

Introduction

Hsinchun Chen

Artificial Intelligence Lab, Management Information Systems Department, University of Arizona, Tucson, AZ 85721. E-mail: hchen@bpa.arizona, http://ai.bpa.arizona.edu

In this era of the Internet and distributed, multimedia computing, new and emerging classes of information systems applications have swept into the lives of office workers and everyday people. New applications ranging from digital libraries, multimedia systems, geographic information systems, and collaborative computing to electronic commerce, virtual reality, and electronic video arts and games have created tremendous opportunities for information and computer science researchers and practitioners.

As the applications become more overwhelming, pressing, and diverse, several well-known information retrieval (IR) problems have become even more urgent in this network-centric information age. Information overload, a result of the ease of information creation and rendering via the Internet and the WWW, has become more evident in people's lives (e.g., even stockbrokers and elementary school students, heavily exposed to various WWW search engines, are versed in such IR terminology as recall and precision). Significant variations of database formats and structures, the richness of information media (text, audio, and video), and an abundance of multilingual information content also have created severe information interoperability problems—structural interoperability, media interoperability, and multilingual interoperability.

The conventional approaches to addressing information overload and information interoperability problems are manual in nature, requiring human experts as information intermediaries to create knowledge structures and/or ontologies (e.g., the National Library of Medicine's Unified Medical Language System project, UMLS; McCray & Hole, 1990). As information content and collections become even larger and more dynamic, we believe a system-aided, bottom-up, artificial intelligence (AI) approach is needed. By applying scalable techniques developed in various AI subareas (and related fields) such as image segmentation and indexing, voice recognition, natural language processing, neural networks, machine learning, clustering and categorization, and intelligent

agents, we could provide an alternative system-aided approach to addressing both information overload and information interoperability.

Digital Libraries, Knowledge Networking, and Semantic Interoperability

The Information Infrastructure Technology and Applications (IITA) Working Group, the highest level of the country's National Information Infrastructure (NII) technical committee, held an invited workshop in May 1995 to define a research agenda for digital libraries. (See <http://Walrus.Stanford.EDU/diglib/pub/reports/iita-dlw/main.html>)

The shared vision is an entire Net of distributed repositories, where objects of any type can be searched within and across different indexed collections (Schatz & Chen, 1996). In the short term, technologies must be developed to transparently search across these repositories, handling any variations in protocols and formats (i.e., addressing structural interoperability; Paepcke et al., 1996). In the long term, technologies must be developed to handle the variations in content and meanings transparently as well. These requirements are steps along the way toward matching the concepts requested by users with objects indexed in collections (Schatz, 1997).

The ultimate goal, as described in the IITA report, is the Grand Challenge of Digital Libraries:

Deep semantic interoperability—the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations. . . . Achieving this will require breakthroughs in description as well as retrieval, object interchange and object retrieval protocols. Issues here include the definition and use of metadata and its capture or computation from objects (both textual and multimedia), the use of computed descriptions of objects, federation and integration of heterogeneous repositories with disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for

© 1998 John Wiley & Sons, Inc.

JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE. 49(0):00-00, 1998

CCC 0002-8231/98/000000-00

/ 8n49\$\$1127

01-08-98 18:36:19

jasli

W: JASIS

1127

e

automatic rating, ranking, and evaluation of information quality, genre, and other properties.

Attention to semantic interoperability has prompted several of the NSF/DARPA/NASA funded large-scale digital library initiative (DLI) projects to explore various artificial intelligence, statistical, and pattern recognition techniques, e.g., concept spaces and category maps in the Illinois project (Schatz et al., 1996), textile and word sense disambiguation in the Berkeley project (Wilensky, 1996), voice recognition in the CMU project (Wactlar, Kanade, Smith, & Stevens, 1996), and image segmentation and clustering in the UCSB project (Manjunath & Ma, 1996).

The ubiquity of online information as perceived by US leaders (e.g., "Information President" Clinton and "Information Vice President" Gore) as well as the general public, and recognition of the importance of turning information into knowledge have continued to push information and computer science researchers toward developing scalable artificial intelligence techniques for other emerging information systems applications.

In the Santa Fe Workshop on Distributed Knowledge Work Environments: Digital Libraries, held in March 1997, the panel of digital library researchers and practitioners suggested three areas of research for the planned Digital Library Initiative-2 (DLI-2): System-centered issues, collection-centered issues, and user-centered issues. Scalability, interoperability, adaptability and durability, and support for collaboration are the four key research directions under system-centered issues. System interoperability, syntactic (structural) interoperability, linguistic interoperability, temporal interoperability, and semantic interoperability are recognized by leading researchers as the most challenging and rewarding research areas. (See <http://www.si.umich.edu/SantaFe/>)

In a new NSF Knowledge Networking (KN) initiative, a group of domain scientists and information systems researchers was invited to a Workshop on Distributed Heterogeneous Knowledge Networks at Boulder, Colorado, in May 1997. Scalable techniques to improve semantic bandwidth and knowledge bandwidth are considered among the priority research areas, as described in the KN report (see <http://www.scd.ucar.edu/info/KDI/>):

The Knowledge Networking (KN) initiative focuses on the integration of knowledge from different sources and domains across space and time. Modern computing and communications systems provide the infrastructure to send bits anywhere, anytime, in mass quantities—radical connectivity. But connectivity alone cannot assure (1) useful communication across disciplines, languages, cultures; (2) appropriate processing and integration of knowledge from different sources, domains, and non-text media; (3) efficacious activity and arrangements for teams, organizations, classrooms, or communities, working together over distance and time; or (4) deepening understanding of the ethical, legal, and social implications of new developments in connectivity, but not interactivity

and integration. KN research aims to move beyond connectivity to achieve new levels of interactivity, increasing the semantic bandwidth, knowledge bandwidth, activity bandwidth, and cultural bandwidth among people, organizations, and communities.

A Scalable Artificial Intelligence Approach to Supporting Semantic Interoperability

Among the artificial intelligence techniques (and the affiliated statistical and pattern recognition fields) that are considered scale and domain independent, the following classes of algorithms and methods have been examined and subjected to experimentation in various digital library, multimedia databases, and information science applications:

Object Recognition, Segmentation, and Indexing

The most fundamental techniques in IR involve identifying key features in objects. For example, automatic indexing and natural language processing (e.g., noun phrase extraction or object-type tagging) are frequently used to automatically extract meaningful keywords or phrases from texts (Salton, 1989). Texture, color, or shape-based indexing and segmentation techniques are often used to identify images (Manjunath & Ma, 1996). For audio and video applications, voice recognition, speech recognition, and scene segmentation techniques can be used to identify meaningful descriptors in audio or video streams (Wactlar et al., 1996).

Semantic Analysis

Several classes of techniques have been used for semantic analysis of texts or multimedia objects. Symbolic machine learning (e.g., ID3, version space), graph-based clustering and classification (e.g., Ward's hierarchical clustering), statistics-based multivariate analyses (e.g., latent semantic indexing, multi-dimensional scaling, regressions), artificial neural network-based computing (e.g., backpropagation networks, Kohonen self-organizing maps), and evolution-based programming (e.g., genetic algorithms) are among the popular techniques (Chen, 1995). In this information age, we believe these techniques will serve as good alternatives for processing, analyzing, and summarizing large amounts of diverse and rapidly changing multimedia information.

Knowledge Representations

The results from a semantic analysis process could be represented in the form of semantic networks, decision rules, or predicate logic. Many researchers have attempted to integrate such results with existing human-created knowledge structures such as ontologies, subject headings, or thesauri (McCray & Hole, 1990). Spreading activation-based inferencing methods are often used to tra-

verse various large-scale knowledge structures (Chen & Ng, 1995).

Human-Computer Interactions (HCI) and Information Visualization

One of the major trends in almost all emerging information systems applications is the focus on user-friendly, graphical, and seamless HCI. The Web-based browsers for texts, images, and videos have raised user expectation on the rendering and manipulation of information. Recent advances in development languages and platforms such as Java, OpenGL, and VRML, and the availability of advanced graphical workstations at affordable prices have also made information visualization a promising area for research (DeFanti & Brown, 1990). Several of the digital library research teams including Arizona/Illinois, Xerox PARC, Berkeley, and Stanford, are pushing the boundary of visualization techniques for dynamic displays of large-scale information collections.

In This Issue

This theme issue consists of five articles that report research in adopting artificial intelligence techniques for emerging information systems applications. The first two articles address large-scale semantic analysis and dynamic searching of textual documents. The third and fourth articles describe experiments on image and video collections, respectively. The last article presents a prototype Web-based collaborative tool. These articles highlight some of the cutting-edge projects that attempt to trailblaze a path to semantic interoperability.

In the "Internet Browsing and Searching" article, the authors describe their experiments in adopting concept space and category map techniques for semantic classification and retrieval of noisy Internet homepages. The underlying techniques include automatic indexing, cluster analysis, and unsupervised neural network learning. The test collection was based on 110,000 entertainment-related Web documents previously extracted by several spider programs. Their experiments showed the scalability of techniques and potential usefulness for searching and browsing of large collections of documents. The demo system, called ET-Map, is available for general access at: <http://ai.bpa.arizona.edu/ent/> Using similar indexing techniques, but relying on a genetic, algorithm-based global, stochastic search method, the second article ("A Smart Itsy Bitsy Spider for the Web") reports the design of a dynamic, customizable Web spider that is built on a Java interface. In a user evaluation experiment, the genetic algorithm spider, acting as an intelligent, personal search agent, outperformed a benchmark best first search spider in recall. The demo system is also available for general access at: <http://ai.bpa.arizona.edu/~mramsey/SPIDER/itsy.html>

While the first two articles describe research for textual

documents, the next two articles report findings for experiments using multimedia documents. In the "Digital Video Library" article by Witbrock and Hauptmann, the researchers applied speech recognition, natural language processing, information retrieval, and image analysis techniques for a digital video library. Their experimental results showed that transcripts generated by speech recognition can make digital video collections more searchable. The "Texture Thesaurus" article by Ma and Manjunath describes an experiment that applied image indexing and segmentation techniques for large-scale aerial photos. A hybrid neural network algorithm was then used to identify visual similarity by clustering patterns in the feature space. Preliminary results demonstrate the potential usefulness of the design in searching over a large collection of airphotos.

Lastly, the "Web-Based Group Support System" article describes the architecture, design, and development of GroupSystems for the Web. The prototype system is currently being used by teams all over the world, while design and development continue. The Web-based collaborative tools provide an alternative graphical, distributed, virtual workspace for teams of knowledge workers.

References

- Chen, H. (1995). Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*, 46, 194-216.
- Chen, H., & Ng, D. T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science*, 46, 348-369.
- DeFanti, T., & Brown, M. (1990). Visualization: Expanding scientific and engineering research opportunities. In *Visualization in Scientific Computing*. New York: IEEE Computer Society Press.
- Manjunath, B. S., & Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8), 837-841.
- McCray, A. T., & Hole, W. T. (1990). The scope and structure of the first version of the UMLS semantic network. In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, November 4-7, 1990, (pp. 126-130). Los Alamitos, CA: Institute of Electrical and Electronics Engineers.
- Paepcke, A., Cousins, S. B., Garcia-Molino, H., Hasson, S. W., Ketcxhpel, S. P., Roscheisen, M., & Winograd, T. (1996). Using distributed objects for digital library interoperability. *IEEE Computer*, 29(5), 61-69.
- Salton, G. (1989). *Automatic text processing*. Reading, MA: Addison-Wesley.
- Schatz, B. R. (1997). Information retrieval in digital libraries: Bring search to the net. *Science*, 275, 327-334.
- Schatz, B. R., & Chen, H. (1996). Building large-scale digital libraries. *IEEE Computer*, 29(5), 22-27.
- Schatz, B. R., Mischo, B., Cole, T., Hardin, J., Bishop, A., & Chen, H. (1996). Federating repositories of scientific literature. *IEEE Computer*, 29(5), 28-36.
- Wactlar, H. D., Kanade, T., Smith, M. A., & Stevens, S. M. (1996). Intelligent access to digital video: Informedia project. *IEEE Computer*, 29(5), 46-53.
- Wilensky, R. (1996). Toward work-centered digital information services. *IEEE Computer*, 29(5), 37-45.