Behind-the-Scenes Data Mining:

A Report on the KDD-98 Panel

George H. John
E.piphany
2300 Geng Road
Palo Alto, CA 94303
gjohn@epiphany.com

ABSTRACT

Successful technology becomes invisible. Few people think much about internal combustion engines while they drive to work in three-thousand-pound hunks of metal powered by them, or electricity while it enables countless parts of their modern lives. Data mining has a long way to go before it succeeds in this way -- or does it? At KDD-98, the Behind-the-Scenes Data Mining panel presented five views on this aspect of the success of data mining. George H. John, Data Mining Guru at E.piphany, moderated the panel which included Usama Fayyad, Senior Researcher at Microsoft; Elliot Fishman, Director of Product Management at DoubleClick; Gerald Fahner, Project Analyst at Fair Isaac & Co.; and Paul DuBose, CTO of Analytika. The story below is a fictional amalgam of their presentations, seen from the viewpoint of an average citizen -- a day in the life of John Q. Record.

Keywords

Data Mining, Applications, KDD-98 Panel Discussion Report, Embedded data mining applications.

1. A DAY IN THE LIFE…

John wakes up at 6:30 to the sound of his clock radio blaring. He got it a couple of months ago at the local Wal-Mart. There were several boxes of this model on the shelf, because the store's sales & inventory forecasting models predicted relatively high weekly sales for this model.

> John has never heard of a "terabyte," and neither had the driver who hauled a truckload of clock-radios and other electronics from a port in San Francisco to Joe's town. But Wal-Mart maintains ten terabytes of point of sale data on an NCR Teradata system, and runs algorithms from NeoVista to forecast demand. [1][2]

After his shower, John takes a hair-loss pill. He's in his thirties and hopes to get back the hair from his twenties in a few years.

> Propecia was first developed to treat prostate enlargement, but it was recently approved by the FDA for treatment of hair loss. After analyzing a study of 1553 patients, they concluded that Propecia promoted hair growth on the scalp, and approved it for this use. Companies like Analytika and Glaxo Wellcome are analyzing large databases of patient information to sytematically discover other beneficial side effects of drugs. [3][4]

He reads the morning paper. The print quality is good, and there aren't any major smears.

> He doesn't think anything about it, but printing is a very complex process, and his newspaper uses a process control method involving data mining to control quality while maximizing speed and minimizing cost. [5]

John has a pleasant drive to work.

> As John drives to work, an onboard neural net is being fed sensor information from his engine, exhaust, and fuel systems 30 times per second. John doesn't know it, but his fuel injector needs cleaning and isn't aspirating the incoming fuel well. His car knows, though, and compensates by recirculating more hot exhaust gas back into the air intake system to promote better combustion. [6]

John gets to work and turns on his computer.

> He's never lost any data to a hard disk crash. John's disk drive was manufactured in 1997, two years after the manufacturer used data mining to find patterns in drives that failed QA tests in the factory. After analyzing the patterns, manufacturing processes were changed and some pieces of equipment were replaced. Now all failed tests are continually analyzed for patterns and causes, and quality has improved. [7]

John is a regional director of sales for a large software company. It's Thursday, and on Thursdays he likes to see how sales are going for the week. He clicks a button to get a profile of sales activity and is pleased that his region is pulling in a higher profit than his peers. However, when he clicks another button to ask for the main differences between his group and other groups, he's disappointed. He discovers that his people aren't doing a good job in selling two products that have low margins but are considered strategic to the company. He clicks another button to ask, among those customers to whom his group has pitched the strategic products, what factors influence whether a customer bought. It turns out there are four different types of customers that seem to be easy targets. He builds a list of these customers and asks his sales staff to focus them for now, and arranges phone calls with the other regional directors to find out more about their sales strategies for other types of customers.

> John knows that he asked how sales were doing and got an answer. He doesn't know that to get the answer, three different data mining algorithms ran against a data mart that pulls data from six source systems within the company. But he just wanted to know about sales -- why should he care about algorithms? Data mining experts at E.piphany built software so that he doesn't have to. [8]

His mother's birthday is coming up, and she's an avid reader, so John logs in to Amazon.com to find something. She mentioned

that she really liked the investment book by Peter Lynch she's currently reading, so he goes to that book's page, and finds a list of other books commonly bought by the same people who bought the Lynch book. He looks through the reviews, picks a couple, and has them sent. He remembers he also wants to send his nephew a gift, and gets a recommendation from the toy store's web site in the same way.

> John has never heard of collaborative filtering or Net Perceptions, but they just made his life easier and his mother happier. He has heard of Microsoft, but didn't know that a company could use its latest web server to automatically recommend and cross-sell products. [9] [10] [11]

Over his lunch hour, he checks on the status of his mortgage application. The loan officer tells him that because of his great credit score of 710, he's eligible for a low rate on a 30-year fixed mortgage.

> Neither John nor the officer really know what a "FICO" score is, and they've certainly never heard of generalized additive models, but that's ok -- the statisticians at Fair, Isaac & Co. know quite a lot about them. Their credit scores are ubiquitous in the financial services industry. [12] [13]

On the way back to work, he stops at a fast food restaurant intending to get a soda, but notices that a lunchtime special value-meal combination is priced only a little higher than the drink alone, and includes a hamburger. He goes for the combo.

> A large fast food restaurant worked with one of the major data mining vendors to understand customer behavior, using time series and market basket analysis. As a result they began a campaign to convert "drinkers" to "eaters."

During a break in the afternoon, he checks his stock prices on StockMaster.com, and is shown an ad for a mortgage broker service on the web. He clicks through and fills out a form to investigate competing mortgage rates.

> That was serendipitous, but not entirely accidental. A company called DoubleClick manages the ad inventory on StockMaster, and their targeting algorithms discovered that the mortgage ad was likely to be of interest to StockMaster visitors, on that webpage, at that time of day. [14]

At 10:31am, 2:03pm, and 4:17pm, John got three different spam e-mails. At 4:17pm, 4:18pm, 4:19pm, and 4:19pm again he got extra copies of the last one, just in case he didn't see it the first time. But John never saw any of them.

> John has never heard of perceptrons or Bayes nets, but Mehran Sahami, Ph.D. grad from Stanford, spent several months at Microsoft teaching them to recognize spam. Microsoft expects the technology to be available in a coming version of Microsoft Outlook. [15]

John and his fiancee meet and spend the evening buying some furniture and home electronics items for the new place. He stops for gasoline, and his credit card isn't accepted. He calls the card company, and they say they just want to be sure it's him using his card.

> Sophisticated fraud detection systems from HNC and other companies keeps an eye on hundreds of millions of accounts worldwide, looking for unusual purchasing behavior. They regularly mine millions of transactions, learning about fraud in general and also learning what type of shopping is normal for each account. [16]

They get home and watch Seinfeld, then stay tuned to NBC through the next show and then ER.

> NBC has a group called Quantitative Programming Research that uses sophisticated data mining software such as S-Plus to understand the audiences of each show, which can aid programming decisions as well as advertising. John might well have data mining to thank for those relaxing Thursday evenings he spends with his fiancee. [17]

Looking through the day's mail before falling asleep, John tosses two credit card applications in the trash.

> He fits the profile of a highly profitable customer, and ends up in the campaigns of major card issuers frequently. [18]

He sleeps peacefully at night, because he saw on the evening news that the city police department had just caught a prolific criminal who frequented his neighborhood.

> ISL's Clementine is being used by one police department to look for unsolved crimes that might have been committed by the same criminal. The discovery that a set of crimes are related might give the department new leads. [19]

Acknowledgments

Many thanks to Usama Fayyad, Elliot Fishman, Gerald Fahner, and Paul DuBose for participating in the panel.

Disclaimer

The indented paragraphs refer to applications of data mining, in many cases making statements about actual companies. Such statements are consistent with the sources cited, but the business world changes too frequently for the author to claim that the statements are accurate at the time this article is published.

2. REFERENCES

[1] NeoVista Software, "Wal-Mart Deploys NeoVista," May, 1997. www.neovista.com/corporate_info/PressReleases/NVWalMartProfileMAY97.htm

[2] NCR Corporation,"Wal-Mart:The Solution," December 1998. www3.ncr.com/product/case_studies/wal-mart/solution.htm

[3] "From a Prostate Drug Comes a Pill for Baldness," *The Wall Street Journal*, 3/20/1997.

[4] Analytika, Inc., "Analytika, Inc. and Glaxo Wellcome Inc. Sign Agreement," December 1997. www.analytika.com

[5] Bob Evans and Doug Fisher, "Overcoming process delays with decision tree induction." IEEE Expert, Vol. 9, No. 1, 60—66. cswww.vuse.vanderbilt.edu/~dfisher/tech-reports/IEEE-Exp-94.html

[6] NASA Jet Propulsion Laboratory, "JPL Neural Network Chip Paves the Way," September 1998. www-techtrans.jpl.nasa.gov/success/stories/FordNeuralchip.html

[7] Chid Apte, Sholom Weiss and G. Grout, "Predicting Defects in Disk Drive Manufacturing: A Case Study in High-Dimensional Classification." IEEE Annual Conference on AI Applications CAIA-93, March 1993.

[8] Data Mining News, "E.piphany has a very good January," January 1999. www.epiphany.com

[9] Net Perceptions, "Net Perceptions Announces… Recommendation Engine", October 1998. www.netperceptions.com/press/release/pr_19981007_3.html

[10] John Foley and Joy D. Russell, "Mining Your Own Business," *InformationWeek*, March 16, 1998. www.informationweek.com/673/73iudat.htm

[11] Alan Saldanha, "Promotions 101," December, 1998. msdn.microsoft.com/workshop/server/commerce/promotions101.asp

[12] Freddie Mac, "Credit Scoring." www.freddiemac.com/corporate/creditscore/creditsc.htm

[13] Fair, Isaac and Co., Inc., "Fair, Isaac's CreditDesk 2.3 Streamlines Operations," January 1998. www.fairisaac.com

[14] DoubleClick, Inc., "DoubleClick Network: StockMaster," www.doubleclick.net/onesheeters/domestic/stockmaster

[15] William Baldwin, "Spam Killers," *Forbes*, September 21, 1998. www.forbes.com/forbes/98/0921/6206254a.htm

[16] HNC Software, "HNC Software Unveils Falcon 4.0," February, 1998. www.hnc.com/news/newpr_021198.html

[17] Director of Quantitative Research at NBC quoted in www.mathsoft.com/splus/spress/roadshow/

[18] "How Winners Do It," Forbes ASAP, August 24, 1998. www.forbes.com/asap/98/0824/088b.htm

[19] ISL, "Data Mining Case Studies: West Midlands Police – Crime Detection." www.isl.co.uk/clementine/wmpolice.html

About the Author

George H. John is the Data Mining Guru at E.piphany, where he led the incarnation of behind-the-scenes data mining into E.piphany's business solutions. He currently manages E.piphany's strategic accounts in the financial services industry. Previously, he worked in IBM's Business Intelligence group, doing data-mining consulting and training for Global 500 companies. He earned a Ph.D. with Distinction in Teaching from the Computer Science Department of Stanford University, where his research was supported by a National Science Foundation fellowship. His dissertation on data mining has been read by over 1000 researchers and practitioners from such institutions as the International Monetary Fund, Fidelity Investments, J. D. Power & Associates, Harvard, and MIT.
(xenon.stanford.edu/~gjohn)