Breakthroughs

The work of University of Minnesota faculty members helps shape the development of data mining

About 10 years ago, Vipin Kumar and several of his colleagues began exploring an emerging area that also introduced them to the earth’s atmosphere, medical records, and biology.

Those disciplines are not the usual ones for computer science and engineering researchers; but their work in data mining has so many diverse applications that it can take them in some unexpected directions.

Along the way, their work also has helped shape the development of data mining and resulted in breakthroughs on many fronts.

Data mining offers tools and techniques to bring the exploding universe of information into practical focus. Data mining makes use of information technology solutions to turn massive amounts of raw data into knowledge and patterns that experts can use to advance their fields.

At the University, a group of talented and experienced researchers are taking data mining to the next level.

And they are compiling an impressive list of results in the growing area of data mining, including preventing attacks on computer networks, bolstering retail marketing and sales, and finding less invasive ways to diagnose and treat disease.

The Digital Technology Center (DTC) offers the University’s data mining faculty members the opportunity to connect with researchers from other disciplines and with industry. Those connections are helping expand the depth and breadth of data mining research. On March 23, the DTC hosted the Data Mining Open House to look at new possibilities for collaboration.

The University of Minnesota is one of the world’s leading research centers in data mining. The University crew includes faculty members Arindam Banerjee, Daniel Boley, George Karypis, Vipin Kumar, Shashi Shekhar, and Jaideep Srivastava. Their accomplishments help demonstrate the potential of data mining in solving a wide range of practical problems.

Web mining

Jaideep Srivastava, computer science and engineering professor, and his group coined the term “web mining” when they

continued next page
wrote about their work in the mid-1990s. Now gathering and analyzing data about browsing and purchasing behavior of internet users are musts for companies that conduct online business.

Srivastava was one of the first experts to take data from multiple sources and make it possible to automatically extract the desired information from those sources.

Today, Srivastava continues his groundbreaking efforts in data mining by collaborating with colleagues and industry partners on diverse projects. He has worked with Intel, IBM, and United Technologies. He also took a leave from the University in 1999 to spend a year at Amazon.com, helping define approaches to offer real-time product recommendations, as well as helping identify the user's role in real time.

Srivastava also has contributed to the MINDS project and to DTC-supported projects that use data mining techniques to study the behavior of chimpanzees and to understand disease (see page 7).

In ongoing research, he continues to explore the next-generation of data mining applications to consumer marketing and customer service, including using those techniques physically in stores. He is also starting to apply data mining techniques to improve the understanding of social networks. For example, by analyzing the e-mails at Enron, Srivastava was able to point out communication patterns among participants.

**Security and NASA**

Head of the computer science and engineering department, William Norris Professor, and DTC faculty member, Kumar recently finished co-writing a textbook on data mining for use by universities worldwide.

With a background in high-performance computing, Kumar began looking at the question of extracting meaningful information from massive data sets as computing capabilities grew. His projects include a NASA-funded initiative that seeks to use satellite and scientific data to detect atmospheric patterns and events. NASA is using the data to discover changes in the global carbon cycle and the climate system.

In another project, the data mining techniques of Kumar and his colleagues are improving security on computer networks. The Minnesota Intrusion Detection System (MINDS) software now alerts cybersecurity analysts to possible computer break-ins and other suspicious activity on networks in real time.

At the Army Research Laboratory’s Center for Intrusion Monitoring and Protection, MINDS helps analyze network traffic from nationwide Department of Defense sites. MINDS software also monitors more than 40,000 computers at the University of Minnesota. The techniques offer promise for additional applications in early identification of health care or financial fraud.

Another recent collaboration with Mayo Clinic involves a project to determine patterns in medical data that may help physicians pursue less invasive diagnostic and treatment methods.

“These are very powerful techniques,” says Kumar. “There are huge opportunities for data mining. We are excited about applying these techniques to new applications.”

**New frontiers**

In his research, George Karypis has developed data mining algorithms to divide large data sets into more manageable and understandable clusters (groupings) of data. His clustering toolkit, CLUTO, has been downloaded by thousands of users.

In the consumer world, cluster summaries offer valuable marketing and sales information. Data mining tools take those volumes of purchase data and translate them into a more precise understanding of buying patterns and behavior.

“The early data mining research focused on the tools for data clustering,” says Karypis, associate professor of computer science and engineering. “Now, with this understanding of the data, we are working on predictive modeling, using the data patterns to draw conclusions about future outcomes.”

Karypis recently started a research project on drug discovery. Data mining can help narrow down the many chemical possibilities in new drug development to those that offer the strongest potential.

For Karypis, the next frontier in data mining involves non-traditional data. He is one of a few to look at the emerging area of graph mining, or data mining of graphs. “We want to first cluster and then classify the patterns in data sets of graphs.” Such a tool can help scientists in fields that rely on graphic representation of data, such as biology and chemistry, tackle new questions.

**Algebra, statistics, and more**

Daniel Boley learned that one of his research areas, linear algebra, applies in many ways to data mining. “I found that I could use some of the methods in data mining problems,” says Boley, computer science and engineering professor. “Linear algebra offers some of the more effective tools for large-scale data mining algorithms.”

He began contributing to data mining when he applied those methods to analyze text documents. In one project, the tools helped researchers examine the alcohol laws of 50 states to see if the laws brought social changes.

His most recent project brings him into biological world, where the combination of mathematics, statistics, and computer science techniques are making it possible to sort through very different data sets. Boley is helping animal
Scientists understand the impact of feed on the biology of cattle and pigs.

He also joined another collaboration with his colleagues and Mayo Clinic to determine patterns in biomedical text. Researchers are searching through clinical notes for any patterns of symptoms indicative of a particular kind of heart disease.

Promising future

The work of data mining faculty has attracted attention from diverse funding sources, including the National Science Foundation, Department of Energy, U.S. Army, Department of Homeland Security, National Institutes of Health, and Department of Defense. With the increasingly pumped-up volume of valuable data, the future of data mining is on the growth track.

Collaboration will remain a critical component to success, and the DTC can play a key role in forging even stronger and more productive collaborations through its work and the consortium concept.

For companies such as Thomson Legal and Regulatory (TLR), which manage large volumes of information, the focus on data mining is a welcome one. “We are now releasing products that analyze information mined from legal documents so we are interested in accessing state-of-the-art tools,” says Peter Jackson, vice president of research and development for TLR, who also serves on the DTC Technology Advisory Team. TLR has licensed software that DTC faculty member George Karypis developed for document clustering.

“Fundamentally, data mining has a strong application component,” says Srivastava. “Interaction with industry opens the data door for researchers to develop new tools and techniques, which benefit both faculty and industry partners. “A consortium offers a place for faculty and industry to gather and work through data mining projects and applications,” he says. “It creates a focal point for interactions.”

In this digital age, the amount of information online far exceeds the amount of information available in our traditional libraries. And each second, more and more data is added to the digital world.

The explosion of such data brings new challenges, including simple overload. With so much available information, how can anybody find what they really need? How can we easily put that data to work for us?

The Digital Technology Center (DTC) is a great place for academic and industry researchers to explore the challenges and questions that are associated with data mining and its applications.

The area of data mining emerged in the last decade and has taken root as it offers tools and solutions to dealing with significant volumes of data. Leading the way for many of those developments is a group of University faculty who are making impressive strides on data mining projects.

The outcomes of faculty work have helped advance e-commerce, strengthen our security, and improve medical efforts. The cover story includes more information about faculty data mining projects and milestones.

The unique pool of talent in data mining here presents some distinct opportunities for new collaborations. Data mining projects typically involve applications to real-life issues in diverse industries that use large amounts of data. Indeed, data mining is a natural area for cooperation among faculty and industry participants.

As a critical part of its mission, DTC acts as a gateway for industry to connect with University expertise, as well as a place for faculty from diverse disciplines to connect with each other and with industry. DTC has developed consortiums in areas where DTC strengths and industry needs intersect. Our consortiums bring DTC faculty and industry representatives together to work jointly on cutting-edge research.

Current consortiums include the DTC Intelligent Storage Consortium (DISC) and the DTC Digital Design Consortium. DISC research has direct relevance to the data mining efforts, and illustrates how DTC helps exploit the synergies offered by its many research programs.

The data mining area already has attracted industry interest, and we believe that a consortium may help bolster existing collaborations and spur new ones. We are beginning to explore the possibilities of a data mining consortium through meetings with industry and our March 23 Data Mining Open House.

DTC is a place where the very act of experts coming together leads to new approaches and new projects. We believe that our faculty strengths combined with industry interest generate momentum for data mining initiatives. We are looking forward to forging new partnerships and leveraging our collective abilities for new breakthroughs.

Andrew Odlyzko is the DTC director, an ADC Professor in the School of Mathematics, and Assistant Vice President for Research. He can be reached at odlyzko@umn.edu.

For information about data mining collaborations, contact Jim Licari, assistant director for industrial liaison, at jpllicari@dtc.umn.edu.
Many Impacts

DTC Initiatives grants result in progress on diverse fronts

In 2004, the Digital Technology Center (DTC) launched its Initiatives in Digital Technology Program, which provides seed funding for projects that promote collaborations in areas related to digital technologies. Summaries of the first 11 projects reveal wide-ranging impacts.

Quantum dots to detect breast cancer

Researchers from mechanical engineering, microbiology, pharmacology, and biomedical engineering are pursuing the use of quantum dots for imaging in the near infrared, which offers an improved tool for biomedical applications that include detection of breast cancer cells.

They succeeded in synthesizing quantum dots that emit near-infrared light and in developing imaging systems to detect them. Their progress is helping them attract additional funding from federal agencies.

“The project is on the cutting edge,” says Allison Hubel, mechanical engineering faculty member. “Our recent imaging studies were our most impressive results to date.”

DTC funds made a difference in launching the project, which also involved Marc Jenkins from the Department of Microbiology, Sundaram Ramakrishnan from the Department of Pharmacology, and Jerry Sedgewick, Biomedical Image Processing Laboratory.

“The real limitation to this project was funds to purchase and assemble necessary equipment,” says Hubel. “DTC funding bridged the gap and permitted us to make equipment and obtain preliminary data.”

Five times the speed

While at DTC one weekend, Jaijeet Roychowdhury from electrical and computer engineering ran into fellow DTC faculty member Wei-Chung Hsu from computer science and engineering. Their conversation about ways to increase speed in shrinking semiconductor technologies led them to collaborate and seek DTC funding for a joint project with impressive results.

Roychowdhury focuses on the chip design for devices. Chips are expensive to build, and his computer programs can model and simulate behavior. “We’re trying to predict what the design will do before building the chip,” he says.

Device models require large amounts of computer power, which means they take time to run. Hsu’s work on cache optimization offered a possible solution. The two developed and successfully implemented a software infrastructure that allows the modeling and simulation programs to run five times faster.
The DTC seed funding is paying off. Industry is very interested in the development, and Roychowdhury has been able to leverage additional funding from the National Science Foundation and SRC Research Corp. for future work. “This would not have happened without the DTC.”

Common language

University of Minnesota Duluth and Mayo Clinic researchers used DTC funds to successfully develop measures that allow them to determine the degree of similarity or relatedness between pairs of biomedical concepts.

“The DTC grant program was particularly well-suited for our work,” says Ted Pedersen, University of Minnesota Duluth faculty member in computer science. “It placed an emphasis on collaborations outside the University and also supported work in biomedical informatics.”

After asking physicians and medical indexers to assign relatedness values to pairs of biomedical concepts, the team concluded that these different types of experts make similarity and relatedness judgments differently.

A tangible outcome of this collaboration is a set of new measures that automatically make judgments consistent with medical experts. Ultimately, such measures are intended to help physicians and others as they search for information about medical conditions that may be different, such as pulmonary edema and congestive heart failure, but are closely related.

The team developed a software package that implements the measures and completed a journal article, currently under review, and continues to develop and refine their measures.

“From the Mayo standpoint, the DTC program represented a unique opportunity to collaborate on a project with the University,” says Serguei Pakhomov of the Mayo Clinic. The involvement also exposed researchers to other joint University-Mayo projects and paved the way for future collaborations.

Powerful networking

Computational biology offers a way for researchers to better understand biological networks and ultimately biological function. To fully benefit, though, it also requires participation by experts from diverse disciplines.

“I was interested in developing a research community of faculty and students in computational biology,” says Claudia Neuhauser, ecology, evolution, and behavior faculty member. She collaborated with colleagues Fumiaki Katagiri and Neil Olszewski from plant biology to plan a seminar series, funded by the DTC grant.

The seminars brought together students and faculty from agricultural bioengineering, bioinformatics, conservation biology, ecology, evolution, and behavior, journalism, mathematics, microbiology, neuroscience, physics, and plant biology.

“The seminar series provided an excellent introduction into the theory of networks and the diverse areas of research where this theory has been applied,” she says. The seminar also inspired Neuhauser to think about applications for her research, including a new direction that involves investigating neural network modeling for food intake.

Online behavior

With state-of-the-art facilities to understand the online user perspective, the Usability Services Lab is a popular resource for University studies, faculty research, student learning, and DTC Affiliate industry members.

Additional eye-tracking equipment was installed in April 2005 to enhance the lab’s capabilities. The DTC grant and financial support from the University’s Office of Information Technology (OIT) funded the purchase of the eye-tracker. The equipment tracks and records eye movements of evaluators as they are looking at the computer monitor.

“The eye-tracker has made it possible for web designers and content writers to understand a user’s experience of a web site, even when the user isn’t describing it aloud,” says Alice de la Cova, manager of Usability Services.

“Simply knowing where users look on a web page provides a lot of insight into the usability of the page,” she says. “Usability evaluation teams
have been very pleased to be able to obtain the user's perspective in this way, in addition to watching and listening to the user."

The eye-tracker fills many needs. OIT makes good use of the eye-tracking equipment in its on-campus usability studies. It helps computer science and other students understand the importance and benefits of in-lab usability studies. The Institute for New Media Studies is taking advantage of the eye-tracker in its research studies, as are other faculty.

“It is a significant piece of infrastructure that has been used as part of grant applications to help show our relative advantage for certain types of research,” says Joseph Konstan, computer science and engineering professor who assisted with the equipment acquisition.

Continued debate

“The Symposium on the Digital Divide,” which received DTC financial support, attracted participants from around the world and resulted in a special double issue of the Journal of the Association for Information Systems based on the symposium’s best papers.

The symposium looked at understanding how and why management and business might react from an economic perspective as efforts come in play to decrease the digital divide by improving online access to all. It attracted interest from diverse disciplines.

“The eclectic nature of the audience helped generate several interesting discussions,” says Fred Riggins, faculty member from information and decision sciences at the Carlson School.

The journal issue helps take those discussions to an even broader audience. “This could not have happened without starting the process at the symposium,” says Riggins. “I believe this is a new stream of literature within the information systems community that promises to be fruitful in coming years.”

Clearer picture

Visualization techniques that geologists use in their studies are proving useful for radiologists as they examine mammograms to detect small and potentially malignant lesions.

The project involved Robert Hoolebeek from the National Digital Mammogram Center in Philadelphia, David Yuen from geology and geophysics, and Witold Dzwinel from AGH University of Science and Technology in Krakow. Hoolebeek and Yuen first connected at a data mining workshop several years ago. Intrigued by the prospect of applying the wavelet method to gain a clearer picture, the team applied for a DTC grant.

“The application of the technique was successful,” says Yuen. “We solved the problem and were able to use the method for better resolution with mammograms.”

The team published three papers based on the project, and Yuen and other University faculty are pursuing additional funding for continued work. In addition, the project also serves as a launching pad for other applications. “Now we are starting to look at blood flow and cancer,” says Yuen.

Educational transformations

The instruction of American Sign Language is changing for the better thanks to a project that applies information technology to education.

The DTC grant helped University researchers Simon Hooper from curriculum and instruction and Susan Rose from educational psychology supplement federal funding to develop an approach where students can learn and demonstrate their skills in signing and where teachers can evaluate and offer improved feedback.

The new system replaces the previous evaluation method of using VHS tapes. University students are currently piloting the new system, where they are able to complete signing exercises through web cameras attached to laboratory computers. Instructors can access student exercises, record comments, and better track and measure student progress and improve instruction.

Already other institutions recognize the benefits of the approach and are interested in project results.
“Not only is the system more efficient than videotapes, but it actually improves the assessments and instruction process and transforms the way American Sign Language is taught at the college level,” says Hooper.

Less invasive treatment

In this project, DTC data mining experts, George Karypis, Vipin Kumar, and Jaideep Srivastava from computer science and engineering, worked with Piet de Groen from the Mayo Clinic to maximize the value of already-captured clinical and molecular information. They are using historical data and data mining techniques to determine patterns and to construct a model that makes it possible for physicians to avoid invasive diagnostic procedures.

“Our goal was to see if we could look at data for liver cirrhosis and be able to predict the biopsy results without the biopsy,” says Karypis. “The project has produced encouraging results.”

It also has strengthened the collaborations between DTC data mining experts and Mayo. Team members have submitted a proposal to the National Science Foundation to support ongoing data mining work in the life sciences.

Chimpanzee behavior

Data mining experts are supplying the tools and techniques that allow researchers to make optimal use of the wealth of data about chimpanzee behavior available at the Jane Goodall Institute’s Center for Primate Studies at the University.

The center contains the research materials from the world-famous studies of chimpanzees in Gombe National Park, Tanzania. The DTC grant supported the development of a searchable digital library of visual material that researchers and scientists can use to advance studies in biology, anthropology, and medicine.

“There are 600 hours of high-quality video recordings of the behavior of chimps, as well as slides and black-and-white photographs,” says Anne Pusey, center director. The visual information can be used by researchers in a variety of fields all over the world, but only if they can easily find what they need.

The collaboration, which also involved Shashi Shekhar and Jaideep Srivastava from computer science and engineering, has resulted in the development of a database prototype, and work continues to incorporate all the visual data in the prototype. In the meantime, Pusey hears from researchers who are anxious to use the data. “It is a fantastic and unique collection.”

Robots at work

In this project, researchers William Schuler and Richard Voyles from computer science and engineering and Jeanette Gundel from linguistics combined their expertise in speech and gesture interfaces in an effort to help non-experts program robots.

As a result of the project, they are developing an interface for programming robots by demonstration. In this approach, users show and tell the robot what to do, similar to how they might “program” another human.

This type of programming interface allows users with limited formal training to quickly implement and refine solutions for quasi-repetitive tasks, such as unloading cargo from a vehicle or putting groceries in the pantry. The interface has many applications in situations where non-experts are using robots for short-run tasks, such as in emergencies, job-shop manufacturing, or elderly assistance.

“The key to our research is an emphasis on task experts rather than programming experts,” says Voyles.

“Many of us understand very well how to get a task done—we just don’t know how to program a robot to complete the task. Yet, we can all teach another human to complete the task. We do this by demonstrating to the human and by coaching as the human practices. Current programming interfaces are designed around the robot. We think programming interfaces should be designed around the casual user.”
DTC faculty members receive McKnight honors

The University recently named Yongdae Kim and Stergios Roumeliotis, both DTC and computer science and engineering faculty members, as 2006-08 McKnight Land-Grant Professors.

The McKnight Land-Grant Professorship Program helps advance the careers of the University’s most promising junior faculty members who show potential to make significant contributions. Kim’s research interests focus on practical security to emerging networked systems, and Roumeliotis works in the area of distributed estimation for networks of mobile robots and sensors.

DTC welcomes new Affiliate and DISC members

Seagate recently joined the DTC Affiliates Program, which promotes interaction between DTC and organizations with an interest in digital technologies. Affiliate members include ADC, Benjamin Moore, Dupont, Engenio, IBM, Industrial Technology Research Institute (ITRI), Seagate, SGI, Sun Microsystems, Symantec, 3M, Thomson, and Unisys.

In addition, ITRI became a new member of the DTC Intelligent Storage Consortium (DISC).

Formed in 2002, DISC brings together academic, industry, and government research interests with the aim of advancing research and development in storage technology. Other DISC members include Engenio, Sun Microsystems, and Symantec.

Check out highlights from DTC events

In October 2005, the DTC hosted the Wireless Cities...Community Context Conference, which explored the use of wireless networks to improve existing community services, and the IEEE Fifth Symposium on Bioinformatics and Bioengineering. In February, the Second Digital Technology Initiative Forum featured a poster session on 2006 funded projects and panel discussions.

For more information about those events, as well as upcoming events, visit www.dtc.umn.edu/seminars.html.