TRN's
Making the Future report
The State of an Emerging Technology and a Look at What Lies Ahead
Report Number 13 April, 2004

Internet Applications: The Emerging Global Computer

Executive Summary

Nearly three-quarters of the U.S. population has access to the Internet, which connects 233 million hosts and contains more than 15 times as much data as the Library of Congress print collection.

The Internet is part data utility, part shared information space, and is becoming a vast distributed computer. It is also a platform for developing new information technologies, a phenomenon to be examined scientifically, and a means of studying human behavior.

This report focuses on research that examines and advances the Internet as an end-user technology.

As a global, distributed interface, the Web is enabling and inspiring better graphics and interface tools, more efficient email, methods of combating viruses, worms and parasitic computing, ways to authenticate people and data, and advanced methods of finding relevant information from vast digital resources.

Key research efforts are aimed at finding better ways to index the Web, enabling users to search the Internet and other repositories of information like databases at once, and finding better ways to visualize the vast collections of data that modern technology makes possible.

Semantic Web software, which enables information indexing and allows search engines and intelligent agents to understand data properties and relationships, and Web services, which allow programs to communicate directly through the Web, are shifting the Web from a publishing and communications medium to an information management environment.

The presence of a global network and the emergence of fast connections is making it possible to assemble virtual ad hoc supercomputers to process large data sets.

It is clear that the Internet is boosting communications abilities, fostering new forms of commerce, enabling new types of research, and increasing political mobilization. It is less clear how conflicts among security, privacy, e-commerce and freedom of expression will ultimately play out.

What to Look For

Privacy and security:

Better-than-human spam filters
Internet-wide virus blocking
Censor-proof Web access
Censor-proof Web publishing

Search:

Effective Internet-based question answering systems
Search-directed browsing
Integrated surface Web/deep Web search
Query-directed Web crawling

Semantic Web:

Semantics-based Web searching
Practical intelligent agents
Embedded systems integrated with the Web

Grid and distributed computing:

Adaptive, on-the-fly Grid computing
Internet-based distributed storage
Shared immersive environments

Wired

In the last dozen years the Internet has evolved from a communications medium for academics to part of the fabric of daily life in the industrialized world. Nearly three-quarters of the U.S. population now has Internet access, according to Nielsen/NetRatings.

The Internet is part data utility, part shared information space, and it is steadily becoming a vast distributed computer. From a research standpoint, the Internet is a platform for developing new information technologies, a phenomenon to be examined scientifically, and a means of studying human behavior.

This report focuses on research that examines and advances the Internet as an end-user technology. Topics include the following:

- Understanding and improving information access
- Preserving privacy and security
- Shoring up email
- Organizing information
- Finding information
- Supporting information visualization
- Tapping distributed computing resources

The Internet is constantly evolving. It is extending into the physical world through wireless access, networked embedded systems and global positioning system (GPS) technologies. Everyday machines like automobiles and refrigerators are beginning to include Internet connections that allow them to access information automatically. Meanwhile, GPS coordinates are making the Internet aware of location.

Researchers are tackling problems like the relentless proliferation of unsolicited email, are implementing new concepts like virtual, ad hoc supercomputers assembled from under used Net resources, and are applying standard concepts like usability measures to determine how well Internet technologies are meeting users' needs and expectations.

The Web experience

The Internet has evolved organically, and even though a significant portion of the world's population uses it every day, scientists are still working out exactly how it is used and how it can be made easier to use.

Researchers from the U.S. Air Force have published a report that draws on established studies of human-computer interaction and information technology to quantify some of what Web applications need to do to provide users with a better experience. They found that users' perceptions of and feelings towards Internet use were predominantly determined by 10 critical quality-of-experience criteria:

- Relevance of information to a task
- Availability of search engine options like multiple criteria
- Amount of information available
- Organization of information
- Ease-of-use of the browser
- Reliability of the connection
- Clarity of directions for Web site navigation
- Whether links are up-to-date
- Availability of information at remote servers
- Perceptions of security and privacy

The Air Force researchers also found that effects like graphics, sound and animation become less relevant if the interface as a

How It Works

Crawling/indexing/querying

The search engine programs that serve up Internet information are really a constellation of three types of software.

A spider, or crawler, finds all the pages on a Web site by mapping out the link structure of pages within a site. Spiders make return trips periodically to find changes.

The index, or catalog, stores information from Web pages found by the spider in a database. Once the information is indexed, it can be accessed by the actual search engine, which looks through the database to find entries that match a query and ranks relevant entries.

Because it takes time to crawl and index Web pages, Web searches are actually only searches of a database of pages found by the crawler at some point — usually days or weeks — in the past.

The deep Web

And you won't find everything contained in the World Wide Web using today's search engines.

Web pages can be excluded from search engines by including appropriate metgatags, and others remain unsearched because search engine spiders can't gain access. Pages that are not indexed by the major search engines are, collectively, the deep Web.

These pages include thousands of specialized databases that can be accessed over the Web. Such databases generate Web pages from search results on-the-fly. The deep Web is probably several hundred times bigger than the surface Web.

The Semantic Web

The Semantic Web initiative is poised to make searches more accurate and enable increasingly sophisticated information services like intelligent agents that can find products and services, schedule appointments and make purchases. The initiative includes a sort of grammar and vocabulary that provide information about a document's components; this information will enable Web software to act on the meaning of Web content.

Semantic Web software includes a special set of Extensible Markup Language (XML) tags that includes Uniform Resource Identifiers (URIs), a Resource Description Framework (RDF), and a Web Ontology Language (OWL).

The Extensible Markup Language tags provide information about a document's components. The Uniform Resource Identifiers contained in the XML tags expand the concept of Uniform Resource

whole doesn't provide a good experience for the user. (See "Study Finds Web Quality Time", page 14.)

In a study that examined how people rediscover information rather than how they find it the first time, Researchers from Virginia Polytechnic Institute and State University showed that people tend to use a two-stage process to re-find information, that they use domain information in context to move closer to a goal, and that annotations make things easier. (See "Web Users Re-Visit in Steps", page 10.)

Researchers from Tel Aviv University and the University of California at Berkeley have examined the nature of the relationship between linguistic patterns and Internet content. The study implies that even though Internet use among non-native English speakers is growing at a faster rate than that of native English speakers, English content will continue to dominate the Internet. (See "English Could Snowball on Net", page 12.)

Web users as science subjects

Just as the Internet has enabled telecommuting, the presence of a global network and the emergence of fast connections have enabled certain types of research. It has made it quick, easy and inexpensive to garner large groups of relatively random users to participate in experiments.

Researchers from the University of Fribourg in Switzerland have tapped the Internet to investigate speculative trading behavior. (See "Web Game Reveals Market Sense", page 16.)

The recommender systems online media sellers use to solicit buyers' opinions about goods like books, movies and CDs is another fertile area of behavioral research.

Researchers from the University of Minnesota have shown that the way recommender systems are set up can affect the opinions they evoke, and that artificially high or low recommendations can raise or lower subsequent recommendations. (See "Recommenders Can Skew Results", page 17.)

Preserving the information commons

The more people use the Internet and the more value it holds for them, the more it attracts fraud, theft and destructive behavior. Researchers are addressing aspects of Internet security that range from authenticating people and data to combating Net viruses.

The Internet allows everyone to be a publisher, which raises the question of sourcing. How do you know that a posted investment history, for instance, is correct and complete? Researchers are working on security schemes that allow users to tell for sure where information comes from so that they have a source to consider when judging information.

Researchers from the University of California at Davis and from Stubblebine Consulting, LLC have come up with a signature scheme that allows portions of documents stored in Extensible Markup Language (XML) databases to be authenticated. Using the scheme, dubbed TruthSayer, an author

Locators (URLs) by adding IDs for objects, concepts and values that are not dependent on location.

The Resource Description Framework is a set of rules for describing objects like Web pages, people and products by their properties and relationships to other objects. There are three elements to a Resource Description Framework object definition: the object, the object's properties, and the values of these properties. For example, the object could be a car that has the property of color with a value of blue. Objects, properties and values are all identified by Uniform Resource Identifiers.

Properties can also be relationships to other objects, like employment and authorship. In the case of a Prof. Stevens who teaches at State University, the object is Prof. Stevens, the property is employment, and the property value is State University, which is also an object that can have its own properties. And in the case of Bill Johnson who is the composer of State University's school song, the object is Bill Johnson, the property is authorship, and the property value is school song, which is also an object.

The Web Ontology Language is a tool for building vocabularies that define specific sets of objects. The vocabularies are expressed and interpreted through the Resource Description Framework.

Semantic Web software makes it possible for an intelligent agent to carry out the request "show me the opticians in the neighborhood" even if there is no explicit list, because it knows that "neighborhood" has the property "location" with the value "Bellevue," and in searching a directory of opticians it knows to skip Dr. Smith, whose location value is "Springfield", but include Dr. Jones, whose location value is "Bellevue."

Comparing text

The DataSpace project and is aimed at making it possible to not only search for, but automatically correlate Web information.

The project includes four major pieces of software that enable this:

- Data Space Transfer Protocol (DSTP) — a protocol for transferring columns of data that includes universal correlation keys, which are analogous to a database's primary key attributes
- Predictive Model Markup Language (PMML), an XML-based language that allows users to define data mining and statistical models, then mark different parts of data according to those models so different sets of data can be compared
- Open source client and server software that allows computers to exchange data

can sign an XML document and give it to someone else to store and post. (See "Data Protected on Unlocked Web Sites", page 22.)

On the other side of the equation, researchers are finding ways to exploit the way the Web works in order to point out where the network needs to be shored up.

University of Notre Dame researchers have discovered a way to use Web server handshakes to compute small pieces of a mathematical problem by disguising the pieces as ordinary Web browser messages. The research shows that the computers connected to the Internet can be used in ways their owners are unaware of. (See "Scheme Harnesses Internet Handshakes", page 23.)

Along with parasitic computing, the Internet has brought computer viruses and worms designed to propagate through a network just as biological viruses spread through a host population. Computer viruses attach themselves to or replace existing software. Worms are separate programs. Both can compromise computers by altering, destroying and sending files.

Researchers from Washington University and Global Velocity have come up with a way to stop computer viruses and Internet worms at the network level. The Field Programmable Port Extender is reconfigurable hardware that can protect an entire network at a time from viruses and worms by scanning every bit of data contained in every packet that passes through a network. The system is fast enough to search for viruses in the wide flow of backbone Internet traffic because it is implemented in hardware rather than software, and because the hardware is reconfigurable it is flexible enough to quickly add virus and worm signatures as they are discovered. (See "Device Guards Net against Viruses", page 24.)

Being seen and being heard

The Internet is becoming an arena for public life, and the awkwardness of this transition is acutely evident in the subject of electronic voting. The convenience of Internet voting has the potential to increase voter participation, but computer scientists are nearly unanimous in warning that today's Internet is far from secure and reliable enough to ensure that every vote is legitimate and counted.

Researchers are working on a range of issues related to public life and the Internet, including privacy and censorship.

Researchers from Virginia Polytechnic Institute and State University, Purdue University, and the University of Minnesota have shown that recommender systems can serve to collect information about groups of people who do not ordinarily interact — connections known as weak ties. The researchers also showed that on the Web, weak ties can be combined with other information to trace individual user's identities. (See "Rating Systems Put Privacy Risk", page 18.)

Massachusetts Institute of Technology researchers and have devised a scheme, dubbed Infranet, that allows users to navigate the Internet using standard hypertext transfer protocol in such a

Who to Watch

Web use

Albert N. Badre, Georgia Institute of Technology
Atlanta, Georgia
www.cc.gatech.edu/gvu/people/faculty/badre.html

Marti Hearst, University of California, Berkeley
Berkeley, California
www.sims.berkeley.edu/~hearst

Andrew Odlyzko, University of Minnesota
Minneapolis, Minnesota
www.dtc.umn.edu/~odlyzko

Cherri M. Pancake, Oregon State University
Corvallis, Oregon
web.engr.oregonstate.edu/~pancake

Privacy/Free speech

Annie I. Antón, North Carolina State University
Raleigh, North Carolina
www.csc.ncsu.edu/faculty/anton

Edward W. Felten, Princeton University
Princeton, New Jersey
www.cs.princeton.edu/~felten

Avi Rubin, Johns Hopkins University
Baltimore, Maryland
www.cs.jhu.edu/~rubin

Security

Ross Anderson, University of Cambridge
Cambridge, England
www.cl.cam.ac.uk/users/rja14

Joan Feigenbaum,Yale
New Haven, Connecticut
www.cs.yale.edu/homes/jf/home.html

Eugene (Gene) H. Spafford, Purdue Unversity
West Lafayette, Indiana
www.cerias.purdue.edu/homes/spaf

Search/Retrieval

Tim Berners-Lee, Massachusetts Institute of Technology
Cambridge, Massachusetts
www.w3.org/People/Berners-Lee

Soumen Chakrabarti, Indian Institute of Technology Bombay
Bombay, India
www.cse.iitb.ac.in/~soumen

Zhixiang Chen, University of Texas - Pan American
Edinburg, Texas
www.cs.panam.edu/~chen

C. Lee Giles, Pennsylvania State University
University Park
clgiles.ist.psu.edu

James A. Hendler, University of Maryland
College Park, Maryland
www.cs.umd.edu/~hendler

way that they cannot be monitored. (See "Scheme Hides Web Access", page 20.)

Researchers from Lancaster University in England have developed a scheme that uses the established technologies of wireless communications, local area networks and Internet access to enable smart spaces while protecting users' privacy. The researchers' system causes screens near a user to display Internet-based information that the user is likely to prefer. It can also be used to control CD players and other media devices. (See "Badge Controls Displays", page 30.)

Researchers from AT&T Labs and New York University have put together an encryption scheme that allows for Web publishing that is both anonymous and difficult to take down. (See "Fault-Tolerant Free Speech", page 19.)

Christopher Olston, Carnegie Mellon University
Pittsburgh, Pennsylvania
www-2.cs.cmu.edu/~olston

Lyle Ungar, University of Pennsylvania
Philadelphia, Pennsylvania
www.cis.upenn.edu/~ungar

Grid

Francine D. Berman, University of California, San Diego
La Jolla, California
www.cs.ucsd.edu/users/berman

Rajkumar Buyya, University of Melbourne
Melbourne, Australia
www.gridbus.org/~raj

Ian Foster, Argonne National Laboratory/ University of Chicago
Argonne, Illinois
www-fp.mcs.anl.gov/~foster

Miron Livny, University of Wisconsin
Madison, Wisconsin that

Email

The Internet has enabled email to rapidly become an important form of human communication. This has spawned research aimed at making email more efficient.

Researchers from Palo Alto Research Center (PARC) have studied the ways people use email in an attempt to pinpoint needed improvements across all the functions it has come to fill — not just communications, but also information organization and workflow management. The study showed a need for better ways of organizing folders, quicker ways to get to recently accessed items such as to-do lists and reminders, ways to track different versions of documents, and the ability to manage URLs. (See "Email Burdened by Management Role", page 27.)

A researcher from Temple University followed the way 38 process-improvement groups in three organizations worked over four years in a study designed to quantify the differences between face-to-face communications and email. The study showed that email communication can produce better results, but requires considerably more cognitive effort. (See "Email Takes Brainpower", page 28.)

The combination of the Internet and email has also spawned research aimed at ensuring secure, private communications.

The classic way to secure email is to use a public-private key scheme. All public key algorithms are based on mathematical formulas like factoring that are easy to solve in one direction, but very difficult to solve in the other direction. For instance, it is easy to find a number given its factors — just multiply the numbers together. But finding the factors of a very large number is very difficult because there are so many possibilities to check. A person's publicly-accessible key is the large number. This is used to encrypt a message bound for that person, but once the message is encrypted only the private key — the factors of the number — can be used to decrypt the message.

Email encryption is somewhat cumbersome to use, however. A recipient must have a public key and a sender must be able to find it in order to encrypt a message bound for that particular person.

Researchers from the University of California at Davis and Stanford University have come up with a way to generate public keys using information contained in email addresses. Using a person's unique email address as a public key makes it possible to send encrypted messages without having to look anything up. (See "Address Key Locks Email", page 25.)

Other efforts are focused on ways to help users control the growing flood of spam enabled by such a cheap, easy means of communication.

The common way to control spam is to use filters that block mail from known spammers who have been blacklisted and that follow general rules for blocking messages including, for instance, those that contain the word "adult" in the subject header. But spammers frequently forge email addresses to dodge blacklists, and keyword-specific blocking can catch legitimate mail as well.

Researchers from the University of Athens and the Demokritos National Center for Scientific Research in Greece have devised a program that creates custom filters that learn to recognize spam by looking through a user's legitimate email and comparing it with a spam library collected by the researchers. Key to the process is that it uses a combination of filters that are based on different learning algorithms. (See "Teamed Filters Catch More Spam", page 26.)

All in the presentation

As a global, distributed interface, the Web has enabled and inspired a variety of graphics and interface tools.

A Cornell University researcher has found a way to identify key characteristics of digital content so the content can automatically be matched with programs that can display it in a browser. The scheme is an alternative to plug-ins because neither the content nor the program that activates it needs to be present on the system that displays the content. Separating the storage and maintenance of digital content from its presentation makes more individualized presentation of data possible, and makes it easier to augment the presentation of data created by others. (See "Content Scheme Banishes Browser Plug-ins", page 33.)

Researchers from the University of Applied Sciences in Germany have developed a scheme that allows teachers to organize digital text, audio and video into databases, then draw from their own and other teachers' databases to compose multimedia lessons. (See "Software Orchestrates Web Presentations", page 34.)

Researchers from the Greek National Center for Scientific Research, the Foundation of the Hellenic World, the University of Edinburgh and the Trentino Cultural Institute have built a system that taps the powers of hypertext, information databases and natural language generation to allow people to more efficiently traverse descriptive text like that used in museums. (See "Software Guides Museum-Goers", page 31.)

Researchers from Georgia State University have made a Web browser that allows people to surf the Web just by thinking. (See "Browser Boosts Brain Interface", page 30.)

Homing in

A large focus of Internet research has been to make it easier for users to find what they're looking for among the vast and growing sea of resources the global network offers. As of January, the Internet contained an estimated 233 million hosts, and in 2002 the Web held about 15 times as much data as the Library of Congress' print collection.

Researchers from Carnegie Mellon University have devised software that shows a user how strongly the links generated by a Web search correlate with the topics she is searching for. The software grades the links a search engine returns and indicates link relevance by, for instance, increasing the font size of links that have more connections to relevant pages. The system is designed to cut down on the links a user must click during a search, and instead make searching more like browsing. (See "Search Tool Aids Browsing", page 36.)

Researchers from the University of Library and Information Sciences in Japan have devised a system that improves Internet search by indexing the Web as an open encyclopedia. Instead of presenting a list of a thousand Web sites that might possibly contain answers, the system extracts information and reference links and organizes it in the form of an encyclopedic entry. (See "Search Tool Builds Encyclopedia", page 35.)

Mapping the information space

Although someday servers may be fast enough to index the whole Web, the Web crawlers that index information for today's search engines generally index only a small percentage of the Internet's vast collection. A key issue is how to decide exactly what sites to index.

Researchers from Contraco Consulting and Software Ltd., T-Online International and Siegen University in Germany have written an algorithm that improves Internet search results by factoring in trends about what people are looking for. If the number of queries about soccer, for instance, is growing in the months before the World Cup, the search engine devotes more time indexing soccer sites. (See "Queries Guide Web Crawlers", page 37.)

Another problem, given the proliferation of intranets, databases, and local files in addition to the vast Internet, is being able to search across these disparate digital resources. Today's search engines are largely limited to the surface Web, which leaves the much greater amount of data held in the deep Web of Internet-connected databases nearly invisible.

Researchers from Birkbeck University of London in England have written software designed to allow users to search across the Internet, a company intranet, local databases, and local computer file systems at the same time. The software could eventually be used to search Internet-connected databases. (See "Web Searches Tap Databases", page 38.)

Researchers from Huazhong University of Science and Technology in China are approaching the same problem from the opposite direction. They have proposed a scheme to integrate basic search functions into the Internet infrastructure, and to assure that the search capability extends to the deep Web. Their scheme, Domain Resource Integrated System (DRIS), calls for individual domains such as universities and corporations to maintain search engines for their domains, for subnetworks such as the China Education and Research Network to maintain indexes and search interfaces to the domain search

engines, and for top-level domains like the Internet in China to implement metasearch interfaces to the subnetworks indexes. (See "Plan Puts Search in Net Structure", page 39.)

Leveraging text

Researchers are also working on ways to better sort and access the textual information in Web pages.

Researchers from Metacode Technologies, Inc. have devised a text-based categorization scheme that uses subject-based metadata to group Web sites the same way a library card catalog groups books. The software ferrets out meaningful text from three distinct sources within a Web site to categorize the site: keywords, page layout instructions that affect the look of the site, and readable text. (See "Software Sifts Text to Sort Web Sites", page 40.)

Project DataSpace, a University, government and corporate consortium, has built an infrastructure designed to do for data comparison what the Web has already done for sharing documents. The project involves four major parts: a transfer protocol for moving columns of data over the Web, a set of tags for marking columns of data so they can be compared meaningfully, and client and server software that allows computers to exchange such data. Version 3.0 of the software was released in March, 2003. (See "Software Sorts Web Data", page 41.)

More than just words

The Internet is gaining other new capabilities.

Semantic Web software organizes Web information so that search engines and intelligent agents can understand properties and relationships. Harvard University, for example, could be defined as an institution of higher education, which is a class of objects that has a set of properties like a population of students.

The World Wide Web Consortium released standards in February 2004 that define the two foundation elements of the Semantic Web initiative: the Resource Description Framework (RDF), which provides a structure for interpreting information, and the Web Ontology Language (OWL), which provides a means for defining the properties and relationships of objects in Web information. (See How It Works, page 2.)

Web services provide a way for software to communicate with each other over the Web, enabling a form of distributed computing in which simple services can be combined to carry out complicated tasks like financial transactions.

Semantic Web software and Web services promise to shift the nature of the Web from a publishing and communications medium to an information management environment.

Getting the picture across

One of the ongoing challenges facing scientists and businesspeople is finding ways to visualize the vast amount of data that modern technology makes it possible to collect so we can more quickly and easily comprehend patterns in data sets like the 3 billion base pairs that make up human DNA, the folding patterns of protein molecules, like years' worth of business statistics, seismic information, and particle accelerator readings.

Sight

Researchers from Sandia National Laboratories have found a way to work with large amounts of data over the Internet in near real-time. Their prototype allows users to manipulate very large sets of data on remote computers while experiencing lag times of less than one-tenth of a second. The scheme calls for transferring the video signal that carries image information from a computer to a monitor rather than transferring the data itself. (See "Remote Monitoring Aids Data Access", page 42.)

Sound

University of Southern California researchers have developed a filtering system that cuts the amount of bandwidth needed to stream surround-sound over the Internet, and allows older recordings to be recast as multichannel sound. The researchers' Virtual Microphone technology maps a concert hall's sound by recording using 10 or 20 microphones around the hall. Once a space has been mapped, the technology can adjust any ordinary recording to match what it would have sounded like recorded through those microphones. The technology could also be used to provide a plug-in that adds multi-channel sound to streaming audio. (See "Virtual Mic Carries Concert Hall Sound Over 'Net", page 44.)

Touch

Researchers from the University of Buffalo have developed a method that enables one person to go through the exact movements of another, including feeling the same resistive forces, over the Internet. The method could eventually be used to capture the touch of a musician, golfer or surgeon and pass it on to someone trying to match that touch. (See "Experience Handed Across Net", page 43.)

Supercomputers on demand

Internet-related research efforts are also aimed at tapping the vast collection of unused resources on the Net. Grid computing identifies resources like idle computers and vacant disk space and puts together virtual computers powerful enough to handle compute-intensive problems like designing drugs and processing the huge amounts of data generated in nuclear physics experiments.

Though the concept of harnessing otherwise unused compute resources distributed around the Net is relatively simple, coordinating these resources to form usable virtual computers is tricky.

University of Melbourne researchers have produced a toolkit that makes it easier to monitor a Grid computer. The toolkit, dubbed Gridscape, allows users to create a Web interface to a Grid computing testbed without programming. (See "Tool Eases Grid Monitoring", page 45.)

Researchers from Monash University and Eliza Hall Institute in Australia have put together a set of tools for Grid-based molecular docking. The tools tap remote databases of chemical structures in order to carry out the molecular matching process. (See "Toolset Teams Computers to Design Drugs", page 46.)

Researchers at Argonne National Labs, the University of California at Berkeley, the University of Chicago, and the Max Planck Institute for Gravitational Physics have developed software that reconfigures a Grid computer on-the-fly. (See "Virtual Computers Reconfigure on the Fly", page 47.)

Researchers from Monash University in Australia and the European Council for Nuclear Research (CERN) in Switzerland have come up with a software architecture and set of policies designed to increase the reach of Grid computing by applying traditional economic models like barter and monopoly to manage Grid resource supply and demand. (See "Tools Automate Computer Sharing", page 48.)

Researchers are also working on non-Grid approaches to using resources distributed around the Internet.

Researchers from the University of California at Berkeley have come up with a scheme, dubbed OceanStore, to archive data across computers world-wide. The scheme constantly updates data, saving it in numerous places so it can be accessed even if some of the computers holding it are lost. The scheme also includes safeguards to keep data safe from unauthorized accessed. (See "Store Globally, Access Locally", page 50.)

The new infrastructure

The Internet has done for information what the highway system did for transportation — it diminished distance as a barrier. This fundamental advance has already brought about widespread changes and is likely to foster further changes that can't be predicted. The highway system boosted interstate commerce. It also increased the mobility of the country's population and fueled the growth of suburbia.

It is clear that the Internet is boosting communications abilities, fostering new forms of commerce, enabling new types of research, and increasing political mobilization. It is less clear how conflicts among security, privacy, e-commerce and freedom of expression will ultimately play out.

Recent Key Developments

Advances in Web Use:

- A study that shows that people re-find information on the Web in stages and that annotations are helpful (Web users re-visit in steps, page 10)

- A method for identifying communities of interest through Web browsing patterns (Net scan finds like-minded users, page 11)
- A study that shows that English is likely to remain the dominant language on the Internet because of its head start (English could snowball on Net, page 12)
- A method for re-finding Web information using mobile devices with voice interfaces, Virginia Polytechnic Institute and State University, November 2001
- A study that shows that the quality of Web experiences are determined by browser ease-of-use and relevance, availability and organization of information (Study finds Web quality time, page 14)
- A study that used a Web-based financial game to show that people have an intuitive sense of markets (Web game reveals market sense, page 16)
- A study that shows that online recommenders systems that display predictions skew results (Recommenders can skew results, page 17)

Advances in Privacy and Free Speech:

- An overview of the inadequacies of today's Internet infrastructure for electronic voting, AT&T Research, December 2002
- A study that shows that anonymous online recommender systems yield information that could be used to identify individuals (Rating systems put privacy at risk, page 18)
- A distributed encryption scheme for anonymous, censor-proof Web publishing (Fault-tolerant free speech, page 19)
- A method for hiding Web access that doesn't generate noticeable traffic patterns (Scheme hides Web access, page 20)

Advances in Security:

- A method for countering email-based distributed denial of service attacks, RSA Security and the University of Iowa, May 2003
- A method for authenticating portions of XML documents on unprotected servers (Data protected on unlocked Web sites, page 22)
- A proof-of-principal demonstration that uses basic protocols to covertly harness processing power from computers connected to the Internet (Scheme harnesses Internet handshakes, page 23)
- A reconfigurable hardware device that filters viruses and worms on Internet backbone links (Device guards Net against viruses, page 24)
- A method for generating public encryption keys using email addresses (Address key locks email, page 25)

Advances in Email:

- A method for combining multiple spam filters (Teamed filters catch more spam, page 26)
- A study that shows that people use email as their primary information organization tool (Email burdened by management role, page 27)
- A study that shows that email takes more cognitive effort than face-to-phase communications (Email takes brainpower, page 28)

Advances in Web Tools:

- A browser that can be controlled using direct neural signals (Browser boosts brain interface, page 30)
- A handheld device that automatically causes nearby displays to show preferred Web information (Badge controls displays, page 30)
- Software that combines hypertext and natural language generation to dynamically supply museum information over the Web (Software guides museum-goers, page 31)
- Software that coordinates content and playback software for displaying the information in browsers (Content scheme banishes browser plug-ins, page 33)
- A system for creating multimedia lessons using content from different sources (Software orchestrates Web presentations, page 34)

- Software that presents search results in an encyclopedia-like form (Search tool builds encyclopedia, page 35)

Advances in Search and Information Retrieval:

- A system that ranks search results by the number of relevant links and differentiates links on returned pages in order to improve browsing (Search tool aids browsing, page 36)
- An agent-based system for providing tailored access to cultural information over the Web, CINECA Supercomputing Centre and the University of Mondena in Italy, March 2004
- An algorithm that uses search queries to direct search engine crawlers to more thoroughly index sites of popular topics (Queries guide Web crawlers, page 37)
- Software that allows search queries to access relational databases along with Web pages (Web searches tap databases, page 38)
- A scheme for integrating search, including database access, into the Internet structure (Plan puts search in Net structure, page 39)
- Software that uses keywords in metadata, HTML tags and readable text to categorize Web sites (Software sifts text to sort Web sites, page 40)
- A project that provides software for comparing textual data over the Web (Software sorts Web data, page 41)

Advances in Visualization and Simulation:

- A system for transmitting computer monitor video signals over the Internet in order to display visualizations of large remote data sets (Remote monitoring aids data access, page 42)
- A method for transmitting a person's hand movements over the Internet (Experience handed across Net, page 43)
- A set of filters that can make audio signals sent over a network sound as though they were recorded using an array of microphones (Virtual mic carries concert hall sound over Net, page 44)

Advances in Grid Computing:

- Software for creating Web interfaces for Grid Computing projects (Tool eases Grid monitoring, page 45)
- An architecture for developing Grid-based natural language processing applications, University of Melbourne in Australia, May 2003
- An algorithm for implementing a distributed neural network using a basic Internet messaging protocol, NEC Europe, February 2003
- A Grid tool for designing drugs that matches pairs of molecules using molecular data in remote databases (Toolset teams computers to design drugs, page 46)
- A system for automatically reconfiguring Grid computers as they run (Virtual computers reconfigure on-the-fly, page 47)
- Software that implements economic models for managing Grid resource supply and demand (Tools automate computer sharing, page 48)
- A global, distributed data storage scheme that uses the Internet (Store globally, access locally, page 50)

Web Use

Web Users Re-Visit in Steps

By Kimberly Patch, Technology Research News
February 11/18, 2004

Half the battle of finding information on the Web is getting back to a page you've already seen.

The Web has long spurred researchers to study how people initially find information, but the tactics people use to get back to previously discovered information remain less understood.

Researchers at Virginia Polytechnic Institute and State University are examining how people relocate information rather than how they find it the first time. "There is evidence that users usually can re-find a page they're looking for, but we hope to guide the development of computer tools that can make the re-finding process easier, especially if the user is mobile," said Robert Capra, a researcher at Virginia Tech.

The researchers' study showed that people tend to use a two-stage process to find information they have seen before, that they use domain information and context to move closer to a goal, and that annotations make things easier.

The results could lead to tools that would help users re-access Web pages more quickly and easily, including

finding the same information using devices ranging from desktop computers to mobile phones. "Imagine that you were preparing to go on a trip and had browsed a number of restaurants, events and attractions that you were interested in," said Capra. "Once you're on the trip, such a system can help you re-find that information on your PDA."

To study how people re-find information, the researchers asked six subjects who were familiar with Web browsing to carry out a set of tasks using the Web, including finding movie theater show times, restaurant phone numbers and addresses, and tourist event information. The subjects also did a freeform search for information. The subjects took notes during the searches using an annotation tool.

In a second session about a week later, the six subjects were each paired with six additional users, or retrievers, with instructions to direct the re-finding process by phone. The retrievers had access to the original subjects' annotations and browser history logs of the searches completed a week earlier, but the original user did not.

The users' naturally broke the re-finding process into steps, and the process often consisted of two stages, said Capra. In the first stage, users tried to get back to a particular page they remembered or thought would be useful. In the second stage, the subjects asked specific questions about the information they were trying to re-find. "At times, [the subject] provided very partial information, allowing the retriever to reach a specific location before making the next request," said Capra.

The subjects used information they remembered to move incrementally closer to the information being sought, said Capra. In more than three-quarters of the re-finding tasks, users tapped waypoints, or Web pages they remembered from along the paths they originally took to find information, he said. Sometimes subjects remembered an exact URL, but other times used a title or just a description.

The research could be applied to Web tools to more easily help users find what they have previously seen on the Web, said Capra. One possibility is software tools that support the approach of making incremental progress toward a goal, he said.

The results are not particularly surprising, but the work helps augment the body of science surrounding the issue of finding information on the Web, and the use of the telephone is interesting, said Marti Hearst, an associate professor of information management and systems at the University of California at Berkeley. "The results are useful in that the study provides another empirical data point to support the assumption that search tasks are done progressively and incrementally, and often involve first locating a source and then navigating that source," she said.

There are two design and analysis flaws in the study, however, said Hearst. The users' are explicitly told that they will be asked to find information again, and the study assumes that people will instruct others to find something in the same way that they would on their own, she said. "It could well be the case that people... break it down into pieces to better help the other person keep track of the different aspects of a task."

The researchers' next step is to run a study that looks at the difficulty of tasks and whether the users' familiarity with a task affects the approach a user takes to re-find information.

They're also planning to build a prototype Web browser add-on aimed at helping users re-find information, said Capra. The prototype could be ready next year, he said.

Capra's research colleague was Manuel A. Pérez-Quiñones. The research was funded by the National Science Foundation (NSF) and IBM.

Timeline: < 2 years
Funding: Corporate; Government
TRN Categories: Databases and Information Retrieval; Internet
Story Type: News
Related Elements: Technical paper, "Re-Finding Found Things: an Exploratory Study of How Users Re-Find Information," posted on the Computing Research Repository (CoRR) at arxiv.org/abs/cs.HC/0310011

TRN

Net Scan Finds Like-minded Users

By Kimberly Patch, Technology Research News
May 7/14, 2003

When you search for information on the Web, chances are you aren't alone—there are like-minded groups of users across the Web searching for the same sorts of things.

Researchers from the University of Chicago have shown that is possible to identify these groups by analyzing browsing patterns, even in networks as far-flung as the Web.

The researchers' method of graphing information across data distribution systems like the Internet shows that, given a large enough sample, computer users can be grouped according to their common interests based only on their requests for data. "One of the first questions we asked was is the group-based collaboration of scientists mirrored somehow in their usage of data," said Adriana Iamnitchi, a researcher at the University of Chicago.

The answer turned out to be yes, across all types of group-based interests. "Communities as heterogeneous as

the Web seem to show this pattern of having users naturally group in interest-based groups," she said.

The information-request graphing method can be used to design scalable, adaptive methods for locating and delivering data, said Iamnitchi. The method could theoretically be used by anyone, including ecommerce vendors, to target communities of interest.

The researchers are working on using the patterns to design more efficient services for resource-sharing environments like Grid computing, Iamnitchi said.

Grid software coordinates a few or even hundreds of computers across networks like the Internet to piece together compute power and resources like databases into powerful virtual computers; the combined resources can speed up scientific and engineering applications like time-consuming equations and three-dimensional simulations.

The researchers found the data-sharing relationship pattern while looking for a way to leverage characteristics of the Grid computing community to make that type of computing more efficient, according to Iamnitchi. "Our idea was to... design mechanisms [that are] able to cope efficiently with large and dynamic numbers of resources—data files, computers, and storage space for results," she said.

One typical characteristic of the community that uses Grid computing is they tend to collaborate, said Iamnitchi. When the researchers analyzed traces of scientific computations from a high-energy physics collaboration that spanned 18 countries and involved 70-odd institutions and thousands of physicists, they found that the patterns of collaboration were mirrored in scientists' data requests.

The researchers looked at the relationships that formed among users based on the data they were interested in. "We captured and quantified these relationships by modeling the system as a data-sharing ... graph whose nodes are the data consumers in that system," said Iamnitchi. Nodes, or people, who requested a given number of the same files within a given time were connected.

In an analysis of six months worth of scientists' requests for data, the researchers found that group-based collaboration is visible in the way information is requested, said Iamnitchi. "Scientists form groups of interest based on the data they used," she said. The researchers found the same pattern in a larger analysis of general Web requests.

The pattern of similar requests shared the small-world characteristic common in many networks, including the way data is arranged in networks like the Internet.

In small-world networks, it is possible to get from one node to any other node by traversing relatively few links. Social networks, with people as nodes and relationships as links, and the Web, with pages as nodes, and links between pages as links, are also small-world networks.

Looking at small-world topologies is not a novel idea, but the method of extracting a graph from an arbitrary data-sharing relationship and using it to study these structures is, said Filippo Menczer, an assistant professor of management services at the University of Iowa.

Data request patterns have been analyzed previously, but in different ways—to examine the popularity distribution of Web requests or to study the most efficient way to cache Internet traffic. In contrast, the Chicago researchers' analysis uncovered relationships between users based on their common interests in data.

The method is potentially useful, especially because a graph can be made from any Web usage log, said Menczer. "Any Webmaster can do this."

The method may be useful for discovering clusters of users who have interest in a certain type of data, Menczer said. "Ecommerce vendors are currently using collaborative filtering techniques, which are related to this," to do so, he said. The method can also be used for distributed caching and broadcasting, similar to the services offered by Akamai Technologies Inc., he said.

The researchers are now making the method more efficient for resource-sharing environments like Grid computing, said Iamnitchi. "We are currently looking... to design mechanisms to locate resources," she said. "The ultimate goal is to provide scalable, adaptive mechanisms [that are] able to deal with variations in resource participation."

The resource location mechanisms could be ready to use within two years, Iamnitchi said.

Iamnitchi's research colleagues were Matei Ripeanu from the University of Chicago and Ian Foster from Argonne National laboratory. The research was funded by the National Science Foundation (NSF).

Timeline: 2 years
Funding: Government
TRN Categories: Internet; Distributed Computing
Story Type: News
Related Elements: Technical paper, "Data-ring Relationships in the Web," posted on the arXiv physics archive at arXiv.org/abs/cs.NI/0302016

TRN

English Could Snowball on Net

By Ted Smalley Bowen, Technology Research News
November 21, 2001

The Internet's ability to connect a wide range of cultures would seem to bode well for diversity of all sorts.

But, while the technology is relatively neutral, the influences of political and economic power have made the Internet a virtual English-language empire.

Researchers from Tel Aviv University and the University of California at Berkeley have teamed up to gauge the nature of the relationship between linguistic patterns and Internet content.

Early returns from the work imply that English content will continue to dominate the Internet, although other studies predict different scenarios.

Currently about 70 percent of Internet content is in English, but only about 44 percent of Internet users are native English speakers. Worldwide, native Spanish speakers outnumber native English speakers, and the number of native Chinese speakers more than equals that of both groups. English dominates online because it was established early on as the lingua franca of the wired world.

The imbalance reflects a first-mover advantage that is common in networks of all kinds, according to Neil Gandal, an associate professor of economics at Tel Aviv University in Israel.

In this case, the language of Shakespeare, Mark Twain, H.L. Mencken, and Yogi Berra benefits from the snowballing effect of a popular medium attracting more users simply because it's popular. The language's popularity spurs more people to learn English, which increases incentives for content providers to cater to an English-speaking audience, which in turn makes it all the more popular.

The researchers examined whether these first-mover effects dictate that English will simply gain momentum and remain the primary online language, prompting even more people to learn it, or whether the demographic and economic realities of a polyglot world will turn the tide.

This question is especially pertinent because Internet use among non-native English speakers is growing at a faster rate than that of native English speakers. By 2003 only 29 percent of Web users will be native English speakers, according to one estimate.

The researchers analyzed the surfing habits of a usefully bilingual population — Canadians in the province of Québec. As of 1996, roughly 5.7 million Québec citizens counted French as their mother tongue, about 600,000 cited English, and about 60,000 listed both.

The researchers looked at users' overall time online and time spent at each of seven types of sites: retail, business and finance; entertainment, news, sports and technology; education; portals, searches and directories; services, including ISPs, careers, and hobbies; government; and adult.

To get a rough breakdown by language of the content surfed, the researchers wrote a spider program that identified the languages of the approximately 40,000 Quebecois URL domains visited.

The researchers compared the overall Internet use of the three linguistic camps by type of sites, regardless of the content language, and then looked at which factors determined the percent of the time devoted to English language sites.

The native English speakers visited English content sites 87 percent of the time and stayed online about 35 percent longer than their French-speaking neighbors. The native French speakers, however, surfed in English a still considerable 64 percent of the time.

The differences also narrowed with age: younger native French speakers looked at more English content than their elders.

The finding that native French speakers are hurdling the linguistic barrier and turning to English sites for content not available in French is evidence that English's first-mover advantage is still snowballing, according to Gandal. These network effects are likely to continue to favor creating content in English and to lower incentives to do so in French, he said.

These preliminary results also indicate that the Internet is increasing the incentive for non-native English speakers to learn English as a second language, which could in turn promote English as a global language, according to Gandal.

In addition, although automatic translation technologies may eventually break down linguistic barriers, they are currently too limited to be a likely influence on the choice of content language, said Gandal. "Translation is very difficult because of the subtlety involved in the use of language," he said.

Computer-generated translation does work well for finding simple information like a train or airline schedule or the location of a particular office, but does not convey more complicated communications like disease diagnosis or an explanation of how to make a retail purchase, said Gandal. "We don't think that they will play a prominent role in the choice of language content in the foreseeable future."

The issue of language representation on the Internet is a contentious one, and is complicated by widespread financial stakes and cultural implications. The researchers' conclusions contradict those of the Foundation for Networks and Development, a private regional development organization in the Dominican Republic.

The current predominance of English on the Internet is largely due to the network's American origins and because the first wave of users worldwide is more likely to speak English as a second language, said Daniel Pimienta, director of the Foundation.

The foundation's statistics show that this is changing, he said. For instance, three years ago 75 percent of Web pages were in English, but that number has dropped to 50 percent today. In addition, the number of English Web pages as a percentage of the population of the world that speaks English as a native or second language is falling relative to Spanish, French, Italian and Portuguese, he said.

As the Internet's population becomes more diverse and an increasing percentage of its users lack English skills, the early predominance of English will continue to fade, he said. "As the Internet evolves toward a more balanced geographical [distribution] and a more balanced socio-economic

distribution, the dominance of English will more and more appear as a transitional phenomenon and the representation of language in the Net will tend to become closer to the natural representation of the language in the world."

As this happens, however, English will retain a special role in bridging communities whose native languages are different, he added. "This is and will remain the case of English, but also of Spanish, French, Arabic and Chinese."

Under this scenario, monolingual native English speakers may be more likely to pick up another tongue, Pimienta said. "The Internet will probably represent a strong asset for the language training industry to add a second language to native English speakers."

The Tel Aviv and Berkeley team's choice of a mostly bilingual population like Quebec's makes it harder to gauge the factors driving the choice of language on the Internet, Pimienta said. That population is able to navigate in English, while 90% of the world population does not understand English, he said.

The Tel Aviv and Berkeley researchers are currently working on a model designed to distinguish among cultural and economic factors driving the spread of English and those effects specific to the Internet, Gandal said.

One goal is finding how closely the use of English online will hew to the demographic and economic realities of English speakers. "The question is whether the percent of Internet content in English will reflect... or... greatly exceed the percentage of native English speakers around the world, weighted by purchasing power," said Gandal.

The researchers plan to delve into data for all of Canada in an effort to quantify factors like the number of Internet pages read or transactions conducted that would justify continued use of and investment in a particular language, Gandal said. "The model will need to distinguish between adults who find it harder to learn a new language... and children who find it easier," and therefore get more out of the experience, he said.

The researchers' updated model will also help quantify the strong network effects favoring development in English and drawing the best bells and whistles to English sites which, at least initially, place non-English sites at a disadvantage.

As more precise language identification software emerges, the researchers will be better able to determine the breakdown of pages visited according to content language, according to Gandal.

Gandal's research associate was Carl Shapiro of the University of California at Berkeley. They presented the work last month at the Telecommunications Policy Research Conference (TPRC) 29th Research Conference on Communication, Information and Internet Policy in Alexandria, Virginia. The research was funded by the UC Berkeley.

Timeline: Now

Funding: University
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "The Effect of Native Language on Internet Usage", Telecommunications Policy Research Conference (TPRC) 29th Research Conference on Communication, Information and Internet Policy, October 27-29, 2001, Alexandria, Virginia

TRN

Study Finds Web Quality Time

By Kimberly Patch, Technology Research News
September 26, 2001

Ever swear at a computer because the two of you were not communicating very well?

The communications interface between computer and human has always been a sore spot, at least for humans. The increasing use of the Web, which has shifted human-computer communications out of the work realm and into our leisure time as well, makes it even more apparent that interfaces count.

Researchers from the U.S. Air Force have published a study that takes some initial steps in quantifying exactly what Web applications need to do to provide users with a better experience. The research could help developers keep to a minimum those experiences that lead to swearing at silicon.

The difficulty of judging the quality of a Web experience is that the possibilities are so broad, according to Jason Turner, a computer network countermeasures engineer at the Air Force Information Warfare Center at the Kelly Air Force Base.

"Unlike a traditional information technology system or network that might be designed or acquired for a specific function [like] banking, office automation [or] academic study, the Internet is simply a collection of connections with little task specificity," he said. "Everyone's experience with the Internet is undoubtedly influenced by many factors: browser, means of access, reason for use, just to name a few."

The researchers drew on established theories of human-computer interaction and information technology to conduct their study. The human-computer interaction theory they used defines ease-of-use as a function of three factors, said Turner. They are "utility — whether the system does what is needed functionally; usability — whether the users can actually work with the system successfully; and likability — whether the users feel the system is suitable," he said. These three factors are all balanced against cost, which includes capital and operating expenses as well as social consequences, he said.

The researchers also took into account regret theory from behavioral science, which places importance on factors that

prevent us from achieving our goals, especially after discomfort arises from a failure to do so.

The researchers surveyed Internet users in order to find what factors mattered to them the most in terms of having a good experience on the Internet. In the study, 148 users of varying ages and Internet experience levels identified key circumstances or issues that they associated with the best and worst experiences they had with the Internet.

The researchers found that users' perceptions of and feelings towards Internet use were predominantly determined by a relatively small, seemingly stable and apparently consistent set of conditions, said Turner. These ten critical quality-of-experience items "may typify the events or factors which influence the relative success with which the sample of Internet users were able to accomplish their goals," he said.

The critical factors were the relevance of information to a task, the availability of search engine options like multiple criteria, the amount of information available, the organization of information, the ease-of-use of the browser, the reliability of the connection, the clarity of directions for Web site navigation, whether links were up-to-date, the availability of information at remote servers, and perceptions of security and/or privacy.

How efficient we believe our Internet experiences will be "center[s] around our assessment of how successful we think we will be at using the Internet to accomplish our goals," said Turner.

The researchers also found that several factors which are often thought to enhance Web experiences fade from view if the interface as a whole doesn't provide a quality experience, said Turner. "Heavy graphics, sound, animation, et cetera might not be so important to the user if the interface doesn't first and foremost help the user achieve his or her goals by providing for those factors which relate to a high-quality usage experience," said Turner.

The researchers' study also suggested that the critical factors do not vary as users gain experience on the systems, said Turner. This means that investment in making systems more usable would apply to both novice and expert users.

Interface developers could probably save time and money by focusing efforts on features that enhance the quality of the interface experience and on those aspects of the interface that helped users attain their goals, said Turner. "It might help to ensure we design ... from the start those ... functions which will ensure [that] the network or application itself is not abandoned or underutilized, regardless of what task or role that network or application will eventually play," he said.

Studies that quantify in some way the quality of users' experiences are badly needed, said Albert Badre, a computer science professor at the Georgia Institute of Technology.

One of the major problems is those studies often focus on usability to the exclusion of user satisfaction, he said. They measure things like the time it takes to perform a task, and make an assumption that "usability in terms of performance — time performance, number of letters and so on — correlates positively with users' pleasantness of experience," said Badre. "That is, in my opinion, a wrong assumption," he said.

In order to test that assumption, "we really need to be able to answer questions like did [users] perform tests the way they expected, were they successful at achieving their goals, and how do they feel as they were performing — do they have a nice life, so to speak," he said.

For instance, "if you provide me with an aesthetically appealing Web site that is exactly the same thing as one that's very dry I will probably enjoy looking at the nice colors and nice pictures even though it might make it a little less efficient because the time to... download graphics is longer," he said.

Usability studies seldom directly measure the user's comfort, preferences, and pleasantness of experience in using any kind of technology, said Badre. This is because before the Web, "the main preoccupation of usability people was business software. And there you do want efficiency because that's what you focus on, productivity," he said.

But productivity is not the main goal when you're sitting at home trying to enjoy yourself surfing the Web, Badre said. "We did a study of that specific issue and [found] that when people have a different purpose they have a different idea of what they'd like to experience. I'm not saying that we don't enjoy good efficient interaction, but it could be there's more to it than simply that. And this is the kind of thing we need to investigate."

The Air Force researchers are "trying to get at quantifying that quality of experience. The objective... is a good one," Badre said.

The research is mainly meant to be a springboard for follow-on research, said Turner. The findings could be used to develop a practical application in two or three years, he said.

Turner's research colleague was Michael Morris, formerly of the Air Force Institute of Technology and now that the University of Virginia. They published the research in the June 1, 2001 issue of the *International Journal of Human Computer Studies*. The research was funded by the researchers.

Timeline: 2-3 years
Funding: Private
TRN Categories: Human-Computer Interaction
Story Type: News
Related Elements: Technical paper, "Assessing Users' Subjective Quality of Experience with the World Wide Web: Exploratory Examination of Temporal Changes in Technology Acceptance," *International Journal of Human Computer Studies*, June 1, 2001

TRN

Web Game Reveals Market Sense

By Kimberly Patch, Technology Research News

The exact workings of the financial markets are a mystery. It is clear that the collective decisions of many traders affect financial markets, but it is less clear how traders make decisions, and how these decisions affect each other.

Researchers from the University of Fribourg in Switzerland have tapped the Internet to investigate speculative trading behavior and found that people tend to employ one of two distinct strategies depending on the complexity of a financial market. The results also show that humans are good at filtering information.

In addition to ferreting out information about markets and human behavior, the method could eventually be used to train financial traders, said Joseph Wakeling, a researcher at the University of Fribourg.

The researchers used a Web-based financial game to gain results from several hundred people playing several tens of thousands of game turns against computer-controlled agents.

Playing the game is very simple, said Wakeling. It provides a market price history and asks players to predict if the next price movement will be a rise or fall.

The underlying mechanism that determines what happens is less simple, Wakeling said. For each person there are 94 computer-controlled players. Each player independently chooses to be a member of one of two groups — those predicting a rise, or those predicting a fall. Whoever is in the smaller of the groups — the minority group — wins that round and gains points. Those in the majority group lose points.

The price movement in the game is the difference in size between the two groups, said Wakeling. "If by the size difference. If the other [group] — sellers — is bigger, the market falls by the difference," he groups will be larger, said Wakeling. "We then assume that [players] will want to join the other group, which they think will be smaller, and so by doing this they affect the actual outcome of the market," he said.

The computer-controlled agents act as controls and make decisions using simple, well-defined strategies. The approach allows researchers to investigate the behavior of a single human in an environment that involves collective actions.

The results showed that human players are "quite good at spotting and exploiting market inefficiencies; they're also good at spotting what information is superfluous and not using more than is necessary," said Wakeling.

When the market complexity is below a certain level most players are able to use a logical, deductive approach to get the better of the market, said Wakeling. As market complexity increases, however, there is an observable limit to humans' ability to cope logically, he said. Beyond this threshold, people have to find other methods of decision-making.

That players' logical capacity should break down like this is not surprising, said Wakeling. What happens next is, however. "People are quite literally repeating the same prediction many times in succession," he said.

More surprising, the strategy performs better than random decision-making, said Wakeling. The open questions are what triggers the behavior change and why the repetitive strategy works.

The researchers have two ideas that may explain the behavior. It may be that as market complexity increases, the number of patterns the player must bear in mind to make a logical decision simply becomes too large to remember, said Wakeling.

Another possibility is that because fluctuations in complex markets are generally very small, it's difficult to try out ideas without actually changing the market situation, Wakeling said. In this case, "an attempt to exploit a pattern can actually destroy it," he said.

Repetitive behavior may outperform random behavior for a similar reason. "Because the market fluctuations are so small, if you change your position, this means that your action decides what the market outcome is," said Wakeling. "So by changing often you can put yourself at a disadvantage."

It could also be that players are picking up a different pattern than the one they use in simple markets. Over any given time period in a market, "there will be a slight bias in one direction — the market is rising overall, or falling overall," said Wakeling. "If you can work out what the long-term trend of the market is, by repeating the same action throughout that period you can exploit that slight imbalance," he said. The process is probably not conscious, but instinctive, he added.

The results also suggest that there is a real limit on the human ability to spot useful information in the markets, said Wakeling. If this is true, "contrary to the propositions of neo-classical economics, there will always be some inefficiencies left behind in the market," he said.

Today's relatively fast Internet connections made the experiment possible, said Wakeling. "Our experiment was able to take place because we now have fast Web browsers which can transmit dynamically-changing graphics at high-speed," he said. This allowed for a graphical interface without users having to download a program, which meant more subjects and thus quicker data for the researchers. "You simply log onto the Web site and you can play — it's all there in your Web browser," said Wakeling.

The researchers used a Web-based C program to do the number crunching and used Flash to construct the graphical interface.

The next step is to do more testing to find out why the transition between deductive and repetitive behavior exists, and why players choose the repetitive strategy rather than something else, said Wakeling.

The researchers' long-term goal "is to have a proper theoretical understanding of how humans make economic decisions, and how those individual decisions add up to the macroscopic behavior we see around us every day," said Wakeling.

A system to train financial traders that is based on the interactive minority game could be developed within three or four years, said Wakeling.

Wakeling's research colleagues were Paolo Laureti, Peter Ruch and Yi-Cheng Zhang. The work is slated for publication in Physica A. The research was funded by the Swiss National Science Foundation.

Timeline: 3-4 years
Funding: Government
TRN Categories: Applied Technology
Story Type: News
Related Elements: Technical paper, "The Interactive Minority Game: a Web-Based Investigation of Human Market Interactions," slated for publication in Physica A and posted at arxiv.org/abs/nlin.AO/0309033

TRN

Recommenders Can Skew Results

By Kimberly Patch, Technology Research News
July 2/9, 2003

Just how accurate are the recommender systems online media sellers use to allow buyers to pass on their judgments about books, movies and CDs to their fellow consumers?

Researchers from the University of Minnesota have shown that the way recommender systems are set up can affect the opinions they evoke, and that artificially high or low recommendations can raise or lower subsequent recommendations.

Displaying a prediction introduces bias, said Joseph Konstan, an associate professor of computer science and engineering at the University of Minnesota. "Lying by [skewing rankings] higher or lower... biases the subsequent rating in that direction," he said. "Even the 'correct' rating led people to select that value more often."

The distortion this chain of events induces may influence consumer buying in the short-term, but adversely affects long-term consumer trust in the system, said Konstan. "While a system can get away with a small degree of lying... in the long run dishonesty erodes trust and satisfaction," he said.

The researchers' work is consistent with a long line of psychology studies showing that people shift opinions to conform to a group, said Konstan. "There's a bunch of psychology research that suggests that people exhibit a desire to conform," he said.

In a 1969 study published in Sociometry, for instance, a research team headed by Serge Moscovici found that about a third of the people in a group would call a blue block green if the researchers planted a couple of vocal people in the group who called the block the wrong color.

The Minnesota researchers conducted three experiments with a total of 536 people in order to see how previous ratings affected the test subjects' recommendations.

They used the Movie Lens recommender system, which includes about 70,000 users, 5,600 movies and around 7 million ratings.

In the first experiment, the researchers asked participants to rate 40 movies the participants had previously rated. The experiment presented lists of 10 movies using four different recommender configurations. The participants used a 1- to 5-star rating scale. One configuration showed no predictions, the second showed predictions equal to the user's original rating, the third showed predictions one star above the original rating, and the fourth showed predictions one star below the original rating.

The results revealed that people were fairly consistent in re-rating movies when there were no other ratings on-screen. Participants gave the movies the same ratings 60 percent of the time, one star below the original rating 20 percent of the time and one star above the original rating 20 percent of the time.

The results also showed that having other ratings on screen, whether they matched the user's original rating or were one star up or down, influenced the second rating the user gave. When ratings were bumped up or down one star, participants rated nearly 30 percent of movies one star above or below the original rating, respectively.

In the second experiment, a group of people rated 48 movies for the first time. The researchers predicted what people's ratings would be, then added or deleted stars in the same way as in the first experiment. They then repeated the experiment with a control group without manipulating the ratings shown to the participants.

The users were again swayed by incorrect ratings. In addition, those shown incorrect ratings were more dissatisfied with the process than the control group, probably because they sensed that the predictions were inaccurate, according to Konstan.

Other research shows that people treat computers socially, similarly to the way they treat other people, said Konstan. "We speculate that this effect may be skewing ratings towards the computer-displayed prediction," he said.

The research did not distinguish between the users' actual preferences and the ratings they entered, said Konstan. "We do not know whether [the rating system] really changes the persons' preference, or just the rating they choose to enter," he said.

Following up on their 1969 experiment, Moscovici's group looked at people's actual preferences in addition to what they said, and showed that even those who did not call the blue block green rated blue-green slides as more green than pretests

predicted they would. The researchers produced similar results after going a step further by asking participants to rate the color of the afterimage they saw after looking at the slide. Afterimages are involuntary artifacts manufactured by the human visual system.

The Minnesota study confirms the line of research that shows that people tend to conform with suggestions, and points out that care is needed to avoid introducing biases in information interfaces, said Konstan.

In a third experiment, the researchers asked users to rate three sets of 15 movies they had previously rated using three different scales: thumbs up or thumbs down, a scale from -3 to +3 not including a zero, or a 0.5 to five-star scale in half-star increments.

This experiment showed that people prefer finer-grained scales, and that finer-grained scales are ultimately more accurate. Participants rated the half-star scale the most satisfactory followed by the plus or minus three scale, and were least satisfied with the binary scale. The finer-grained scales are more accurate because people tend to give borderline movies the benefit of the doubt when forced to rate on a coarse scale, according to Konstan.

To evoke recommendations that are as independent as possible, recommender systems should give consumers an environment that allows them to provide ratings without having to see previous ratings, Konstan said. And the system should provide fine-grained rating scales rather than simpler thumbs up, thumbs down ratings, he said.

The Minnesota researchers are ultimately aiming to better understand how interfaces, social and economic structures, and other design factors influence people's participation in and use of recommender systems, said Konstan. The design implications of the current results can be used immediately to improve recommender sites, he said.

Konstan's research colleagues were Shyong K. Lam, Istvan Albert and John Riedl. They presented the results at the Association of Computing Machinery (ACM) Computer-Human Interaction conference held in Fort Lauderdale, Florida April 5-10, 2003. The research was funded by the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "Good Ratings Gone Bad: Study Shows Recommender Systems Can Manipulate Users' Opinions," presented at the Association of Computing Machinery Computer-Human Interaction (ACM-CHI) Conference, Fort Lauderdale, Florida, April 5-10, 2003; "Influences of a Consistent Minority on the Responses of a Majority in a Color Perception Task," Sociometry 32 Moscovisi & Personnaz, 1980

TRN

Privacy
Rating Systems Put Privacy at Risk

By Ted Smalley Bowen, Technology Research News
July 25, 2001

The Internet has given us new ways of carrying out activities as diverse as shopping and political agitation, and many of these new modes share a strong dependency on the medium's shaky guarantees of privacy and anonymity. This uncertainty has led to a variation on the trap of guilt by association: the threat of exposure by indirect association.

The chance you take when you use a Web recommender is typical of this new jeopardy, which researchers at three U.S. universities have quantified into basic equations of risk and benefit.

A Web recommender, or recommendation system, is a consumer rating system popular with online buyers of books, movies, and other items whose merits are a matter of taste.

A Web recommender may, for example, suggest to a person who has rated only books about baseball that he might also like a book about ballet. The recommender would have this information if another person had rated books on both topics. The recommendation system could unearth this connection using a nearest neighbor algorithm, which searches for the query point, or data point nearest the reference.

In this example, the recommendation system, while supplying a form of advice, has also showed the baseball fan a weak tie, which in social network theory is a connection between groups that don't ordinarily interact. A malicious user could exploit this seemingly incidental piece of information, according to the researchers.

On the Web, weak ties can be combined with other information to trace individual users' identities. Such tracing robs users of the option to act anonymously, and can be used to mine personal, financial, political and other information and affiliations.

Even though the risks are intuitively apparent, it's difficult to quantify the odds of weak tie exposure.

Toward that end, a group of computer scientists from the Virginia Polytechnic Institute and State University, Purdue University and the University of Minnesota has analyzed the risks of exposure by mapping the types of connections users make — often unconsciously — when participating in recommendation systems.

"Our main goal was to quantifiably assess the benefits and risks," said Naren Ramakrishnan, a professor of computer science at Virginia Tech. Everybody talks of risks in terms of 'don't disclose credit card', 'don't disclose age and address'. But we hope to identify more subtle forms of risk involving seemingly harmless information," he said.

The researchers did this using graph-theoretic models, which show relationships and connections among entities in a way similar to family trees, highway maps and organization

charts. By mapping exposure risk, the researchers quantified the risks and benefits of recommendation systems in general.

"In our case, we use a graph-theoretic model to represent the connections between people and the artifacts they rate," Ramakrishnan said. Recommendation systems make connections between people based on their common recommendations. Such connections, or jumps, move beyond the common items to the people who rated them.

These jumps can be represented as social network graphs, which depict people and how they are related.

Recommender graphs go a step further and include the artifacts, or items that people have rated in common. With this information, it's possible to find the connection between a user making a query and one who has rated the item of interest, according to Ramakrishnan.

Although it's laborious, a user could game the system and sift for connections that can be traced back to individuals, said Ramakrishnan. "By varying the ratings, you might notice that the recommendations change," he said. "In addition, you might notice that a particular recommendation of book X happens only for some specific values for ratings. If you know something about the algorithm behind the recommender system, then you could reverse-engineer the rating by inspecting the behavior of the algorithm."

To calculate the risk and benefit inherent in a given recommendation system, the researchers drafted a rough formula: benefit = w/l2, where w is a connection or connections between people who have rated the same item or items and l is a sequence of such connections.

"The... higher the w, the higher the benefit. The lower the l, the higher the benefit. The "squared" is there to make the second statement a little stronger than the first," Ramakrishnan explained.

This formula applies to any recommender system that works by making connections, which is how most of today's e-commerce recommender systems work, said Ramakrishnan. "Its limitations are that it might have to be adjusted for individual domains. The formula as it stands is a good qualitative measure, nevertheless," he said.

The key is presenting risk in terms of how a person relates to the larger social context of a recommender system, he said. "Thus, the same person with the same ratings may not be at risk in a recommender system where he is just like everybody else; it is his uniqueness [within a given system] that is posing the risk."

The risk equation can be likened to the way an individual can be singled out in a crowd, said Ramakrishnan. "If you look like everybody else, nobody can single you out. If you wear crazy clothes, you will be immediately spotted. Similarly, if you rate like everybody else, sure you get along and there is no danger," he said. "If you rate crazily, on the one hand you provide a lot of benefit to the recommender, but then you are at risk."

The researchers are aiming to demonstrate the risks inherent in such rating systems and broaden the context in which they are considered, said Ramakrishnan. "We're still studying this area," he said. They are looking into the causes of weak links, looking for other ways of quantifying benefit and risk and are looking to derive new ways to manage recommendation systems, he said.

The use of social network theory to study Web dynamics is compelling, although the seriousness of these risks is debatable, said David Madigan, a professor of statistics at Rutgers University.

"Making the connection with the social network literature is fascinating. [But] is the privacy threat real? I don't think so," Madigan said. The researchers' example of identifying someone through their ratings seems "far fetched in the context of large-scale e-commerce," he said.

A more likely threat comes from old-fashioned violations of privacy agreements, according to Madigan. "While I might trust, say, Amazon.com, a less trustworthy e-tailer might try my name and password on lots of other sites and get a complete picture of all the stuff I buy," he said.

Ramakrishnan's colleagues were Benjamin J. Keller and Batul J. Mirza of Virginia Tech, Ananth Y. Grama of Purdue University, and George Karypis of the University of Minnesota.

Timeline: Now
Funding: University
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "When being Weak is Brave: Privacy Issues in Recommender Systems," posted on the Computing Research Repository at xxx.lanl.gov/abs/cs.CG/0105028

TRN

Fault-Tolerant Free Speech

By Kimberly Patch, Technology Research News
July 12, 2000

Free speech on the World Wide Web sometimes lasts only as long as it takes a secret police agent, judge, or corporate lawyer to swing into action.

In an effort to strengthen free cyber speech, two scientists from AT&T labs and a New York University grad student have put together software that allows for Web publishing that is both anonymous and difficult to remove.

The software, called Publius after the pen name of the Federalist Papers authors, encrypts a file and publishes it on many sites. Because it is encrypted, however, the sites carrying it cannot read it. Publius then breaks up the encryption key and sends single pieces, or shares, around to the sites. But the publishing sites still can't read the encrypted document.

"The Web servers have no idea what's being stored at their site because they only get one share and the encrypted file — they don't get information on where the other shares are," Avi Rubin, a research scientist at AT&T Labs.

To read a Publius document, a person must download a copy of it, then download a certain number of shares to reconstruct the key, then use the key to decrypt the document. The publisher can set the number of key shares needed to reconstruct the key. The default is three of twenty available shares.

This way, shutting down one, or even most of the servers carrying the document will not preclude people from continuing to access it. "If somebody shuts down 15 of the servers as long as there are still five or even three [available], then the key can be reconstructed," said Rubin.

Rubin said the researchers goal is to make the Web more censor resistant. "There are numerous examples where there might be pressure on someone to take down a Web site that somebody else doesn't like," Rubin said. "Imagine a very powerful chemical company that's dumping chemicals a river," said Rubin. "If you want to make people aware of it... but you don't want any retribution and you don't want [the company] to be able to take it down... then you might want to publish it on Publius," he said.

The researchers are making the software available for a two-month Internet trial starting July 28. It can be downloaded at cs.nyu.edu/waldman/publius.

After the trial, the researchers will make adjustments to the software and either continue the trial or do a new deployment, said Reuben. "Our goal is to have this thing existing and widespread on the Web," he added.

Avi Reuben's colleagues in the research are AT&T Labs Research Scientist Lorrie Cranor and New York University Ph.D. student Mark Waldman. They are presenting a paper on the subject at the Usenix Security Symposium in August.

The project was funded by a Usenix student grant awarded to Waldman.

Timeline: Now
Funding: Association; Corporate
TRN Categories: Internet; Computers and Society
Story Type: News
Related Elements: Anonymous Publishing Website

TRN

Scheme Hides Web Access

By Ted Smalley Bowen, Technology Research News
October 2/9, 2002

The ringing declaration that information wants to be free often bounces off a hard reality — the free flow of information can attract interference. The reality online is that censorship and surveillance are widespread and growing.

The everyday flow of ordinary Internet traffic, however, could provide cover for political dissidents, whistleblowers, or anyone else who wants to access censored information online without the activity being recorded or blocked by others.

Researchers from the Massachusetts Institute of Technology have come up with a scheme that could guarantee users access to data in such a way that their actions would not be monitored.

The development follows an age-old pattern. Strictures on communication traditionally provoke workarounds, from prisoners tapping on cell bars to con men gaming early telegraph systems to get the jump on stock market or horse race results.

Latter-day examples have played out on the Internet for years. Proxy software allows users to surf anonymously, covering virtual tracks by masking Internet Protocol addresses and other personal information; and the Web's hypertext transfer protocol — HTTP — allows users to encrypt requests for information. But these solutions have not proved watertight.

Proxy software, which serves as an intermediary to let people access Web pages anonymously, can draw attention and be blocked by censorship software. Common security protocol software can also fail to protect users' identities, and it can be stymied by firewall software.

The MIT researchers' scheme, dubbed Infranet, allows Internet users to navigate using standard hypertext transfer protocol without being noticed.

The key to the scheme's ability to allow users to avoid monitoring is that it handles covert communications without adding a conspicuous amount of traffic. To be useful, a covert Internet communications system needs to cloak transmissions well enough to foil most would-be detectors, but must also be efficient enough to permit reasonably speedy browsing.

Infranet consists of software for Web servers and browsers. The scheme's responder software runs on public Web servers that store or are able to access data that is blocked or banned for some parts of the Web. Its requester software runs on systems seeking secure access to that data.

The software employs a transmission cloaking method, tried-and-true public-private key and shared session key encryption mechanisms, and existing data-hiding schemes.

Public-private key encryption allows anyone to use a receiver's freely-available public key to encrypt a message so that only the receiver's private key can decrypt the message and access its contents.

A shared session key is a single key that can be used to decrypt the messages it was used to encrypt.

To gain access to blocked data using Infranet, the requester begins a session by sending a shared session key using a responder's public key. "As long as either the requester or

responder know how to communicate with the other initially, they can come to agreement on the session key," said Nick Feamster, a researcher at MIT's Laboratory for Computer Science.

The responder then uses the session key to send code to the requester that translates hypertext transfer protocol traffic into a kind of alphabet that will allow the requester to hide ensuing transmissions to the responder within ordinary requests for non-censored Web pages.

This coded alphabet is made up hypertext transfer protocol requests for pages on the responder's Web site, and the code is different for each requester. A request for a covert Web page consists only of a series of requests for permissible Web pages on the server.

The order and timing of the requests for openly available pages determines the covert request. "If the requester and responder agree on how visible HTTP traffic maps to hidden messages, then everything works," said Feamster.

The responder uses the shared session key to encrypt the requested information, uses separate data-hiding techniques to embed the encrypted information in non-censored material, and sends that material to the requester as ordinary hypertext transfer protocol traffic.

The scheme currently calls for hiding the data served to the requester in JPEG's, one of several types of image files that can be transferred using the hypertext transfer protocol. In theory, responders can hide data in many types of files served up by Web computers, including MPEG video streams, said Feamster. "Our basic philosophy is to leverage existing steganography and data hiding techniques for the downstream communication," he said. In downstream communication served to the requester, "we're dealing with a pretty traditional data hiding problem," he said.

Although the researchers chose to conceal the requested information in JPEGs, and embed requests in the order and timing of hypertext transfer protocol requests, the method could work with any number of bi-directional communications, said Feamster. "Many possibilities exist: instant messaging, news feeds, stock tickers, satellite radio, online games, just to name a few," said Feamster.

The main qualification of a suitably innocuous scheme is that the communications be largely unidirectional, with more downstream than upstream traffic. The cloaked requests need only contain small amounts of information, while the responses pack the censored data into larger, more ordinary files that are openly sent to the requester. This fits well with the uneven nature of most Web communications: requests for data typically require much less bandwidth than serving up that data.

The researchers tested Infranet by subjecting it to passive attacks by monitors that logged all transactions and packets passing through a given segment of the Internet, and to active attacks by detection schemes that mimicked Infranet systems.

The process of covertly requesting and then serving up data hidden within other files turns out to be reasonably efficient. Half of the researchers' tested requests fit in six or fewer served files, and 90 percent of the requests required ten or fewer files. The requested files could be concealed in typical Web images by adding about 1 kilobyte of hidden data to each ordinary transmission, which typically range between 5 and 50 kilobytes.

One potential drawback of with this type of scheme is that users might suspect that the scheme itself is a surveillance tool. This can probably be addressed by including existing mechanisms that ensure that users can trust downloaded software, Feamster said.

Another issue is how to conceal the initial download of the Infranet software, a problem the researchers are currently addressing, said Feamster. Physically distributing the software via disks is one way to minimize the risk of disclosure.

For a scheme like Infranet to succeed, the responder software would have to be installed on a considerable number of public Web servers. "We're thinking of starting with something on the order of 50 to 100," Feamster said. If the responder software were bundled with a Web server like Apache, active participants would be much harder to detect, according to Feamster. The researchers' requester prototype is an Apache module.

"The trick is that you need to allow clients to discover the responders," Feamster said. "But if it's too easy to discover all of them, the censor can simply block them. Thus, we have to have enough to make it difficult for the censor to keep up with where all of the responders are."

In the cat-and-mouse contest that pits censorship and surveillance against the free flow of information, time works against such schemes, according to Avi Rubin, a secure systems researcher at AT&T Labs. "[It] illustrates an arms race. Once the adversary, in this case, a censoring government, knows about Infranet and how it works, they can attempt to detect and block it," he said.

Infranet is an impressive, novel scheme, said Rubin. "This is a big step forward towards evading that kind of censorship," he said. "It's actually going to be a bit of work for the censoring bodies to counter this, so it forces them to put in some additional effort, thus raising the cost of censoring."

Infranet could probably be optimized to allow more information to be exchanged without detection, Rubin said. "They could eventually develop high-bandwidth covert channels," he added.

Feamster's MIT colleagues were Magdalena Balazinska, Greg Harfst, Hari Balakrishnan, and David Karger. The researchers presented the work at the 11th USENIX Security Symposium in San Francisco, August 5 through 9, 2002.

Timeline: < 6 months
Funding:

TRN Categories: Computers and Society; Computer Science; Cryptography and Security; Internet
Story Type: News
Related Elements: Technical paper, "Infranet: Circumventing Web Censorship and Surveillance," Proceedings of the 11th USENIX Security Symposium, San Francisco, California, August 5-9, 2002, and posted at www.usenix.org/publications/library/proceedings/sec02/feamster.html

TRN

Security

Data Protected on Unlocked Web Sites

By Chhavi Sachdev, Technology Research News
December 19/26, 2001

Looking up information on the Internet is easy, but is it sometimes too good to be true? How do you know that a posted investment history, for instance, is correct and complete?

Existing technology allows an author to use a digital signature to authenticate a document. The author signs the document using a private key program, which performs a mathematical calculation on the document. To view the signature, the reader downloads the authors public key, which can be posted in a publicly available place.

But existing signature schemes only work with specific sets of data. To request the last two years of that investment history, for example, you might have to download the entire record to get an authenticated copy.

A team of researchers has come up with a signature scheme that allows portions of signed documents that are stored in Extensible Markup Language (XML) databases to be retrieved and authenticated. "The existing XML signature standard won't let you do that. You can only authenticate an entire document, not parts of it," said Premkumar Devanbu, an associate professor of computer science at the University of California at Davis.

Using the researchers' TruthSayer scheme, an author can also sign an XML document and give it to someone else to store and post, said Devanbu. In other words, the author would not have to be the publisher in order to authenticate the material. This means that anyone, from a government agency to the Mafia, could have a Web site that published authenticated data from multiple sources, and the receiver would be able to verify the origin of the documents, Devanbu said.

When the originator of the data uses the scheme to sign a document, the system processes the data involved, including its indexes, which are pieces of software that handle queries from clients and speed up searches, said Devanbu. "Typically, only a tiny fraction... of these indexes need to be looked at to answer the client's query. It is actually this index that is digested in a special way, to compute the database signature in our scheme," he said.

The secure data is then sent to an untrusted publisher. "When the publisher gets a signed [answer] from the owner, he checks to see if that's right using the owner's public key," said Devanbu. When anyone queries the data, the publisher provides the response and a verification code to prove that the accompanying answer is accurate and complete, he said.

When an untrusted online site gets a client query, it searches through the indexes, keeping track of which parts of the index were searched, and returns those parts along with the answer, Devanbu said. "The client now runs a [verification] program over the answer [and] the returned parts of the index."

The verification program compares the publicly available author's key with the publisher's certificate. "The critical thing about the verification [code] is that it doesn't depend on any keys at all. It uses a... digesting operation to prove that the answer that was sent by the publisher was the same as the answer the owner would have given," said Devanbu.

If the comparison proves a match, the client knows the data has not been compromised. If there is a discrepancy, she knows the data has been changed by someone other than the author.

"If a bad guy replaces a publisher's copy of the owner's public key with a forged public key, then the bad guy can make the publisher trust an invalid root hash value, and deceive the publisher into publishing bad data," said Devanbu. "But as long as the clients have the correct copy of the owner's public key, they won't believe this deceived publisher."

To digest documents, the signature system uses the Merkle hash tree mathematical function. The function starts with a set of data and computes until there is only one root value left, which is the key the author uses when he signs a document, said Devanbu.

The scheme could be used to retrieve authenticated portions of published data, from traffic citations and court proceedings to Freedom of Information Act requests, "all of which are either already or soon will be in XML," said Devanbu. In short, "any situation where correctness of data and efficiency of access is important."

"Suppose the government signs a large XML document containing all discussions within the Department of Labor on some topic, and gives it to another agency to handle responses to FOIA queries," said Devanbu. "Someone in the Department of Labor who wanted to hide something might try to coerce the person at the agency handling FOIA queries to hide some details in responses to queries. With [Truthsayer,] a false or incomplete answer to queries on the XML document would be detected immediately," he said.

Another advantage of this encryption scheme is that the owner of the data does not have to be online. "If the owner is physically disconnected, he cannot be hacked, and no one can steal his private key. So his signature is not forgeable,"

said Devanbu. This type of system is called an 'air gap' and is used by many Defense Department systems, he said.

This work is elegant and efficient and could spur further developments in this area, said Andrew Odlyzko, a professor of mathematics and the director of the Digital Technology Center at the University of Minnesota.

The most important feature of this scheme is that it could "provide authenticated information access through untrusted intermediaries," Odlyzko said. People might, however, opt for simpler solutions than this one because the threat the authors scheme guards against is probably not all that serious, he said.

The researchers are getting ready to test the scheme with a realistic, open-source database system, said Devanbu. It could be ready for practical use in 4 to 6 years, he said.

Devanbu's research colleagues were Michael Gertz, April Kwong, Chip Martel, Glen Nuckolls, and Philip Rogaway from the University of California at Davis, and Stuart G. Stubblebine of Stubblebine Consulting, LLC.

They presented the research at the 8th ACM Conference on Computer and Communications Security held in Philadelphia between November 5 and 8, 2001 and is scheduled to be published in the Computer Security Journal, 2001. The research was funded by the National Science Foundation (NSF), and the Defense Advanced Research Project Agency (DARPA).

Timeline: 4-6 years
Funding: Government
TRN Categories: Cryptography and Security; Internet; Databases and Information Retrieval
Story Type: News
Related Elements: Technical paper, "Flexible Authentication of XML Documents," in the 8th ACM Conference on Computer and Communications Security in Philadelphia, November, 2001; Technical paper, Authentic Re-Publication by Untrusted Servers: A Novel Approach to Database Survivability," presented at the Third Information Survivability Workshop 2000, October 24-26, 2000, in Boston

TRN

Scheme Harnesses Internet Handshakes

By Eric Smalley, Technology Research News
September 12, 2001

Whenever you click on a link on a Web site, your computer sends a message to the site's Web server and the server responds. Billions of such network handshakes take place on the Internet every day. Although individually these handshakes are insignificant, large numbers of them can add up to an impressive amount of computer processing power.

A team of researchers at the University of Notre Dame has figured out a way to use Web server handshakes to compute small pieces of a mathematical problem by disguising the pieces as ordinary Web browser messages.

The researchers' parasitic computing scheme uses the processing power of unwitting Web servers by exploiting one of the most basic operations carried out by all computers connected to the Internet, the Transmission Control Protocol (TCP) checksum, said Vincent W. Freeh, an assistant professor of computer science and engineering at the University of Notre Dame.

TCP breaks messages from one computer to another into small pieces, or packets, sends them over the network and reassembles them on the receiving end. The TCP checksum adds up the number of bits in the message and attaches the result to the message. On the receiving end, the computer adds up the number of bits received and compares it to the checksum number to make sure the message arrived intact.

The researchers performed their experiment with a type of math problem that can only be solved by examining each possible solution until the right solution is found. They encoded each candidate solution as a Web browser request for a web page so that the TCP checksum was actually checking to see if the message contained the correct solution.

The Web servers that received requests treated the messages that contained failed solutions to the math problem as corrupted messages and discarded them. The Web servers treated messages that contained the correct solution as a request for a Web page that did not exist and sent the standard 'page not found' error message to the researchers' computer.

Although the parasitic computing scheme demonstrates a principle, it is not a useful tool because the amount of computer resources used to implement the scheme far exceeds the amount that would be needed to solve the problem on the researchers' computer by itself, said Freeh.

"For a general communication protocol, I think the probability [of developing an efficient version of the scheme] is very remote," he said. "By design, the receiver doesn't have to do that much. However, I think people are [already] exploiting specific Web sites."

Web servers that run interactive applications and process forms are good candidates for this kind of scheme, said Freeh. "This is where lots of host cycles can be gotten," he said.

The parasitic computing scheme raises the possibility that computers on the Internet can be used in ways their owners are unaware of, which raises ethical and legal issues about the use of publicly available computer resources.

Though the Web server resources used in the Notre Dame implementation were barely measurable, if the scheme were used aggressively it could have a similar effect to denial-of-service attacks in which one or more Web servers are flooded with messages and effectively shut down, said Ian Foster, a computer science professor at the University of Chicago and a senior scientist at Argonne National Laboratory.

Each tiny message in the parasitic computing scheme is by itself indistinguishable from any other Web page request, said Freeh. "The way to tell is by seeing many such messages and deducing what is happening," he said. The researchers have configured an intrusion detection system to detect their parasitic computing scheme and they are working on configuring the system to detect variations of the scheme, he said.

"The instance of parasitic computing that [the researchers] demonstrate... is totally inefficient, returning a minuscule amount of computation for great effort," said Foster.

The question is whether there are more efficient versions of such a scheme, he said. "Within the Internet infrastructure [it] seems very unlikely to me, given its fundamental simplicity." It's more likely, though still doubtful, that someone could develop an efficient scheme to exploit peer-to-peer networks like Gnutella, he said. "I wouldn't discount it totally, especially as these infrastructures evolve."

Even if computationally efficient versions of the scheme can be developed, it remains to be seen if it can perform useful work, said Miron Livny, a computer science professor at the University of Wisconsin.

"It's a creative idea [but] it's not clear to me how it will work if you really care about the result," said Livny. The the problem is the scheme counts on messages that do not generate a reply to indicate that the message did not containing the correct solution, but failing the TCP checksum is not the only reason a message might not be returned. "The biggest challenge in distributed systems is to understand why somebody is not responding, because there [are] many, many reasons why you didn't hear back," Livny said.

The researchers tested the reliability of their scheme by repeatedly sending out the correct solution. They got the correct answer back at rates that varied from about 99 out of 100 to about 16,999 out of 17,000 times, according to Freeh.

Freeh's research colleagues were Albert-Laszlo Barabsi, Hawoong Jeong and Jay B. Brockman of Notre Dame. They published the research in the August 30, 2001 issue of the journal *Nature*. The research was funded by the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Internet; Distributed Computing
Story Type: News
Related Elements: Technical paper, "Parasitic Computing," *Nature*, August 30, 2001

TRN

Device Guards Net against Viruses

By Kimberly Patch, Technology Research News
December 17/24, 2003

Keeping a computer safe from viruses usually means installing virus-catching software and keeping it running and updated. Not everyone takes the trouble to do this, and viruses spread because there are enough unprotected machines to propagate them.

Researchers from Washington University and Global Velocity have come up with an alternative way to stop computer viruses and Internet worms.

The Field Programmable Port Extender is reconfigurable hardware that can protect an entire network at a time from viruses and worms. Information sent over the Internet is broken into packets that are reassembled at the data's final destination. The Field Programmable Port Extender scans every byte of data contained in every packet that passes through a network and stops packets that contain an Internet worm or computer virus signature.

Computer virus and worm software is designed to propagate throughout a network, just as biological viruses spread through a host population. And like biological viruses that can sicken hosts, computer viruses can damage computers by altering, destroying or sending files. Viruses attach themselves to or replace existing software. Worms, which are less common, are separate programs.

Because the Washington University system stops viruses and worms at the network level it has the potential to eradicate them more thoroughly than software running on end-user's computers, according to John Lockwood, an assistant professor of computer science and engineering at Washington University and co-founder of Global Velocity. "It could be used to instantly stop the spread of a virus," he said.

The system is fast enough to search for viruses in the wide flow of backbone Internet traffic because it uses hardware rather than software.

Hardware is faster than software, but is generally less flexible. By using reconfigurable hardware, however, the researchers were able to construct a system fast enough to filter data going through high-speed network backbones and flexible enough to add virus and worm signatures quickly as they are discovered. The researchers' device filters data at 2.4 billion bits per second, said Lockwood. "Software-based systems don't operate even close to fast enough to be usable on high-speed network backbones," he added.

The hardware generates a large number of customized circuits that each scan data for a certain type of virus or worm. The researchers developed a Web-based interface for the system that allows a network manager to easily add new worm or virus signatures, according to Lockwood.

The device is the result of several different ideas, said Lockwood. The concept of using reconfigurable hardware to

selectively block data from passing through a network came first. Next, the researchers had to work out how a custom hardware machine could be built and used to scan, modify and take action on data. Then they had to figure out how to scan for thousands of signature strings of data simultaneously.

And to make the device practical, the researchers had to build the protocol processing circuits that could examine Transmission Control Protocol/Internet Protocol (TCP/IP) traffic at very high speeds and identify viruses and worms even when the bits of malicious software are broken up among multiple packets and interleaved among multiple traffic flows, according to Lockwood. TCP/IP is the software used to direct Internet traffic.

The system is ready for practical use now. "We have a working prototype of the platform running," said Lockwood. "We're working with partners to deploy systems into remote networks now," he said.

Lockwood's research colleagues were James Moscola from Washington University and Matthew Kulig, David Reddick and Tim Brooks from Global Velocity. They presented the work at the Military and Aerospace Programmable Logic Device (MALPD) conference in Washington, D.C. September 9 through 11, 2003. The research was funded by Global Velocity.

Timeline: Now
Funding: Corporate
TRN Categories: Cryptography and Security; Internet
Story Type: News
Related Elements: Technical paper, "Internet Worm and Virus Protection in Dynamically Reconfigurable Hardware", Military and Aerospace Programmable Logic Device (MALPD) conference, Washington D.C., September 9-11, 2003 and posted at www.arl.wustl.edu/~lockwood/publications/MAPLD_2003_e10_lockwood_p.pdf

TRN

Address Key Locks Email

By Chhavi Sachdev, Technology Research News
October 31, 2001

When you pay for a book online or check stock quotes from your mobile phone, your password and credit card number are kept secure by an encryption scheme; one of the most widely used ways to spy-proof transactions is to use encryption keys.

In this type of encryption, each party has two keys: one to lock, or encrypt, messages and the other to unlock, or decrypt, them. If I wanted to send you a confidential message, I would look up your public key, use it to encrypt the message and then send my message to you. The only way to decipher the coded message would be your private decryption key.

Looking up a public key takes time and requires the receiver to first set one up. A pair of researchers has made the process easier with a scheme that automatically generates public keys using something most people have already made publicly available: an email address.

Using a person's unique email address as a public key makes it possible to send encrypted messages without having to look anything up.

Common encryption schemes like RSA can also use names to generate public keys, but not everyone can get a key based on a name because only one John Smith can use the John Smith name key; also, like getting a phone number that spells your name on a phone key pad, there is a certain amount of overlap. Using unique email addresses solves this problem.

"What we have tried to do is to create a new public key encryption scheme... designed so that... every user will get a valid key," said Matthew Franklin, an acting associate professor of computer science at the University of California at Davis.

All public key algorithms are based on difficult mathematical problems, said Franklin. The security of RSA, for instance, depends on a mathematical problem that is closely related to factoring large numbers. Two factors multiply together to produce a number. For example, 3 and 5 are factors of 15. Finding the particular factors of a really large number is very difficult because there are so many possibilities. RSA uses the large number as the public key and the two factors make up the private key.

The researchers' algorithm uses mathematics based on the Weil Pairing, a mathematical function that takes as input two points on an elliptical curve. Although the mathematics is different, "the speed of encryption and decryption and [the] size of keys and ciphertexts for our scheme is comparable to... popular public key encryption schemes such as RSA and ElGamal," said Franklin.

To send an encrypted email message, the sender would use an email program that incorporated the encryption scheme and could automatically generate the public key using the email address of the recipient, said Franklin.

The system's drawback is that it requires a central administrator who authenticates users and assigns private keys, he said. "When the recipient gets the encrypted email, she won't be able to decrypt it until she registers with the proper authorities to get her private decryption key," said Franklin.

Registering is a one-time burden for the recipient. "Once she has her private decryption key installed in her mail program, she can read any encrypted email that comes to her from any sender," he said.

The catch to having a central administrator is that someone would be privy to all encrypted email. The master key, however, could be split among several parties. "The functionality of the master key can be split among many parties — geographically distant, mutually suspicious —

which greatly decreases the chances that its power will be abused," Franklin said.

The work is novel and potentially useful, said Andrew Odlyzko, director of the Digital Technology Center at the University of Minnesota. The researchers have provided "a clean solution to a famous problem... that has been open for a long time," he said.

"Key management is a very complex problem with conventional cryptosystems, and public key cryptography was invented largely to solve its difficulties. However, it turns out that public key systems also have their own... difficulties. The authors' system is a nice solution," Odlyzko said.

The reliance on a central authority means identity-based crypto systems are not an easy sell, however, and any new scheme is not likely to be accepted quickly, he said. "Known public key systems tend to be preferred, and new ones are slow to be accepted."

Although identity-based cryptography has been proposed before, this research is excellent, said Ronald Rivest, one of the creators of the RSA encryption scheme and a professor of computer science at the Massachusetts Institute of Technology.

While there are no technical barriers to implementing the proposal immediately, "it would be prudent to give the cryptographic community more time to assess the strengths and weaknesses of our proposal," Franklin said.

Franklin's research colleague was Dan Boneh of Stanford University. They presented the research at the 21st Annual International Cryptology Conference held at the University of California at Santa Barbara from August 19 to 23, 2001. Boneh was funded by the Defense Advanced Research Projects Agency (DARPA) and the Packard Foundation; Franklin was funded by the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Cryptography and Security; Internet
Story Type: News
Related Elements: Technical paper, Identity-Based Encryption from the Weil Pairing," presented at the 21st Annual International Cryptology Conference, University of California at Santa Barbara, August, 2001; Demo: crypto.stanford.edu/ibe

TRN

Email

Teamed Filters Catch More Spam

By Chhavi Sachdev, Technology Research News
August 22/29, 2001

How much unsolicited email do you find cluttering your inbox every morning? Even when Internet service providers block junk email, spam creeps in, disguised, for example, as innocuous messages from people who seem to have only first names. At the same time, spam filters sometimes block legitimate messages.

A group of researchers in Greece has come up with a method that could solve both problems.

Most spam filters work by doing two things: they block known spammers who have been blacklisted, and they follow general rules, such as blocking messages that contain the word 'adult' in the subject header, said Ion Androutsopoulos, a research fellow at the Demokritos National Center for Scientific Research (NCSR) in Greece.

But spammers frequently forge email addresses to get around the blacklists, and those filters that use keyword-specific blocking might also nix that funny anecdote about your brother's kids that contains the word 'nude.'

The NCSR spam program creates custom filters for each user that learn what is spam and what is not, said Androutsopoulos. The filters learn to tell the two apart by looking through a user's legitimate email and comparing it with lots of spam collected by the researchers, he said.

The key to the process is using several filters that work together. The researchers found that they could bolster accuracy by combining filters based on different learning algorithms that individually made different types of errors, said Androutsopoulos.

The program analyses the user's existing mail using Natural Language Processing algorithms to build the set of anti-spam filters, said Androutsopoulos. It calculates the probabilities of certain words appearing in spam versus legitimate messages and classifies incoming messages by comparing them with previously analyzed email.

"The individual filters are treated as members of a committee presided [over] by a higher-level classifier, which is trained to learn when to trust each of the members," said Androutsopoulos. When a new message arrives, the committee members cast their votes on whether the message is spam. "The president of the committee then makes the final decision by taking into consideration the opinions of the members, the message itself, and its previous experience regarding when to trust each member," he explained.

The stacked spam filter is more accurate than keyword-based spam filters, Androutsopoulos said. It identifies about 90 percent of junk email accurately, and mistakes a legitimate email for spam about 1 percent of the time, he said. The accuracy could be increased further by returning messages classified as spam to their senders and asking them to change the address, he said. If the email is legitimate, the originator can send it again to a different, unfiltered address.

"Training the filter takes a few minutes per user, depending on the number of training messages. Classifying an incoming message is almost instantaneous," said Androutsopoulos. When the filter is configured separately for each user, it could be installed either on the end user's desktop or on the ISP's

server. "In the latter case, the ISP would run the user's filter on behalf of the user before downloading the messages to [a] desktop, saving bandwidth wasted by spam messages," he said.

The same configuration of the filter can be applied to all users on a network, said Androutsopoulos, "but I would expect the accuracy of the filter to be worse than when using filters especially configured for each user." The training time will also go up because more training messages would be needed for a pan-network filter, he said.

Better spam filters are definitely needed, said Ben Gross, a visiting scholar at Berkeley, and a coordinator of the Digital Libraries Initiative Phase Two for the National Science Foundation (NSF). "Spam remains a nearly intractable problem for most users [and] better Natural Language Processing techniques for spam could certainly improve the current state of technology," he said.

An important variable the researchers did not discuss, which may bear on the scheme's use in large networks, is time. "For a system to be viable for large scale deployment with email it must be highly efficient," said Gross. Still, if a spam filter's performance were to prove inadequate, it could be deployed at the users' desktops, he said.

The stacked spam filter could be used by firewall makers, listserve moderators, newsgroups, ISP's and individual users, said Androutsopoulos. It will be ready for such use within a year, he said.

The researchers' next step is to improve the filters by evaluating more thoroughly how the filters work and improving the system's learning algorithms, according to Androutsopoulos. The researchers would like to make the system's training period faster, he said.

Androutsopoulos's research colleagues were George Sakkis and Panagiotis Stamatopoulos at the University of Athens and Georgios Paliouras, Vengelis Karkaletsis, and Constantine D. Spyropoulos at the Demokritos National Center for Scientific Research. They presented the research at the 6th Conference on Empirical Methods in Natural language Processing held in Pittsburgh, PA on June 3 and 4, 2001. The research was funded by the universities.

Timeline: >1 year
Funding: University
TRN Categories: Natural Language Processing; Internet
Story Type: News
Related Elements: Technical paper, "Stacking Classifiers for Anti-Spam Filtering of E-mail," in the Proceedings of 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001), at Cornell University on June 2001

TRN

Email Burdened by Management Role

By Ted Smalley Bowen , Technology Research News
January 2, 2002

By now, our computing lives were supposed to be anchored in personal information management software, groupware, Web browsers, or any of a number of specialized packages that would organize our time and information better than was possible off-line.

In a case of make-do evolution, however, the humble email program has become a primary organizer, haphazardly covering a far broader range of functions than it was originally designed for.

Email has taken on this information-central role largely because it is a primary communications conduit. Computer users tend to stretch the functions of email programs to organize information where it accumulates rather than shifting it to more specialized software.

Because of this, email's practical uses have mushroomed to include organizing and transferring files, managing work flow and scheduling, and even organizing to-do lists. People also share Web pointers using email, making it a point of departure for going online.

A pair of researchers at the Xerox Palo Alto Research Center are taking a look at how people use email in an attempt to improve it to meet these greater demands.

The researchers conducted case studies at three organizations: their own 300-person office, a 150-employee multimedia production firm, and a group of six design consultants.

They interviewed 28 people who had used email for an average of 11 years. The subjects had used their current email programs an average of three years, and sent an average of seventeen messages a day, and received 42.

Not surprisingly, the researchers found that email tends to be overextended.

People use it to organize reams of data and transfer file attachments even though its relatively primitive file management abilities do not include the ability to specify different versions of files, said Xerox PARC research associate Nicolas Ducheneaut, who is also a graduate student at the University of California at Berkeley.

These file transfers are often slower than if they had taken place over a network. People also use email to schedule meetings, organize project details, and even do audits, even though it is not specifically designed to handle those tasks, said Ducheneaut.

The researchers found that 72 percent of the interviewees sent to-do lists to themselves, and 83 percent of the group said they left messages in their in boxes as reminders.

Despite their computer experience and heavy use of email, the sample group did not customize the software to suit their needs, he said. "It seems email users are strongly influenced

by the first interface they are presented with and then stick to it. Options [that] are frequently buried deep inside menus [do not] encourage experimentation."

More experienced emailers tended to use more folders. The most common number of folders was 27, and the average 91. One inbox bulged with more than 400 folders. Previous research has found that using more than 30 folders to organize messages is of debatable benefit, however, said Ducheneaut. Simply counting folders did not reveal how many were in common use, he added.

Workers generally categorized folders by organization, project and personal interests. They also sorted through messages in order to find something rather then using a search tool, probably because it's faster to sort than to start up a search engine, said Ducheneaut.

Sixty percent of the users eschewed filters either because they were not accurate enough or they were too difficult to use. "Sorting and foldering were used more than searching, but there is still space both for traditional foldering and searching in email," Ducheneaut. Searching, however, should be more closely modeled on users' practices, he said.

Even an email package that addressed these limitations wouldn't automatically reach the masses, Ducheneaut said. About 40 percent of the interviewees were required by their organizations to use a certain email package, and training was generally not provided. This institutional inertia will not necessarily be solved simply by building a better mousetrap, he said.

Lack of training could explain why people do not tend to experiment with email options. "We know very few computer users look at their software's documentation, so there is a strong tendency for them to use only the tip of the features' iceberg, so to speak," said Ducheneaut.

The study found that email use is only somewhat affected by the physical layout of an organization's workspace. Although some of the office spaces offered ample opportunities for face-to-face exchanges, people still preferred to use email to exchange documents and URLs, probably because it serves as an electronic paper trail, said Ducheneaut.

The study also showed that email patterns reflect some basic organizational structures.

For instance, Xerox PARC mail folders tended to be organized by project, multimedia production firm folders by department, and design consultant folders as either personal or professional.

People at Xerox PARC sometimes toted laptops to meetings and tapped out email during the proceedings, but used email less to assign tasks. The consulting firm relied less on email to document activities, which the researchers took to reflect the nature of a small start-up.

And the two larger organizations did more email broadcasting.

Email use also varied according to which rung on the corporate ladder a person occupied. Managers, for example, tended to fire off more meeting agendas than other workers.

Given email's critical role as communications hub, it is likely to continue to gain uses, and should be re-tooled with this in mind, said Ducheneaut.

The case studies showed a need for better ways of organizing folders, quicker ways to get to recently accessed items such as to-do lists and reminders, ways to track different versions of documents, and the ability to manage URLs. The research also raised the issue of tailoring user interfaces to individuals' roles within an organization.

The study is interesting and the findings sound, said Ned Kock, a professor of information systems at Temple University.

But studying the effectiveness of email also means considering video and virtual environments, Kock said. "The way to really make email better is to make it more face-to-face-like without giving up on its advantages." One way to do that would be to "allow people to share a virtual context, so cognitive effort is reduced," Kock said.

Using video as that shared context, or even exchanging video clips, is difficult today for many reasons, including the large amount of bandwidth required and the lack of uniform video and audio standards, Kock added.

The practical short-term solution is to improve people's ability to organize and exchange pointers to online content, Ducheneaut said.

Ducheneaut's research colleague was Victoria Bellotti of Xerox PARC. They published the research in the September/October issue of the Association for Computing Machinery magazine *Interactions*. The research was funded by Xerox.

Timeline: Now
Funding: Corporate
TRN Categories: Applied Computing
Story Type: News
Related Elements: Technical paper, "Email As a Habitat, An Exploration of Embedded Personal Information Management", the Association for Computing Machinery magazine *Interactions*, September/October, 2001

TRN

Email Takes Brainpower

By Kimberly Patch, Technology Research News
October 17, 2001

Should you send email or set up a face-to-face meeting? It's not a trivial question.

The two modes of communication are different in many ways, including how many words you use and how hard you have to think in order to come up with appropriate answers, according to a researcher from Temple University who has

tapped the principles of evolution to explain why we communicate the way we do.

According to evolution theory, organs are optimized over many generations because the animals that benefit from random genetic changes to their organs are more likely to survive and pass on their genes. Another principle of evolution says that the body and the brain that guides it must evolve together.

Over the five or six million years that it took for us to evolve from small-brained primates into loquacious Homo Sapiens, we communicated face-to-face, said Ned Kock, a professor of information systems at Temple University. "Our biological communication apparatuses as well as our brains were optimized for face-to-face communication. When we move too far away from face-to-face communications... extra cognitive effort is required," he said.

This doesn't mean that face-to-face communications is always better than email, said Kock. But it does go a step toward toward quantifying why the two types of communication feel different. Taking into consideration our natural predilections in communications can also help us improve electronic communications, he said.

Kock studied the way 38 process-improvement groups in three organizations worked over a little more than four years as the groups used either face-to-face meetings or email to do their jobs.

The groups that communicated via email produced slightly better results, according to the perception of the participants. The cognitive effort required, however, was much greater because people are not as fluent in written language as they are in spoken language, said Kock.

Put simply, it is more difficult to write a paragraph then to speak one, especially if the ideas involved are complicated concepts or descriptions. This is easily illustrated by accounting for how much time it takes to write versus how much time it takes to speak. "Say you have a certain number of ideas and you need a certain number of words — say 600 words — to explain those ideas. If you used email, chances are that it's going to take you more than an hour to convey those 600 words. Over a face-to-face meeting you'll probably be able to convey the same number of words... over maybe five or ten minutes," said Kock.

The upshot is it can be more than an order of magnitude more difficult to communicate electronically versus face-to-face. "If you use words-per-minute as a surrogate of cognitive effort... it is between 10 and 20 times more time-consuming, more cognitively demanding to communicate over email the same number of ideas than it is to communicate face-to-face," said Kock.

The effect grows as the communication becomes more complicated. "If the communication is very simple... say I'm giving you my phone number... you won't see the decrease in fluency because the communication is not complex enough," he said.

So why did the groups that used e-mail to communicate about the complicated subject of process improvement produce not only acceptable but slightly better results?

The process improvement groups adapted to the differences, said Kock. While the groups that met face-to-face communicated in meetings that averaged two hours, the email groups spent the same amount of time per person communicating over 40 days. The email communications generated less than half the number of words per person, but those words were more focused, said Kock.

Part of the extra effort in composing written messages was also balanced out on the other end. It is more efficient to read email then to listen to speech, said Kock. "Reading emails is probably about two times faster than having to listen to contributions face-to-face because... you can jump from one part of a contribution to the other." Although the written nature of email also allowed the groups using that medium to reread contributions, they didn't tend to do so, said Kock.

The more focused contributions of email ultimately proved an advantage. "Online you do have the opportunity to prepare a focused and bigger type of contribution and therefore you can condense more information into one contribution than [you can] face-to-face. [The email] focused on the topic at hand... and therefore they used fewer words, and achieved better results by using fewer words," he said.

Although email was a workable solution in a business environment where employees were motivated to adapt, there are many situations like customer relations where the extra cognitive effort required may scuttle communications, Kock said. "If [an online] interaction requires more cognitive effort from the customer... they will be less satisfied with that communication or interaction and therefore the probability that they will move to another provider... will be higher," he said.

Another place where online communications has proven more difficult than first imagined is online learning. "The amount of cognitive effort and therefore the amount of time required for [online] instruction is much higher than face-to-face-like instruction... even if you factor in transportation-related time, et cetera. Nearly all faculty and students that I have talked to support this," Kock said.

Recognizing why electronic communications require more effort could go long way toward making electronic communications more natural, and therefore easier, he said.

Until the last hundred years or so our natural communications always involved colocation, or holding a conversation in the same physical space, and synchronicity, or talking in real-time. We also naturally use the tone of voice, facial expressions and body language to add information to our speech. Given these extra, contextual channels of information, our brains don't have to work so hard to extract what is meant from the words alone.

It is possible to add some of this natural context to certain types of electronic communications to make them faster and

easier, Kock said. Using video clips in certain situations, for example, would add tone of voice, facial expressions and body language to electronic communications. Using chat-type communications would add synchronicity.

The research was carefully done, said Carrie Heeter, a professor of telecommunication at Michigan State University via San Francisco, and director of MSU's Virtual University. Although email is presumably a worse way to generate new collaborative ideas, it is probably a better way than face-to-face communications to mine the knowledge of each individual and group, Heeter said. One thing that may help asynchronous online discussions to be more natural is to encourage shorter posts, which are more like face-to-face conversations, she added.

Although in his paper on the research, Koch quotes a participant as saying that sometimes things are left hanging with email communications because people can have different interpretations of the same message, there may also be ambiguity in face-to-face communications, said Heeter. "I wonder, in face-to-face [communications], whether there is less perceived ambiguity, but in actuality perhaps even more disparate perceptions of what has been said. There is no recording of face-to-face [communications] other than each individual's memory. When I read minutes from a meeting I have attended, I'm often surprised," she said.

Kock is currently working on more finely quantifying the relative importance of natural contextual communications, he said.

He is also working with a psychologist to apply the research to a medical problem. Gradually increasing the naturalness of communications with other people can be used to help people who have social anxiety, said Kock. This is needed because many current treatments for this type of illness, which brings on panic attacks in those who have it, involve desensitization, he said.

Kock published the research in the April, 2001 issue of *Information Systems Journal*. The research was funded by the Department of Defense (DOD) and the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Computers and Society; Internet
Story Type: News
Related Elements: Technical paper, "Asynchronous and Distributed Process Improvement: the Role of Collaborative Technologies," *Information Systems Journal*, April 2001. "The Ape That Used Email: Understanding E-munication Behavior through Evolution Theory," Communications of the Association for Information Systems (AIS), February, 2001

TRN

Web tools

Browser Boosts Brain Interface

Technology Research News, June 4/11, 2003

Georgia State University researchers have come up with a Web browser that allows people to surf just by thinking.

Previous research has shown that it is possible to move a cursor by controlling neural activity. The researchers' BrainBrowser Internet software is designed to work with the limited mouse movements neural control allows.

The browser window is divided into an upper section that resembles a traditional browser and a lower control section. Common controls like "Home", "Refresh", "Print" and "Back" are grouped in the left-hand corner and provide feedback. When a user focuses his attention on a button, it becomes highlighted, and when the user successfully focuses on clicking the button, it emits a low tone.

The right side of the control section displays links contained in the current Web page. This allows the user to more easily scan and click the links.

The researchers are working on a virtual keyboard with word prediction technology that will allow users to enter URLs.

The technology will be ready for practical communications applications in two to five years, according to the researchers. They presented the work at the Association of Computing Machinery Computer-Human Interaction (ACM-CHI) conference in Fort Lauderdale, Florida, April 5-10, 2003.

TRN

Badge Controls Displays

By Eric Smalley, Technology Research News
January 28/February 4, 2004

Rooms and public spaces that sense human presence, notice where attention is focused, and recognize gestures and spoken commands promise to make interacting with computers as natural as interacting with people.

Technology that enables smart spaces — computer vision, gaze tracking and speech and gesture recognition — is available. These cutting-edge components are expensive and difficult to combine into a smoothly functioning system, however. They also introduce privacy issues.

Researchers at Lancaster University in England have achieved some smart space capabilities by instead combining several more established technologies: wireless communications, local area networks and Internet access. The scheme also promises to protect users' privacy.

The system causes screens near a user to display Internet-based information that the user is likely to prefer. It can also be used to control CD players and other media devices.

The system hinges on the pendle, a device that can be worn around the neck. The pendle contains a computer, wireless transmitter, acceleration sensor and touch sensor. It automatically transmits user preferences to nearby computers and can also be taken in hand to perform command gestures, said Nicolas Villar, a research associate at Lancaster University.

The other components of the system — wireless receivers, display devices and a computer that locates appropriate content — are connected to a local area network. Simply by wearing the device, a user is able to inform computers in the environment of his preferences so that the environment can best tailor its behavior to match the user's interests, said Villar.

The pendle stores lists of keywords and Internet addresses chosen by the user and transmits these to receivers placed around the room. The nearest receiver forwards the information along with its own identification to a computer that locates the display device nearest the user.

If the information transmitted by a user's pendle is an Internet address, the display locates the address and shows the page or clip. If the pendle transmits a list of keywords, the computer searches the Internet for relevant information and forwards an appropriate Internet address to the display device.

When the user picks up the pendle, its touch sensor switches the device to command mode. Holding it up causes the nearest display to access the next Internet address stored in the pendle. Shaking the pendle removes the current information from display.

The system can also be used to control sound clips. Shaking the pendle to the right, for example, can signal a device to advance a CD to the next track, said Villar.

The system makes it unnecessary to use computers that see and hear to determine a person's context, said Villar. "The pendle provides an easier way for the environment to make a guess at [a user's] intentions by providing a defined list of the user's preferences," he said. This also allows a user to determine what information to make available to the system.

Though pendles require unique ID's to allow users to issue commands, the IDs do not have to be linked to users' identities, according to Villar.

Similar technologies exist, including pendants for gesture control and smart badges that track a person's location so that preferred information can be displayed on nearby devices. The Lancaster researchers' system is different because it combines these capabilities to give people both passive and active means of personalizing display information, according to Villar. The result is a way to proactively display preferred information but also allow a person to override the system and explicitly control the display.

The initial prototype of the system displays Web-based information served up by Microsoft's Internet Explorer and Media Player programs. Information generated from keywords is displayed for two minutes, which keeps information relevant but also makes screen shifts infrequent enough not to distract users, according to Villar. Information generated from explicit commands remains until the user removes them or leaves the vicinity, at which point the system returns to displaying keyword-generated information.

The system can be set to display information related to the keywords for one or a few users, or to the pool of keywords from a wider number of users. One avenue for improving the system is using more advanced algorithms for finding relevant information, according to Villar.

The pendle system could be used in practical applications in two to five years, said Villar.

Villar's research colleagues were Albrecht Schmidt, Gerd Kortuem and Hans-Werner Gellersen. The work appeared in the December 2003 issue of Computers & Graphics. The research was funded by the European Union.

Timeline: 2-5 years
Funding: Government
TRN Categories: Human-Computer Interaction
Story Type: News
Related Elements: Technical paper, "Interacting with Proactive Community Displays," Computers & Graphics, December 2003

TRN

Software Guides Museum-Goers

By Kimberly Patch, Technology Research News
June 12/19, 2002

Reading the written material that goes along with museum exhibits is always a little tricky. If you're the type who has to read every word, you're likely to see the same background information over and over again, and if you're the type who likes to dip in and out of the text, you'll probably end up missing at least some of background material.

Researchers from Europe have built a system designed to tap the powers of hypertext, information databases, and natural language generation to allow people to go as deeply or as quickly as they wish through the written material in museum-type settings without repeating or missing much. "It occurred to me that... these problems can be addressed by using natural language generation technology," said Jon Oberlander, a reader in cognitive science at the University of Edinburgh.

The information can be displayed in several forms in physical places like museums and virtual spaces like the World Wide Web. "The same information server and generator can dynamically supply information to wireless handhelds in a real museum gallery, or drive synthetic speech over a mobile phone, or build Web pages on-the-fly to describe a virtual

gallery," said Oberlander. The system is also designed to work with any language.

There are several inherent problems with museum labels, according to Oberlander. First, they are generally designed to be accessed in any order. This means they must each represent all the relevant information about their object, which can mean overly wordy and redundant descriptions. "Small differences between two objects may be submerged in a sea of similar details," he said. Using traditional labels, the only way to avoid massive redundancies is to force visitors to read the descriptions in a certain order "and that's not great for their sense of freedom," he said.

"Secondly, there's no guarantee that [visitors] will actually find what they need," he added. In contrast, a live curator can find out what museum-goers want, present options, and, if necessary, steer them to objects they were not aware of, said Oberlander.

The researchers' system addresses those problems by generating answers to visitors' questions on-the-fly. It keeps track of what a visitor has seen in order to tailor the descriptions appropriately.

Someone visiting via the Web would start from a page of icons showing a gallery of objects, and when the visitor clicked on a particular icon, a new page would be generated, with a larger image, a title, a description and a list of links to related objects. "At this point they can return to the main page and choose another object, where they can follow one of the suggested links, or they can ask for more information about the current object. Either way a new page is generated for the chosen object [and] the description of the page will take into account what other descriptions have been generated so far, tailoring both content and form," he said.

Under the hood is software that includes four key components: a content potential module, a text planner, a surface realizer and a module that chooses the best presentation for the generated description.

The content potential module keeps track of, and links together, facts extracted from museum databases and curator interviews. It also places different values on each fact, depending on how important the curator judges it to be and how interesting and familiar it is expected to be to the visitor. This familiarity value changes throughout the course of a visit.

When a visitor requests information, the text planner module selects a subset of facts from the content potential module. "It starts from the... selected object, and includes all the facts which are nearby and sufficiently interesting, important and unfamiliar," said Oberlander.

The module takes into consideration the number of facts available for the current type of user, and organizes the information into a coherent order that signals explicitly how the facts fit together, Oberlander said. "The text structure built up this way is still essentially independent of the language which is used to express the information," he said.

The surface realizer takes this abstract information and chooses the best way of expressing it using grammatical constructions, words and connectives. "This is also where the system takes into account the different ways we refer to objects when we mention them the first time, [than] on subsequent occasions," said Oberlander. For example, the first time you mention a designer, you might say 'a British designer named Jessie M. King', then later refer to her as 'Jessie M. King', 'King', or 'she'.

The final module decides whether to wrap the textual description in HTML with live links, send it as pure text, or put it through a speech synthesizer.

In theory, the software can work with any language. The researchers are currently working with English, Italian and Greek. "One of the key challenges in the current project has been to cleanly separate the parts of the system that are independent of English, Italian or Greek from the parts that have to rely on knowledge of the particular language," he said.

In some ways, English is the easy language, Oberlander added. "The sophistication of the system [had] to be considerably increased for languages with complex word-information rules like Greek. But once you've done Greek, Italian is relatively easy," he said. In the end, it shouldn't cost much to add a new language, he said.

As part of the project, the researchers and a partner, the Foundation of the Hellenic World in Athens, have constructed an immersive view of the ancient city of Miletus using the software.

The researchers are also looking to use the software to mine many types of existing textual information, including online catalogs. "It will work with almost any kind of online catalog and in customer relationship management," said Oberlander. The researchers are also planning on using the system for tutoring, he said.

The software combines work in several different areas in a very interesting way, said Paul Aoki, a research scientist at the Xerox Palo Alto Research Center. "They're able to [make] previous technologies really deployable," he said. "You can imagine that typical audio guide content like overviews, jokes and dramatic stories would be tough to generate on-the-fly, but something like [this] could be used to weave pre-recorded pieces together with dynamic factual content."

The overall approach of generating text from a database of descriptive elements could have many uses, Aoki said. "There are many different... scenarios where this kind of technology can be applied — walks through historic districts, botanic gardens, historic houses. Another example might be an audio restaurant guide that knows you care about parking and price... and gives you natural-sounding descriptions that are tailored to those preferences," he said.

Oberlander's research colleagues were Ion Androutsopoulos and Aggeliki Dimitromanolaki of the Greek National Center for Scientific Research in Greece, Vassiliki

Kokkinhai of the Foundation of the Hellenic World in Greece, Jo Calder of the University of Edinburgh, and Elena Not of the Trentino Cultural Institute in Italy. They presented the research at the 29th Conference on Computer Applications and Quantitative Methods in Archeology held in Gotland, Sweden, April 25 to 29, 2001. The research was funded by the European Union.

Timeline: Now
Funding: Government
TRN Categories: Human-Computer Interaction; Databases and Information Retrieval
Story Type: News
Related Elements: Technical paper, "Generating Multilingual Personalize Descriptions of Museum Exhibits — The M-PIRO Project," presented 29th Conference on Computer Applications and Quantitative methods in Archeology in Gotland, Sweden, April 25-29, 2001

TRN

Content Scheme Banishes Browser Plug-ins

By Ted Smalley Bowen, Technology Research News
April 17/24, 2002

There's not much guesswork involved in pulling a book off the shelf. Little has changed since Johann Gutenberg came up with movable type — provided you read the language, you can just crack the cover and you're on your way.

By contrast, the babble of data formats represented on the present-day Internet considerably confuses the process of accessing digital information. Your basic Web browser can only handle so many data types, and the prospect of searching for and adding the right plug-in can be laborious even when successful. And as with all things digital, nothing stays the same for long.

In order to display what you want, your browser must be able to make sense of the relationships within groupings of digital files so it can, for instance, show the correct graphic with a block of text, accommodate both the thumbnails and larger views of a set of pictures, or synchronize a video with lecture slides. The problem involves finding and coordinating the right programs to display the various types of text, image, or sound files scattered throughout the Internet.

A Cornell University researcher has found a way to identify key characteristics of digital content in order to match content with programs that can display it in a browser.

The scheme involves a modification of existing software for storing and displaying digital content files that separates the two processes, according to the researcher, Naomi Dushay.

The scheme is especially useful for displaying content from the Internet because neither the content nor the program that activates it needs to be present on the system that displays the content.

The context broker software at the heart of the scheme is a set of Java programs that generates Web pages. The context broker acts as a go-between for the repositories that store digital content, the programs that act on the content, and the browsers that display the results.

By separating the storage and maintenance of digital content from its presentation, the scheme could foster more specialization throughout the digital community, said Dushay. "Digital content providers might choose to specialize in content only, or also put up context brokers and go after both presentation as well as content," she said.

The scheme also opens up the possibility of more individualized presentation of data, or for augmenting the presentation of data created by others, said Dushay. "For example, the Cornell University Library might have some whizzy presentation, rendering mechanisms targeted for members of the Cornell University community. These might be made available via... context brokers."

The method could also allow "searching, categorizing sites such as Google or Yahoo [to] provide a context broker so users want to access resources via their sites," she said.

The context broker gains information about the content from the metadata contained within content files. Metadata is data about data, and can include, for example, descriptions of the contents of a file or groups of files, or administrative details related to the data.

The scheme uses this structural metadata to identify the appropriate program for presenting the data. The programs are listed in a behavior registry, which also includes information about how a playback program can be accessed, which data structures the program can handle, and what effects each program can produce.

The context broker ties digital content to the behaviors these playback programs can produce. By changing the behaviors listed in the behavior registry, content behaviors can be changed without modifying the content itself.

Dushay tested the scheme using the Cornell Digital Library research group's Fedora, a repository that stores agglomerations of different types of data drawn from multiple locations.

To use the scheme, a user looks through a list of playback effects available for a given piece of content in the repository and requests that a certain program present the content in question.

The software matches the content's structural metadata and access points for assigning behaviors to the content with the appropriate playback program, then loads the program and uses it to access the content and display it in a Web browser.

A lot of content is now created with the kind of explicit structural metadata the scheme calls for, said Dushay. In addition, objects lacking it could be assigned metadata by

inferring the information or by "using some sort of fuzzy pattern matching on structural access points required by behavior mechanisms," she said.

Although the context broker model does not require control over the content or playback programs, end-users will need direct or indirect authorization to access the content. Metadata access could be made separate from access to the data itself, said Dushay. "It's possible to expose structural metadata without exposing the content itself. It's possible to determine the potential for [playback] behaviors with only the structural metadata, though eventually that content will be required to actually perform those behaviors," she said.

Dushay is also planning on making the scheme work with other content repositories. The scheme will eventually use more sophisticated pattern matching as a means of sorting through structural metadata, and there may be ways to add more detailed descriptive information to that metadata, she said.

She also has plans to tailor the context broker's playback for individual users, to allow differences in spoken language or language proficiency to condition how each user receives the data.

To bring the scheme beyond the proof-of-concept stage the amount of computing and network resources needed to pull the various pieces together could become an issue, Dushay noted.

The network resources needed to provide access to structural metadata "could get costly, but perhaps this could be alleviated with caching or mirroring of frequently used data and mechanisms at context broker sites," she said.

The format of metadata and the behavior mechanism input requirements will also impact performance, she said. "If it's possible to index the input requirements [and] structural metadata for fast look up, great. But if they're "fuzzy" matches, then this may become a performance issue."

Dushay's work was funded by the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Databases and Information Retrieval; Internet
Story Type: News
Related Elements: Technical paper, "Using Structural Metadata to Localize Experience of Digital Content", posted on the Physics Archive at arXiv.org/ftp/cs/papers/0112/0112017.pdf

TRN

Software Orchestrates Web Presentations

By Ted Smalley Bowen, Technology Research News
April 3/10, 2002

Like television before it, the Internet was supposed to bring learning to anyone near the right screen at the right time. While the Web holds more promise as an educational vehicle, timing issues still make it difficult to coordinate the kind of multimedia presentations that can convey lessons in living color.

Although Web technologies are evolving to better handle streaming files that have strict timing requirements, pulling together different types of files drawn from multiple sources into a single, coherent presentation still takes a lot of work.

With the classroom in mind, a research team from the University of Applied Sciences in Germany has developed a scheme that allows teachers to organize digital text, audio and video into databases, then draw from their own and other teachers' databases to compose multimedia lessons.

The scheme allows teachers to pull together digital teaching material without having to rely on programmers or having to become programmers themselves, according to Thomas Schmidt, director of the computer center at the University of Applied Sciences.

The researchers' Media Object Model software is a framework for composing multimedia lessons for a classroom or the Web. The framework uses metadata within the files, like information about formatting, authorship, and access privileges, to smooth the sharing process.

The framework includes an object model, database, lesson planning toolkit and interface that a teacher can use to assemble master documents, or presentations, according to Schmidt.

Each database includes a reference list of its constituent parts and a set of active references, or actions that can be performed on other teachers' databases specified in the reference list. The framework also accommodates annotations.

To pull together a presentation from disparate databases, teachers can specify actions based on the type of metadata contained in each database. They can also draw information from different databases based on the active references contained within the data; the software keeps everything synchronized and spatially coordinated, according to Schmidt.

There are already document models for assembling and presenting teaching materials via the Web, and there are also several standards initiatives aimed at adding time-sensitivity to the Web to allow animation, video and other multimedia files to be streamed efficiently to users' browsers.

The researchers' framework, however, allows teachers to combine and reuse these types of media, and coordinates complicated interactions, said Schmidt. "The database allows

for a context-sensitive file system view. An author will experience objects in the specific [presentation] context, even though the same object may appear in a completely different context, as well. This eases the authoring process of complex structured presentation enormously," he said.

The Media Object Model includes the Media Information Repository database that stores information and keeps track of how the data is organized using a modified form of structured query language (SQL).

The model's Web authoring tool uses a screenplay motif and shows a graphical view of each presentation's spatial arrangement and playback sequence. The tool includes a set of methods and a programming interface that allows extensions to be added so it can access other applications, according to Schmidt.

Although the database, object model and toolkit are specific to the researchers' scheme, it also uses standard Web technologies, such as Extensible Markup Language (XML) and streaming media protocols, according to Schmidt. XML is a common coding scheme used to create Web pages. Streaming media refers to time-sensitive materials like video and audio.

A key component of the scheme is reordering media files from their semantic data storage grouping into the playback sequence on the viewer's end, a task handled by a flow generator, said Schmidt.

"Our data structures on the storage layer are organized in a semantic tree. Time, however, in our lives, is linear, so there has to be a resolver, which requests the right data in time and reorganizes temporal instructions in a linear fashion," he said.

To view sequential presentations, users must have Java virtual machine software installed on their computers.

While the researchers tested their model using custom-written data objects, adapters could be written to allow the software to handle existing files, Schmidt said. "The information scheme we use is encoded in XML. So there is no principal difficulty in providing in/out filters for [other] content," he said.

The researchers are working on adding graphical tools that will allow users to edit the XML code, arrange presentation views and timing, and edit the interactions between objects, Schmidt said.

Parts of the scheme are ready for classroom use, while others are still prototypes. The model will be completed in 9 to 12 months, he said.

Schmidt's research colleagues were Bjoern Feustel, Andreas Karpati, Torsten Rack. It was funded by the University of Applied Sciences.

Timeline: < 1 year
Funding: Government
TRN Categories: Internet
Story Type: News

Related Elements:Technical paper, "An Environment for Processing Compound Media Streams," initially presented at the 7th International Conference of European University Information Systems at Humboldt University in Berlin, March 28-30, 2001

TRN

Search Tool Builds Encyclopedia

By Chhavi Sachdev, Technology Research News
August 1/8, 2001

The best part about the Internet is having so much information at your fingertips. You type in a word or phrase, hit "search" and wait for your hits. Then you hope for the best as you click on a description to see if the site contains what you need.

A pair of researchers at the University of Library and Information Sciences in Tsukuba, Japan has come up with a system that winnows down the process of an Internet search by indexing the Web as a sort of open encyclopedia. Instead of seeing a list of a thousand Web sites that might possibly contain answers, the system extracts the information and its reference links and organizes it in the form of an encyclopedic entry.

"The interface has two fundamental modes: keyword and concept input," said Atsushi Fujii, a postdoctoral research assistant at the University. If you type a word such as 'pipeline,' which could be either a means of conveying liquids and gases or a computer processing method, the application distinguishes between the two usage domains, and then shows the various entries describing each usage, Fujii said. The resulting page looks much like it came out of a paper dictionary or an encyclopedia, except each description has a hyperlink to its source page.

In the concept input mode, users can type in sentences rather than keywords, such as, 'What infects computer files by way of e-mails?' Fujii said. To answer the question, the system generates a list of candidate keywords, such as 'microvirus' and 'computer virus,' he said. Users select one of the keywords to see its description page, essentially switching back to the keyword input mode.

The system culls entries from Web pages and stores them in a database. Because the system uses the Google search engine to generate sites, the raw material the system works with is what anyone would get from searching on a term like microvirus.

The system deletes layout information and links and retains only the sentence fragments surrounding a key term. It uses a statistical language model and a morphological analyzer to prevent the output from resembling garbled strings of words. The morphological analyzer segments the input sentences into words; the statistical language model is "a set of probabilities that each word appears in a given context," Fujii said.

Using two preceding words as contexts, the statistical language model extracts three-word patterns, or tri-grams, such as "go to school" that are inherent in term descriptions. "Given a fragment extracted from a Web page, our method extracts all the possible tri-grams from the fragment, and computes a combined probability for them," Fujii explained. The result is very readable, and quite accurate, he said.

To test accuracy, the researchers generated an encyclopedia from 96 test terms collected from the Japanese IT Engineers Examinations. The method generated appropriate descriptions for 90 percent of the test terms. The answers from the generated encyclopedia were comparable to an existing hand-compiled computer encyclopedia, said Fujii.

The system is better than encyclopedias and dictionaries that are unable to keep up with new developments and information, said Fujii. "Our method facilitates searching the Web for encyclopedic knowledge related to input terms. Consequently, users can easily obtain knowledge associated with new or technical terms unlisted in existing encyclopedias," he said.

Once an encyclopedia has been generated for a search term, it is stored in a database. The database is updated periodically, Fujii said. If the search term has already been indexed, it takes only a few seconds to find an entry. Terms that are not indexed in the encyclopedia are processed in real-time, which can take up to a couple of minutes, he said.

"On the whole it is promising, but the current system is too premature to be practically interesting just yet," said John Prager, a research staff member at IBM's T.J. Watson Research Center. If a user wanted to research a technical subject, "this could be an interesting front-end to a traditional search engine such as Google, but as a Question-Answering system it is well below the state of the art," he said.

Its drawbacks are that it only deals with "what-is" questions of a multiple choice nature, for which the correct answers are already supplied. Its performance on these questions is also no better than existing systems, Prager said.

The researchers are planning to use a parallel PC cluster to speed up the process since each description can be processed independently, Fujii said. They also plan to expand the system to answer "how" and "why" questions along with "what" questions, he said.

The system is currently used for Japanese text only, but it could be used for several other languages, according to Fujii. It will be ready for practical application in two years, he said.

Fujii's research colleague was Tetsuya Ishikawa. They presented their research at the 39th Annual Meeting of the Association for Computational Linguistics (ACL2001), held in Toulouse, France from July 6-11, 2001. The research was funded by the University of Library and Information Science, Tsukuba, Japan.

Timeline: > 2 years

Funding: University

TRN Categories: Natural Language Processing; Databases and Information Retrieval; Internet

Story Type: News

Related Elements: Technical paper, "Organizing Encyclopedic Knowledge based on the Web and its Application to Question Answering," scheduled to be presented at the 39th Annual Meeting of the Association for Computational Linguistics (ACL2001), July 6-11 2001, Toulouse, France

TRN

Search

Search Tool Aids Browsing

By Kimberly Patch, Technology Research News
March 10/17, 2004

Many research teams are working on the problem of how to make finding information on the Web faster and easier. Researchers from Carnegie Mellon University have devised a scheme that gives existing search engines some extra help.

The software, dubbed ScentTrails, shows a user how strongly the links generated by a Web search correlate with the topics she is searching for. The software grades the links a search engine returns by increasing the font size of links that have more connections to relevant pages.

Like conventional search engines, the software uses content cues on a page to determine how useful that page is in relation to a user's query. But the scheme takes searching a step further by showing the user how many other relevant pages a given link is connected to. "A very strongly highlighted hyperlink indicates that many nearby pages match the query closely," said Christopher Olston, an assistant professor of computer science at Carnegie Mellon University.

The software guides the user toward information that matches his search criteria in a way that allows for continuous browsing, said Olston. For example, if a user is looking for a photocopier that copies at least 75 pages per minute, pages that contain links for fast printers will appear larger. This serves to guide the user to select links that are likely to take him closer to his goal rather than links that go cold; this cuts down on the number of times he must interrupt the browsing process to go back and check another link on a search results page.

The idea is to provide "search-driven guidance as people browse the Web, which they are free to follow or ignore as they see fit," said Olston. "The goal is to provide a happy medium between unassisted browsing, which can be tedious, and standard keyword search, in which it is difficult to remain oriented."

The researchers' prototype works within a single site, but could eventually be used in Web-wide searches, and could be used in tandem with existing search engines, said Olston.

The researchers took their prototype through a limited test run using 12 subjects. The tests showed that users were able to consider browsing cues and search cues simultaneously, and the statistics the researchers collected from the study showed that the software allowed the users to locate information more quickly than by either searching or browsing alone, said Olston.

The subjects were given eight tasks. The first three were relatively straightforward: finding a photocopier with recyclable toner, photo support or glossy print capability.

The other five involved finding combinations of functions: a digital, black-and-white copier capable of rotating copies, a copier with remote diagnostic technology that prints at least 80 copies per minute, a 400-dots-per-inch copier with a counterfeit deterrence system, a black-and-white copier that scans, faxes, prints and collates, and a 5-to-20-copies-per-minute machine that has photo support.

The users were trained on the software for five minutes, then asked to carry out each task once. Ten of the subjects said they preferred the ScentTrails interface to both browsing and searching for these types of tasks; the two other subjects said they could not tell which was easier.

The main challenge in building the software was making it fast enough. "A naive implementation would take several minutes to determine highlighting gradations for a single Web page, even for a relatively small Web site," said Olston. To make it usable, however, the researchers needed to make the process happen in less than a second, he said.

To do so, the researchers made sure as much information as possible was computed in advance. The software computes the connectivity scores for all pairs of pages on the site ahead of time, said Olston. Match scores, which represent the degree of relevance of pages to queries, are largely computed ahead of time."When a user issues a search query or views a new Web page, highlighting can be performed very quickly by looking up the appropriate set of connectivity and match scores and combining them." Olston said.

The researchers are currently working on three major issues that need to be solved before the software can be used practically across multiple sites, said Olston.

"First, we need to find good methods for highlighting hyperlinks that can be applied across a diverse variety of Web pages, and do not unduly impact content readability," he said. "If link highlighting is too annoying or obtrusive, people will turn it off."

The researchers are also working on improving the software algorithms to make it possible to apply them to very large Web collections, said Olston. "Ultimately I hope ScentTrails or some variant can be used on the Web as a whole, not just on a site-by-site basis."

Third, the researchers are working on finding good starting points for users' queries. "In other words, if someone [has] a partially-formed search query, where should the computer suggest that they begin browsing to find the answer," he said. This would make it easier to apply the method to the entire Web, said Olston.

The method could be used to search within a Web site in one to two years, and for Web-wide searching in two to five years, according to Olston.

Olston's research colleague was Ed H. Chi. The work appeared in the September, 2003 issue of the Association of Computing Machinery (ACM) *Transactions on Human-Computer Interaction*. The research was funded by Palo Alto Research Center (PARC).

Timeline: 1-2 years, 2-5 years
Funding: Corporate
TRN Categories: Internet; Databases and Information Retrieval
Story Type: News
Related Elements: Technical paper, "ScentTrails: Integrated Browsing and Searching on the Web," Association of Computing Machinery (ACM) *Transactions on Human-Computer Interaction*, September, 2003

TRN

Queries Guide Web Crawlers

By Kimberly Patch, Technology Research News
October 22/29, 2003

Only a small percentage of the Internet's vast collection of information is indexed by search engines, which makes it important to improve the way search engines find what they do index.

Researchers from Contraco Consulting and Software Ltd., T-Online International and Siegen University in Germany have written an algorithm that improves Internet search results by factoring in what people are looking for. The researchers took their cue from the audience analysis that drives format and programming changes in television.

The algorithm, dubbed Vox Populi, picks up trends by analyzing patterns in people's Web searching behavior, then directs search engine crawlers to more thoroughly index relevant sites, according to Andreas Schaale, a partner at Contraco Consulting and Software. For instance, "if we see that the amount of queries about soccer is growing before entering the World Cup, this algorithm would give more resources for... soccer sites," he said.

The algorithm analyzes the queries people use to ask for information to find those that represent what the average user is searching for, sends these to the Web crawler component of an existing search system with instructions to give the relevant domains more Web crawler resources. The algorithm determines how much more attention each domain should gain. Web crawlers travel around the Web making the raw indexes of Web pages that search engines use.

Internet searching has gone through several changes in the past decade. The first search engines, like AltaVista, ranked purely on relevancy. Today the major search engines use static rank algorithms, which also consider domain popularity. Google introduced this method in 1997.

Web crawlers have evolved as well. Focused crawlers index pages related to specific topics, and adaptive crawlers reorder their lists of uncrawled pages based on the relevancy of the pages they have crawled.

Vox Populi also takes into account the subjects the average user is searching for. The algorithm "answers the question 'What are most of the people searching for?'" Said Schaale. Vox Populi does not replace the existing ranking algorithms, which retrieve their results from an index, he said.

The need for directing crawlers based on feedback from queries is driven by economics; data storage and handling is a growing cost, said Shaale. "A shop owner orders his products [depending on] what his customers ask for," said Schaale. "Vox Populi does basically the same," he said. This type of ranking is only necessary because search engines are not nearly powerful enough to crawl all Internet content in real-time, he said. The Google crawler, for instance, does its main crawl to update its index of the Web about once a month.

The researchers' scheme also includes methods to suppress spam, or unwanted content. Spam suppression is especially important in this method because in "most wanted" topic areas like free downloads, adult content, and shopping, the amount of spam is clearly above-average," said Schaale.

The main challenge to making the method work is not related to the algorithm, but the filtering, Schaale added. "The spammers and the search engine optimizers... adapt fast to new methods of filtering. This is a challenge for each search engine," he said.

The basic idea of improving searching by incorporating user context, including queries, has a lot of potential and is an active research area, said Filippo Menczer, an associate professor of informatics and computer science at Indiana University. The researchers' idea of improving a search engine by modifying its crawling and ranking algorithms to capture the preferences inferred from user queries is interesting, but its mathematical framework is incomplete, he said.

The researchers' algorithm can be used in combination with the ranking methods used by search engines, according to Schaale. It could be used in vertical information systems that search by subject and personalized searches that take into account a user's topics of interest, he said.

The method could be ready within a year, said Schaale.

Schaale's research colleagues were Carsten Wulf-Mathies from T-Online International AG in Germany and Sönke Lieberam-Schmidt from Siegen University in Germany. The research was funded by Contraco Consulting and Software.

Timeline: > 1 year
Funding: Corporate

TRN Categories: Internet; Databases and Information Retrieval
Story Type: News
Related Elements: Technical paper, "A New Approach to Relevancy in Internet Searching - the "Vox Populi Algorithm", posted in the Computing Research Repository (CoRR) at arxiv.org/abs/cs.DS/0308039

TRN

Web Searches Tap Databases

By Kimberly Patch, Technology Research News
September 24/October 1, 2003

Although the computer has made it possible to quickly search through documents and databases, sifting through a series of sources — like a local database, a bunch of text documents, and the Web — still means using different programs and different searches.

Researchers from Birkbeck University of London in England have written software designed to allow users to search for something without having to know where it might reside.

The search method makes it possible to search different types of sources at the same time, said Richard Wheeldon, a researcher at the Birkbeck University of London. "Think of how difficult it is to search a company's intranet, file system and databases at the same time," he said. "With a few alterations to our technology, it could be made incredibly simple."

The key to the method is software, dubbed DbSurfer, that permits free-text searches on the contents of relational databases. Data stored in relational databases is ordinarily accessed using queries structured to match the organization of the database.

DbSsurfer enables free text relational database searches through a modified version of the trails method used to organize the links contained in hypertext, said Wheeldon.

A trail is a sequence of connected pages. As far back as 1945, computer pioneer Vannevar Bush wrote about the concept of a web of trails. Hypertext and the World Wide Web take advantage of this concept, but databases do not, according to Wheeldon. "Trails have often been used in hypertext systems, but never in relational database systems," he said.

Relational databases organize information using tables subdivided by fields like columns and rows. Individual pieces of information, or records, reside in cells delineated by columns and rows. Relationships between records are determined by fields that the records have in common.

The researchers' software automatically constructs trails across tables in relational databases, according to Wheeldon. The software treats each database row as a virtual Web page, and builds links according to database settings, he said.

When presented with a free text database query, DbSurfer's navigation engine calculates scores for each database row, and the best scores are used to construct trails. The scheme uses a probabilistic best-first algorithm to select the most relevant trails. A probabilistic best-first algorithm assigns more promising alternatives higher probabilities. The researchers' Best Trail algorithm does this in two ways — proportionally according to the score assigned to the trail, and decreasing exponentially according to rank.The program presents the results to the user as a navigation search interface.

The researchers have also used the same basic system to search the Web, a group of Java documents, program code, and Usenet newsgroups, said Wheeldon.

"Theoretically, it could also be used in virtual environments or as a search application at the operating system level," he said.

The method uses standard keyword searches of data sources, and is easily customized, said Wheeldon. Data is represented in the Internet's extensible markup language (XML), and this means "the look of the pages can be changed in many different ways," Wilson said.

The technical challenge to building the software was being able to construct trails efficiently, said Wheeldon. "Trail construction is now typically performed in a few hundredths of a second," he said.

The current prototype won't scale to very large databases, but this is not a fundamental limitation, said Wheeldon. "Anything more than a few tens of millions of rows and the system will choke [but] this is easily fixed in theory," he said.

The software does have a downside — it is not secure enough for highly sensitive data, Wheeldon said.

The next steps in developing the system are linking the DbSurfer indexer to a Web robot, optimizing the indexer for common databases, and adding software that will enable the entire index and trail structure to be accessed from within the database interface, according to Wheeldon.

The software could be ready for deployment in less than a year, said Wheeldon.

Wheeldon's research colleagues were Mark Levine and Kevin Keenoy. The research was funded by the UK Engineering and Physical Sciences Research Council (EPSRC).

Timeline: > 1 year
Funding: Government
TRN Categories: Databases and Information Retrieval; Internet
Story Type: News
Related Elements: Technical paper, "Search and Navigation in Relational Databases," posted in the Computing Research Repository (CoRR) database at arxiv.org/abs/cs.DB/0307073

TRN

Plan Puts Search in Net Structure

By Kimberly Patch
April 7/14,2004

There is a wealth of information available on the Internet and in separate repositories like university library databases. It would save time to be able to search all digital resources using one interface.

Researchers from Huazhong University of Science and Technology in China have come up with a distributed information retrieval system that promises to help. The system could eventually become part of the Internet infrastructure as an extension of the domain name service.

The Domain Resource Integrated System (DRIS) allows a user to search for information whether it resides on Web pages or in databases, said Wang Liang, a researcher at the University. "Hundreds of databases have been introduced in many libraries of universities, and there are many more free information resources on the Internet," he said. The researchers' prototype provides a unified search for the Web, FTP resources, and databases at the researchers' university.

The software organizes resources on three levels — an individual domain like a university or company, a sub network like the China Education Network (CERNET) that includes all the universities in China, and a top-level domain like the Internet in China, said Liang.

At the individual domain level, the domain organization such as a university or corporation would implement a standard search engine that crawls pages, creates an index and provides a search interface.

At the subnetwork level, the search function would include an index and search interface but no Web crawler. Instead, the data for the domains within the subnetwork would be supplied by the domain-level search engine databases. And at the top-level domain, which is often a country domain, the search function would be a distributed metasearch, or search interface to the indexes at the subnetwork level.

The scheme provides the basic ability to search across resources. This search infrastructure could be used by other software programs to provide more elaborate search abilities, according to Liang.

The researchers are building a Domain Resource Integrated System that integrates all the information resources in Chinese universities, according to Liang. "Our first testbed will be built on [the] China Education Network," he said. "If our first testbed proves that DRIS is a successful system, we can extend it to the whole Internet in China with the help of government."

The testbed will include a standard distributed, platform-independent search interface, a collection descriptions standard, an information retrieval protocol, a

standard metadata harvest system, and a standard public pages search system, according to Liang. The system will also take into account the ability of the next-generation Internet standards including Internet Protocol Version 6 (IPV6), especially the ability to assign priorities to different types of data flows, he said.

The ultimate goal is to make the Domain Resource Integrated System software an information retrieval system built into the Internet, said Liang. "The basic idea is that search should be [an] internal function of [the] Internet and everyone should have his own personal intelligent search engine," said Liang.

Connecting personalized front-end search engines to an integrated Internet search function would be convenient in several ways, said Liang.

The integrated search function would cut down on irrelevant and outdated search results, would cover more of the Internet as well as university and other databases, and would cut down on traffic from many search engines periodically crawling the Web to build page indexes.

A personalized search engine would allow a user to better tailor queries. In contrast, every user obtains the same results for a query from general search engines that can't express special interests, he said.

The information retrieval system in the film *Time Machine* was impressive, and is not far from reality, said Liang. Many researchers are working toward it, he said.

The researchers are planning to complete the experimental system built on the China Education Network by summer, 2004. Liang's research colleagues were Guo Yi-Ping and Fang Ming. The research was funded by the China Academic Library and Information System (CALIS).

Timeline: < 2 years
Funding: Government
TRN Categories: Internet; Databases and Information Retrieval
Story Type: News
Related Elements: Technical paper, "Make Search Become the Internal Function of Internet," posted on the Computing Research Repository (CoRR) at arxiv.org/abs/cs.IR/0311015

TRN

Software Sifts Text to Sort Web Sites

By Ted Smalley Bowen, Technology Research News
February 21, 2001

Although the World Wide Web is a multimedia network, the job of classifying sites is still largely a matter of interpreting textual information. In an attempt to make that process quicker and more accurate, a research team has developed a method of automatically sifting and categorizing the various forms of text found on the Web.

The researchers devised a spider program that crawls, or examines the various types of text that make up a Web site, and classification software that organizes the information the spider finds, creating a sort of card catalog system for parts of the Web.

Although sophisticated methods of searching pictures, video and audio are under development, text-based categorization promises a more immediate improvement in traversing and making sense of the Web, said John Pierre, a member of the technical staff at Interwoven, Inc. Pierre and his research colleague, Bill Wohler, developed the categorization system while employed by Metacode Technolgies, Inc.

They have used the software to group English language sites into business categories. The scheme could apply to other languages and categories as well, Pierre said.

The software ferrets out meaningful text from three distinct sources within a Web site to categorize the site: the words that make up the site's Hypertext markup language (HTML) meta tags, the words within its HTML body tags, and the readable text on the site.

HTML meta tags are key words contained in the hidden code of a Web site that summarize the type of information a Web site contains. HTML body tags, also hidden, are page layout instructions that affect the look of the site.

The term metadata, which generally refers to information about information, has many different meanings in computer science. Pierre's scheme uses subject-based metadata, or data about subject categories to group Web sites in much the way a library card catalog groups books. "You can think of a library catalog card, where you have author and title, and number of pages in the book, but you would also have other fields like subject or Dewey decimal code," said Pierre.

Other organizational systems could be tapped to make use of other types of metadata fields as well, he said. "This type of system, in the larger picture of metadata creation, can really serve as a driver for processing [content] beyond keyword searching to more reason-based searching," he said.

The drawback to using descriptive metadata, however, is there is not enough of it. "People are a little resistant or lazy about deploying metadata. It's a tedious task that nobody really wants to do," he said.

Web developers must enter meta tags into their sites in order for Web search engines to use them in ranking search results. Despite this incentive, however, less than a third of Web sites use meta-tag key words and descriptions.

The program also looks for text in HTML body tags, which usually contain page layout commands but often also include useful information like Web addresses.

Based on an examination of 19,195 Web domains, Pierre found that while most had words in title tags, which allow Web browsers to title each page, the information was of

limited use in classifying sites because there were few words and they often consisted of generic terms like "homepage."

The scheme also addresses pages without words — some pages have only frame sets, images or software plug-ins, and do not lend themselves to accurate classification, Pierre said. Frame sets are organizational elements that divide the browser's window into multiple frames. Software plug-ins are programs that give a larger program additional functions, like the ability to play movies.

The spider program searches first for text in HTML meta tags and titles, then follows links for frame sets and hyperlinks. The program searches body text only if no meta tag information is found, according to Pierre.

Once the spider gathers the information, it passes it to a Latent Sematic Indexing (LSI) information retrieval engine, which identifies matches based on concepts rather than single words, but does not have the heavy computational requirements as a true natural language processor, he said.

"It provides a certain level of concept-based matching, without any specialized knowledge base or rules, and it does that along with a complete framework for matching terms in documents — similarity matching," he said.

Next, the information is fed into a classification engine, which, for the sake of performance, uses shallow parsing, according to Pierre.

"It's a way of understanding and extracting some limited subset of the data without worrying about the complete structure of it. For example, you could assign and extract proper names in sentences without having to understand every word, diagram all parts of speech, and understand the full meaning of the sentence," he said.

The scheme works most accurately using meta tags as the only source of text, while classifications based partly or entirely on body text are less accurate, according to Pierre.

The categorization system pulls together various ways of retrieving information on the Web, said Jon Kleinberg, and assistant professor of computer science at Cornell University. "In a sense, it's collecting a sequence of techniques which have been widely used in the information retrieval and machine learning community [and] grouping them into a single architecture for [Web] classification tasks.

The big question for the system is whether metadata will ultimately be more widely developed, said Kleinberg. "It remains to be seen to what level metadata will be adopted, and whether there's a standard that will somehow achieve widespread use" since metadata is not visible to end-users and can be labor-intensive to create," he said.

The scheme is also competing with other general classification schemes that build automatic taxonomies of Web pages and existing, more focused classification systems, Kleinberg added.

In general, research schemes like this need to be more accurate. "Ultimately, we need to develop a better method to combine a natural language processor and statistical [analysis software]," said Pierre.

Some elements of the classification system are already in commercial use. The spider program could be in wide use in three to five years, according to Pierre. The research was funded by Network Solutions.

Timeline: Now, 3-5 years
Funding: Corporate
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper, "On the Automated Classification of Web Sites," posted at xxx.lanl.gov/abs/cs.IR/0102002

TRN

Software Sorts Web Data

By Kimberly Patch, Technology Research News
September 20, 2000

A research consortium is putting the finishing touches on a set of software programs designed to do for data comparison what the Web has already done for sharing documents.

Today, comparing sunspot activity with heart attack data, or cross-referencing tire safety data with ZIP codes is possible, but it's not a quick, point-and-click type of task. It involves finding relevant sets of data, scraping them into a program, making conversions to comparable units, and futzing with the data to make it readable.

For the past four years, project DataSpace, a university, government and corporate consortium, has been building an infrastructure to change all that. "The idea is there's a lot of data out there but it's not always easy to see how it's related to other data," said Robert Grossman, director of the Laboratory for Advanced Computing at the University of Illinois, CEO of Magnify Inc. and DataSpace project leader.

The DataSpace project involves several pieces of software. The four major parts are DataSpace Transfer Protocol (DSTP), a protocol for moving columns of data over the Web; Predictive Model Markup Language (PMML), a set of tags for marking columns of data; and open source DSTP client and server software, which allow computers to exchange such data. DSTP is the data equivalent of Hypertext Transfer Protocol (HTTP) and PMML is the data equivalent of and Hypertext Markup Language (HTML).

Data needs its own markup and transfer protocols because two columns of data must share common units in order to be compared meaningfully, said Grossman. The DataSpace format is like email versus a fax, he said. "If you send a fax, you see the same image but you can't manipulate it. When data is in HTML it cannot be immediately manipulated even though the information is visually apparent. You need some

protocol that has some format that is understood by your application," Grossman said.

The DataSpace protocol addresses the problem by adding universal keys to columns of data. A user can compare data that has keys in common regardless of its location and format. For example, with ZIP code and date keys in common, "you can compare diabetes related deaths per ZIP code with average income and average education per ZIP code," said Grossman. The DataSpace project includes many standard keys and also allows users to find more.

One of the four pieces has been completed each year of the project. With the advent of the transfer protocol, DataSpace is ready to be driven, said Grossman, who compared the project to a car that now has four working wheels.

"It's ready to be used. The [PMML] language is being supported by [Microsoft, IBM, Oracle and NCR]. The protocol is well-defined. We have shown through Java applications that we're giving out on the Web that it's useful and easy to use and we're encouraging people to... adapt them to needs of their own," Grossman said.

The scientists are continuing to stress test the system. One big road test is slated for November, when 10 applications involving about 50 scientists and engineers will debut at the Supercomputing tradeshow in Dallas. "They'll be [showing] everything from looking for patterns in genomic data to looking for patterns in business data to looking for patterns in engineering data to looking at the Firestone tire data [related to] fatalities," said Grossman.

The project has involved about 50 scientists and engineers a year from various universities, companies and government labs, said Grossman. The most recent work included a testbed at the University of Illinois with funding from the National Science Foundation and input from a dozen other entities, said Grossman. Software, demonstrations and a full list of participants are available at www.dataspace.net.

Timeline: Now
Funding: Government, University, Corporate
TRN Categories: Internet
Story Type: News
Related Elements: Project DataSpace web site: www.dataspaceweb.net

TRN

Visualization and simulation

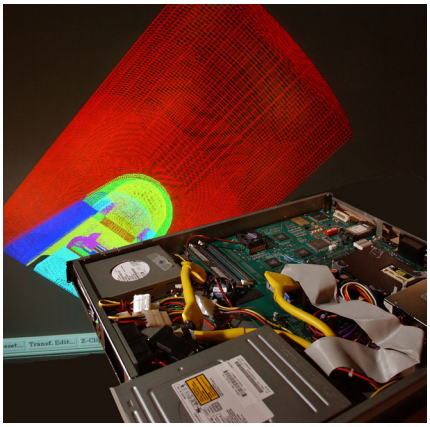
Remote Monitoring Aids Data Access

By Kimberly Patch, Technology Research News
January 15/22, 2003

One of the ongoing challenges facing scientists and business people is how to access and visualize the vast amounts of data modern technology allows us to collect.

Researchers from Sandia National Laboratories have found a way to work with large amounts of data over networks in near real-time. The researchers' prototype uses the Internet to give people access to very large sets of data stored thousands of miles away and allows them to manipulate the data with a lag time of less than one tenth of a second.

Remote access schemes tend to focus on moving data, said John Eldridge, a principal member of technical staff at Sandia National Laboratories. "Either they send the data set in its entirety or they transmit the image geometry so that the remote computer can render and display the image through its own video adapter," he said.



Source: Sandia National Labs

This device sends video signals rather than data files over the Internet; this makes it possible to manipulate graphical representations of large, remote data sets nearly in real-time.

The Sandia method doesn't transfer data at all, but instead transfers the video signal that normally carries image information from a computer to its monitor. "The video card is designed to put out a video signal to a local monitor... we extend the signal," said Eldridge.

The approach can speed remote access, said Eldridge. "Where large data sets... are involved, it may be more efficient to move the video signal rather than the data," he said.

The interactive remote-visualization hardware could allow doctors to view and manipulate very large images, like magnetic resonance imaging (MRI) files, remotely, according to Eldridge. It could also allow people who work in other fields that involve very large amounts of data — like geophysical modeling and financial services — to view and manipulate data remotely, he said.

Lag time makes sharing and manipulating remote data difficult, said Eldridge. "The simplest example of this is in trying to coordinate the movement of the computer's mouse with a mouse pointer on the display," he said. "As the delay between the action and the apparent response increases, the interactivity and usability decreases; as the processing delay approaches 0.1 seconds, [it] noticeably affects a user's interactivity," he said.

To decrease lag time, the group took advantage of today's graphics cards for video games, which render two-dimensional and three-dimensional images very quickly. These images are typically fed to nearby monitors. The researchers found a way to instead move the video signal across the Internet.

The researcher's encoder/decoder hardware attaches to a computer's video card adapter. It digitizes the video signal, compresses the digitized data stream, and then formats the data stream into standard network protocol packets of data, said Eldridge. The card then sends the packets to a Gigabit Ethernet interface card, which transmits the packets across a network.

At the remote location, the researchers' hardware receives the packets, rebuilds the data into a video stream, and translates the video signal for a locally-attached video monitor, said Eldridge.

The main challenge to speeding things up was to perform the encoding and decoding process in near real-time, Eldridge said.

To do this the researchers changed the way the data was stored in memory, and the way the hardware performs frame differencing, said Eldridge. Because each frame of a video looks a lot like the frame before, it saves a lot of time to only transmit image changes. The system performs frame differencing by looking at differences between successive video frames; it then transmits only those differences across the network.

The time required to capture frames for the differencing process is responsible for most of the response-time delay, said Eldridge. For video screens that refresh at 60 hertz, or times per second, the encoder/decoder hardware completes the frame differencing step in about 32 milliseconds, he said.

The researchers used reprogrammable logic chips to process and compress the video image. "Since the processing is performed in logic hardware it is very quick," said Eldridge. "The hardware extracts a great deal of performance from each clock cycle," he added.

The signal can be compressed further by removing redundant timing information and even reducing the frame update rate, if necessary, said Eldridge.

The raw information rate from a computer's video display is about 2.5 gigabits, or billion bits, per second for a screen resolution of 1280 by 1024 and a 60 hertz refresh rate. The researcher's prototype achieved network transfer rates of between 70 and 800 megabits per second, said Eldridge.

The idea of transferring just the video signal has the nice property that the required bandwidth, though high, is limited and doesn't go up with the complexity of what is being visualized, said Peter Schröder, a professor of computer science and applied and computational mathematics at the California Institute of Technology. "However, one would have to do a careful bandwidth analysis to see where that cutoff point is," he said.

The technology's potential usefulness hinges on its cost, Schröder added. "The commodity hardware business is very unforgiving of custom solutions with only a few potential customers," he said.

The researchers are aiming to transfer the technology to a partner who can commercialize it, said Eldridge. They are also aiming to make the scheme work with multi-tiled displays.

The prototype could be packaged into a commercial product in 6 to 12 months, said Eldridge.

Eldridge's research colleague was Lyndon Pierson. The research was funded by Sandia National Laboratories.

Timeline: 6-12 months
Funding: Government
TRN Categories: Networking; Graphics; Internet; Data Representation and Simulation
Story Type: News
Related Elements: None

TRN

Experience Handed Across Net

Technology Research News, July 16/23, 2003

Teaching the subtleties of a good golf swing or a precise surgical method electronically is just not the same as showing someone in person. But it's getting closer.

Researchers from the University of Buffalo have developed a method that enables one person to go through the exact movements of another, including feeling the same forces, over the Internet. The method could eventually be used to capture the touch of a musician, golfer or surgeon and pass it on to someone trying to match that touch, according to the researchers.

The system involves a glove that captures force and transmits it through the Internet to the receiver, who uses a combination of a sensing tool to feel the forces, and the act of following a point on a computer screen to recreate the movement of the other person's hand. The method differs from haptic techniques that allow users to feel the movement of another person's hand from the outside, or allow one user's hand to be pulled in the same direction as the other user's.

A practical system could be available within three years, according to the researchers. The researchers are scheduled to present the work at the American Society of Mechanical Engineers (ASME) International Mechanical Engineering Congress in Washington in November, 2003.

TRN

Virtual Mic Carries Concert Hall Sound over 'Net

By Kimberly Patch, Technology Research News
November 29, 2000

Today's best audio systems use multiple channels to surround listeners with sound. It sure sounds great, but has some practical drawbacks.

First, in order to get true surround sound, rather than just the same sound coming at you from several directions, the original recording process must use a separate microphone for each channel. Second, if you want to stream all those channels over the Internet, you need a whole lot of bandwidth — about a megabyte per second for each channel of uncompressed sound, or 200 kilobytes per second compressed.

Researchers from the University of Southern California (USC) have developed a filtering system that addresses both problems and also allows older recordings to be recast as multichannel sound.

The Virtual Microphone technology allows the researchers to map a concert hall once, recording sound from 10 or 20 microphones set around the hall, then adjust a recording to match what it would have sounded like recorded through the microphones.

"We've taken 1948 recordings and converted them and it's pretty amazing, the hall opens up around you," said Chris Kyriakakis, assistant professor of electrical engineering at USC.

The problem sounds easier than it is. "If you're trying to model what happens to the acoustics inside a large hall, the problem is that you can't do it for all possible frequencies and wavelengths, it's too complicated," said Kyriakakis.

Instead, the researchers used reference measurements, a signal processing algorithm, and some key intervention with the human ear to find the differences between the sound recorded by a standard front mic and the other mics scattered throughout the hall, to make a filter that will alter a recording accordingly.

Sound from a microphone in the back, for instance "has been modified by every surface in the room and every delay, and early reflections and late reflections. We don't know what they are, but they're in the signal. Our filter modifies the front signal to sound like the back signal, without having to solve all the equations," he said.

To make the filters, the researchers run each microphone through an algorithm based on adaptive filter theory. "The recording in front... is my reference, and a recording in the back of the same music is the goal I am trying to reach. And then I just let the filter iterate and say, 'keep changing this front thing until it sounds like the back thing,'" he said.

The process takes about eight hours on an 800 MHz computer. "When it's done we listen to it, and then we compare it to the real thing, and say 'it's not quite there,' and then we have to make some intelligent guesses as to why. Then we make some tweaks," to the algorithm's parameters, and run the sound through again. Each channel, or microphone takes three or four cycles and several days before it is complete. Once all 10 or so microphones are done, the filters for that hall are complete and can be used on any recording.

There are four major types of changes the algorithm is making to the sound to essentially throw it to a virtual microphone. These are changes, or cues that the human brain uses to determine where sound is coming from. They include early reflections, reverberation, high frequency attenuation and height.

One of the first clues to how the structure of the room affects sound are early reflections, which bounce off sidewalls and give us a sense of width, said Kyriakakis.

Next, we gain clues about the size of the room from reverberation, a phenomenon that happens after sounds bounce around for a few thousands of a second, he said. "Psychoacoustically it gives our brain a sense of distance. The ratio between the direct sound and the reverberant sound is what you can fool with to give the apparent distance of a source."

As microphones get further away from the source, high frequencies fade faster than low frequencies because the air absorbs them at a higher rate. This effect becomes important in a large concert hall said Kyriakakis.

The sense of height is also very important, said Kyriakakis. "It sounds a little strange because of course no instruments are out there, but we placed microphones 70 feet above the orchestra, hanging from the ceiling. When we play that back — appropriately filtered over... loudspeakers that are hanging from the ceiling pointed down the listener — if you close your eyes the sense of depth of the room increases dramatically. It gives you the sense that the stage is bigger," he said.

The filters are also useful for transmitting multichannel audio over a network, because if only one channel is transmitted, along with the filters, the other channels can be recreated on the other side. In addition, the filters for any given concert hall only need to be sent once.

Currently, it takes about as much time to create the multiple channels as the original recording was long. For instance, a 10 minute recording would require a 10 minute delay before all the channels were ready to play. The researchers are working on a more efficient, real-time version of the software that would allow for streaming, multichannel audio. "We're not there yet, but the idea would be you could have it as a plug-in so... as the music comes in it goes through these filters before going through your speakers," said Kyriakakis.

Further out, the researchers plan to tackle the problem of using the filters on a recording that wasn't originally recorded in that concert hall. "The difficulty is that it involves one

additional, difficult step of completely removing the acoustics of the existing hall from the recording," said Kyriakakis.

"It's a nice application," said Angelos Katsaggelos, professor of electrical and computer engineering at Northwestern University. "It's a challenging problem just being able from one file to generate the sound as if it were coming from different directions or recorded from different mics," he said.

"From a practical point of view... there are products out in the market like audio DVD players that support multichannel recordings but [recording] with 10 and 16 mics... is hard and not done routinely," said Katsaggelos. In addition, "it is a contribution in sync with the direction of major developments in the area of multimedia processing and immersive reality," he said.

A professional version of the software for creating multichannel recordings could be technically feasible within six months, said Kyriakakis. It will take a year to 18 months to produce a real-time version of the filters, he said.

Kyriakakis' colleague in the research is Athanasios Mouchtaris. They presented their results at the International Conference on Multimedia in New York in July, 2000. The research is funded by the National Science Foundation (NSF).

Timeline: 6 months; 1-1 ½ years
Funding: Government
TRN Categories: Signal Processing; Applied Computing
Story Type: News
Related Elements: Technical paper, "Virtual Microphones for Multichannel Audio Applications," presented at the International Conference on Multimedia, July, 2000 in New York City

TRN

Grid computing

Tool Eases Grid Monitoring

By Kimberly Patch, Technology Research News
December 31, 2003/January 7, 2004

Grid computing takes advantage of Internet connections and unused resources connected to the Net — like idle computers and vacant disk space — to put together virtual computers powerful enough to handle compute-intensive problems like processing huge amounts of scientific data.

Although the concept of coordinating otherwise unused computers distributed around a worldwide network is relatively simple, the coordination takes a lot of effort. As Grid computing becomes more commonplace, researchers are developing tools that simplify the practice.

University of Melbourne researchers have produced a toolkit that makes it easier to see how a Grid job is going. The tool allows users to create a Web interface to a Grid computing testbed without having to do any programming.

The tool, dubbed Gridscape, includes a template that allows users to plug in information like a testbed name, logo, information about the computers being used in the testbed, and a geographical map, said Rajkumar Buyya, a lecturer of computer science and software engineering at the University of Melbourne in Australia.

Most simply, Gridscape provides users with a holistic view of a testbed that shows user application jobs running on different Grid nodes, said Buyya. The tool can also be used to search for resources like computers that have certain attributes, to check the attributes of a given resource, or to check which resources are currently on-line, he said. "The status of Grid resources is displayed on a geographic map [of] the testbed [that] can be queried further for detailed information."

The tool dynamically creates Web content using Java JSP, Servlets and interactive client side JavaScript. "Once the user has finished customizing [the] portal, the changes may be viewed on-line immediately," said Buyya.

To create the tool, the researchers began with another Grid tool: the information services components of the Globus Grid toolkit. "It provides us with a standard interface for gathering Grid resource information," said Buyya. "Without those protocols a generalized tool like this really would be difficult to produce," he said.

The researchers' aim was to make a simple, widely accessible tool that would enable more rapid development of Grid testbed portals, said Buyya. Existing portal development kits provide programming interfaces to a lower-level Grid framework that requires an a significant amount of programming effort, he said. Gridscape makes it possible to rapidly develop testbed portals without programming, instead offering "a more generic solution," he said.

The Web-based interface is interactive, dynamic and widely accessible, and the client-side portion of the tool is light-weight, meaning it does not take a lot of computing resources, said Buyya. The drawback to the tool is it may not be specific enough for some testbed requirements, he said.

In general, Grid computing is allowing scientists to tackle very large problems that require large amounts of computer resources, said Buyya. "Distributed computing is... allowing us to deal with data and compute intensive problems which we previously thought were unfeasible," he said.

When this type of computing is further developed, "Grid computing power [will] become analogous to our current electrical power grids," said Buyya. People will be able to elect to consume computing power as a utility, as they do with electricity, gas and water, he said.

The first version of Gridscape is in regular use, and the researchers are getting ready to release an open source version, according to Buyya.

The researchers also have plans to integrate Gridscape with their G-monitor Web portal tool. G-monitor allows users to monitor, control and steer the execution of Grid applications, said Buyya.

They are also working on extending Gridscape to support handheld devices and mobile phones, he said.

Buyya's research colleague was Hussein Gibbons. The research was funded by the University of Melbourne, the Victorian Partnership for Advanced Computing, and Sun Microsystems.

Timeline: Now
Funding: Corporate, University
TRN Categories: Distributed Computing
Story Type: News
Related Elements: Technical paper, "Gridscape: A Tool for the Creation of Interactive and Dynamic Grid Testbed Web Portals," Research Report from the GRIDS Lab at The University of Melbourne, July 2003, posted in the Computing Research Repository (CoRR) at arxiv.org/abs/cs.DC/0307052

TRN

Toolset Teams Computers to Design Drugs

By Ted Smalley Bowen , Technology Research News
January 16, 2002

Computational grids provide the raw material for assembling temporary, virtual computers from sometimes far-flung resources connected to the Internet or private networks. They came about because researchers often require processing power, storage, and bandwidth far beyond the scope of their own systems.

This type of distributed computing, which can also include scientific instruments, makes the means to tackle complex applications available on an ad hoc basis, and allows researchers to draw on widely-dispersed stores of information.

The molecular modeling programs used to design drugs are especially data-hungry and computationally intensive applications. Designing a drug involves screening massive databases of molecules to identify pairs that can be combined, and figuring out the best way to combine them to achieve a certain affect. The molecules could be enzymes, protein receptors, DNA, or the drugs designed to act on them.

During this molecular docking process, researchers try to match the generally small molecules of prospective drugs with the larger biological molecules they are designed to affect, such as proteins or DNA. These searches can entail sifting through millions of files that contain three-dimensional representations of the molecules.

A group of researchers in Australia has put together a set of software tools to perform molecular docking over a computational grid. The tools tap into remote databases of chemical structures in order to carry out the molecular matching process.

Grid computing software finds and accesses resources from networked computers that can be physically located almost anywhere. It coordinates scheduling and security among systems that may be running different operating systems, to combine, for example, the processing capabilities of half a dozen Unix servers and a supercomputer with databases stored in a collection of disk drives connected to yet another computer.

The researchers adapted a molecular docking program to work on a grid configuration by having it run several copies of a molecular matching program on different systems or portions of systems. The software performed many computations at once on different subsets of the data, then combined the results. This type of parallel processing, also known as a parameter sweep, enabled the grid application to work through the matching process more quickly.

The complexity of each molecule record and the scale of the database searches involved in molecular docking put such applications beyond the reach of most labs' conventional computing resources, according to Rajkumar Buyya, a research scientist at Monash University in Australia. "Screening each compound, depending on structural complexity, can take hours on a standard PC, which means screening all compounds in a single database can take years."

Even on a supercomputer, "large-scale exploration is still limited by the availability of processing power," he said.Using a computational grid, however, researchers could feed extensive computing jobs to a coordinated mix of PCs, workstations, multiprocessor systems and supercomputers, in order to crunch the numbers simultaneously.

A drug design problem that requires screening 180,000 compounds at three hours each would take a single PC about 61 years to process, and would tie-up a typical 64-node supercomputer for about a year, according to Buyya. "The problem can be solved with a large scale grid of hundreds of supercomputers in a day," he said.

To run the docking application on a computational grid, the researchers developed a program to index chemical databases, and software for accessing the chemical databases.

To speed the scheme, the researchers replicated the chemical database so that more requests for database information could be processed at once. To further speed the process, the researchers wrote a database server program that allowed computers to field more than one database query at a time.

The researcher's scheme compensates for the uneven bandwidth, processing speeds, and available resources among grid-linked systems by mapping the location of files and selecting the optimal computer to query, according to Buyya. "The data broker assists in the discovery and selection of a suitable source... depending on... availability, network proximity, load, and the access price," he said.

Because the performance of database applications suffers over network connections, the researchers generated indices for each chemical database, including references to each record's size.

This allowed each computer to respond to queries by first checking the index file for the record's size and location and then accessing the record directly from the database file, rather than sequentially sifting through the database, said Buyya.

The application requirements and the tools used to meet them are specific to molecular docking, but similar software would speed compute-intensive tasks like high-energy physics calculations and risk analysis, according to Buyya.

The researchers tested the scheduling portion of their scheme on the World Wide Grid test-bed of systems in Australia, Japan and the US, and successfully estimated the time and cost required to run the applications in configurations optimized for speed and for budget, Buyya said.

Using the test bed, they screened files of 200 candidate molecules for docking with the target enzyme endothelin-converting enzyme (ECE), which is associated with low blood pressure.

The researchers' use of grid computing tools to automate molecular docking is "an excellent application of grid computing," said Julie Mitchell, an assistant principal research scientist at the San Diego Supercomputer Center. Features like "deadline- and budget-constrained scheduling should make the software very attractive to pharmaceutical companies" and to companies interested in such computationally demanding applications as risk analysis, scientific visualization and complex modeling said Mitchell. "There's nothing specific to molecular biology in their tools, and I imagine they could be applied quite readily in other areas."

The researchers also handled the process management aspects of adapting the applications to grids well, she added.

"The [researchers'] approach is obviously the way to go for those type of applications on the Computational Grid," said Henri Casanova, a research scientist in the computer science and engineering department of the University of California at San Diego. "The notion of providing remote access to small portions of domain-specific databases is clearly a good idea and fits the molecular docking applications," he said.

The economic concepts underlying the scheduling and costing of grid applications are still immature, Casanova added. "The results concerning application execution are based on a Grid economy model and policies that are not yet in place. There are only vague notions of "Grid credit unit" in the community and the authors of the paper assume some arbitrary charging scheme for their experiments. This is an interesting avenue of research, but...there is very little in terms of Grid economy that is in place at the moment," he said.

The data access and computation techniques are technically ready to be used in practical applications today, according to Buyya.

Buyya's research colleagues were Jon Giddy, and David Abramson of Monash University in Australia and Kim Branson of the Walter and Eliza Hall Institute, in Australia. The research was funded by the Australian Cooperative Research Center for Enterprise Distributed Systems Technology (EDST), Monash University, the Walter and Eliza Hall Institute of Medical Research, the IEEE Computer Society, and Advanced Micro Devices Corp.

Timeline: Now
Funding: Corporate; Government
TRN Categories: Distributed Computing; Applied Computing; Supercomputing
Story Type: News
Related Elements: Technical paper, "The Virtual Laboratory: Enabling On-Demand Drug Design with the World Wide Grid," posted on the computer research repository (CoRR) at xxx.lanl.gov/abs/cs.DC/0111047

TRN

Virtual Computers Reconfigure on the Fly

By Ted Smalley Bowen, Technology Research News
November 28, 2001

Grid computing, which pieces together temporary, virtual computers from resources on the Internet, is in theory a good way to handle tough number-crunching tasks that change over time.

Grid software combines the muscle of a few or even hundreds of computers by coordinating scheduling and security across the different types of systems. These combined resources are needed to speed up scientific and engineering applications that frequently involve complicated equations and elaborate graphical simulations.

But early efforts have only been able to handle simple, relatively predictable programs, rather than the complex, custom programs run by scientists, engineers, and their ilk.

What is lacking is a steady mechanism for maintaining sufficient levels of compute power for the duration of a virtual Grid computer's tasks. Grid applications need to be able to monitor the resources and performance of the systems that fuel them and switch to other appropriate systems when the original contributors fail to meet their requirements.

Toward this end, a group of researchers at Argonne National Labs, the University of California at Berkeley, the University of Chicago, and the Max Planck Institute for Gravitational Physics have developed software that reconfigures a virtual grid computer on-the-fly in order to keep it humming.

This adaptive approach is designed to help existing Grid computing software address the compute power problem. "Grid computing must be adaptive, because... one is required to operate in an environment about which one has imperfect knowledge and that has dynamically varying characteristics," said Ian Foster, a professor of computer science at the University of Chicago, and an associate director in the Department of Computer Science at Argonne National Laboratory in Argonne, IL.

The researchers' software uses notification and event services to determine when things change, said Foster.

To create the system, the researchers started with the Cactus set of Grid computing tools, which allow programmers to run groups of calculations in parallel across multiple computers that can range from PCs to supercomputers. The researchers also used the Globus toolkit to provide Grid resource discovery, access, location, migration and security functions.

The researchers added programs for adapting applications to run on different types of computer systems, for detecting drops in performance, for finding appropriate resources, and for handling the migration process.

They also added software that keeps tabs on the progress of a given program through a series of checkpoints in order to carry that information over to new systems as they are recruited.

The checkpoints save a snapshot of the computation in a form that permits the job to be shifted to another system, even one that has a very different architecture and operating system, or different amounts of disk space and memory, said Foster.

The various systems involved in a virtual Grid computer using the researchers' software must perform to the standards of a contract between the user and the systems providing the compute power. If a contract is broken, the software finds other resources and reconfigures the virtual computer.

The researchers evaluated their software on several Grid testbeds, loading down virtual Grid computers with more and more tasks until performance dropped by more than 10 percent. They set the software so it found alternative resources that gave the bogged-down virtual computer more compute power after three such drop-offs.

The researchers' system currently requires the operators to monitor this performance manually. "We obtain per-time-step measurements, and monitor according to a user-specified definition of what forms a contract violation. Future plans have us doing this automatically," Foster said.

The experiment involved no scheduling software, although eventually computers participating in Grid applications will be subject to random use and will need to prioritize their resources. "So far, we assume no scheduling technology: applications discover unloaded servers, and initiate computation there if authorized," said Foster.

The researchers plan to add asynchronous notification of resources, meaning an application can begin at a lower speed or fidelity, and improve if and when more resources become available, he said.

The software is powerful and generic enough that it can accommodate many different variables for determining application migration, said Henri Casanova, a researcher in the Grid computing lab at the University of California San Diego. It is likely to "motivate Grid application developers to architect their applications in ways that will support migration," he said.

In doing so it will open up the interesting questions of how to decide whether to trigger migration, and when and where to do it, Casanova said.

The work also opens the way for a more detailed exploration of Grid computing issues like scheduling, resource selection, and application adaptability, he said.

In general, the scheme is best suited for large applications that must run over long periods, said Casanova. In large scientific simulations that consume large amounts of tightly coordinated resources, migration will be useful if the cost of migrating is not greater than the cost of running the application on potentially sub-optimal resources, he said.

Foster's colleagues in the study were Gabrielle Allen, Gerd Lanfermann , Thomas Radke and Ed Seidel of the Max Planck institute, David Angulo and Chuang Liu of the University of Chicago, and John Shalf of Lawrence Berkeley National Laboratory. The work is slated to appear in an upcoming issue of the *International Journal of Supercomputer Applications*. The study was funded by the National Science Foundation (NSF).

Timeline: Now
Funding: Government
TRN Categories: Distributed Computing; Applied Computing; Supercomputing
Story Type: News
Related Elements: Technical paper, "The Cactus Worm: Experiments with Dynamic Resource Discovery and Allocation in a Grid Environment", slated for publication in November in the *International Journal of Supercomputer Applications*

TRN

Tools Automate Computer Sharing

By Ted Smalley Bowen, Technology Research News
September 12, 2001

How many economists does it take to inaccurately forecast a recession?

To answer that impertinent puzzler, you might try tapping into a grid of computing power made up of spare cycles from sources as disparate as a university supercomputer across town, a cluster of servers in another state, and a scattering of

workstations around the world. You'll also need some stray disk storage and spare networking resources to tie it all together.

Grid computing started as a response to scientific users' need to pull together large amounts of computing power to tackle complex applications. These ad hoc assemblages of distributed resources are coordinated by software that mediates different computer operating systems and manages things like scheduling and security to create sophisticated, virtual computers.

Grid computing, still generally confined to the research community, is one manifestation of utility-style data processing services made possible by the Internet. Peer-to-peer computing, which allows disparate users to dedicate portions of their computers to cooperative processing via the Internet, is a related phenomenon used mostly by consumers and businesses.

Both models harness a potentially vast amount of computing power in the form of excess, spare or dedicated system resources from the entire range of computers spread out across the Internet. The University of California at Berkeley, for example, coordinates one popular scientific example of grid computing — an Internet community application that uses background or downtime resources from thousands of systems, many of them home PCs, to analyze telescope data for the search for extraterrestrial intelligence (SETI) project.

A group of researchers at Monash University in Australia and the European Council For Nuclear Research (CERN) in Switzerland has proposed a scheme that has the potential to increase the reach of grid computing by applying traditional economic models - from barter to monopoly - to manage grid resource supply and demand.

The researchers have built a software architecture and mapped out policies for managing grid computing resources; these could also work with peer-to-peer applications, according to Rajkumar Buyya, a graduate student in the computer science department at Monash University.

The methods could facilitate a broad range of computing services applications, said Buyya.

"They can be used in executing science, engineering, industrial, and commercial applications such as drug design, automobile design, crash simulation, aerospace modeling, high energy physics, astrophysics, earth modeling, electronic CAD, ray tracing, data mining, financial modeling, and so on," he said.

Although peer-to-peer and grid computing are not new, there hasn't been an overarching scheme for handling the massive amount of bargaining and staging required to carry out such on-demand jobs with reliable levels of quality, and pricing to match, Buyya said.

The researchers' scheme is aiming to fill that gap, he said. "We are focusing on the use of economics as a metaphor for management of resources and scheduling in peer-to-peer and grid computing, as... a mechanism for regulating supply-and-demand for resources depending on users'... requirements."

The researchers scheme allows consumers and computing service providers to connect and hammer out pricing and service levels. It would allow the parties to agree on one price for quick delivery of services during times of peak demand, and another for less urgent delivery, for example.

Resource brokering/sharing tools analogous to Napster will eventually handle the trade in access to computers, content, scientific and technical instruments, databases, and software, Buyya said.

"With new technologies, the users need not own expensive [computer] resources. Resource brokers [can] lease services that are necessary to meet... requirements such as deadline, spending limit, and importance of the work. Our technologies help both resource consumers and providers to manage the whole scenario automatically," he said.

In a grid computing scheme, consumers usually enlist brokers to procure computing resources for a given project. Grid service providers make their systems available by running specialized applications and resource trading services. A grid market directory links brokers and providers.

The researchers' grid architecture goes a step further, using standard economic pricing models, such as commodity market, posted price, bargaining, tendering and auctions, to hash out the terms of broker-provider deals.

The researchers' tools, Nimrod-G Computational Resource Broker, DataGrid broker, Grid Trading Services, Grid Market Directory, and Grid Bank, work with existing grid middleware like the Globus toolkit.

The researchers have tested the tools on the World Wide Grid (WWG), a global network testbed of different types of computers including PCs, workstations and servers.

Two types of tests simulated brokering, scheduling and execution computing jobs, and emphasized speed and cost, respectively. The tests used a commodity market pricing, or fixed-price model. One application scheduled computations needed for a drug design application that screened molecules, he said.

The researchers used Nimrod-G to aggregate the systems resources as they were needed. "The resource broker automatically leases necessary resources competitively, depending on the [users'] requirements, such as deadline and budget constraints," Buyya said.

Using a more common systems-centric approach would make it more difficult to provide service levels that can vary from user to user and application to application, depending on the importance of the problem at the time of execution, he said.

As the tools get established, they could be deployed for use in production systems such as Australian Partnership for Advanced Computing (APAC) and Victorian Partnership for Advance Computing (VPAC) resources for routine use, said Buyya. "Depending on market forces, we believe that it will

take two or three years for widespread use of economic models for Grid and [peer-to-peer] computing," said Buyya.

The researchers plan next to test the methods' scalability, improve scheduling algorithms, and update the Nimrod/G broker software to handle more sophisticated task allocation and management, Buyya said.

The study makes a good start at hashing out ways in which disparate computing resources can be made available and consumed, according to Lee McKnight, a professor at Tufts University's Fletcher School of Law & Diplomacy.

The researchers' contribution is "imagining and testing a standards or protocol-based framework through which computing resources may be accessed or shared on the basis of one of a variety of different models for brokering or trading resources," he said.

But the way the researchers used the models is artificially limited to narrowly defined grid computing resources and doesn't address networked computing services like application hosting and bandwidth brokering, and quality controls like service level agreements, said McKnight.

The work "is but one element of a yet-to-be defined economic model of pervasive computing and communications environments," he said. "The 'data economy' as the authors call it will ultimately include both [peer-to-peer and] a variety of other interaction and resource access modes."

Buyya's research colleagues were Jonathan Giddy and David Abramson of Monash University and Heinze Stockinger of CERN.

The work was funded by the Australian Government, Monash University, Cooperative Research Center (CRC) for the Enterprise Distributed Systems Technology (DSTC), and the Institute of Electrical and Electronics Engineers (IEEE) Computer Society. Heinz Stokinger's work was funded by CERN and the European Union.

The researchers are scheduled to present their work at the International Society for Optical Engineering (SPIE) International Symposium on The Convergence of Information Technologies and Communications (ITCom 2001) in Denver, August 20-24, 2001.

Timeline: 2-3 years
Funding: Government, University
TRN Categories: Internet
Story Type: News
Related Elements: Technical paper "Economic Models for Management of Resources in Grid Computing," Preceedings of the International Society for Optical Engineering (SPIE) Conference on Commercial Applications for High-Performance, August, 2001

TRN

Store Globally, Access Locally

By Ted Smalley Bowen, Technology Research News
January 31, 2001

Keeping track of information will likely become more difficult in the not-too-distant future of innumerable computing devices of all sizes churning out and navigating unfathomable volumes of information.

Researchers at the University of California at Berkeley are charting a data storage scheme that will cater to a world where computers are part of almost everything — from smart shoes to smart buildings — and are nearly always connected to a network.

Instead of keeping data only in a central place, like a single hard drive, the OceanStore scheme will archive data across computers world-wide. This will make it more readily available to people who use several computing devices to access to the same data and to groups of users sharing information.

The scheme constantly updates data, saving it in numerous places so it can be accessed even if some of the computers holding it are lost, and keeps it safe from unauthorized access.

"The original motivating factor was the idea that Moore's Law growth in storage is really almost a liability, because it encourages huge pools of data to be [stored] on little tiny devices," said John Kubiatowicz, assistant professor of computer science at Berkeley. "If you have a terabyte storage in a little pen, or something — which is not all that far off — and you run over it with your SUV, you've just destroyed a terabyte of storage."

The key elements of OceanStore will be software that routes, organizes, and encrypts information on the Internet, data searching and recovery tools, and programming interfaces that will allow other programs to access the scheme.

In order to do this, the system will have to support more than 100 trillion files, according to Kubiatowicz.

OceanStore's routing software, or service protocol, will augment Internet Protocol (IP), which controls the flow of traffic on the Internet. The service protocol will route information to data repository computers, or servers, around the Internet.

The data will be encrypted, tagged with global unique identifiers (GUIDs), split in pieces and dispersed among servers in different geographical areas.

This dispersed network of servers will be treated as 'untrusted,' meaning they will be able to read the global unique identifier tags, but not the underlying data. Some servers will also be able to interpret protocols for maintaining data consistency, although they still will not be able to read the underlying data.

This scheme assumes that users might access data from anywhere, and so will allow frequently accessed data to be

cached, or temporarily stored in an easily accessible place, in order to speed up access. "You can decide to move data close to you that's important, and you can decide to move data you don't care about far away from you," said Kubiatowicz.

When users search for data, OceanStore will first use a probabilistic algorithm, which looks in the most likely place first. The algorithm will pass the request from one system to neighboring systems in a given vicinity. If the probabilistic search fails, a wider ranging hierarchical search will begin. Using the two types of search algorithms in this way will allow for faster access to cached data.

Instead of anchoring to a centralized control scheme, individual systems will keep track of their available storage and computing resources and communicate that information to the system at large.

The local systems will also track the flow of and interactions among the data traffic. They will automatically tune these interactions, adjusting the placement, number and location of objects, to, for example, cluster related files.

"The infrastructure itself observes patterns of access that you make and may decide that whenever you access file A you typically access files B and C," said Kubiatowicz. "If you go to Europe... and you access file A, it can know to immediately start getting files B and C somewhere close to you. That's called clustering, and that's just one of many [possible] optimizations."

Individual systems will communicate tracking information to a parent node, which will coordinate a hierarchical local arrangement of servers. If, for example, a single system cannot handle the volume of requests for a given set of data, it will communicate that to a parent node, which will trigger the creation of a replica of that data on another system.

If a number of servers fail, taking with them a percentage of a given set of data, OceanStore will use error correction codes similar to those used in computer memory to recover the original information.

Data that has been fragmented and distributed among multiple servers in this scheme can be recovered from as few as one quarter of the fragments, said Kubiatowicz.

Software developers will be able to write programs that make use of the scheme and modify older programs to gain limited access using application programming interfaces (APIs).

The researchers plan to demonstrate the scheme by the end of the year, according to Kubiatowicz. This proof-of-concept version will be written in the Java programming language, and will include UNIX system interfaces and a read-only proxy for the World Wide Web.

There are "many public companies and start-ups that are focusing on developing solutions for providing storage over the Internet," said Jehoshua Bruck, professor of computation and neural systems and electrical engineering at the California Institute of technology.

OceanStore is an interesting scheme, he said. "It is an integration of a number of existing concepts into a system. [But] I think the [researchers] underestimate the complexity of building the distributed computing/network side."

Given the scope of OceanStore, a business arrangement similar to cooperative utilities, or cell phone network partnerships would be needed to run it, according to Kubiatowicz. A limited version of the scheme could be adopted within three years, but a full implementation would take five to ten years, he said.

Kubiatowicz's research colleagues were David Bindel, Yan Chen, Steven Czerwinski, Patrick Eaton, Dennis Geels, Ramakrishna Gummadi, Sean Rhea, Hakim Weatherspoon, Westley Weimer, Chris Wells, and Ben Zhao.

The scheme is described in "OceanStore: An Architecture for Global-Scale Persistent Storage," published in Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000) in Cambridge, Massachusetts, November, 2000.

The research was funded by Defense Advanced Research Projects Agency (DARPA), the National Science Foundation (NSF), EMC, IBM, and Nortel.

Timeline: 3 years, 5-10 years
Funding: Government, Corporate
TRN Categories: Internet; Data Storage Technology
Story Type: News
Related Elements: Technical paper, "OceanStore: An Architecture for Global-Scale Persistent Storage," Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000), Cambridge, Massachusetts November, 2000

Index

TRN's Making The Future Report is published 10 times a year by Technology Research News, LLC. Each 20- to 40-page package assesses the state of research in a field like biochips, data storage or human-computer interaction.

Single reports are $300 to $500. A one-year subscription is $1,600. To buy a report or yearly subscription, go to www.trnmag.com/email.html.

We welcome comments of any type at feedback@trnmag.com. For questions about subscriptions, email mtfsubs@trnmag.com or call (617) 325-4940.

Technology Research News is an independent publisher and news service dedicated to covering technology research developments in university, government and corporate laboratories.

Kimberly Patch
Editor
kpatch@trnmag.com

Eric Smalley
Editor
esmalley@trnmag.com

Ted Smalley Bowen
Contributing Editor
tbowen@trnmag.com

Chhavi Sachdev
Contributing Writer
csachdev@trnmag.com