



Engineering **ENTERPRISE**

THE ALUMNI MAGAZINE FOR ISyE AT GEORGIA INSTITUTE OF TECHNOLOGY

Fall 2004

**Unleashing the Power of the Internet:
The Art of the Search**

**Knowledge Mining:
Taking Control of the Information Age**

**Text-Mining:
The Engineer's Approach to Literature**

REAL WORLD EDUCATION FOR WORLD-CLASS EXECUTIVES



TODAY'S COMPANIES NEED TO:

- Increase supply chain efficiencies
- Move to a global supply chain strategy
- Groom their rising stars
- Expand collaborative relationships

LEARN to improve supply chain efficiencies by grooming your executives in Georgia Tech's Executive Master's in International Logistics (EMIL) Program.

EXPERIENCE real-world results by learning best practices from the world's leading experts in EMIL's five 2-week residences at key locations around the globe.

BUILD a team that can deliver measurable results by linking finance with global supply chain management.

For more information, visit <http://www.emil.gatech.edu> or call 404.385.2538



WWW.GTBN.ORG

SPONSORED BY THE
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

KNOWLEDGE



CONNECTION



COMMUNITY



Knowledge Mining



by **William B. Rouse**

PUBLISHED BY

ISyE
Lionheart Publishing Inc. Georgia Institute of Technology
John Llewellyn, President

EDITORIAL

Managing Editor Ruth Gregory
ISyE / Georgia Tech
Atlanta, GA 30332-0205
Tel: (404) 385-2627
Fax: (404) 894-2301
ruth.gregory@isye.gatech.edu

Contributing Editor Sarah Banick
sbanick@mindspring.com

ART DIRECTION & PRODUCTION

Art Director Alan Brubaker, ext. 218
albrubaker@lionhrtpub.com

Publication Designer Donna Browning, ext. 226
donnam@lionhrtpub.com

SALES & MARKETING

Advertising Sales Marvin Diamond, ext. 208
marvin@lionhrtpub.com

CIRCULATION

Circulation Manager Maria Bennett, ext. 219
bennett@lionhrtpub.com

ADVERTISING OFFICE

LIONHEART PUBLISHING INC.
506 Roswell Street, Ste. 220
Marietta, GA 30060, USA
(770) 431-0867
Fax: (770) 432-6969
E-mail: marvin@lionhrtpub.com

Engineering Enterprise is published quarterly by Lionheart Publishing Inc. and ISyE, Georgia Institute of Technology. Editorial contributions including manuscripts, news items, and letters to the editor are welcome. Unless stated otherwise, articles and announcements reflect the opinions of the author or firm and do not necessarily reflect the opinions of *Engineering Enterprise*, Lionheart Publishing Inc., ISyE, its advertisers, or sponsors. Yearly subscriptions (four issues) are available for \$18 (U.S.), \$22 (Canada & Mexico). Payable in U.S. funds.

Copyright © 2004 by Lionheart Publishing Inc. and ISyE. All rights reserved. No portion of this publication may be reproduced in any form without the written permission of the publisher. Printed in the USA.

The many faculty, staff, and students in the School of Industrial and Systems Engineering are adept at creating knowledge and gaining skills in the multiple subdisciplines within the School. They publish what they learn in a wealth of academic journals associated with these subdisciplines. In this way, ISyE makes numerous contributions to advancing many subdisciplines.


Unfortunately, this wealth is not easily accessible by those in other disciplines and practitioners who are addressing the many complex problems our society faces today. This difficulty is not only a problem for our School. It is endemic to all the disciplines across Georgia Tech and academia in general. The extent of our specialization has reached the point that only specialists can penetrate the mysteries of this knowledge.

Of course, this is not only the case for knowledge generated by universities and research institutes. The Internet has enabled access to immense bodies of information on organizations, products, services, genealogy, and so on. We can literally find millions of items – “hits” – on any topic. Digesting this information – that is, converting it to knowledge relevant to your intentions when posing the query – can be a daunting task.

This issue of *Engineering Enterprise* focuses on knowledge mining. We are very fortunate to have David Seuss among our many highly accomplished alumni. David is CEO of Northern Light, a leading provider of custom search solutions. David’s interview begins with how an ISyE graduate found his way from North Avenue to chemical plants to search engines, and ends with an insightful view into what is possible when mining the web.

Prof. J.C. Lu, one of the senior members of our rapidly growing statistics faculty, considers knowledge mining from a statistician’s perspective. His concern is with how best to unearth reliable patterns in large data sets. This goal is common among organizations ranging from retailers to intelligence agencies. Sometimes you can get the feeling that various folks are trying to make sense of everything you do, at least everything you do on your computer.

Alan Porter, a recently retired ISyE professor, brings us back to all those journal articles that I mentioned earlier. He and his colleagues have developed methods for mining this literature. He is not content, however, to just find these articles. He uses patterns among articles, e.g., citations, to assess emerging technologies in terms of their maturity and related metrics.

Georgia Tech is in the knowledge business – creating it and imparting it. We create more value than we can deliver through traditional channels. Archival articles and classroom lectures are great, but there is much more additional value that can be provided by also adopting less traditional channels. ISyE is not only adopting such channels – we are creating them. 

William B. Rouse is the H. Milton and Carolyn J. Stewart Chair and Professor in the School of Industrial and Systems Engineering at the Georgia Institute of Technology.

Interview with an EMIL Alumnus: Educational Experience Helps Intel to Meet Global Supply Chain Goals

By Terri Herod, EMIL Managing Director

An increasingly impressive alumni roster serves as testimony that our Executive Master's in Logistics (EMIL) program has captured the imagination of the world's leading companies. And with last year's 30 percent jump in enrollment, coupled with ongoing features in *Frontline Solutions*, *Information Week*, and *Inbound Logistics* – and research on hot topics like Chinese logistics and returns management, EMIL is firing on all cylinders. This means we're even better positioned to create value for program participants and their company sponsors through real-world supply chain education and solutions.

As the world's premier supply chain master's degree program, EMIL has garnered so much interest from Fortune 500 companies that, with new applications now being accepted for the class of 2005-06, Georgia Tech anticipates 50

percent more qualified applications than there are actual positions available.

Applicants selected for this new class will complete an 18-month real-world, supply chain curriculum, taking part in five EMIL residences:

Residence I: Consumers	Atlanta	May 15 – 27, 2005
Residence II: European Logistics and Trade Issues	Europe	October 2 – 14, 2005
Residence III: Asian Logistics and Trade Issues	Asia	February 19 – March 3, 2006
Residence IV: Transportation and Logistics in the Americas	Latin America	May 14 – 26, 2006
Residence V: Manufacturers	Atlanta	September 10 – 22, 2006

INTERVIEW WITH CINDIE BLACKMER, INTEL CORPORATION

Q: Can you cite some specific examples of what you have learned at EMIL?

A: We're not just following what other organizations are doing but taking the best ideas from companies like Dell, Ford, and Schneider and blending them with our own experience to see how they fit with the Intel model. I have applied these ideas in managing inventory, transportation costs, and optimization of warehouse distribution locations. Specifically, we are reviewing our warehouse flow processes for optimum output, and we are using what we've learned from network optimization to understand all supply chain decision tradeoffs.

Q: What is your global supply chain project for EMIL?

A: It focuses on enhancing the Intel extended warranty program. Our goal is to provide replacement parts within three to five days of a customer's order. With focus on emerging market countries, we are working to determine the optimum locations for inventory and the best transportation providers to meet our customers' needs.

Q: How is EMIL helping you solve this challenge?

A: The Latin American residence was extremely valuable to me, especially as it relates to specific logistics issues, such as the transportation network, security, and customs clearing. Through the residence, I immediately achieved an understanding of these areas that I could never get sitting in my office in Chandler, Arizona. Now I know what's within our control and what's not. Learning the special requirements of Brazilian documentation has helped me better understand what I need from a service provider in this market. EMIL exposed me to a mix of industry leaders in South America and worldwide that I could have never accessed otherwise.

Q: Is this a good time to be investing in the EMIL program?

A: Given current economic conditions, Intel is only funding core business activities. However, Intel has chosen to invest in EMIL because it prepares us for the future. It positions us to be world class in supply chain logistics, so we can take full advantage of changes in the

An increasingly impressive alumni roster serves as testimony that our Executive Masters in Logistics (EMIL) program has captured the imagination of the world's leading companies.

The early application deadline for the 2005-06 class is January 15, 2005; the late deadline is March 15, 2005.

Focusing on the new class and the diverse residences that this educational experience offers, we took this opportunity to interview Cindie Blackmer, an Intel Corporation executive and EMIL alumnus, to hear how her company is leveraging the program to optimize its global supply chain (see sidebar below). Blackmer serves as Intel's worldwide transportation manager. [e](#)

**Class Starts: May 2005
Application Deadlines:
January & March 2005**

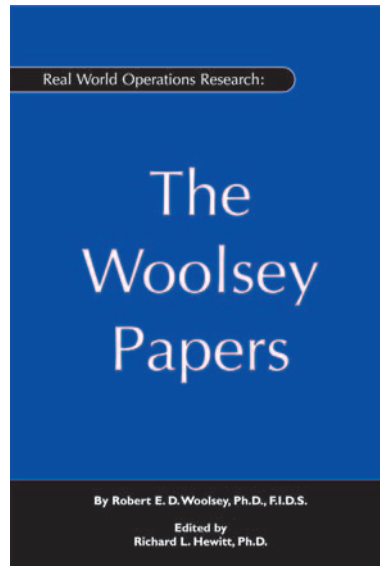
economy and respond quickly to any and all opportunities.

Q: Overall, how has EMIL been received at Intel?

A: Intel has sent three people through the program, and our experiences have caught the attention of others. The idea of participation is spreading upward within the company. Everyone, including managers, wants to participate. People recognize that there is a lot of new information in logistics and supply chain management, and they want it! EMIL provides avenues into this information that just don't come across your desk on a daily basis. [e](#)

Real World Operations Research: The Woolsey Papers

Edited by
Richard L. Hewitt, Ph.D.



Real World Operations Research: The Woolsey Papers is a collection of the diverse writings of one of OR's most outspoken and controversial figures, Gene Woolsey. Whether he's the man you love to hate or hate to love, Woolsey's humorous and practical writings leave little wonder as to his venerable status in the field.

"Gene Woolsey is unique, provocative, insightful and entertaining. This collection of some of his articles is an important and thought provoking read for anyone in the field of OR. Hopefully, this book will motivate and guide the behavior of those in the profession to successfully apply OR to resolving real problems that matter and to insure that the solutions are actually used."

— Tom Cook,
Chairman and CEO,
CALEB Technologies Corp.
President (2003), the Institute for
Operations Research

Real World Operations Research: The Woolsey Papers

By Robert E. D. Woolsey, Ph.D., F.I.D.S.
Edited by Richard L. Hewitt, Ph.D.

\$19.95 • 164 pages • 6 x 9 • paperback •
ISBN: 1-931634-25-4

Published by **Lionheart Publishing, Inc.:**
506 Roswell Street, Suite 220, Marietta,
Georgia 30060,
(888) 303-5639, ext. 214
fax: (770) 432-6969
e-mail: lpi@lionhrtpub.com

Order online at:
www.lionhrtpub.com/books

This collection contains 33 articles published from 1972 to 2003, covering a broad spectrum of subject matter relevant not only to OR/MS professionals, but also educators, managers and corporate administration. To accompany his writings on operations research, chapters also cover topics from communication in the corporate world to handling labor disputes, getting promoted to getting fired. Through creative storytelling and down-to-earth advice, Woolsey provides readers with the knowledge and philosophical mindset to conquer operations and management situations in all settings.

EPICS Works with New One Semester Format

EPICS (Engineering Projects in Community Service) is changing its format slightly this year to keep up with changes in the Senior Design classes.

“This is the first group of EPICS projects that are being conducted under the new one-semester senior design project,” says Associate Professor Faiz Al-Khayyal, who coordinates the EPICS project, a national program that places teams of undergraduate engineering students in partnerships with local community service agencies and institutions. Students will receive the same amount of credit as they did for the past two-semester projects. Projects will be completed in December, and a new round of projects will be chosen for Spring Semester. Following is a short summary of the projects for Fall Semester:

DeKalb Medical Center: Decreasing Disposition Time

Disposition, the last stage of patient care within the Emergency Department (ED), is a process by which a physician determines the next step in a patient’s care. There are three possible outcomes: discharge, admittance, and transfer. This project is focusing on the discharge portion. Disposition times at DeKalb Medical Center (DMC) are higher than industry standards, and decreasing these times should improve the ED’s patient capacity and allow the DMC to send fewer patients to other emergency hospitals.

The goal of the project is to develop an Arena model that simulates the current operations and then modify the processes within the model to determine the effects on disposition times. The variations in the model will provide several cost/benefit scenarios, which will offer insight as to whether the hospital’s current disposition process should be modified, and if so, which aspects.

DeKalb Medical Center: Emergency Department Space Utilization and Layout Redesign

DeKalb Medical Center’s (DMC) 24-hour Emergency Department (ED), one of the busiest in Metro Atlanta, treats approximately 75,000 patients a year. The ED is faced with strained resources and limited space, and is planning to expand by adding more space. The architectural additions are not scheduled to take place in the immediate future; therefore the team has been instructed to make the necessary changes to fully utilize the available space.

The current situation is running at 106 percent capacity and has appreciable overflows, such as bed-ridden patients waiting in hallways due to the lack of space. Goals include decreasing the average time for a patient in the system by reallocating the current ED space configuration, while staying within a given budget, and allowing for a seamless transition between the current emergency center and the planned future expansion. A detailed analysis of the current layout will be completed, and areas of space will be identified that could either be re-assigned or reconfigured. The team will also look into developing an optimization model that will determine which types of rooms should be added.

Children’s Miracle Network: Impact of Promotions on Fundraising

Children’s Miracle Network (CMN) plans to implement an inventory management system with a standard set of documented procedures that will help them keep track of their office supplies, promotional items, and fundraising supplies. They are developing a database system through the use of software in the EPICS lab. They also plan to develop a documentation system that keeps track of contact/sponsor information,

and brief descriptions of all correspondence. These will allow the CMN staff to remember what they have and what they have not discussed with a particular client. In addition, with the developed system, CMN will be able to perform cost benefit analysis concerning promotional items; specifically, which promotional items generate the most donations, and whether or not greater quantities of promotional items would have generated more monetary donations. This analysis will help CMN more effectively use its resources when trying to solicit donations.

Atlanta Housing Authority: Resident Services Survey

For the past seven years, the AHA has distributed a resident satisfaction survey to assess the quality of life of its residents. Unfortunately, past surveys have proven to be ineffective. This project involves the design and development of a pilot of AHA’s new, reliable survey tool that will collect statistically valid data and provide meaningful analysis. One of the most challenging aspects of the assignment is the population studied in this survey. At least 30 percent of the approximately 8,500 heads-of-household have less than a seventh grade education and a significant portion are partially or completely illiterate. Past surveys did not receive complete responses due to length, wording, lack of incentive, and perceived lack of anonymity. The low response caused by these and other design constraints was one of the main reasons that little or no meaningful analysis of the data could be performed. In addition, AHA has requested the design team provide staff members with statistical background in order to fully understand and utilize the new survey tool. [e](#)

FACULTY NEWS

Georgia Tech's *Traveling Salesman Problems* webpage received a mention in the August 27 edition of *Science* magazine. The brief blurb, titled "You Can Get There from Here," describes the history of the problem and offers a link to www.tsp.gatech.edu/index. The page is the work of several academic and industry experts in combinatorial optimization, including ISyE Professor **William Cook**, who holds the Russ and Sammie Chandler Chair.

Dr. Augustine Esogbue, ISyE Professor and Director of Intelligent Systems and Controls Laboratory, was recently elected Fellow of the Institute for Operations Research and the Management Sciences (INFORMS). Esogbue was cited for his "outstanding contributions, achievements, and service that have advanced the profession of operations research and the management sciences." Esogbue is also a Fellow of AAAS, IEEE, and the Nigerian Academy of Science.

Professor C. John Langley, Jr. has been named as one of the nation's top five logistics professionals by the Executive Master's in International Logistics program at Georgia Tech. Langley is The Logistics Institute Professor of Supply Chain Management and director of the Supply Chain Management Executive programs. His selection was based on the ability to endure during unusually challenging economic times, as well as his ability to create value in a highly competitive market environment.

ISyE Associate Professor Eva Lee was featured in the Spring/Summer edition of *Research Horizons*. Lee uses mathematical optimization techniques originally developed for the industrial world to help doctors produce the best results from radiation therapy. Through externally-applied beams or "seed" implants, radiation therapy provides a valuable tool for treating cancer. But its effectiveness depends on the ability to target cancer cells with appropriate radiation doses while sparing healthy tissues. To read the full article, go to: <http://gtresearchnews.gatech.edu/reshor/rh-ss04/c-radiation.html>.

ALUMNI NEWS

Clive E. Hardy, BIE 1967, of Perry, Georgia, recently incorporated Hardy Homes, Inc., to develop residential property and build upscale spec and custom homes.

Elizabeth "Betsy" Higgins, BIE 1991, has been named chief financial officer of Oglethorpe Power Corporation. She previously served as senior vice president and group executive for the Finance and Planning Group at Oglethorpe. Before joining Oglethorpe in 1997, Higgins worked for sev-

eral consulting firms, including Southern Engineering, Deloitte & Touche, and Energy Management Associates. She graduated with honors from Georgia Tech and also earned an MBA at Georgia State University.

Floyd A. Peede, Jr., BIE 1948, is the author of three books of poetry, known as *Georgia Boy I, II, and III*, which he hopes to soon publish. Peede, who is retired, lives in Americus, Georgia.

Tripp Rackley, BIE 1992, and **Bird Blitch, BIE 1997**, co-founders of BroadSource, Inc., have raised \$5.7 million for their start-up company, housed in Georgia Tech's Advanced Technology Development Center. BroadSource develops software that checks for and prevents errors in telecom bills.

STUDENT NEWS

Doctoral candidates **Abhyuday Mandal, Zhiguang Qian**, and **Andrew Smith** led Georgia Tech to the Team Championship in the ASA Stat Bowl, held at Joint Statistical Meetings in Toronto. Smith was also the Individual Runner-Up. Learn more in the article on our website. [e](#)

GAMS**OPTIMIZATION**

The General Algebraic Modeling System (GAMS) is a high-level modeling system for mathematical programming problems. It consists of a language compiler and a stable of integrated high-performance solvers. GAMS is tailored for complex, large scale modeling applications, and allows you to build large maintainable models that can be adapted quickly to new situations.

GAMS Development Corporation

1217 Potomac Street, N.W.

Washington, D.C. 20007, USA

Tel.: +1-202-342-0180 • Fax: +1-202-342-0181

sales@gams.com • <http://www.gams.com>

UNLEASHING THE POWER OF THE INTERNET:

THE ART OF THE

SEARCH

C. David Seuss, BSIE 1972, is chief executive officer of Northern Light, an Internet search engine company in Cambridge, Massachusetts, which provides search and content integration solutions for enterprises and individuals. Northern Light won three *PC Magazine* Editors' Choice awards in a row for being the best Web search engine, earned a position as a "Top 100" company in *eContent* magazine, was designated "Best of the Web" by *U.S. News and World Report*, was picked for the "Top 100 Companies That Matter," by *KMWorld* magazine, and was designated "Best of the Web" and "a professional researchers dream" by *Forbes* magazine, among other awards. Mr. Seuss was also founder and CEO of Spinnaker Software Corporation, which he led from inception to a public company with \$65 million in revenue and 280 employees. He also holds an MBA from Harvard Business School.

EE: How did you get from Georgia Tech to the search engine business?

Seuss: It was a long trip, even though each piece is connected. After Georgia Tech, I worked as an industrial engineer down in the bowels of chemical plants. Chemical plants don't have much to do with the things that we normally study in industrial engineering. For example, all the material handling is done in pipes. There is no labor. A person performing this task watches dials on a monitor. Is it working? Is it fully loaded? Who knows? So the labor costs, at least in the traditional sense, aren't very amenable to traditional engineering analyses.

EE: So, it's hours and hours of boredom with moments of sheer terror?

Seuss: Yes, if you make a mistake, you blow the plant up. You may start talking about how two operators in the control room can do what three are doing now, perhaps with some impossible-to-define increase in the risk of blowing the plant up. There is not as much interest in material handling or labor costs or the products in the chemical plants, things like chlorine, or nitric acid, or hydrazine, so you don't get into assembly and things you typically do as an industrial engineer in some sense.

One thing you might work on is the maintenance environment. Maintenance is an extraordinarily complicated information-processing problem at a chemical plant. You have to bring the process down and make sure it is tied off properly. If it is a unionized plant, you have to have a wheelwright and a welder, and plumber and a carpenter, all arriving at the site at the same time. They are all requested through their organizational structures, of course. And they all have to have the right tools and equipment with them, and they have to have a document that explains what it is they are supposed to do, which is probably right about 85 percent of the time. The other 15 percent of the time you expect them to be well trained enough to figure it out and do the right thing, as opposed to what the document says. Then you have to unwind that whole process.

In one of my plants, we had 400 people in the maintenance department. It was the largest single discretionary cost. The information flow — how's a maintenance ticket generated, how do you make all these things happen — is a perfect industrial engineering problem. Consequently, I spent four years working as an industrial engineer on information systems for manufacturing plants.

Then I went off to Harvard Business School. I became a management consultant where I actually practiced industrial engineering. We had clients who were losing money in a division and we would reorganize the entire manufacturing system of the division. Many of our assignments at Boston Consulting Group were industrial engineering on a big scale.

Somewhere along the way, looking at my software roots, I started looking at software ventures. I got the urge to start a business. I was looking at, actually, maintenance problems... software that would address maintenance issues back in my industrial engineering days. Through a series of analyses of business opportunities, I eventually came to a software idea that I liked a lot and started a software company.

I think my partner and myself basically invented the home education software market.

That company, Spinnaker Software, was focused on the home education market. I think my partner and myself basically invented the home education software market. We were the early leader in that market, then we went public and we sold it, and I made an attempt to retire in 1994. I was a miserable failure at retirement. I went from the hard charging CEO of a public company to a guy who would get up in the morning and no one would ask my opinion about anything. I found that to be very challenging and, after a while, I started looking for a job. The Internet was happening, so in 1996 I joined a startup called Northern Light. I've been running Northern Light mostly since 1996.

What is it about the search engine business that attracts me? I think, from the viewpoint of computers and networking, we have really solved a problem. Every computer in the world can be linked to every other computer in the world at this moment. There is no development of science, engineering, or network infrastructure, hardware or software, required from this point out in order to accomplish the task of linking every computer to every other computer. It's a done deal. That exists and it is called the Internet.

What that means is we can query every piece of content in the world that is available anywhere in the world, at least in theory. Turns out in practice, though, that part of the equation is a complete and miserable failure most of the time because the search solutions are not up to the task. On the one hand, we have tremendous opportunity to find anything, anywhere; and on the other hand, it is the most frustrating problem that most of us spend hours and days and weeks struggling with in our business lives. So current problems, not solved, amenable to technological innovations — that is my fascination, or I guess the basis of it.

EE: Remember a few years ago using Ask Jeeves and Yahoo and other kinds of search engines, then it seemed everybody gravitated to Google. Why do you think that happened?

Seuss: I think it's a couple of things. First of all, we're talking about web search when you're talking about those search engines. The corporate environment is enormously more complicated than web search. We certainly had a moment in time when everyone believed Google had better relevance than

other folks on a web search. This was because they took the idea of citation analysis from the Internet community and applied that to web search.

EE: Did this make Google better or faster?

Seuss: Google doesn't do that anymore. They use a different system now, which is much closer to what other search engines use. And in research tests of public web search engines, the conclusion was that all of them are quite poor. Now, they're all about the same. Brands establish a position. That's the current thing in the consumer-marketing channel; they rely on the brand. So at this point, it's more or less a marketing issue rather than anything on the technical side.

EE: The simple screen of Google is also attractive. With most of the other choices, you get all kinds of other things in your face while you're trying to search.

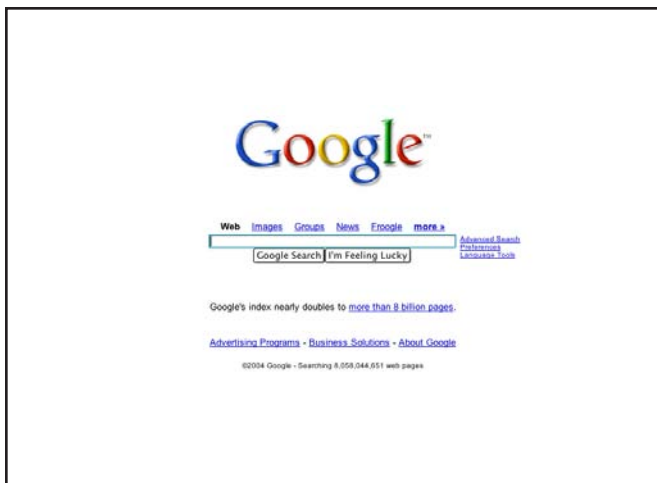
Seuss: I personally prefer lots of information. Different people prefer different things. I'm more attracted to the Google news page than I am to the Google first page. It's more individual than anything. In a consumer market, you battle and segment. There are three rules of marketing. The first one is segmentation. The second one is segmentation, and the third one is segmentation. I'm a market segment and you're a different one.

The corporate environment is really different, however. If you've got to apply citation analysis within the corporate environment, it cannot deliver. "Page rank" is the Google phrase for it. Page rank seems to work because the content relevance is not very good due to the diverse sites you don't control. You know nothing about the content. If you compare that to the corporate intranet market or content management systems, where the amount of content is much smaller, you know a lot about this type of content. You can know that in this file location, we keep RFPs. If the document has this code on it, that means it's an engineering requirements document for the new product line. And, so on. You know a lot more, at least in theory, about what the content might be, in the corporate market.

Combining internal and external content can leverage the power of this knowledge of content. For example, in a market research setting, you might want to license content from 20 or 30 or 50 vendors, or you might want to have a news feed. You might want to include a web crawl of your competitor's website — sort of an alerting system, every time they put an announcement on their website. And you might want to have competitive intelligence reports from the field. All of that mixed into one database. So the challenge in corporate applications is often content integration as much as it is service. That's the first challenge, content integration. In a web search



Screenshot of www.northernlight.com



Screenshot of www.google.com

market that is not an issue. You just get a list of hits according to page rank.

EE: How does this change the way a search engine operates, a Google vs. Northern Light?

Seuss: In the corporate intranet market, the first difference is the role of content integration. Northern Light has years and years of experience developing databases that are custom to a corporate application. Our clients can have any content located in the world on any computer in any format in possession of any organization, as long as they have legal rights to that content to put it on a web page.

EE: Customized in terms of the sources and the content?

Seuss: First of all customized in terms of the sources. You can have external content, you can have web content, and you

can have internal content. You can have internal content from any place in the company, you can have licensed feeds, you can have anything you imagine would help people in this setting be put into a database for you. That's the first difference.

The second difference is the role of classification. As more and more content is added to a database, being able to classify it becomes increasingly important. You might want to know what the document is about. Northern Light has 17,000 known subjects that encompass all human knowledge, and a trained classification engine that can take arbitrary unstructured content and classify it to the right subject — automatically.

You can also classify content by type. The subject is what the document is about. Type in "what is the document?" Is it an RFP for a new business proposal? Is it an engineering document? Is it a usability study report? Is it a market research report? You can also classify by source, both geography and language. Northern Light is the only search engine designed from scratch from day one with the idea that we would use classification, performed at query time, to filter and organize, and sort search results. The fact that such classification is available to the query search enables fundamental changes of the search application.

EE: Is this built into your search technology, rather than something you rebuild for each customer?

Seuss: The platform is general, but it gets expressed differently for each customer. One customer has one set of types, a different customer has a different set of types. One customer has 27 market research firms; another customer has five news feeds and a competitive web crawl. Another customer only wants a search engine that searches the 1,600 sites of the laminating, coating, and metalizing industry. The platform is always there, and the platform includes the classifications, the taxonomy, and the ability to use these things. It's expressed for each customer, depending on what that customer's needs are.

EE: Are there things that your customers ask you for that you don't know how to do yet, but you see on the horizon?

Seuss: The big thing is in collaboration. We are getting ready to launch our collaboration tool. Lots of people work in groups, a lot of people share search results, people also want a way to organize their research life. Search engines are not designed to be personal research work stations. You can point it, first of all, at any arbitrary search engine: Northern Light, the Google public web search, you point it at internal search engines, you can point it at a Northern Light enterprise search engine running on your corporate intranet, point it at a competitor, or whoever. Point it at whatever the collection of search resources are that you use. Then, when you do a search, you have the option of declaring that search persistent, and put the

Collaborating is a big deal, and that's the next wave, I think, in the corporate search world.

results in the folder. You have the option of declaring it persistent, and if it is persistent, then that results list is preserved for you. So that's the first step, it is persistent search results.

Next thing you do is rank the documents as on target or not on target. One's that are not on target are deleted. For one's that are on target, we determine how we can get 10 or 20 or 30 of those together. Then, from a technical viewpoint, we have a pretty good idea of what you're looking for. We now have much more information than we had when you first expressed the query. We now know how to find additional documents that are very likely to meet your criteria. This augmented search now runs in the background, looking for documents like the ones you rated as highly relevant.

With our collaboration tool, you can also define a collaboration group and admit other people into it. All of those people then have access to all the documents that are in the folder. You can define who will be able to share these documents, contribute to the folder, or just have read-only privileges. So now your workgroup can share the results of their research. Collaborating is a big deal, and that's the next wave, I think, in the corporate search world.

EE: Are there some things people request that you have no idea how to do?

Seuss: Importance is a really hard one. For example, "Who's the most important son of Genghis Khan?" That is a really hard question, unless you can fake it by looking for articles entitled "Who was the most important son of Genghis Kahn?" or "Fred was the most important son of Genghis Khan," but that is cheating. That's only looking for a word match.

Importance is really difficult. We have no starting point in the search engine for how to deal with that. There is science that has not been developed, and we are nowhere close to being ready to do engineering. That kind of question is actually less prevalent in the corporate world. I'll give you a corporate example of that. "Is this document relevant to Sarbanes-Oxley?" "Is this document important to my Sarbanes-Oxley compliance efforts?" The relevance is a statistical calculation of match to the query. Importance is different. I can search an e-mail system and come back with a million documents that can relate to compliance issues. But where are the ten that are important — that is a really hard problem. And we've had customers that have asked us how to approach that and we don't have any idea.

EE: You mention collaboration. You have gone from an traditional organization format, for instance with Spinnaker, to much more of a virtual organization. Can you talk about how that works?

Seuss: We have one office that is 10' by 10'. We rent space, and as part of that deal we have an option to use the conference rooms. We have around 55 full-time and part-time employees that revolve around a hub in one way or another. We mostly work with e-mail and telephone conferences and meetings in our conferences rooms at our office.

In a way, it feels very similar to a normal office environment — you walk down the halls. I've found whenever any of the companies I've been running exceeded 100 employees, I couldn't remember their names. I still have room for names in my brain, but in order to learn a new one, I have to forget an old one. You walk down the hall, you see people you don't know, and if you need to have a meeting, you schedule it in your conference room. That's the way it works in our office as well.

EE: What's the geographic distribution of your employees?

Seuss: They're kind of all over New England; we have people in Florida, some Midwest employees, and then we have employees in St. Petersburg, Russia. We're scattered across nine time zones, I think.

EE: What kind of challenge does that present for you as the leader?

Seuss: You certainly need a lot more trust. You can't see what people are doing. You really rely on people to accomplish their task. I'll have to say it works better with employees whom you've known for a very long time. Our tenure is years and years and years. I don't have to worry what these folks are doing today. I know they are doing the right thing. We have some things like priority lists and we do a lot with shared databases, for example, directing our crawlers. We have a URL database and people from any location can enter URLs in the database they are working on to crawl and fill in the relevant content and then the crawler configuration. So you can be the analyst in Cincinnati entering URLs to crawl in this database. The crawlers are running in Massachusetts and are managed by a database hosted in Cambridge. But the operator of that database and the crawlers are located in Russia. After the analyst determines that this is a good URL for this project, then that analyst enters it into the database and the crawler configuration is automatically generated, and it is assigned to a crawler, and the crawler crawls it. Automated processes and systems, shared databases, those kinds of things become more important.

EE: Do you have any employees you haven't met?

Seuss: I do. We have people in St. Petersburg that I've never met.

EE: Do you have employees that no one has ever met?

Seuss: Now that's an interesting question. I would say that all of the folks that we deal with remotely are either one or two categories: they have worked with Northern Light before or for a long time, or they have worked with someone who is a Northern Light employee.

EE: So there is only one degree of separation?

Seuss: There is only one degree of separation. I do find that we have a bias toward hiring family members. When someone recommends a family member, they only recommend the really good family members. So there is a higher quality to the references. Now references in this day and age are virtually useless...and dangerous from the legal standpoint, if you actually tell the truth about an ex-employee. So no one will tell you anything in a reference check. References checks don't have any value in hiring. But family members will tell you the truth. They'll say, "I've got two brothers; one of them, I wouldn't put him in this company if he paid me \$1 million to do it. The other one is really great, and we ought to use him for this..." Family members are one way to recruit employees that exist in a virtual company. Another way is people you've worked with for a long time.

EE: How do you know there aren't people in your network that you don't realize are there?

Seuss: Well, you know, we've joked about that. When you're dealing with offshore employees, where the wages are very different, the people offshore could recruit people and you'd never know it. We do know that our St. Petersburg office has occasionally hired people for a song to accomplish tasks for us. I can actually say I am 100 percent sure that has happened.

EE: There is also a possibility of people being in your network, monitoring your technology, and taking advantage of it and no one in the company knows about it.

Seuss: Not really. We have source code control systems that record who checks out what. In our St. Petersburg office, as an example, I think our security is much better than it is here. We don't have CD-ROM burners on the computers there. We have no personal e-mail at your workstation. We follow the local practices in the software development community there. All the Internet connectivity is for business purposes only.

I worry a lot more in the United States about somebody walking into our office and getting into the server


area, or walking out with a computer under their arm. In our office in Cambridge, we had to tackle a guy once walking toward the elevator carrying two of our computers. One of the employees took him down. I'm more worried about that.

Also software code does not exist in isolation from the capabilities of the organization. Northern Light can take our technology and do wonderful things with it. Someone else would be less successful with our technology because they won't have the years of experience building these applications for corporations. We've built more than 200 enterprise search applications. It's more than the code; it's the skills of the organization. Also the brand name.

EE: How scalable do you think the organization model is? Can there be 500 or 1000 Northern Light employees?

Seuss: I don't know. It depends a lot on personal relations, and that doesn't feel like it's scalable to 500 people to me. Is it 100? Sure. 200? Maybe...500, I seriously doubt it. And I actually think as we roll ahead, the structure of the economy is going to change. There are technologies like web services...the whole environment of corporate software application development is shrinking from macro applications to micro applications. Software development projects that were millions of dollars are now hundreds of thousands of dollars and will soon be tens of thousands of dollars to execute. So I actually believe the growth in the industry is in the small firms. We can use the web now to market, distribute, and support the products. That changes fundamentally the organizational scale issues. Before you might need a full sales force, which meant you had to be a big company, with lots of revenue and high software prices. You had to have a big engineering staff, which meant a bigger company with higher software prices. Now you need a website. Websites download a lot of stuff, and people will seek you out and we have hundreds of leads a day coming in over our websites. I think we're coming into a period when smaller organizations will be where the action is.

EE: What do you see Northern Light being like five or ten years from now?

Seuss: Oh, I think we'll have a couple of hundred employees. I'd like to keep us on a size level where we're able to maintain this organizational structure. It's very satisfying to work in. Our average commute to work is 18 seconds or something? That makes for a very efficient workday. 

For more information about Northern Light search engines, visit www.northernlight.com



TAKING CONTROL OF THE INFORMATION AGE

By Dr. J.C. Lu, Professor, School of Industrial and Systems Engineering

The Information Age can be overwhelming. Thanks to computers, we have learned how to collect and process massive quantities of data, then dress it up in color-coded pie charts and pass it around in bound folders at meetings. In the last two decades, we have let our computers and their sorted data guide the direction of our markets and strategic visions. Yet we are only beginning to understand where data can take us.

Knowledge mining (also known as data or text mining) is a hot topic in virtually every industry and academic discipline. But, by far, no one is as intrigued as the business community. In a highly competitive marketplace, a thorough understanding of consumer trends and habits can make or break a product. In the world of business, data mining can help target specific products and services that customers are more likely to buy, and determine credit card users' buying patterns to more accurately predict their future purchases. It can also be used for identifying stolen credit cards.

Knowledge mining, defined for industrial engineers, is a process of extracting systematic patterns or relationships and other information from databases for improving people's decision ability.

In the area of product design and development, it can be used to understand the relationships between customer needs and design specifications, and systematically identify the factors affecting a project's success or failure based on past projects. In manufacturing, it brings the potential for fault diagnosis and prediction of the amount of product defects in manufacturing; and operational manufacturing controls such as intelligent scheduling systems, which learn the dynamic behavior of process outcomes and generate control policies.

Knowledge mining, defined for industrial engineers, is a process of extracting systematic patterns or relationships and other information from databases for improving people's decision ability. It has a wide range of applications including business intelligence gathering, drug discovery, product design, intelligent manufacturing, supply chain management, logistics, and even research profiling.

Enterprise Miner and Intelligent Miner, released by SAS and IBM, respectively, are some of today's more popular knowledge mining software. Some of the companies utilizing knowledge mining tools successfully in their operations today include: Fleet Financial Group (for customer characteristics analysis), Ford (for harshness, noise, and vibration analysis), Boeing (for post-flight diagnostics), Kodak (for data visualization), Texas Instruments (for fault diagnosis), and Motorola (for customer data management and analysis).

The knowledge mining process is iterative and consists of the following main stages:

- understanding problem goals,
- data selection,
- data cleaning and preprocessing,
- discovering patterns,
- analysis and interpretation,
- reporting, and
- using discovered knowledge.

Pattern discovery is a crucial step. There are several approaches to discover patterns, including classification, association, clustering, regression, sequence analysis, and visualization. Each of these approaches can be implemented via one of the following competing and yet complementary techniques, such as statistical data analysis, artificial neural networks, machine learning, and pattern recognition. This article will not go deeply into the core of the above methods. Instead, we present a real application of pattern discovery using a machine learning technique called *CART* (Classification and Regression Trees) in financial analysis.

There are three main reasons for the popularity of tree-based methods used in CART. First, they decompose a complex

problem into a series of simpler problems (e.g., binary decisions). Second, the tree structure resulted from successive decompositions of the data often provides a great understanding of the complex problem. Third, the methods generally require a minimal set of assumptions for solving the problem.

Example 1 (Customer Characteristics): Fleet Financial Group, a Boston-based financial services company with assets of more than \$97 billion, is currently redesigning its customer service infrastructure, including a \$38 million investment in a data warehouse and marketing automation software. To profit from this repository of valuable information on more than 15 million customers, Fleet's analysts are using data mining tools to learn about their customers and to better target product promotions, such as home equity lines of credit. In order to target the mailing list for the company's third quarter home equity product promotion, Fleet wants to develop a model to estimate each prospect's probability of responding to the mailing, as well as estimate the expected profitability of respondents. Based on this expected profitability, the database would be segmented by scores that identify which prospects should receive one of several home equity marketing pieces and which should not receive a mailing at all.

By hybridizing CART and logistic regression techniques, Fleet can use each methodology's strength to complement the others. The first step in the modeling process is to gather the historical data for creating a prediction model. Fleet selected a sample of approximately 20,000 customers with response records; included were 100 percent of past profitable respondents, as well as two percent of past non-respondents. The data set was then transferred into CART to display the interaction of the data. The resulting effects were incorporated into a logistic regression model that illustrated the overall and local landscape of the data.

When the data were fed into CART, the software automatically generated a decision tree whose branches and nodes showed the hierarchy of binary data-splits and displayed the data set's myriad variables and their interactions. This hierarchy distilled nearly 100 predictor variables into a more manageable amount of approximately 25. In addition, the CART nodes provided probability ratios that were used to understand why one segment would be more responsive than another.

The CART model illustrated certain characteristics of "best" respondents by predicting the expected balance they would carry on the credit line, as well as how much they might transfer from another line. In addition, the CART results painted a portrait of the principal characteristics of the least responsive customers. These prospects would either not likely respond to a Fleet product offer because they do not have a need for a large

line of credit, or — equally of concern — they would respond but their subsequent credit line usage and/or likely loss would not be profitable for the bank.

Fleet’s project team is cautiously taking into consideration other factors, such as the mailing’s time of year and the number of other financial product offerings that customers have received recently. “Test and control groups are needed to validate the efficiency of our targeting with this predictive model,” said Fleet’s project manager. “We are, however, very confident that Fleet will achieve a high response rate with this mailing. Our customers have many more dimensions than the previous mailing model could encapsulate for predictions. Creating a hybrid model using CART and our other data mining and statistical tools was a more sophisticated approach that painted a very descriptive portrait of our prospects, enabling us to increase the probability of their response.” See www.salford-systems.com for more details.

Other examples show the wide-range applicability of knowledge mining tools across fields:

Example 2 (Product Design): Through the advance in computer and communication systems, process sequences of engineering product designers (or analysts) working together to develop products can be recorded in databases. Data and text mining techniques start to become useful in capturing design knowledge and make significant impact in shortening design cycles and meeting customer needs. In a project (Ishino and Jin, 2001) designing a “double-reduction” gear system consisting of four gears, three shafts, a few bearings, and a case, pattern matching and dynamic programming procedures were used to select one design satisfying the evaluation criteria and being competitive to the other seven alternatives at all design sequences.

Gabowski, Lossack, and Weibkopf (2001) described the use of geometric object classification systems to search databases of parts for automatically locating the part matching the requirements of a design. Machine learning tools (e.g., C4.5 and CART) were used in Schwabacher, Ellman, and Hirsh (2001) to select the best prototype and predict which design goals are achievable. Specifically, given a set of constraints including wind speed and racecourse, a combination of eight design parameters of racing yachts was selected to minimize the race time.

Example 3 (Research Profiling): Text mining is useful in discovering intelligence in a large body of electronic text sources such as publications, patents, and grant-award abstracts. Applying the text mining techniques to review the literature in “mining information in large databases,” Porter, Kongthon, and Lu (2002) spotlighted 15 ways for their research profiling procedures to be useful. They include:

- Depict the research *context* to target our research efforts wisely (1)
 - observe related topics (techniques) within our “large database mining” domain (2)
 - observe related techniques beyond our “large database mining” domain (15)

- gain a “big picture” perspective on the research activity (in our case, in data splitting, analysis, and integration) (7)
- at the “big picture” level find intersecting interests (8)
- Understand the research *community*
 - identify a range of information sources (9)
 - use the distribution of research literature in its own right to gain insight into how innovation is progressing (10)
 - find active organizations and individuals whose research relates to one’s own interests, particularly those working in different disciplines or research domains (15)
- Explore *topics* (techniques) (11)
 - map (graphically represent) topical interrelationships for a whole research area (4)
 - examine how our target technique fits with other approaches (6)
 - generate research opportunities in combining techniques – through mapping (12) or in-depth probing (13)
 - examine trends to ascertain which topics are “hot” (3)
 - zoom in to examine promising topics in depth (5)

(Numbers in parentheses refer to the 15 ways illustrated in the case analyses of their paper.)

Furthermore, Kongthon (2004) developed text-based association-rule mining procedures to capture parent-child hierarchies and sibling relations among related terms. Application of the procedures to abstracts of Thailand’s science and technology publications provided several insights for improving strategies of managing the Thai government’s R&D funding policies.

Methodological Studies of Knowledge Mining Methods

Although data mining procedures are successful in many applications as illustrated in examples given above, most of them are not effective in dealing with large size data. If there is an effective way of reducing the size of data, the commonly used data mining procedures could then be useful. Lu (2001) outlined several data-reduction ideas. They include: splitting data sets into smaller pieces; sampling representative data; and using summarizing, modeling, and transformation techniques to reduce the size of data. However, special procedures are required for some data with complicated nonstationary structure such as the antenna testing signals as shown in Figure 1 on the next page. The following are a few examples from on-going research in ISyE.

Example 4 (Wavelet-based Data Reduction): Professors Lu, Vidokovic, Huo, and their students have developed a data reduction method for complicated functional data based on wavelets. Six testing curves and two real-life data sets were used to compare their methods with procedures for data compression in the signal processing field and for data denoising in the statistical modeling field. Jeong *et al.* (2004a) showed that their

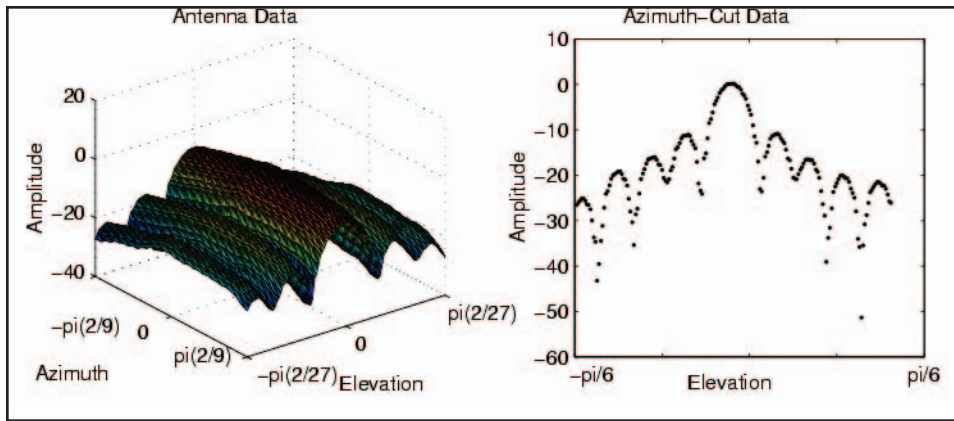


Figure 1: Data Signals from Nortel's Antenna Manufacturing Process

methods are more aggressive in reducing the size of data without sacrificing much modeling accuracy. In their experiment, some of the data cannot be analyzed by commercial software due to the large size of data, but are easily handled by the same data mining tool (e.g., cluster analysis) when the data reduction methods were used. Examples in detecting process problems in manufacturing applications showed the potential of the developed methods.

Example 5 (Data Reduction for Multiple Curves): Most of the data reduction, data compression, and data denoising procedures were developed for a single curve. In Jung and Lu (2004) a wavelet-based random-effect model was proposed to characterize the behavior of multiple curves collected in antenna testing and in sheet-metal stamping process for automobile manufacturing. Then, “vertical energy” based thresholding procedures were developed to reduce the size of data. Real-life and simulation examples showed that the proposed methods have excellent data reduction properties and the modeling accuracy is very satisfactory. Data mining (e.g., hierarchical clustering) based on the reduced-size data provided the same amount of decision power as analyzing the original large-size data.

Example 6 (Statistical Process Control (SPC)): SPC techniques are very popular in monitoring quality measures in manufacturing and service industries. Examples include automobile manufacturing’s quality monitoring (Lawless *et al.*, 1999), checking potential shifts in signals provided by mass flow controllers in semiconductor manufacturing (Kang and Albin, 2000), and detecting changes in acoustic emission signal patterns in nano-manufacturing (Ganesa *et al.*, 2003). Just as monitoring testing results from antennae produced from Nortel’s manufacturing process, these examples all discussed problems of monitoring linear and nonlinear curve data using the traditional SPC procedures. Moreover, in some applications (e.g., nano-manufacturing) the data size could be very large for typical SPC procedures to handle. Jeong and Lu (2004b) developed SPC procedures to select a few data quantities well representing data change-patterns such that the process monitoring can be efficient and effective. A series of investigations based on simulation and real-life example studies showed that

the proposed methods perform better than procedures extended from the literature.

Novel Applications

There are several ongoing projects in ISyE for developing novel applications of knowledge mining tools. The following show a few examples:

Example 7 (Supply chain and Logistics Management): Lu and his graduate students, Wang and Mangotra, are working on multi-scale modeling

techniques to summarize the information collected from supply chains’ logistics network (including distribution centers (DCs) and store locations and demands) for strategic decision-making purposes. Statistical spatial models such as Kriging (Cressie, 1993), spatially correlated Poisson, and conditional autoregressive Gaussian (Besag *et al.*, 1991) are used to model the large-size data at different scales of details. Generalized estimating equations are developed to link the summary information at various levels. With these approximated models representing the logistics network, decisions at different scales, such as the locations and capacities of global-level DCs serving stores in larger regions (covering many states), local DCs serving a few counties and city-level supply-routes can then be evaluated. A combination of data mining techniques, statistical modeling methods, game theory, and optimization methods is used to derive sound decisions for “coordinately managing” inventory at different levels of DCs and retail-stores’ shelf-space for meeting various space and time constraints.

Example 8 (Supply Chains’ Acceptance Sampling Plans): There are many reasons that product shipments could contain certain amount of “damages” when they arrive at DCs and stores. In Kim, Lu, and Kvam (2004), “product damages” include missing components, broken packages and goods, or even wrong orders. These problems resulted in numerous unsold products at stores. If there is no inspection for the in-coming products at DCs or stores, it will be very expensive to deal with these damaged products at the final destination of the supply chain, the retailing store. In working with one of the largest retailers in the world selling more than 5,000 products, Dr. Lu’s team discovered that the amount of “damaged” products could be as high as 20 percent, causing serious “stockout” problems at stores. Thus, it is critical to dig into product damage information across the entire supply chain collected at various locations and in several time points for developing a cost justifiable, efficient (keeping the logistics flow smooth), and effective (weeding out most of defects) acceptance sampling plan. Moreover, extending the SPC procedures studied in Example 6 to the supply chain data collected at multiple stages can detect anomalies in the logistics processes. This research could also be very important in third-

party oriented “just-in-time” manufacturing, where parts are shipped through the supply chain for making half-finished products and moving to various levels of assembly centers for making the final products at the destination.

Example 9 (Support Vector Tree): Professors Huo and Tsui, together with their graduate students, are working on several projects for extending data mining’s applicability. In particular, they have explored the idea of using Support Vector Trees (SVT) to extend the applicability of a population classification tool, Support Vector Machine (SVM). One problem with the current version of SVM is that it only considers two classes, i.e., the responses are binary. Extending the SVM method with a tree-structured hierarchical model, Huo *et al.* (2002) introduced SVT for handling responses with multiple classes. See their paper for possible applications in areas such as logistics, target recognition, and automatic control.

Example 10 (Beamlets): Knowledge discovery requires great computational efficiency in data mining algorithms. In analyzing pattern detection problem for image analysis, Arias, Donoho, and Huo (2003) show that by using a multi-scale method – beamlets – they can achieve the statistical and computational optimality simultaneously. Their work has been used in improving Lockheed Martin’s Automatic Target Recognition (ATR) systems.

Example 11 (Security Intelligence Mining): Professors Huo and Wu have a project supported by the National Science Foundation for locating “sparse” patterns in large-size stream data. Currently, their challenge is to identify a few events not occurring often (called “sparse” events) from monitoring large-size Internet traffic data for exploring potential terrorist activities. Huo and Chen (2004) have developed a special decision tree (cascade) to efficiently locate sparse events. Application to a fast-rate network traffic data shows the potential of this research.

Dr. Lu’s group is working on “scenario”-based intelligence mining methods. The motivation is from the following statements made by Director Tony Tether of DARPA on May 6, 2003 before several members of the House of Representatives. “The current commercially available data mining tools are ineffective for identifying patterns and building predictive models to understand terrorist activities. ‘Fishing expeditions’ through massive amounts of personal data is a wholesale invasion of Americans’ privacy.” Their research follows DARPA’s approach of developing attack “scenarios,” useful in locating specific patterns that could indicate terrorist activities. The following outlines a few key steps in their ongoing research.

1. Focus on one selected scenario. The analysts would start with suspects’ names (or

profiles) and use the object-oriented association rule mining tools in the link discovery system (see Figure 2) to evaluate if there is any evidence of linking this suspect with other suspects or suspicious activities.

2. Convert suspicious activities of this suspect into a terrorism intensity profile similar to the plot in the right side of Figure 1, where the conversion will build on experts’ past experience and empirical evidence.
3. The intelligence analyst uses “feature selection” techniques guided by the link discovery system to reduce the size of data and search for “explanatory variables” describing the reasons why the terrorist’s intensity-curve has sharp changes.
4. If there are other suspects in the same “group” for planning a specific terrorism activity (discovered by link analyses), their intensity curves will be tested against each other to evaluate the *probability* that these suspects are indeed in the same group. All information pertinent to this scenario will be collected and organized through the link discovery system. A risk assessment model will be called to evaluate the urgency of its potential impact on the nation’s homeland security.
5. If there is a potential high impact of this investigation case, more evidence will be collected from continuous monitoring (approved by legal departments) of suspects’ daily activities. This new evidence will be added to the intensity measures for predicting if a high-impact event will occur in the near future. Proper actions will be suggested to combat this terrorism.

Example 12 (Data Mining in Bioinformatics and Cancer Research): The National Institutes of Health (NIH) supported many data mining research projects in bioinformatics and computational biology fields. Dr. Lu’s group is extending their research experience in the above examples to these fields. For example, they are exploring patterns of association in gene expression, protein, and other genomic data linked to disease such as breast cancer. In most studies the genomic and disease data are transformed into typical statistical measurements for solving biological and medical research problems. Then,

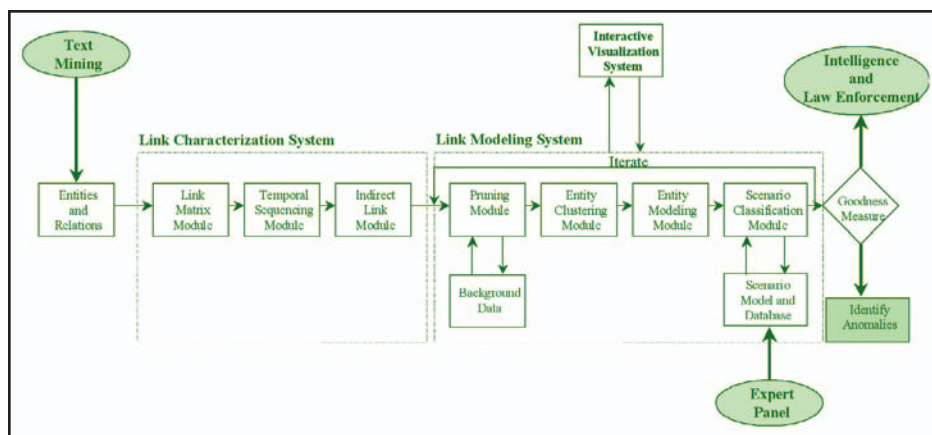



Figure 2. Link Discovery System

data mining tools can be used to look for patterns and trends that help understand and solve the problems. For instance, singular value decomposition and regression methods are used to study profiles of gene expression for understanding the uncertainties in classifying breast cancer tumor types (Spang *et al.*, 2002). Clustering analysis is used to identify regulatory binding sites where genes are transcribed into mRNA (Olman *et al.*, 2003). Since tremendous amounts of data in bioinformatics and cancer research are available online, system engineering research in these areas could be easier than many studies in the manufacturing fields where most of data are proprietary. However, the key challenge is to work with biologists and medical researchers to understand scientific problems, data insights, and significance of knowledge discovered.

Concluding Remarks

This article provides several examples showing that knowledge mining becomes more important in a wide range of applications. When data collection instruments become more advanced and people rely more on computers to discover “intelligence,” the opportunity for the knowledge mining field to grow is endless. Although many commercial data and text mining software are available, there is still much room for research and education activities to expand for advancing data mining techniques and for showing their potential in several emerging fields such as computation-based biological and medical studies. Hopefully, with the exposition of this article, more ISyE students, faculty members, and alumni will use knowledge mining tools in their studies, research investigations, and business activities. 

References

- Arias, E., Donoho, D. L., and Huo, X. (2003), “Asymptotically Optimal Detection of Geometric Objects by Fast Multiscale Methods.” Submitted manuscript. <http://www-stat.stanford.edu/~donoho/Reports/2003/MultiScaleDetect.pdf>.
- Besag, J, York J, and Mollie A. (1991), “Bayesian Image Restoration with Two Applications in Spatial Statistics,” *Annals of the Institute of Statistical Mathematics*, 43, 1-59.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data* (2nd edition). John Wiley: New York.
- Gabowski, H., Lossack and Weibkopf (2001), “Automatic Classification and Creation of Classification Systems Using Methodologies of Knowledge Discovery in Databases,” Chapter 5 (pp. 127-144) of *Data Mining for Design and Manufacturing: Methods and Applications* edited by D. Braha, Kluwer Academic Publishers: New York.
- Ganesan, R., Das, T. K., Sikdar, A., and Kumar, A. (2003), “Wavelet Based Detection of Delamination Defect in CMP Using Nonstationary Acoustic Emission Signal,” in review with *IEEE Transactions on Semiconductor Manufacturing*, 16(4), to appear.
- Huo, X., and Chen, J. (2004), “Building a Cascade Detector and Applications in Automatic Target Recognition,” *Applied Optics: Information Processing*, 43(2): 293-303.
- Huo, X., Chen, J., Wang, S., and Tsui, K. (2002), “Support Vector Trees: Simultaneously Realizing the Principle of Maximal Margin and Maximal Purity.” Research report can be obtained at <http://www.isye.gatech.edu/research/files/tsui-2002-01.pdf>.
- Ishino, Y., and Jin, Y. (2001), “Data Mining for Knowledge Acquisition in Engineering Design,” Chapter 6 (pp. 145-160) of *Data Mining for Design and Manufacturing: Methods and Applications*, edited by D. Braha, Kluwer Academic Publishers: New York.
- Jeong, M. K., Lu, J. C., Huo, X., Vidakovic, B., and Chen, D. (2004a), “Wavelet-based Data Reduction Techniques for Process Fault Detection,” accepted by *Technometrics*. This paper can be obtained at <http://www.isye.gatech.edu/apps/research-papers/>.
- Jeong, M. K., and Lu, J. C. (2004b), “Statistical Process Control Charts for Complicated Functional Data.” This paper can be obtained at <http://www.isye.gatech.edu/apps/research-papers/>.
- Jung, Uk, and Lu, J. C. (2004), “A Wavelet-based Random-effect Model for Multiple Sets of Complicated Functional Data.” This paper can be obtained at <http://www.isye.gatech.edu/apps/research-papers/>.
- Kang, L., and Albin, S. L. (2000), “On-Line Monitoring When the Process Yields a Linear Profile,” *Journal of Quality Technology*, 32, 418-426.
- Kim, H., Lu, J. C., and Kvam, P. (2004), “Product-order Decisions Considering Uncertainty in Logistics Operations.” This paper can be obtained at <http://www.isye.gatech.edu/apps/research-papers/>.
- Kongthon, A. (2004). *A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management*. Unpublished Ph.D. Thesis, The School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA.
- Lawless, J. F., Mackay, R. J., and Robinson, J. A. (1999), “Analysis of Variation Transmission in Manufacturing Process-Part,” *Journal of Quality Technology*, 31, 131-142.
- Lu, J. C. (2001), “Methodology of Mining Massive Data Set for Improving Manufacturing Quality/Efficiency,” Chapter 11 (pp. 255-288) of *Data Mining for Design and Manufacturing: Methods and Applications* edited by D. Kluwer Academic Publishers: New York.
- Olman, V., Xu, D., and Xu, Y. (2003), “CUBIC: Identification of Regulatory Biding Sites Through Data Clustering,” *Journal of Bioinformatics and Computational Biology*, 1(1), 21-40.
- Porter, A. L., Kongthon, A., and Lu, J. C. (2002), “Research Profiling – Improving the Literature Review: Illustrated for the Case of Data Mining of Large Datasets,” *Scientometrics*, 53(3), 351-370.
- Schwabacher, M., Ellman and Hirsh (2001), “Learning to Set Up Numerical Optimizations for Engineering Designs,” Chapter 4 (pp. 87-126) of *Data Mining for Design and Manufacturing: Methods and Applications* edited by D. Braha, Kluwer Academic Publishers: New York.
- Spang, R., Zuzan, H., West, M., Nevins, J., Blanchette, C., and Marks, J. R. (2002), “Prediction and Uncertainty in the Analysis of Gene Expression Profiles,” *Silico Biology*, 2, 33-44.

Text Mining:

The Engineer's Approach to Literature

Call it an engineer's approach to literature. That's how Alan Porter, professor emeritus in ISyE and the School of Public Policy, describes text mining. The method of counting words, instead of reading them, is finally catching on. Porter thinks that it is about time.

Porter, who also co-directs Georgia Tech's Technology Policy and Assessment Center (TPAC), jointly housed in the Public Policy and ISyE schools, is responsible for developing leading edge text-mining software, including VantagePoint. In 2003, VantagePoint, a technology-transfer success story developed between Georgia Tech and Atlanta-based Search Technology, Inc., generated its first royalties. While it was a modest amount — less than \$10,000 — royalties are expected to increase significantly with future versions of the software.

VantagePoint is a text-mining tool that allows technical-intelligence managers to quickly analyze search results from bibliographic databases and R&D literature. It produces summaries, charts, and graphs that help people spot patterns and relationships in massive amounts of data, enabling them to extract relevant information and make better decisions.

Competitive technical intelligence is the name of the game, says Porter. "Today it is critical to have the right technology at the right time," he explains. "Companies want to keep an eye on competitors so they don't drop the ball by introducing a new

product or technology too late. For example, General Motors looks to see what is published by and about Toyota — and more importantly, what it is patenting, because that shows what Toyota is really interested in."

VantagePoint grew out of an approach known as Technology Opportunities Analysis (TOA), which was developed by Porter and his colleagues at TPAC in the 1990s. Its roots go back to the "bibliometrics" (counting bibliographic activity) movement in the 1970s. "The gist of the approach is to exploit the amazing compilations of research and development information electronically available," says Porter. "In the past, such databases were used only to locate a few good references, track them down, and read a couple of pertinent papers. Text mining helps do that more effectively, but it also enables one to see the big picture. What's the overall pattern of R&D on a particular technology?"

Learning to Count

The Defense Advanced Research Projects Agency (DARPA) found TOA's text-mining research interesting enough to fund in 1994. Grants from the U.S. Army, the Office of Naval Research, the Department of Education, National Institutes of Health, and the National Science Foundation followed. The TOA project evolved into an ongoing industry-academic part-

An unusual partnership between Georgia Tech, industry, and government, VantagePoint's roots trace back to 1993. That summer, Porter and his son Doug, then a computer science student at Virginia Tech, were looking at ways to improve forecasting of technology trends. Later, Porter continued the project at TPAC.

The resulting software tool led to an invention disclosure and caught the interest of Search Technology, located in Norcross, Georgia. Search Technology licensed the technology from Georgia Tech in 1996, and the partnership began refining the text-mining software for the military. The commercial version was officially launched in 2000.

The commercialization of university research is typically an uphill battle. According to a University of Pennsylvania study, less than one percent of invention disclosures generate significant royalties for the respective universities.

Dr. Alan Porter's continued involvement has been another key factor in VantagePoint's success. Kevin Wozniak, former associate director of Georgia Tech's Office of Technology Licensing, says, "Faculty must make an effort to actively participate in the process. They can't just file an invention disclosure, then walk away and expect commercialization to happen".

Porter, who has been with ISyE since 1975, received his Ph.D. in Engineering Psychology from the University of California-Los Angeles in 1972. In the early 1990s, he also served as acting director of Georgia Tech's Management of Technology program. His research and teaching has long focused on technology forecasting and assessment and management of technology. In recent years, he has helped J.C. Lu in his efforts to spearhead "that other IE—information engineering."

In 2003, Search Technology introduced a new version of VantagePoint: Derwent Analytics, which is designed especially to use with the Derwent World Patent Index from Thomson Scientific, a subsidiary of information powerhouse, Thomson Corporation. Because of Thomson's vast distribution channels, this new release has the potential to reach an even larger number of people than the original software. "With that in mind, we're anticipating that VantagePoint will become one of the larger generators of royalties for Georgia Tech," Wozniak added.

nership with two commercial partners with Georgia Tech connections, Search Technology and IISC, signed on to commercially develop and market the software. Search Technology (www.searchtech.com), founded in 1980 by ISyE chair Bill Rouse, and now led by Paul Frey, led the software development. IISC (www.iisco.com), founded by Nils Newman, BME 1989, MSTASP 1993, led in the application of TOA on studies for commercial and government clients, as well as marketing.

"The text-mining approach is closely related to numerical data mining," says Newman. "It seeks to uncover relationships through a combination of analytical means. TOA relies most heavily on statistics to relate concepts, cluster associated terms, and ascertain trends in activity. It employs natural language processing to rip apart text sequences. It uses some 'AI-lite' (artificial intelligence) in the guise of rules for text recognition and fuzzy matching. All this is served up in an MS Windows setting to allow users to make lists with search and group capabilities, cross-compare two lists to see co-occurrence patterns, and to map technologies."

"As engineers, we apply TOA to understand technological change," Porter continues. "We create time series of activity patterns in research projects, publications, patents, and so forth to help forecast technological development. We capture linkages with contextual factors to create innovation indicators. Most of our applications focus on competitive technical intelligence and projection of emerging technologies' development paths. Benchmarking against other companies' development activities is a popular extension. Technology road mapping goes on to relate technological changes to organizational product development prioritization."

Porter says the software has followed an interesting migration in form, name, and capability. What is now known as VantagePoint was originally programmed in Pascal, migrated to C for UNIX operating systems, and then to C++ for Windows. The first "government use only" version of the text-mining software, called TOAS, was released in 1997. Today this is called TechOASIS, and it is used by the U.S. Army, U.S. Navy, and U.S. Air Force, along with other government organizations. After seven years of development, the commercial version was released in the spring of 2000. VantagePoint is being used at some 22 sites in a dozen countries, including large global firms, foreign government organizations, and consulting firms.

Information Overload

In today's world, the amount of information can be overwhelming. Because of this, few people have a handle on the whole picture of any given field, especially in the business world. Porter and his colleagues researched the challenges in getting technology managers and professionals to make use of empirical technology analyses. "Somewhat surprisingly, we found that technology managers are among the least quantitative," he says. "They tend to be considerably more intuitive than their production manager counterparts (think statistical quality control), financial managers (accountants count), or even marketing types (real-time tabulation of campaign effectiveness for fast adjustment)."

Studies by the National Science Foundation, the National Institutes of Health, and Georgia Tech took a look at how information is utilized in the decision making process. "Information really isn't used that much, which was one of the interesting things we found in the study," says Newman, one of Porter's former graduate students and the founder of IISC. "That's because of the inability to access information and the easy ability to be overwhelmed by that information. A manager who is responsible for making technology choices is more likely to resort to expert opinions to make a decision, rather than accessing readily available information."

"Let's say you're a collection of investors and you're trying to decide if you want to spend \$50 million building a new facility with new technology," Newman explains. "It's going to take you 10 years to build the plant. The question arises, is that technology even going to be competitive in five years? The way you handled that problem in the past is you went around and you talked to experts and said, 'Well, what do you think of this technology?' No one really sat down and looked at all the information that was available. So in the end they make this \$50 million decision off of three or four pieces of paper and the opinions of several dozen people. It really is just an opinion as opposed to a comprehensive assessment of information."

"There is a case with one of my clients," continues Newman, "Literally, they had a staff of 40 people out monitoring technology, and these people read reports about where technology was going. I asked them how this information was utilized. I was told that in the end, the head of R&D decided what technology to pursue based on what's in the technology section of the Sunday paper. Literally, that's how they do it. There are utilization issues when it comes to information. These are the things that ISyE is working on. How do you present information in a format that is palatable and understandable and usable by managers?"

A New Generation


Porter and Newman have taken "a lot of hard knocks over the years" about their efforts to push and improve text mining. "We thought if you present people with a lot of new information, they would be happy, but we realized later that people didn't actually utilize the information they already had available. It would be a hard sell." Both researchers believe that the newer and future generations of managers will not be as timid about taking advantage of technologies like text mining.

"We are finding that the younger cadre of technologists and technology managers are used to working in a richer information environment," says Newman. "One of the things that we determined over time is the information distance between the decision maker and the information. In the mid-1990s, you could find as many as five individuals between who needs the answer and the information source that contained the answer, because the difficulty of getting the information was such a barrier. You had to be a professional searcher if you wanted to go into a database."

He continues, "All that changed in the 1990s, with the World Wide Web becoming useful in a business environment, and then with improvements in information technology. Some are still unable to access information because they were never actu-

ally trained to. But you look at the cadre of engineers and managers who are coming out of school now, and they can work in an information-rich environment. They are much more savvy about it than a manager who was trained 10 years ago."

Newman compares this technological savvy to the skill of typing. "To point out how old I am, I actually took typing in high school," he says. "I was horrible at it; I could maybe scrape out 15 words a minute. Look at what I do today: I basically type for a living. At that time, the only people who could type were professional typists. So in a period of less than 10 years, typing went from a very specialized skill being used by few people to being essentially ubiquitous. They don't teach typing in school; they just pop the PCs in front of the kids."

"I see the same analogy with people's ability to utilize information," he adds. "Information was such an arcane art in the corporate world up until about 1995. I had customers in the late 1990s who were not allowed to use the Web! Another five or six years from now, and you're going to have a whole group of junior managers who are used to going out and looking at information first. You only have to look at the information skills of the current crop of Georgia Tech Management and ISyE students and compare them to students a decade ago to see that things have profoundly changed." 

For more information, visit the TPAC website at <http://tpac.gatech.edu>. Dr. Alan Porter may be reached at alan.porter@isye.gatech.edu, and Nils Newman may be reached at newman@searchtech.com.



ORMS TODAY

Subscribe to **ORMS Today**, your source for Operations Research and the Management Sciences.

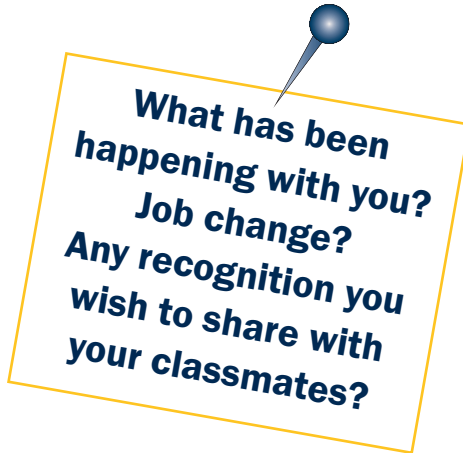
Visit us on the web: www.orms-today.com or call Maria Bennett: 770.431.0867, ext. 219 for more information

ALUMNI NEWS

Please take a minute to complete this form,
and mail or fax it to the school.

Please send to:

Engineering Enterprise
School of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive, Atlanta, GA 30332-0205
or fax to 404.894.2301



Name _____

Degree/Year _____

Home Address _____

City _____ State _____ Zip _____

Home Phone (____) _____

Title/Company Name _____

Business Address _____

City _____ State _____ Zip _____

Business Phone (____) _____

E-mail Address _____

Your news _____

Other IE topics you would like to read about in *Engineering Enterprise* _____

