

The Interspace: Concept Navigation Across Distributed Communities



Within the next decade, computing technology will transform the Internet into the Interspace, an information infrastructure that supports semantic indexing and concept navigation across widely distributed community repositories.

Bruce R. Schatz

University of Illinois
at Urbana-
Champaign

Information sharing has always been one of the Internet's most popular functions. Although large archives for home shopping and scientific databases dominate Web traffic, personal sites for small communities are exploding in number. This trend is similar to what the US Defense Department's Arpanet experienced in the 1970s, when interpersonal communication superseded remote computation as the network's primary service. The Internet is approaching the limit of its effectiveness as a knowledge-exchange mechanism and must evolve as the Arpanet did. The ability of average users to create and maintain their own homepages has created a glut of such sites, making it increasingly difficult to retrieve desired information through Web browsers or search engines.

With billions of documents now online, fixed links are no longer an effective navigation tool. Searches return so many results that, to identify relevant documents, the ability to dynamically create links during user sessions is necessary. This requires a quantum leap in functionality—from looking for words in texts to automatically identifying documents containing related concepts.

The Internet is evolving into a new information infrastructure capable of serving widely distributed communities. Unlike the Internet, the *Interspace* will directly support interaction with abstraction.

Using technologies that go beyond searching individual repositories to analyzing and correlating knowledge across multiple sources and subjects, the Interspace will offer distributed services to transfer concepts across domains, just as the Arpanet used distributed services to transfer files across machines and the Internet uses distributed services to transfer objects across repositories.

Standard protocols for the emerging information infrastructure will support searching knowledge collections maintained and indexed by specialized communities and residing directly on users' personal machines. These protocols will automatically interconnect related logical spaces, letting individuals navigate across community repositories rather than searching for interlinked objects within physical networks. *Concept navigation* will become a standard function in the Interspace just as document browsing is in the Internet.

At the Community Architectures for Network Information Systems (CANIS) Laboratory, we have developed a working Interspace prototype (<http://www.canis.uiuc.edu/interspace-prototype>) that uses scalable technologies for concept extraction and concept navigation. We have successfully tested these technologies, which compute contextual frequency of document phrases within a community repository, on discipline-scale, real-world collections.

Evolution of the Global Information Infrastructure

The Interspace represents the third wave in the ongoing evolution of the global information infrastructure, driven by rapid advances in computing and information technology during the past 35 years.

As Figure A indicates, the first wave involved advances in accessing and transmitting data. It began in 1965 with the US Defense Department's Arpanet and continued during the next 20 years as large-scale distributed file systems evolved. Researchers focused on transparently transferring data packets—raw bits and files—from one machine to another, efforts that culminated in the development of e-mail and file sharing.

Technological breakthroughs in organizing and retrieving information ushered in the second wave. This phase was characterized by a transition from packets to objects, which contain display and interaction software for groups of packets.

The concept of document browsing on the Internet began with distributed multimedia information networks such as the telesophy system, conceived in 1984.¹ By 1989, the telesophy prototype had demonstrated the technical feasibility of using interlinked objects to transparently navigate across the global information network, foretelling the coming of worldwide information spaces. That same year, the World Wide Web prototype was conceived at the European Organization for Nuclear Research.

In 1994, Mosaic, the first widely distributed graphical browser for the Web, was developed at the University of Illinois's National Center for Supercomputing Applications. Mosaic demonstrated that the telesophy paradigm could be implemented efficiently enough to become the mass standard for a new information infrastructure. Coupled with Mosaic, telesophy technology made

worldwide information spaces an everyday reality.

This second wave took a decade to peak in functionality, and it has continued during the past five years with the consolidation of document browsing technologies in commercial distribution.

The transition about to occur in the third wave will involve concept navigation, a radical new paradigm for network information retrieval. Concepts, which contain indexing and meaning for groups of objects, are useful for analyzing content and correlating knowledge. During this phase, information retrieval will move beyond searching individual repositories to analyzing heterogeneous data across sources and subjects.

Just as the telesophy prototype led to Mosaic, the Interspace prototype will lead to a widely used system with stan-

dard protocols that support direct interaction with community knowledge. Each specialized community will maintain its own repositories, indexing these collections on their own machines. The Interspace will interconnect all these knowledge spaces, enabling switching across communities by navigating concepts across repositories.

Within five years, the Interspace will be incorporated into the information infrastructure. Within a decade, concept navigation will be a ubiquitous commercial service on the global network.

Reference

1. B. Schatz, "Telesophy: A System for Manipulating the Knowledge of a Community," *Proc. IEEE Global Comm. Conf. (Globecom 87)*, IEEE Press, Piscataway, N.J., 1987, pp. 1181-1186.

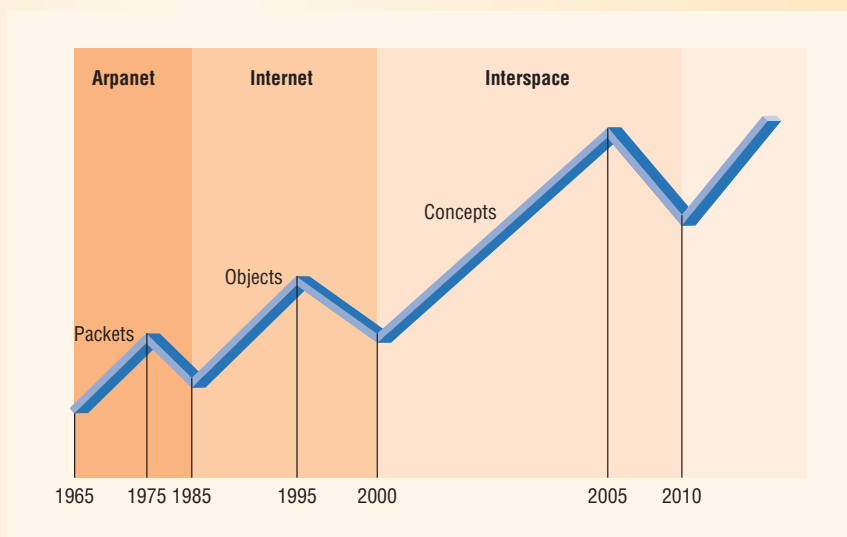


Figure A. Evolution of the global information infrastructure. The technological progress of knowledge exchange—from e-mail in the Arpanet to document browsing in the Internet to concept navigation in the coming Interspace—has occurred in three waves, each building on the previous one. The wave pattern roughly describes four distinct phases of functionality: fundamental research (trough), development of prototype systems (ascent), emergence of commercial systems (crest), and mass propagation (descent).

SEMANTIC INDEXING

As the "Evolution of the Global Information Infrastructure" sidebar indicates, the Interspace will radically transform how we interact with knowledge. Today, most online information is stored in archival centers containing large collections indexed by trained professionals. However, as information retrieval becomes standardized, independent communities will find it easier and more

effective to develop and maintain their own collections rather than transferring them to a central archive. In the near future, small collections maintained and indexed by individual communities will contain most online information.

Because nonprofessionals will classify most sources, the Interspace must provide interactive support for *semantic indexing* of community repositories. Amateur searchers rather than professional

Extracting concepts requires a semantic parser that can retrieve the appropriate units from documents of any subject domain.

librarians will likewise dominate information retrieval, so the Interspace must also provide interactive support for concept navigation. To enable true concept navigation, semantic indexing must incorporate five robust technologies that are readily adaptable to many applications and purposes:

- document representation,
- language parsing,
- statistical indexing,
- peer-to-peer networking, and
- concept switching.

Some of these technologies are already being widely implemented while others are relatively immature. The Interspace will become a reality once they all mature into commercial components and become part of the information infrastructure.

Document representation

The ability to store documents in a single format made global information retrieval possible. Prior to the widespread adoption of the Hypertext Markup Language (HTML), collections were limited to what a single central organization could administer, such as Dialog for bibliographic databases of journal abstracts and Lexis/Nexis for full-text databases of magazine articles.

Standard protocols implied that a single program—what became Web browsers—could retrieve documents from multiple sources. Developed by the National Center for Supercomputing Applications (NCSA), Mosaic—which became the forerunner of commercial browsers—combined streamlined standards and flexible interfaces to attract millions of users to information retrieval for the first time.¹ As the number of online documents increased exponentially, identifying the initial document for hypertext browsing became a major problem.

Search engines began to dominate browser interfaces as the primary interface to the global information infrastructure. However, these engines demonstrated the weakness of syntactic searches, such as word matching within Dialog, and increased the demand for semantic indexing.

Web designers responded by extending markup languages from formatting to typing. HTML has thus evolved into the Extensible Markup Language, which enables syntactic specification of many types of units, including custom applications. XML is the Web version of the Standard Generalized Markup Language (SGML) used in the publishing industry to tag document structures such as sections and fig-

ures. SGML has also been used for many years to tag phrase types in scholarly documents—for example, to indicate whether names in the humanities literature represent a person, place, book, or painting.

The markup languages currently under development² will let authors specify metadata describing semantics within a document's text for standard use in search engines and other software. For example, the Resource Description Framework builds a data model on top of XML as the first step toward creating what the World Wide Web Consortium calls the *semantic Web* (<http://www.w3.org/2001/sw/>). Many of these languages will support ontologies, which formally define the important concepts in a given domain. However, the languages' dependence on author reliability and accuracy, which is also problematic in SGML, has fostered the development of automatic tagging techniques to augment manual tagging when possible.

Language parsing

Because document standards eliminate the need for format converters for each collection, a syntax parser can universally extract words. Extracting concepts, however, requires a semantic parser that can retrieve the appropriate units from documents of any subject domain. Multiword noun phrases are the most discriminating unit for information retrieval in text documents.

The key to context-based semantic indexing is identifying the right-size units to extract. Over the years, the emergence of increasingly powerful computers has made concept extraction more precise. Initially, heuristic rules used stop words and verb phrases to approximate noun-phrase extraction. Simple noun-phrase grammars for particular subject domains followed. Finally, statistical parsing technology became accurate enough to compute extraction without explicit grammars. These parsers can automatically extract noun phrases for general texts after being trained on sample collections.³ It is even possible to determine the type of noun phrase, such as person or place, with high precision.

Semantic parsing is part of an overall trend in statistical pattern recognition in many areas. For example, in computational linguistics, the best noun-phrase extractors no longer have an underlying definite grammar but instead rely on neural nets trained on typical cases. The Tipster Text program (http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/) was a \$100 million 1990s effort by the Defense Advanced Research Projects Agency (DARPA) to extract facts from newspaper articles

for intelligence purposes. Its initial phases relied on grammars; however, in its final phases, the project used statistical parsers.

Once the parser extracts the units, such as noun phrases, they can be used to approximate meaning by computing their frequency across the collection. Just as noun phrases represent concepts, the contextual frequencies represent meanings. Today's computers let statistics replace rules, so that context becomes a practical substitute for semantics.

Statistical indexing

Semantic indexing can be accomplished by computing the statistical frequency of extracted noun phrases within each document in a collection. Frequencies for each phrase form a space where each concept is related to every other concept by co-occurrence. The search process uses this space to generate related concepts, then retrieves other documents containing these concepts. The space consists of the interrelationships between the concepts in the collection, which users can interactively navigate.

Researchers first used algorithms for computing statistical co-occurrence in the 1960s,⁴ but the computations were confined to collections of only a few hundred documents. Retrieving information from real-world collections of millions of documents relied instead on exact matching of text phrases, as in a full-text search. Even the largest computer could not semantically index the smallest collection.

In recent years, computers have become capable of extracting concepts from real collections and retrieving concept relationships. Community machines can perform semantic indexing on their repositories offline. For example, it takes a small laboratory server an hour to index 1,000 documents for a space with 10 people and a large departmental server three hours to index 10,000 documents for a 100-person community.

As computing power increases, indexing time will decrease from batch to interactive, making semantic indexing feasible on dynamically specified collections, such as those selected during a user session. Within the next decade, indexing barriers for all real-world collections will fall—the smallest computer will be able to semantically index the largest collection.⁵

Efforts are already under way to develop statistical indexing technology mature enough to incorporate into the information infrastructure. In 1992, the National Institute of Standards and Technology organized the Text Retrieval Conference (<http://trec.nist.gov>) as part of the DARPA Tipster Text program. TREC is now an annual public indexing com-

petition in which international teams use statistical software to generate semantic indexes for gigabyte document collections.

Peer-to-peer networking

The Web is fundamentally a client-server model, with few large servers and many small clients. The clients are typically user workstations that prepare queries for processing at archival servers. As the number of servers increases and the size of collections decreases, the global information infrastructure will evolve into a peer-to-peer (P2P) model in which each user machine is both a client and server at different times, directly exchanging data with other machines that contain the same protocols.

Simple P2P protocols already exist that let specialized communities share data sets without the intervention of central authorities. The best-known example is Napster, an extremely popular music-swapping service that enabled individuals to make specially formatted digital files accessible to other peer machines via a local program that supported the sharing protocol. Napster was eventually shut down because it lacked a searching capability to filter out copyrighted songs, but current anonymous file-swapping services such as GnutellaNet (<http://gnutella.wego.com/>) and Freenet (<http://freenet.sourceforge.net/>) let users download software that supports a sharing protocol onto their personal computers so that they can directly exchange files.

Scientific communities also use P2P protocols to facilitate research. These programs implement a simple service on a volunteer's machine where downloaded software performs small computations on limited data. Then a large computation on the complete database combines the results from many PCs at a central site. For example, millions of PCs around the world run SETI@home software (<http://setiathome.ssl.berkeley.edu/>) as a saver, each machine computing radio telescope survey results from a different sky region. The PCs send these results to a central repository for merged data on the search for extraterrestrial intelligence across the entire universe. In another example, the Intel Philanthropic Peer-to-Peer Program (<http://www.intel.com/cure>) uses the commercial version of this software to conduct public-service medical research using the PCs of more than one million volunteers.

The growth in community databases is spurring development of true data-centered P2P protocols, which are necessary for global search of local repositories. In these protocols, each PC computes

The growth in community databases is spurring development of true data-centered P2P protocols for searching local repositories.

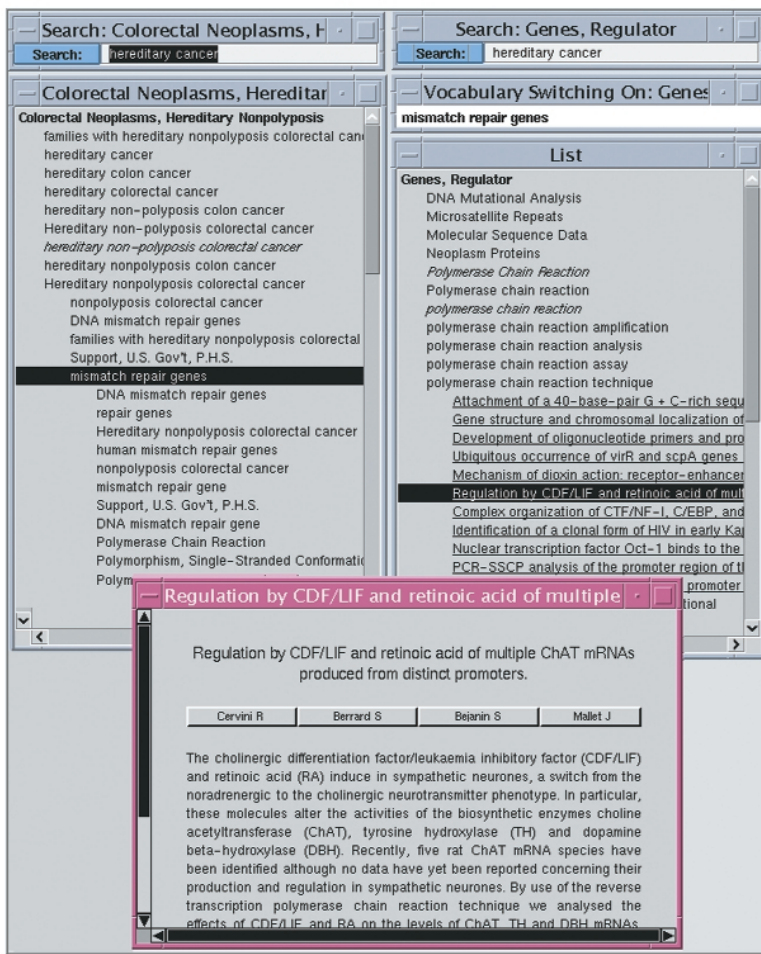


Figure 1. Concept switching in the Interspace prototype. By starting with the broad term “hereditary cancer” and using common terms as bridges, the user can locate the article displayed at the bottom without doing an explicit search. The source subject domain is “Colorectal Neoplasms, Hereditary Nonpolyposis” and the target subject domain is “Genes, Regulator.” When a straight text search for “mismatch repair genes” in the target subject domain returns no hits, the prototype invokes concept switching from the source to the target domain.

a locally administered database on its own, and the central site only merges the database computation from each local site.

Current Internet indexing projects such as Open Directory (<http://www.dmoz.org>) rely on distributed subject curators to index Web sites within assigned categories, the entries being entire collections. In contrast, using Interspace technology such as automatic subject assignment,⁶ distributed community curators will be able to index the documents within the collections. The technology will then aggregate these local indexes to provide global indexes.

Concept switching

The Internet only supports searching for objects, such as matching phrases within documents. Because specialized communities composed of small groups of distributed individuals will dominate future information sources, the Interspace’s infrastructure must explicitly support correlation

across communities through *concept switching*. Just as switching gateways in the Internet reliably transmits packets from one machine to another in a different location, concept switches in the Interspace effectively map concepts in one community repository to those in another community repository. Users will navigate transparently from concept to concept, within and across multiple repositories. Even if they do not know the specialized terminology beforehand, users will be able to identify relevant concepts from related research and then search for a desired object.

The difficulty in locating related research lies in extracting related concepts from distributed repositories that use their own terminology. Each community must identify the concepts within its own repository and index the concepts generically in a way suitable for mapping to concepts in other community repositories.

Concept-switching technologies are still immature. Vocabulary switching,⁷ a specialized form used since the 1970s, maps concepts across multiple subject thesauri, with relationships that human indexers specify. For example, the Unified Medical Language System (UMLS) Metathesaurus,⁸ developed at the US National Library of Medicine (NLM) (<http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html>), uses vocabulary switching to relate biomedical concepts. Vocabulary switching is expensive to maintain because it requires human tracking of thesaurus concepts by experts knowledgeable about both sides of the vocabulary map. The promise of automatic methods is effective mapping at viable cost.

Concept switching scenario

Figure 1 illustrates a scenario in which a biologist uses relationships within the concept spaces to navigate across two Medline community repositories. Within the subject domain “Colorectal Neoplasms, Hereditary Nonpolyposis,” the user enters “hereditary cancer” as a search term, which returns a list of all concepts that are lexical permutations. Indented levels in the display indicate the co-occurrence list’s hierarchy, so that the user can move from “hereditary nonpolyposis colorectal cancer” to the related “mismatch repair genes” in the concept space.

The user then looks for this term in another domain repository—“Genes, Regulator”—to understand how genes regulate colon cancer. A straight text search returns no hits, so the prototype invokes concept switching to transfer concepts from one domain to another across their respective concept spaces. The concept switch intersects “mismatch

repair genes” and all related terms from its indented co-occurrence list in the source concept space for “Colorectal Neoplasms” into the target concept space for “Genes, Regulator.”

After syntactic transformations, the concept switch produces a list of concepts computed to be semantically equivalent to “mismatch repair genes” within “Genes, Regulator.” In this implementation, the equivalence computation is based on the source space’s list of related concepts. Switching occurs by bridging across community repositories using the shared term “Polymerase Chain Reaction,” which represents an experimental technique common to both subject domains. Navigating the concept space down to the object level locates the document displayed at the bottom, which discusses a “leukaemia inhibitory factor” related to colon cancer.

INTERSPACE PROTOTYPE

The Interspace prototype is at the same state of technological maturity as the Internet research prototypes were just a few years before the development of Mosaic. We are running a fully operational prototype in our laboratory that has semantically indexed large-scale real-world collections and supports concept navigation across multiple sources. The technology is ready for widespread deployment that will catalyze the Interspace for concept navigation, just as Mosaic catalyzed the Internet for document browsing.

The product of more than a decade of research, the Interspace prototype was conceived in the late 1980s during an analysis⁹ of the telesophy project, an early system for supporting worldwide information spaces. The Worm Community System (1990-1994),¹⁰ which implemented a complete analysis environment for an international molecular biology community, provided the model for the prototype architecture. The algorithms for semantic indexing were developed as part of the Illinois Digital Library project (1994-1998).¹¹ We developed the Interspace prototype itself from 1997-2000 with support from the DARPA Information Management program.¹²

Services and indexes

The Interspace prototype consists of:

- a suite of indexing services with protocols that support semantic indexing of community collections; and
- an analysis environment, which utilizes these indexes to navigate within and across collections at abstract levels.

The indexing suite reproduces automatically, for any collection, equivalents to standard physical library indexes. These indexes represent abstract spaces for concepts and categories above concrete collections of units and objects. Services in the prototype generate the semantic indexes for all collections within the spaces. The indexing services extract a common set of concepts and use these concepts to create concept spaces and category maps.

A *concept space* records the co-occurrence of units within objects, such as words within documents or textures within images. The technology operates generically, independent of subject domain. Interactive navigation of the concept space is useful for locating related terms relevant to a particular search strategy.

Like a subject thesaurus, a concept space lists alternative words for the user to consider while searching—if the specified word doesn’t retrieve the desired documents, the user can try another word that appears together with it in a different context.

Experience indicates that augmentation, not automation, of human performance is technologically feasible for semantic retrieval. Because concept spaces provide semiautomatic categorization, they are useful for interactive retrieval, with machine suggestion but human selection.

A *category map* records co-occurrences between objects within concepts, such as two documents with overlapping concept words. Like a subject classification, a category map can identify clusters of similar objects for browsing and locate which subcollection to search for desired items.

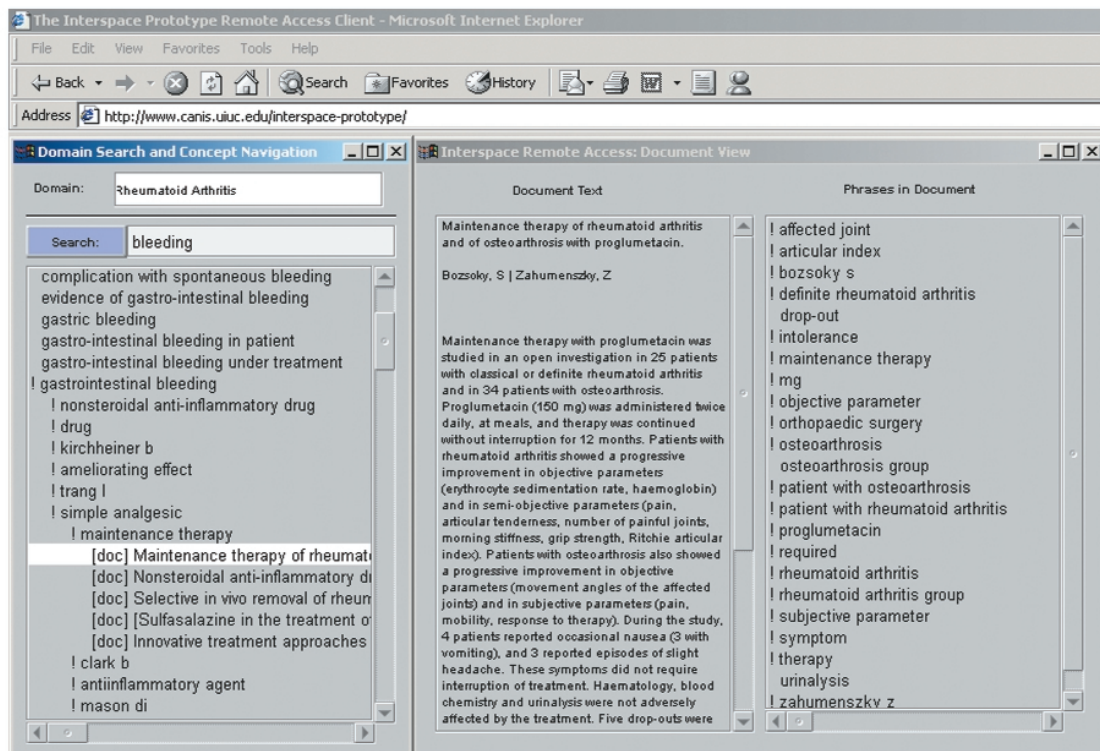
Scalable semantics

The Interspace prototype supports concept navigation using *scalable semantics* to index arbitrary collections. Scalability refers to the breadth of subject domains the technology can index; semantics refers to the depth of underlying meaning the indexing can capture. While traditional technologies have been either broad but shallow (full-text search) or narrow but deep (expert systems), scalable semantics strives for a balance between the two extremes for shallow parsing of large collections and deep parsing of small collections.

The parsing function extracts generic units from the objects, while the indexing function statistically correlates these units uniformly across sources. For text documents, the generic units are noun phrases, while the statistical indexes record co-occurrence frequency—how often each phrase occurs with every other phrase in a document within the collection.

The prototype technology is ready for widespread deployment that will catalyze the Interspace for concept navigation, just as Mosaic catalyzed the Internet for document browsing.

Figure 2. Concept navigation in the Interspace prototype. The upper bar in the left window indicates the subject domain being searched. After the user enters a phrase in the search bar, the system displays related concepts. Indented subentries represent a further navigation within the concept space. The window on the right displays a document located during the concept navigation.



Scalable semantics makes concept mapping viable for community-scale collections. It can support cluster-to-cluster—rather than term-to-term—concept switching by automatically parsing all concepts and computing all relationships. Cluster-to-cluster switching generates an equivalence class for related concepts in a particular situation, then maps the equivalence class from one space into the most relevant classes in other spaces. Full concept switching in the future will likely use neural net technology, such as spreading activation across related concepts in related documents.

IMPLEMENTATION

The Interspace prototype enables navigation across different levels of spaces for documents, concepts, and categories. Users can navigate spaces from concept to concept in an integrated analysis environment where they can transparently navigate individual indexes of the various collections. For example, users can easily move from a category map to a concept space, to concepts, to documents, then back up again to higher levels of abstraction from concepts mentioned in the document.

The interface is implemented in VisualWorks Smalltalk, but users interact via an emulator created with Applied Reasoning's ClassicBlend. This software dynamically transforms Smalltalk graphics into Java graphics, enabling Web browser invocation.

Concept navigation scenario

Figure 2 illustrates concept navigation within a community repository. In this scenario, the user is a physician looking for an analgesic drug that reduces

arthritis pain without causing gastrointestinal bleeding. After selecting the subject domain “Rheumatoid Arthritis” from a list of semantically indexed medical collections, the physician searches for all noun phrases containing “bleeding.” Under “gastrointestinal bleeding,” related concepts include general terms (“drug”), artifacts (“ameliorating effect”), names (“trang l”), and specific concepts (“simple analgesic”). Navigating in concept space from “simple analgesic” to “maintenance therapy” locates the displayed document, which discusses a new drug called proglumetacin that does not “adversely affect” a patient’s “haematology, blood chemistry, and urinalysis”—does not cause bleeding.

This document would have been difficult to retrieve by a standard text search on the NLM’s Medline database (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>) by someone unfamiliar with the terminology actually used. To learn more, the user can start with phrases such as “proglumetacin” to navigate directly to other related concepts within this collection or to other documents containing those concepts.

Concept extraction

The Interspace prototype comprises an analysis environment across multiple indexes of multiple sources. The analysis processes each source separately, but all sources are available within the environment. Processing a source involves uniformly extracting the concepts from each object. With documents, the prototype parses and records the position of every noun phrase. It then uses these noun phrases to compile a series of indexes at different

levels of abstraction. Having a common set of phrases implies that multiple indexes can uniformly reference a single phrase. Because phrases represent concepts, users can transparently navigate between concepts across indexes and sources.

We developed the prototype's current concept extractor using standard components for noun-phrase extraction over general text documents. After experimenting with several research and commercial systems, we created an effective parser from public domain source code. The noun-phrase extractor is based on the Brill tagger¹³ and NPtool's noun-phrase identification rules. The parser itself has several major stages.¹⁴ It first extracts words from the documents and canonicalizes them; then builds and canonicalizes phrases; and finally tags the parts of speech, identifying the noun phrases.

We chose this software because it is generic. The trained lexicon was derived from several sources, including the *Wall Street Journal* and Brown corpora, and thus offers fairly general coverage of the English language. The lexicon can be applied across subject domains without further customization while maintaining a comparable parsing quality. In our studies, a noun parser enhanced with the UMLS lexicon performed slightly better than the generic version on a collection of 630,000 Medline abstracts, but the difference is not statistically significant. A similar parser also works well on certain classes of gray-scale images, specifically aerial photographs, using texture density as the extracted units.

Our research indicates that the parser's generic nature without requiring customization enables a full range of noun-phrase parsing. In a careful evaluation using the complete biomedical literature, the automatic concept extractor parsed 45 million unique noun phrases from 10 million Medline abstracts.¹⁵

Context analysis

We have used concept space algorithms in numerous experiments to generate and integrate multiple semantic indexes. The space consists of the interrelationships between the concepts in the collection. Interactive navigation of the concept space is useful for locating related terms relevant to a particular search strategy. Creating a concept space involves using a noun-phrase parser to find the context of terms within documents and then using co-occurrence analysis to compute term or noun-phrase relationships.

Co-occurrence analysis computes the contextual relationships between noun phrases within the collections. The analysis processes the documents in

the collections one by one, relating two concepts when they occur together within the same document. Multiple-word terms receive heavier weights than single-word terms because they usually convey more precise semantic meaning. The relationships between noun phrases reflect the strength of their context associations within a collection. The analysis ranks co-occurring concepts in decreasing order of similarity.

SIMULATING THE INTERSPACE

We have performed several discipline-scale experiments using high-end supercomputers to simulate the Interspace. In these experiments, we partitioned a large existing collection into many community repositories to model typical situations. We performed semantic indexing on each community repository to investigate strategies for concept navigation and switching within and across repositories.

In 1998, following a computation on engineering abstracts carried out as part of the Illinois Digital Library project,¹¹ we processed the complete literature for medicine, the largest scientific discipline. Medspace (<http://www.canis.uiuc.edu/projects/medspace/>) was a semantic index of all Medline records,¹⁶ the back files of which comprise 10 million abstracts in a broad and deep database. Using the NLM Medical Subject Heading classification, we partitioned the collection into approximately 10,000 community repositories and computed concept spaces for each repository. Using multiple classifications for each abstract expanded the total number of abstracts by a factor of four.

We used the 128-node, 64-Gbyte SGI/Cray Origin 2000, implementing parallel algorithms for improved performance. The complete Medspace included 400 million phrase occurrences within 40 million abstracts—an order of magnitude greater than the engineering computation for the same computing time of two days. Figures 1 and 2 show concept navigations across sample repositories from this computation, used experimentally by physicians and biologists.

In less than a decade, users will routinely perform semantic indexing on their personal computers.

Regardless of indexing style, discipline-scale collections will be processed in real time on desktop computers and within minutes on handheld devices. Ordinary people will be able to reproduce the supercomputer experiments of the 1990s on their watches.

We have performed several discipline-scale experiments using high-end supercomputers to simulate the Interspace.

Eventually, the information infrastructure will support path matching, in which a user's navigation patterns are placed within the context of all available knowledge.

Semantic indexing will later extend beyond concepts and categories to *perspectives*, which relate concepts within categories, and *situations*, which relate categories within collections. Increasing levels of computing power will enable larger units to be grouped together across multiple relationships. Eventually, the information infrastructure will support path matching, in which a user's navigation patterns are placed within the context of all available knowledge. These more abstract semantic levels will lead to a closer matching of the meanings in the user's mind to the world's objects. ■

Acknowledgments

The DARPA Information Management program, under manager Ron Larsen, provided financial support for the Interspace prototype through contract N66001-97-C-8535, with the author as principal investigator. Charles Herring at the University of Illinois and Hsinchun Chen at the University of Arizona were co-principal investigators. Herring, Bill Pottenger, Kevin Powell, Dorbin Ng, Dmitri Roussinov, Marshall Ramsey, and Kris Tolle served as technical leads. The primary programmers were Conrad Chang, Les Tyrrell, Yiming Chung, Dan Pape, Qin He, and Nuala Bennett. Duncan Lawrie and Bob McGrath evaluated computer power for semantic indexing.

References

1. B. Schatz and J. Hardin, "NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet," *Science*, 12 Aug. 1994, pp. 895-901.
2. D. Fensel et al., "The Semantic Web and Its Languages," *IEEE Intelligent Systems*, Nov./Dec. 2000, pp. 67-73.
3. T. Strzalkowski, "Natural Language Information Retrieval," *Information Processing & Management*, vol. 31, 1996, pp. 397-417.
4. P. Kantor, "Information Retrieval Techniques," *Ann. Rev. Information Science and Technology*, vol. 29, 1994, pp. 53-90.
5. B. Schatz, "Information Retrieval in Digital Libraries: Bringing Search to the Net," *Science*, 17 Jan. 1997, pp. 327-334.
6. Y. Chung et al., "Automatic Subject Indexing Using an Associative Neural Network," *Proc. 3rd Int'l ACM Conf. Digital Libraries*, ACM Press, New York, 1998, pp. 59-68.
7. R. Niehoff, "Development of an Integrated Energy Vocabulary and the Possibilities for On-line Subject Switching," *J. Am. Soc. Information Science*, Jan./Feb. 1976, pp. 3-17.
8. D. Lindberg et al., "The Unified Medical Language System," *Methods Information Medicine*, vol. 32, 1999, pp. 281-291.
9. B. Schatz, *Interactive Retrieval in Information Spaces Distributed Across a Wide-Area Network*, tech. report 90-35, Dept. Computer Science, Univ. of Arizona, Tucson, 1990.
10. B. Schatz, "Building an Electronic Community System," *J. Management Information Systems*, vol. 8, winter 1991-1992, pp. 87-107.
11. B. Schatz et al., "Federated Search of Scientific Literature," *Computer*, Feb. 1999, pp. 51-59.
12. B. Schatz, "High-Performance Distributed Digital Libraries: Building the Interspace on the Grid," *Proc. 7th Int'l Symp. High-Performance Distributed Computing (HPDC-7 98)*, IEEE CS Press, Los Alamitos, Calif., 1998, pp. 224-234.
13. E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing," *Computational Linguistics*, vol. 21, 1995, pp. 543-565.
14. K. Tolle and H. Chen, "Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools," *J. Am. Soc. Information Science*, Mar. 2000, pp. 380-393.
15. N. Bennett et al., "Extracting Noun Phrases for all of MEDLINE," *Proc. 1999 Am. Medical Informatics Assoc. Symp. (AMIA 99)*, AMIA, Bethesda, Md., 1999, pp. 681-688.
16. Y. Chung et al., "Semantic Indexing for a Complete Subject Discipline," *Proc. Int'l 4th ACM Conf. Digital Libraries*, ACM Press, New York, 1999, pp. 39-48.

Bruce R. Schatz is director of the Community Architectures for Network Information Systems Laboratory and a professor in the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, where he served as principal investigator of the Illinois Digital Library project. He is also a senior research scientist at the National Center for Supercomputing Applications, where he served as the scientific advisor for information systems during the development of Mosaic. His research interests include information infrastructure, digital libraries, health-care infrastructure, biomedical informatics, and scientific communities. Schatz received a PhD in computer science from the University of Arizona, where he later served as principal investigator of the National Collaboratories project that built the Worm Community System. He is a fellow of the American Association for the Advancement of Science. Contact him at schatz@uiuc.edu.