languages, visit the site for the Advanced Search Facility.

Check them out!

Advanced Search Facility • **asf.gils.net/semantic-map.html**
SemioMap • **semio.com/demo/index.htm**

Bottom Line

A personal ontology manager could be the next "killer app" for desktops. We hope entrepreneurs somewhere are paying attention, because we desperately need an information management tool to keep us afloat. ■

REFERENCES

1. G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Univ. of Chicago Press, 1987.
2. M. Lattig, "Getting Business Smarts," *InfoWorld*, 2 Aug. 1999, pp. 32-33.
3. M.N. Huhns and M.P. Singh, "Ontologies for Agents," *IEEE Internet Computing*, Vol. 1, No. 6, Nov.-Dec. 1997, pp. 96-98.
4. I. Greenberg and L. Garber, "Searching for New Search Technologies," *Computer*, Vol. 32, No. 8, Aug. 1999, pp. 4-7.

Michael N. Huhns is a professor of electrical and computer engineering at the University of South Carolina, where he directs the Center for Information Technology. He is associate editor for *IEEE Intelligent Systems* and *ACM Transactions on Information Systems.*

Larry M. Stephens is a professor in the Department of Electrical and Computer Engineering at the University of South Carolina. His current research interests are multiagent systems and ontologies. He earned a BS in electrical engineering from the University of South Carolina and an MS and a PhD in electrical engineering from Johns Hopkins University.

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 online at http://computer.org/internet/. Articles are peer-reviewed for technical merit and copyedited 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. Annual subscription: $32 in addition to any IEEE Computer Society dues, $48 in addition to any IEEE dues; $58 for members of other technical organizations. Nonmember subscription rates are available on request. Back issues: $10 for members, $20 for nonmembers. This magazine is also available on microfiche.

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