Newsletter

of the INFORMS Computing Society

Volume 22, Number 1 Spring 2001

Contents

Data Mining Emerges as A New Discipline in a World of Increasingly Massive Data Sets

By James Case

Reprinted, with permission, from SIAM News,
Volume 32, Number 10

On May 12, the day shared by the Sixth SIAM Conference on Optimization and the 1999 SIAM Annual Meeting in Atlanta, Olvi Managasarian of the University of Wisconsin gave a joint invited talk, "Optimization in Machine Learning and Data Mining."

Unlike machine learning, which has occupied the AI community for years, data mining is a discipline of relatively recent origin. The term refers to a collection of techniques for extracting useful information from large data sets –too large, in many cases, even to be loaded into main memory. In a world where even the simplest daily task—like making a phone call, using a credit card, or buying hardware and groceries—leaves an electronic footprint, an increasing number of increasingly important data sets fit that description.

Automated data collection devices, capable of generating terabytes (a.k.a. terrorbytes) and even petabytes of information at rates measured in gigabytes per hour, are

Data Mining: continued on page 4

ICS Officers

Newsletter Co-Editors

S. Raghavan (suraghav@rhsmith.umd.edu)
Decision & Information Technologies Dept.
Robert H. Smith School of Business
University of Maryland
College Park, MD 20742-1815
(301) 405-6139

Tom Wiggen (wiggen@acm.org)
Dept. of Computer Science
University of North Dakota
Grand Forks, ND 58202-9015
(701) 777-3477

The ICS Newsletter is published semiannually by the INFORMS Computing Society (ICS). Manuscripts, news articles, advertisements and correspondence should be addressed to the Editors. Manuscripts submitted for publication will be reviewed and should be received by the Editor three months prior to the publication date. Requests for ICS membership information, orders for back issues of this Newsletter, and address changes should be addressed to:

INFORMS Computing Society Business Office
901 Elkridge Landing Road, Suite 400
Linthicum, MD 21090-2909
phone: (410) 850-0300

Message from the Editors

We hope you'll enjoy this new issue of the ICS newsletter. Once again, we apologize for not getting this issue out earlier. As you can see, it's titled the spring 2001 issue, so we'd better hope that old news is still good news for you.

Data Mining is emerging as one of many fruitful areas of research, related to marketing and e-commerce, in the OR/CS community. Our feature article is on this topic, and is reprinted, with kind permission, from the SIAM News of December 1999.

Another article (starting on page 14) was contributed by Karla Yale (an ICS member) and her staff at Yale Systems, Inc., and describes their speech recognition engine and its application in a peoplebot robot to implement a physical assistant for stroke victims. For those who are not familiar with speech recognition processes, the article contains a number of details about the design and characteristics of Yale System's speech recognition engine.

Our featured member is Sebastian Ceria, President and CEO of Axioma. He has a very interesting background, both in academia and now in industry, and we hope you will enjoy reading about him.

Congratulations to Harvey Greenberg and his colleagues at the U of Colorado-Denver for formally launching their center for computational biology (see page 19), and to John Hooker on the publication of his book (see page 12). If you are reading this newsletter, we invite you to describe your research work in the newsletter. (This is free advertisement of your research!!). If we could get two ICS members to write these descriptions for each issue, we'd have material for the next 150 years! [If you do a quick calculation in your head, you'll infer that we have about 600 ICS members.]

Remember (if you've missed any past issues and are dying to get a hold of them) that back issues of this newsletter are online. You can reach them through the ICS web page at http://computing.society.informs.org/. Even more important, however, is the fact that all Journal of Computing issues are now online. Thanks to JOC editor David Kelton for his work in accomplishing this goal. His announcement is cleverly hidden in this newsletter.

Finally, we've been enjoying our role and stint as editors of the newsletter. However, we feel that it is now time for us to make way for a new editorial staff. We will be discussing with the ICS officers a timeframe over which to step down and transition to a new editorial team. If you have any interest in getting involved with the newsletter, do contact them.

Tom and Raghu

Data Mining: continued from page 1

rendering existing inference methods obsolete. The biggest data warehouse in the world—the Wal-Mart system, built on an NCR platform called Teradata—reportedly contained 11 terabytes (11×10^{12} bytes) of information as of 1996 and has doubtless continued to expand. The satellites of NASA's Earth Observing System are capable of generating more than a terabyte of data per day. In such a world, even the simplest browsing operation can result in an avalanche of useless and irrelevant data.

Imagine a program for deciding whether two rows of data differ in more than "a few fields." While such "find similar" problems appear simple and can be approached in various ways, executing any one of them on a massive data store is by no means trivial. More complicated questions could require the analysis of millions of data points residing in a space of a thousand dimensions. Who can do that?

From Raw Data to "Documented Knowledge"

Among the more predictable results of the emerging discipline has been a new wave of multiletter acronyms. Perhaps the best known are OLAP, for On-Line Analytical Processing, and KDD, for Knowledge Discovery in Databases. As used at Microsoft, the former term refers to a database capable of responding to queries more complex than those handled by the standard "relational" databases of the 1970s and 1980s, while the latter refers to a newer and even more versatile generation of software. OLAP has been prolific in its output of related acronyms, including MOLAP (multidimensional OLAP), ROLAP (relational OLAP), HOLLAP (hybrid OLAP), and most recently DOLAP (desktop OLAP).

OLAP systems rely heavily on precomputed aggregates—obtained after a single pass through the data, by summing or averaging over particular indices and groups of indices—as well as projections onto lower-dimensional subspaces of the high-dimensional space in which raw data so often reside. Because the number of potential aggregates increases exponentially with the number of dimensions, much of the work in OLAP systems involves deciding which aggregates to precompute, and how to derive (or estimate) additional aggregates from those that have been precomputed.

OLAP exploration is guided by user-supplied instructions regarding the histograms to be created, the variables to be plotted against one another, and the level of detail employed. Inference and modeling are left to the user, who is expected to recognize patterns of interest via visualization in lower-dimensional subspaces, and to formulate testable hypotheses concerning the reduced data sets furnished by the system. KDD, in contrast, combines methods from database theory, statistics, pattern recognition, AI, high-performance computing, and the like, in an effort to discover patterns hidden within the data and model the behavior that produced them.

A pattern might be a simple data summary, a data segmentation, or a model of dependencies (a.k.a. links) within the data. KDD, intended to lead all the way from raw data to "documented knowledge," proceeds in several steps. Because the raw data generated by industrial processes like manufacturing, telephone switching, and

customer billing are often recorded at remote locations in arcane formats, KDD cannot even begin until all have been assembled, cleaned up, and organized into what is called a "data warehouse." From that, a subset of the data that is relevant to a given project must be extracted—rather as a student might borrow relevant books from a library before writing a term paper—followed by the formation of natural aggregates between which causal relationships can be expected to exist.

The mere construction of a data warehouse is often enlightening. One firm, for instance, reportedly discovered that its records contained 27 separate and distinct spellings of the name K-Mart. Others have uncovered missing fields, conflicting reports, and all manner of other debilitating flaws.

Even if each step of the KDD process "runs" successfully, the results will not necessarily be informative. A group at Silicon Graphics was amused to learn from an early version of SGI's own data mining tool MineSet—that "with 99.7% certainty, all individuals who are husbands are also males." While obvious to humans, such conclusions are every bit as intriguing to a computer as patterns that flag double billing by physicians in Australia, an incipient epidemic in Maine, or a "softening" of U.S. demand for Sport Utility Vehicles.

A famous result of first-general KDD analysis was the observation that the most frequent late-night purchases in supermarkets are diapers and beer. As a result, strategically located aisles are frequently blocked after 10 PM to route diaper purchasers past the beer cooler on their way to checkout. A new generation of data mining tools now emerging, although not yet simple enough to accommodate corporate end-users, are able to detect patterns of the foregoing sort while being significantly easier to use than the options previously available as add-ons to traditional statistical analysis packages.

Applications

The most promising applications of KDD are in fields requiring high-payoff knowledge-based decisions in rapidly changing and information-rich environments. Perhaps the most persuasive business application is "database (i.e.,

DIMACS Summer School Tutorial on New Frontiers in Data Mining

DIMACS Center, Rutgers University, Piscataway NJ, 08854-8018
August 13 - 17, 2001
http://dimacs.rutgers.edu/Workshops/MiningTutorial/

Organizers: Dimitrios Gunopulos, Univeristy of California at Riverside (dg@cs.ucr.edu), Nikolaos Koudas, AT&T Labs - Research

Local Arrangements: Jessica Herold, DIMACS Center, jessicah@dimacs.rutgers.edu, 732-445-5928

Short Description:This "summer school" tutorial program is aimed at providing background, vocabulary, and theoretical methodology to non-specialists in data mining and to others who wish to explore this field and at bringing together students, postdocs, and researchers working on algorithms for data mining with those working in various applications areas. More specifically, we aim to introduce the attendees to the fundamental theoretical/algorithmic issues that arise in data mining and its applications.

mail-order) marketing," which relies on customer information to tailor offers to particular segments of the market. Affluent suburban customers with expensive tastes, for example, will receive the "sneak preview issue" of ABC Outfitters' autumn catalogue some weeks before those with less promising credentials and/or purchase records receive the ordinary issue, and months before proven misers receive the "remainders" issue. Needless to say, the prices decline from issue to issue.

Because scientific users typically know their data in intimate detail, it may be easier to develop KDD applications in science than in retailing, finance, or other areas of commerce. A case in point is the 2nd Palomar Observatory Sky Survey, which took more than six years to complete and gathered more than 3 terabytes of image data, in which an estimated 2 billion sky objects are "visible." The 3000 photographic images were scanned into digital format, with 23,040 x 23, 040 pixels per image and a resolution of 16 bits per pixel.

An automatic method was needed for identifying and then classifying (as star, galaxy, quasar, black-hole accretion disc, and so forth) the many objects in each digital image. The majority of objects are faint, visible to the naked eye but impossible to categorize by visual means. To meet this need, a team from the Jet Propulsion Laboratory developed the Sky Image Cataloging and Analysis Tool (SKI-CAT). Once the basic image segmentation was complete, 40 attributes deemed important by astronomers were measured for each object identified. SKI-CAT then discarded 32 of the 40 measurements and devised a classification scheme exploiting the remaining eight.

To test SKI-CAT's accuracy, a small sample of the classified objects were reclassified by a far more expensive method. The two classifications agreed on 94% of the sample. SKI-CAT subsequently helped astronomers to discover—in record time—16 new high-red-shift quasars. Such objects, among the most distant (and therefore oldest) in the universe, are extremely difficult to find. But they provide rare and valuable clues about the early history of the universe.

KDD methods have also been used to count and locate the mountains on Venus (from the Side Aperture Radar data obtained by the Magellan spacecraft during its five years in orbit around that most Earth-like of planets); to identify individual genes in the human genome; to detect and measure tectonic activity in the Earth's crust (from satellite data); and to estimate the duration and strength of tropical cyclones (from massive amounts of electronically gathered atmospheric data). As more sensitive data collection methods become available along these and other avenues of inquiry, ever more sophisticated methods of analysis will be called for.

Computational Considerations

The focus in Mangasarian's Atlanta talk was on computational methods. In his first two papers on pattern recognition (1965 and 1968), he told the audience, he had proposed the repeated use of a linear programming routine, called as a subroutine by a main program designed for a more specific purpose. It was a novel suggestion at the time. Underwhelmed by the response, he shifted his focus to other matters. Another twenty years were to pass before a chance dinner-party conversation led him to a genuine "killer ap" for his long-dormant work on pattern recognition. His description of both the application and the method appeared in the September 1990 issue of

SIAM News.

The original problem Mangasarian considered was the separation of two finite and disjoint sets of points in R^n – think of them as *x*'s and *o*'s drawn on a blackboard of arbitrary (but again finite) dimension—by means of piecewise-linear discriminant functions. If the convex hulls of the two sets are disjoint, the problem is trivial: A single hyperplane separates the convex hull of the *x*'s from that of the *o*'s, and (signed) distance from that hyperplane serves as a (linear) discriminant function. Moreover, solution of a single linear program suffices to determine the separating hyperplane.

If the convex hulls intersect, however, no single hyperplane suffices. In that case, Mangasarian demonstrated (in a 1965 paper) the utility of a method that finds the thinnest closed hyperslab H_1 (a subset of R^n bounded by parallel hyperplanes) for which (a) the intersection of the two convex hulls is contained in H_1 and (b) each of the open halfspaces complementary to H_1 contains only *x*'s or only *o*'s. H_1 can also be determined via the solution of a single linear program. The *x*'s and *o*'s exterior to H_1 are then regarded as "already separated," and a hyperplane separating only the remaining *x*'s and *o*'s is needed. If one is found, the process stops with all the *x*'s separated from all the *o*'s by a piecewise-linear discriminant function; otherwise, the process continues.

Eventually, the process terminates with the *x*'s and *o*'s separated by a piecewise-linear discriminant function. (The reader may find it instructive to construct the required function for three *x*'s and three *o*'s situated at alternate vertices of a regular hexagon.) In rare degenerate cases, the exterior of the most recently computed hyperslab contains neither *x*'s nor *o*'s, necessitating the insertion of an anti-degeneracy step between successive linear programs.

Even in the presence of degeneracy, the process can always be carried to completion, given only that no point of R^n is both an *x* and an *o*. It then makes sense to infer that any uncategorized point for which the discriminant function assumes a positive value is another *x*, while those for which the discriminant function assumes negative values are *o*'s. If consideration of additional *x*'s and *o*'s shows that assumption to be erroneous, the larger sets of *x*'s and *o*'s can be used to "retrain" the classification scheme. More recently, it has been demon-strated that Mangasarian's "hyperslab method" constitutes an effective method for training neural networks with partially preassigned weights.

Beyond the Killer Ap

The killer ap brought to Mangasarian's attention at the fateful dinner party was breast cancer diagnosis. Other applications have continued to surface, and he alluded—albeit telegraphically—to a number of them in his Atlanta talk. A particularly interesting one concerned the authorship of the disputed Federalist papers. The original method has been generalized in numerous directions, including separation into more than two sets, separation by nonlinear hypersurfaces, and the classification of many as a million points in . As formulated by Mangasarian, all such classification problems can be reduced to the minimization of concave (piecewise-linear) objective functions subject to (numerous) linear constraints.

Such problems, relatively routine in spaces of low dimension, become unwieldy when the

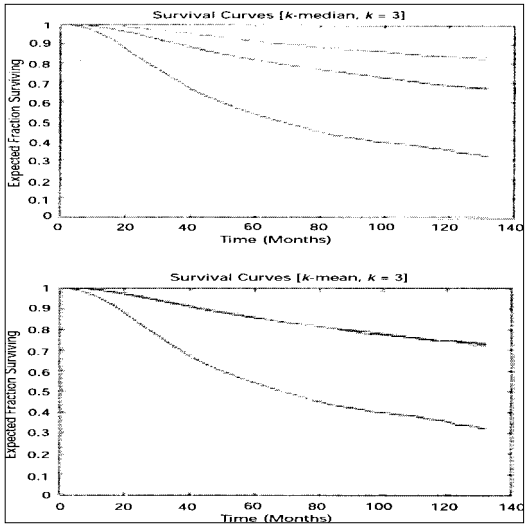


Figure 1: Visual evidence that k-medians (top) can be more useful than k-means (bottom) for practical purposes. The rival methods were used to separate 21,960 data points from a National Cancer Institute database into three clusters; survival frequencies for each cluster were then plotted against time.

constraint matrix exceeds main memory capacity. To avoid that difficulty, he and his co-workers have developed a "chunking" method for solving such problems subject to a small, randomly chosen subset of the given constraints. The constraints found to be inactive are discarded, and the problem is solved again subject to the union of all currently active constraints with another small random subset of the given ones; these steps are repeated until convergence. Convergence was achieved after 25.9 hours for a problem with 5×10^5 constraints, and after 231.3 hours for another involving 10^6 constraints.

Two other important problems in data mining, the so-called *k*-mean and *k*-median clustering problems, constitute alternative ways of identifying the most densely populated subregion of the region in which given data points reside, along with the second most densely populated subregion, ….., up to and including the *k*th most densely populated subregion. The *k*-median problem, Mangasarian pointed out, can also be solved by minimization of a piecewise-linear concave objective function subject to linear constraints.

He produced visual evidence that *k*-medians are sometimes more useful than *k*-means for practical purposes: Using the rival methods, he separated 21,960 data points in a National Cancer Institute database into three clusters and plotted survival frequencies for each cluster against time (Figure 1). The survival characteristics of the three groups identified by the *k*-median method ($k = 3$) differ markedly, while those of two of the three groups identified by the *k*-means algorithm do not. Prospects for the high-risk groups identified by the two methods, on the other hand, are virtually identical.

It was evident to all in the audience that much has been accomplished in this important emerging field, that much remains to be done, and that Mangasarian, his students, and his various other co-workers are leaders in the field.

James Case is an independent consultant who lives in Baltimore, Maryland

Call for Papers

Special Issue of Annals of Operations Research on METAHEURISTICS - Theory, Applications and Software

Metaheuristics provide decision-making managers with robust tools that obtain high-quality solutions, in reasonable time horizon, to important applications in business, engineering, economics and science. In recent years, there have been significant advances in the theory and the application of metaheuristics to the approximate solution of hard optimization problems. Metaheuristics are iterative master processes that guide and modify the operations of subordinate heuristics to efficiently produce high-quality solutions. They may manipulate a complete (or incomplete) single solution or a collection of solutions. The subordinate heuristics may be high or low level procedures including, e.g., local search approaches. The family of metaheuristics includes, but is not limited to, adaptive memory procedures, tabu search, ant systems, evolutionary methods, scatter search and their hybrids. Annals of Operations Research is one of the most renowned journals in the field. A special issue on metaheuristics provides an excellent opportunity to advance the knowledge of this interesting and evolving field of research.

AREAS OF INTEREST: For the special issue, we expect original research papers of high quality providing theoretical and/or computational results or showing the application of metaheuristics to real-world problems. We invite submissions of high-quality papers that focus on:

- general theoretical aspects of metaheuristics (e.g. analysis, convergence)
- successful practical applications (e.g. in finance, supply chain management, transport)
- software (e.g. general components, integration into ERP systems)
- hybrid approaches (e.g. with constraint programming, branch and bound)
- new approaches

REVIEWING: The submitted papers will be peer-reviewed in the same manner as any other submission to a leading international journal. The major acceptance criterion for a submission is the quality and originality of the contribution.

SUBMISSION: The deadline for submission is October 31, 2001. Manuscripts must be written in English and should be submitted in electronic form in a platform-independent format such as postscript or pdf. Please send your submissions to metaheuristics@tu-bs.de. Detailed instructions for authors can be found on the Annals homepage http://www.baltzer.nl/anor/anor_authinstr.asp. 25 reprints of each article published in this journal will be supplied free of charge. Authors who make use of the Baltzer Style File for their article will be entitled to 50 reprints free of charge. The Baltzer Style File is available on the Baltzer WWW homepage. The URL is http://www.baltzer.nl/auth.inst.html.

GUEST EDITORS

Prof. Dr. Ulrich Derigs
Universtaet zu Koeln,
seminar fuer Wirtschaftsinformatik und Operations Research,
Pohligstraße 1,
D-50969 Koeln, Germany
phone: ++49-221/4705327, fax: ++49-221/4705329
E-mail: derigs@informatik.uni-koeln.de

Prof. Dr. Stefan Voss
Technische Universitaet Braunschweig,
Institut für Wirtschaftswissenschaften / Informationsmanagement,
Abt-Jerusalem-Straße 7,
D-38106 Braunschweig, Germany
phone: ++49-531/3913210, fax: ++49-531/3918144
E-mail: stefan.voss@tu-bs.de

Call for Nominations

2000 INFORMS COMPUTING SOCIETY PRIZE

Nomination Deadline: August 15, 2001

Nominations are invited for the INFORMS Computing Society (ICS) Prize for the best English-language paper or book on the Operations Research/Computer Science interface. The objectives of the prize are to:

- Promote the development of high-quality work advancing the state of the art in the operations research/computer science interface,

- Publicize and reward the contributions of those authors/researchers who have advanced the state of the art, and

- Increase the visibility of excellent work in the field.

To be eligible, a nominated work must be:

- Published in the open literature,

- Pertinent to the operations research/computer science interface, and

- Written in English.

The prize committee consists of David Shanno (Chair, Rutgers), Julie Higle (Arizona) and Benjamin Melamed (Rutgers).This prize will be awarded November 5, 2001, at the INFORMS Fall Meeting in Miami Beach, Florida. The award is accompanied by a certificate and a $1,000 honorarium.

Nominations must include: the title, author's name, place and date of publication, and a copy of each nominated work (in quadruplicate, please, if it is not both easy and legal to photocopy). If you wish the nomination materials to be returned after the review process, so indicate.

Nominations must be received by August 15, 2001, and should be sent to the following address, with a cover letter justifying the nomination:

```
Professor David Shanno
RUTCOR, Rutgers University
640 Bartholomew Road
Piscataway, NJ 00854-8003
732-445-4858
shanno@rutcor.rutgers.edu
```

For further information regarding the INFORMS Computing Society Prize, see the ICS home page at: http://computing.society.informs.org/

Report on the INFORMS Journal on Computing

The year-end statistics for 2000 are as follows: there were 87 new papers submitted, we accepted 30 papers, there were 44 papers rejected or withdrawn, and at the end of the year there were 85 papers in process. For the first quarter of 2001, there were 33 new papers submitted, 5 accepted, 13 rejected, and we now have 94 papers in process.

The 2001 volume (Volume 13) of *JOC* began a new page layout and look, intended to improve readability, streamline both the manuscript-preparation process for our authors and the production process for our printer, and to bring *JOC* into conformance with the emerging "standard" style already in use by several other INFORMS journals. Authors are encouraged to visit the Instructions for Authors section at the *JOC* website (http://joc.pubs.informs.org/) for specifics and downloadable templates to aid in manuscript preparation.

We have two Special Issues of *JOC* in the works, and for which complete Calls for Papers are on the *JOC* web site. "Mining Web-based Data for e-Business Applications" is being put together by Associate Editors Louiqa Raschid and Alex Tuzhilin. "The Merging of Mathematical Programming and Constraint Programming" is being developed by Area Editor John Chinneck. Both of these Special Issues are on timely topics of intense interest, and I encourage you to visit the web site to learn more.

Another addition to our web site is something that I hope will be of broad use to the community, and perhaps save a lot of trips to the library and a lot of seeking correct change for dim, balky copiers. Since early this past summer we've been working hard at scanning the entire history of *JOC* prior to 1998 to electronic format — complete, full-text papers, along with all introductory and index material. The project is now complete. This material is freely available to everyone in the Back Issues section of the *JOC* web site, which is organized by volume/year, then issue, then individual papers in the order in which they originally appeared in the print journal. This rather massive project was managed and carried out by Assistant Editor Susan Norman with great skill, not to mention considerable patience. With the subscription-based service from INFORMS PubsOnLine (http://pubsonline.informs.org/) for volumes from 1998 up to the present, everything ever published in *JOC* is now available online in one way or another. We are now in the process of compiling a subject-based index to the back issues of *JOC*, and will put it on our web site when it is completed.

There have been two changes in Area Editorships. After more than seven years of fine service, Anant Balakrishnan of Penn State has decided to step out of his very active role as Area Editor, first for Telecommunications, and more recently as Area Co-Editor for Telecommunications and Electronic Commerce. We thank Anant for his diligent work in shaping this critical Area of the journal. Anant is being replaced by Prakash Mirchandani of the University of Pittsburgh, who has been a *JOC* Associate Editor for several years, a role from which he now steps down. We welcome Prakash and thank him for his willingness to take on this important service to the Telecommunications community. And after many years of excellent service, Mike Taaffe has decided to step aside as Simulation Area Editor, and I thank him sincerely for his selfless service to *JOC* and INFORMS in this role. The new Simulation Area Editor is Susan Sanchez of the Naval Postgraduate School. We welcome Susan with thanks for agreeing to take this on.

There are also some changes among the Associate Editors. At this time I'd like to thank Arie Segev and Sudha Ram as they have decided to rotate out. And we welcome new Associate Editors Alexander Bockmayr, Rudolf Müller, Alexander Tuzhilin, Rema Padman, Marvin Nakayama, and Vijay Mookerjee.

Finally, we welcome several new Institutional Sponsors: OptTek Systems, Inc. of Boulder, Colorado, LINDO Systems, Inc. of Chicago, Haverly Systems, Inc. of Denville, New Jersey, and GAMS Development Corporation of Washington, DC. We thank them for their support.

Respectfully submitted by W. David Kelton, Editor-in-Chief, April 16, 2000

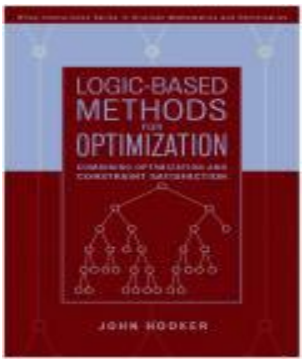
Book Announcement

Logic-Based Methods for Optimization

John Hooker

Wiley, 2000, ISBN 0-471-38521-2, 495 pp.

There has been much recent interest in combining optimization and constraint programming. This book uses logic as a key to unifying the two in a natural and systematic way. It then goes a step further and shows how extending the role of logic in optimization can lead to more flexible modeling and more effective solution techniques.



It presents methods in the first 15 chapters that are ready for implementation or have already been proved in practice, plus others in the last 6 chapters that are still in development. The book provides tutorials in constraint satisfaction, constraint programming, and logical inference. Specific topics include a modeling framework for hybrid methods, the logic of 0-1 inequalities, cardinality clauses, classical boolean methods, domain reduction for and continuous relaxation of global constraints, logic-based Benders decomposition and outer approximation, branching heuristics, relaxation duality, inference duality and logic-based sensitivity analysis, generalizations of partial-order dynamic backtracking, nonserial dynamic programming, and discrete relaxations. The book should be accessible to practitioners and academics in operations research or artificial intelligence.

INFORMS Computing Society: 2000 Year End Financial Statement Summary

	Audited 2000	Audited 1999	Audited 1998	Audited 1997
REVENUES	**$3,229.52**	**$3,502.09**	**$5,272.18**	**$4,306.39**
dues (individuals)	$2,619.00	$2,903.00	$1,434.00	$3,480.00
dues/contributions (corporate)			-$500.00	
meetings receipts	-$11.04		$3,385.85	
interest	$621.56	$599.09	$677.33	$826.39
other			$275.00	
EXPENSES	**$5,043.37**	**$4,664.78**	**$8,395.19**	**$3,610.94**
newsletter - HQ labor	$73.50	$128.69	$13.68	
newsletter - production			$136.06	
newsletter - printing	$1,287.99	$2,014.35	$3,377.42	$960.92
newsletter - postage/mailing	$723.31	$1,029.93	$2,701.67	$349.95
balloting	$57.25		$192.94	$757.48
national meeting (receptions)	$1,142.15	$351.03	$670.51	$747.24
award	$1,000.00	$1,116.08	$1,302.91	$594.73
supplies/copying/postage	$259.17	$24.70		$200.62
contributions to journals	$500.00			
Ending Balance	**$12,050.56**	**$13,864.41**	**$15,027.10**	**$18,150.11**

Message from the Chair

Irv Lustig

I have just returned from the first INFORMS Practice Meeting, Optimizing the Extended Enterprise in the New Economy, held in San Diego, California, May 20-22, 2001. This bold new experiment by INFORMS was deemed to be a success by all who attended (except for one academic who showed up at the registration desk expecting the usual INFORMS meeting format and fees!). The INFORMS Board has approved doing the Practice meeting for two more years and it is anticipated that attendance at the meeting will increase next year.

All speakers at the meeting were invited by members of the meeting Advisory Council, of which I was a member. In addition, the Edelman award competition was held in conjunction with the meeting. The sole focus of the meeting on the practice of Operations Research combined with the multiple networking opportunities was appreciated by many. It is clear that the connections with computer science and computing are a key contribution to the success of operations research practice. Almost nobody is successful without doing some form of computation.

As people interested in the connections between Operations Research and Computer Science, we have an opportunity to continue to contribute to the growth of good applications of Operations Research techniques in practice. Our sponsorship of the INFORMS Journal on Computing, sessions at the Fall INFORMS meeting, and our regularly scheduled conferences all provide an opportunity to foster this growth. Our society aims to brings together those people interested in this connection.

We are now planning the next INFORMS Computing Society conference. The meeting will be cochaired by Nong Ye of Arizona State University and Hemant Bhargava of Penn State University. The tentative dates and location are January 8-10, 2003 in Tempe, Arizona. Nong and Hemant are both starting to work on various aspects of the meeting. Please contact them at nongye@asu.edu and bhargava@computer.org if you are interested in helping with the organization of the meeting.

Our next ICS Business Meeting will be at INFORMS Miami in November. Dave Woodruff has put together 11 sponsored sessions plus 1 tutorial for this meeting. At this meeting, we will be electing new officers and awarding the ICS Prize. I hope to see you there.

A Speaker Recognition Engine and its Applications

Jihiui Lu, jlu@yale-systems.com
Yale Systems, Inc.

Editor's Note: This article in interview format, contributed by Karla Yale and Jihiui Lu of Yale Systems Inc., describes features of Yale System's independent speaker recognition engine, VID8.8™ and some of the applications intended for it, including their Physical Assistant Robot (PAR™) application.

Q. What is a speech recognition system? What does it do?

An independent speech recognition system converts voice wave files to the word or words being spoken regardless of gender, age, dialect, accent, or language. This is what VID8.8™ does.

Q. Where does a speech recognition system's input come from? What are the possibilities?

Input is from a microphone or set of microphones, depending on the application. In the deployed system, the input is the verbal string derived from the wave file obtained from the microphone set.

Q. How does a speech recognition system produce outputs? What form do these outputs usually take?

The output is the identification of what the verbal string is. The identification is what causes the commanded action to occur.

Q. Can speech recognition systems identify long and short "sounds" equally well? How can they distinguish words, sentences and paragraphs if the input is an audio stream?

Yes, long words and one syllable words are equally identifiable. The input wave file is parsed first and there is actually a moment of silence between syllables that is recognized.

Q. What speech recognition applications is Yale Systems developing?

The biggest uses are in telecommunications systems. Order entry is another big application. The objectives of a proposed project [a Physical Assistant Robot or PAR™] are to enhance the vocabulary set available from the VID8.8™ speech recognition engine to include the words, logical or unique, for enhancing the physical life of stroke victims and their care givers. The interfaces of these commanded words will be put onboard the PeopleBot robot (*Editor's note: According to* ***ActivMedia Robotics,*** *PeopleBot™ is the first affordable intelligent mobile robot designed especially for human interaction. Its specifications can be seen at http://www.activrobots.com/ROBOTS/index.html*). There will be immediate applications of this work to improve acoustic-phonetic performance for various tasks in Automatic Speech Recognition (ASR).

The benefits of the collaboration with ActivMedia and the use of the PeopleBot mobile robot are listed in this abbreviated scenario showing the number of stroke victim activities supported by the PeopleBot mobile robot. The detailed scenario is

available upon request.

TOP LEVEL PAR™ CAPABILITIES	Stroke Victim Activities	PeopleBot
Ambulatory	6	X
Eating	1	X
Drinking	1	X
Taking medicine	1	X
Answering the door	1	X
Answering the phone	1	X
Accidents	6	X

Several strategies will be integrated to achieve this objective including the time sequencing of the phoneme phase and out of vocabulary elimination.

Q. What hazards make speech recognition difficult to implement? How are you dealing with those hazards?

Let us take the example of a Physical Assistant Robot, PAR™. Since PAR™ responds by some voice activated prompting, what if the patient is aphasic? For the purposes of this development, there are 2 kinds of speech disturbance. One is the type where the patient always says the same wrong word for a given word, like always saying purple for people. The other is when the patient either says the same nonsense for everything or never says the same wrong word. PAR™ would be excellent at helping with the first case, even if instead of one wrong word, there is a set of wrong words. In the latter case, PAR™ cannot help.

The preliminary scenario is developed for PAR™. The requirements specifications are finalized for a limited vocabulary development utilizing the VID8.8™ engine.

Factors such as ambient or background noise, speaker consistency, and speaker compliance with the training routine will affect the results. For PAR™, the TV might be on, a radio might be on, music might be playing, the phone might ring, beepers could go off, a fish tank might be burbling, the speaker might have a cold. Therefore, the minimum number of spoken words for training can be large enough to average out transient or ambient noise.

Classifications of word sets will allow for larger vocabulary development. Commanded action based on the word sets will be unique to the implementation scenario.

Q. Describe the technical features and capabilities of Yale System's speech recognition engine.

Most vocabularies require about 720 samples that are in sound wave files for training the VID8.8™ engine. The samples have to contain all of the words in the command vocabulary. The recordings must be made in the environment in which the product will be deployed. For example, if a medical application is going to always be speaking in a "quiet room" into a microphone, that is how the 720 samples need to be recorded.

The VID8.8™ speech recognition engine is composed of 2 parts:

- Parser
- Recognizer

The Parser takes any sentence or verbal string which is represented by sound wave files. It has a table that drives the noise level setting for a particular application. For

example, ambient noise is different in a car versus on a telephone. The Parser separates the verbal string into syllables. Each syllable is put into a separate file.

The separate syllable files are then read by the Recognizer. The Recognizer has 2 parts:

- Builder
- Identifier

The Builder reads the syllable file and converts the phoneme peaks and word envelop geometrics into 67 features. These features are transformed and normalized into a representation acceptable to the Identifier.

The Identifier is a set of heuristics which call neural networks with the subset of the 67 features that are required for a particular syllable. The word is formed and the verbal string or sentence is formed, which causes the system to act.

VID8.8™ is gender, accent, and language neutral. VID8.8™ is speaker independent. This can be changed to speaker dependent via weighting of the training samples.

Underlying the VID8.8™ metrics are signal processing based mathematics. The metrics include the use of chaos theory, fractals, and multidimensional integral calculus.

Q. How well does your speech recognition engine work?

Preliminary data shows that VID8.8™ gives above 95% recognition for a sample of 6 out of 11 one and two syllable words, based on 80 untrained speakers, 40 of whom are male and 40 are female. All USA dialects are represented. The entire vocabulary set needs to be above 95% level of recognition. The PAR™ application is easier because the commands will only come from the patient. This means that the independent speech capability is reduced to dependent speech. However the OOV, out of vocabulary faults, must be tested.

Q. What does the future hold for the Physical Assistant Robot application? Who will use it? What will it be able to do to enhance life?

With respect to marketing issues, PAR™ will be available on a per unit basis either as a buy or lease. Institutions all over the world might want to buy several units for training purposes. Families will want to lease.

With respect to broad societal issues, we hope that PAR™ will enhance extended life spans while minimizing stress felt by the family, particularly the caregiver. Development of this robot will allow more stroke victims to live independently and reduce public funded expenses for nursing home care. A good nursing home costs $75,000 annually per stroke patient. Use of PAR™s could help to avoid these costs for many stroke victims.

An improvement to human health is the result we seek. Approximately 200,000 people die from stroke every year and 2 million survivors continue to suffer its aftermath. One out of 10 families is touched by stroke. PAR™ will provide a better quality of life for stroke patients and their caregivers.

For the stroke victims, PAR™ will allow them to remain independent for a longer time period than otherwise. For the family and caregivers, it will help towards freeing them from the stress of constant care and allow them to pursue their own lives.

There are 600,000 stroke victims annually in the US and many more overseas. Thus, the size of the potential market for PAR™ is substantial. PAR™ may also have applications for helping the blind, handicapped, and Alzheimers patients.

Commercialization efforts are underway. The unit would be sold primarily to institutions and available for lease to private families. The leasing price will be determined. To facilitate commercialization, Yale Systems has identified the Tommy Nobis Center, a multimillion-dollar rehabilitation facility, and IU Medical School as ideal locations for field trials. We have also established connections to the American School of Psychology, the Head Injury Association of Georgia, and the Emory University Center for Rehabilitation for development and testing.

The need for this product exists in the marketplace. Stroke is the third leading killer in the USA, killing 160,000 annually. Only 20 to 30 percent of stroke victims are aphasic to such severity that they could not use PAR™. Worst-case scenario is around 308,000 potential users. However, managed care is tight fisted and reluctant to provide things like medical treatment and expensive medical devices. So, let us say that the market is only 154,000 annually. It may be higher than that because stroke victim's longevity is more than one year.

Another way to cast the market size is by the number of accredited medical schools and hospitals. There are 125 medical schools and 4,900 hospitals in the USA. If half ordered 3 PAR™'s to lease out to their patients, that is 7,350 PAR™'s.

ICS Member Profile: Sebastian Ceria

Dr. Sebastian Ceria has been a member of INFORMS since 1988, during which time he has held a leadership position within the optimization chapter and served on the Lanchester Prize Committee. Today, he is the President of Axioma, Inc., a company that develops, markets, and implements optimization modeling and decision support software.

As an undergraduate studying applied mathematics, for Sebastian, the appeal of mathematical models and algorithms was in the resulting software that lay hidden behind the numbers. It was this fascination with mathematics and problem solving that led him to pursue doctoral studies at Carnegie Mellon University's Graduate School of Industrial Administration. He studied under advisors Egon Balas and Gerard Cornuejols - renowned experts working together for the first time. Sebastian recalls, "It was great working with these leaders in the field of integer programming. They knew where optimization could go next."

After completing his PhD, he joined the faculty at Columbia Business School, where he introduced optimization technology to MBA students. Recognized by his students as the "best core teacher," Sebastian was also honored with the Career Award for Operations Research, which is annually given to the two best researchers and teachers in the area. Interest in optimization was developing, and Sebastian recognized the opportunity to answer practical business needs with sophisticated decision-making tools. He began consulting for several business contacts and before long, former students were also bringing back potential optimization engagements.

In 1998, Sebastian recognized that a company dedicated to optimization would best serve business needs. Technological improvements had made the business environment fertile ground for decision support software and optimization tools. "I wanted to fill the gap between business needs and their tools, and bring optimization technology to practitioners." In response, he co-founded Axioma with other experts who recognized that optimization was "ready for the big leagues."

As President and CEO of Axioma, Sebastian believes that technological talent is critical to developing a competitive advantage, as it not only creates real value for the client, but also raises entry barriers. He says, "Fortunately, my access to academia allowed me to identify the top optimization experts and build the organization. Axioma develops sophisticated optimization technology that is really state-of-the-art." However, he cautions, "Technical talent is not enough; a company needs the ability to sell the products and serve the customer. Effective optimization is not just about developing the fastest algorithms but about tailoring the tools to the needs of the end users."

According to Sebastian, the future for optimization is bright. "As time goes by, companies are realizing that their success depends on the ability to make the right decisions within the time limitations that a complex competitive environment presents. Optimization-based tools can deliver measurable value and intelligent decision support."

Editor's Note: Further information on Axioma, Inc. can be found at www.axiomainc.com

JOC Backissues are online

Full papers from all issues of the INFORMS Journal on Computing from Vol. 1 (1989) through Vol. 9 (1997) are now freely available via the Back Issues link at the JOC web site, http://joc.pubs.informs.org/. Papers and abstracts from later issues are at INFORMS PubsOnLine, http://pubsonline.informs.org/. So now the full text of everything ever published in JOC is available online from one of these two sources.

The 6th INFORMS Conference on Information Systems and Technology

(CIST-2001) will be held November 3-4, 2001 at Miami Beach, Florida. The theme of the conference is "IT Challenges in the E-Commerce Era".

Emphasis will be placed on research contributions that deal with technical, behavioral as well as economic issues encountered while developing IT solutions for E-commerce applications. Detailed call for paper announcement can be found at http://www.mgmt.purdue.edu/faculty/kemal/cist_MIAMI.htm. Submission Deadline: July 16, 2001

University of Colorado at Denver, in association with the UC Health Sciences Center, launches

Center for Computational Biology

Director: Harvey J. Greenberg, CU-Denver Mathematics Dept
UCD Associate Director: Krzysztof (Krys) Cios, Chair of Department of Computer Science and Engineering
HSC Associate Director: Dennis Lezotte, Department of Preventive Medicine and Biometrics
http://www.ucdenver.edu/ccb/

The Human Genome Project has transformed molecular biology into an information science. A science that was once data poor now has so much data that new methods of computation are needed to obtain useful information from the data banks that have emerged. Content-based searches for proteins, genes, and other elements require large-scale modeling, analysis and algorithm design. To meet this new challenge, CU-Denver, in association with the UC Health Sciences Center, has launched the Center for Computational Biology (CCB).

This is an interdisciplinary structure, bringing together researchers in biology and other natural sciences, medicine, computer science, mathematics and statistics. Its first mission is to engage in projects already in the Health Sciences Center, bringing expertise in computer science, mathematics and statistics. The CCB has a second mission: to create courses and programs in computational biology, drawing from resources at CU-Denver and the Health Sciences Center.

While the primary missions are research and education, the CCB approach will foster unification in at least three dimensions. First, esearch and education will be integrated, giving new opportunities to students and faculty. Second, UCD and HSC will strengthen their existing ties by forming this partnership and working collaboratively, bringing complementary strengths to the projects. Third, connections with industry and government will unify efforts to share knowledge with those who can bring research results to people.

901 Elkridge Landing Road, Suite 400

Linthicum, MD 21090-2909