SEMANTIC ISSUES FOR DIGITAL LIBRARIES

Hsinchun Chen

INTRODUCTION

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, 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 World Wide Web, 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). 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 apply-

ing 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 can provide an alternative system-aided approach to addressing both information overload and information interoperability.

FEDERAL INITIATIVES: DIGITAL LIBRARIES AND OTHERS

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/iitadlw/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 search transparently 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 automatic rating, ranking, and evaluation of information quality, genre, and other properties. (p. 5)

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 et al., 1996), and image segmentation and clustering in the UCSB project (Manjunath & Ma, 1996).

The ubiquity of online information as perceived by U. S. 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 described in the KN report (see http://www.scd.ucar.edu/info/KDI/).

The Knowledge Networking 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, and in mass quantities—"radical connectivity." But connectivity alone cannot assure: (1) useful communication across disciplines, languages, and cultures; (2) appropriate processing and integration of knowledge from different sources, domains, and nontext media; (3) efficacious activity and arrangements for teams, organizations, classrooms, or communities working together over distance and time; or (4) a 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 semantic bandwidth, knowledge bandwidth, activity bandwidth, and cultural bandwidth among people, organizations, and communities.

SEMANTIC ISSUES FOR DIGITAL LIBRARIES

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 libraries, 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 extract automatically 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 (Wactler 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, multidimensional scaling, regressions), artificial neural network-based computing (e.g., back propagation 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. Spreading activation-based inferencing methods are often used to traverse various large-scale knowledge structures (Chen & Ng, 1995).

HUMAN-COMPUTER INTERACTIONS 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 expectations 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.

ILLINOIS DLI SEMANTIC RESEARCH: AN EXAMPLE

In this section, we present an example of selected semantic retrieval and analysis techniques developed by The University of Arizona Artificial Intelligence Lab (AI Lab) for the Illinois DLI project. For detailed technical discussions, readers are referred to Chen et al. (1996, 1998). A textual semantic analysis pyramid was developed by The University of Arizona AI Lab to assist in semantic indexing, analysis, and visualization of textual documents. The pyramid, as depicted in Figure 1, consists of four layers of techniques, from bottom to top: noun phrase indexing, concept association, automatic categorization, and advanced visualization.

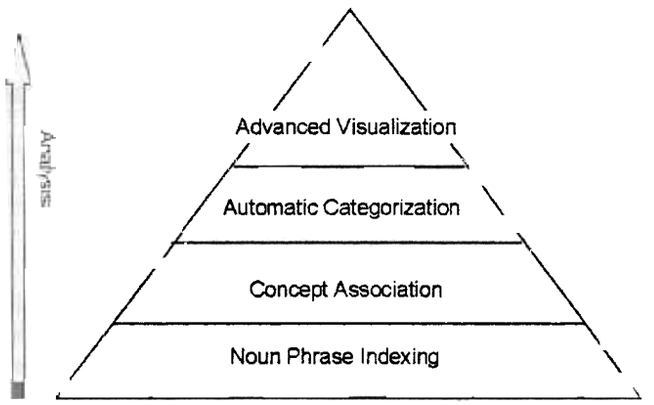

Figure A Textual Semantic Analysis Pyramid

Noun phrase indexing aims to identify concepts (grammatically correct noun phrases) from a collection for term indexing. It begins with a text tokenization process to separate punctuation and symbols. It follows by part-of-speech-tagging (POST) using variations of the Brill tagger and thirty-plus grammatic noun phrasing rules. Figure 2 shows an example of tagged noun phrases for a simple sentence (the system is referred to as AZ Noun Phraser). For example, "interactive navigation" is a noun phrase that consists of an adjective (A) and a noun (N).

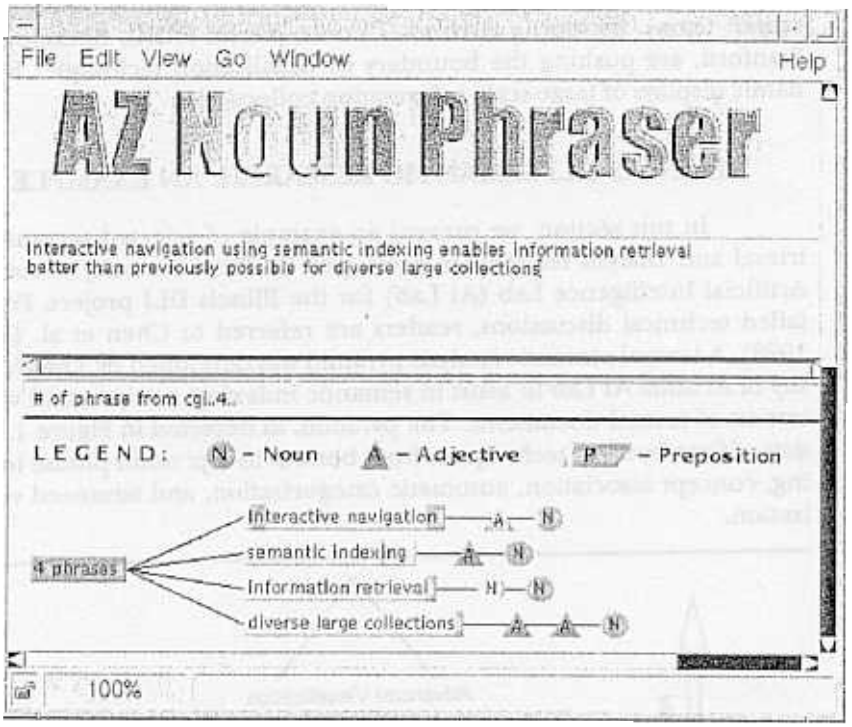

Figure 2. Tagged Noun Phrases

Concept Association: Concept association attempts to generate weighted, contextual concept (term) association in a collection to assist in concept-based associative retrieval. It adopts several heuristic term weighting rules and a weighted co-occurrence analysis algorithm. Figure 3 shows the associated terms for "information retrieval" in a sample collection of project reports of the DARPA/ITO Program–TP (Term Phrase) such as "R system," "information retrieval engine," "speech collection," and so on.

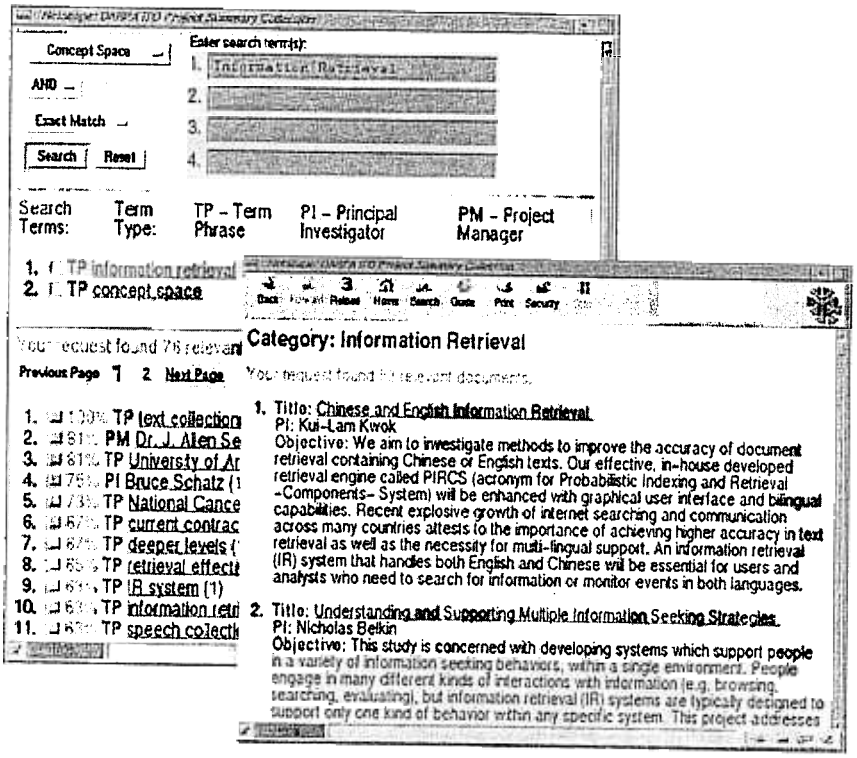

Figure 3. Associated Terms for "Information Retrieval"

Automatic Categorization: A category map is the result of performing a neural network-based clustering (self-organizing) of similar documents and automatic category labeling. Documents that are similar (in noun phrase terms) to each other are grouped together in a neighborhood on a two-dimensional display. As shown in the colored jigsaw-puzzle display in Figure 4, each colored region represents a unique topic that contains similar documents. Topics that are more important often occupy larger regions. By clicking on each region, a searcher can browse documents grouped in that region. An alphabetical list that is a summary of the 2D result is also displayed on the left-hand side of Figure 4—e.g., Adaptive Computing System (thirteen documents), Architectural Design (nine documents), and so on.

Advanced Visualization: In addition to the 2D display, the same clustering result can also be displayed in a 3D helicopter fly-through landscape as shown in Figure 5, where cylinder height represents the number of documents in each region. Similar documents are grouped in a

Figure 4. Category Map

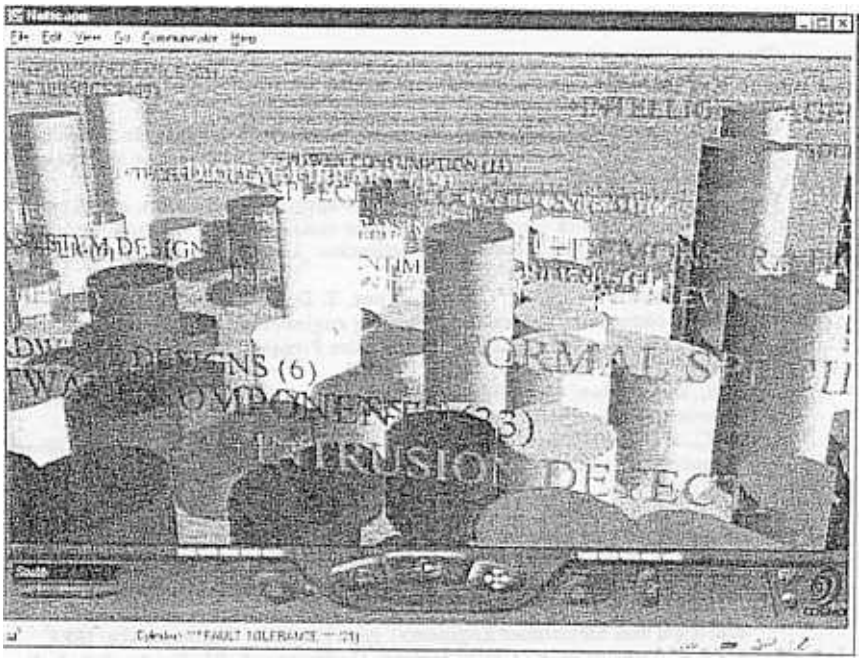

Figure 5. VRML Interface for Category Map

same-colored region. Using a VRML plug-in (COSMO player), a searcher is then able to "fly" through the information landscape and explore interesting topics and documents. Clicking on a cylinder will display the underlying clustered documents.

DISCUSSIONS

The techniques discussed above were developed in the context of the Illinois DLI project, especially for the engineering domain. The techniques appear scalable and promising. We are currently in the process of fine-tuning these techniques for collections of different sizes and domains.

ACKNOWLEDGMENT

This project was funded primarily by: (1) NSF/CISE "Concept-based Categorization and Search on Internet: A Machine Learning, Parallel 6 Computing Approach," NSF IRI9525790, 1995-1998, and (2) NSF/ARPA/ NASA Illinois Digital Library Initiative project, "Building the Interspace: Digital Library Infrastructure for a University Engineering Community," NSF IRI9411318, 1994-1998.

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*(3), 194-216.

Chen, H.; Houston, A. L.; Sewell, R. R.; & Schatz, B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science, 49*(7), 582-603.

Chen, H., & Ng, D. T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound versus connectionist Hopfield net activation. *Journal of the American Society for Information Science, 46*(5), 348-369.

Chen, H.; Schatz, B. R.; Ng, T. D.; Martinez, T. D.; Kirchhoff, A. J.; & Lin, C. (1996). A parallel computing approach to creating engineering concept spaces for semantic retrieval: The Illinois Digital Library Initiative Project. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 771-782.

DeFanti, T., & Brown, M. (1990). *Visualization: Expanding scientific and engineering research opportunities.* New York: IEEE Computer Society Press.

Lynch, C., & Garcia-Molina, H. (1995). *Interoperability, scaling, and the digital libraries agenda.* Unpublished report on the May 18-19, 1995 IITA Digital Libraries Workshop, August 22, 1995.

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 14th Annual Symposium on Computer Applications in Medical Care* (held November 4-7, 1990, Los Alamitos, California, Institute of Electrical and Electronics Engineers) (pp. 126-130). Waltham, MA: IEEE.

Paepcke, A.; Cousins, S. B.; Garcia-Molino, H.; Hasson, S. W.; Ketchpel, 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 Publishing Company, Inc.

Schatz, B. R. (1997). Information retrieval in digital libraries: Bring search to the net. *Science, 275*(January 17), 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. P.; & 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.