Integrating Data Mining into Vertical Solutions: Problems and Challenges

Panel organizers

Ronny Kohavi  
Director, Data Mining  
Blue Martini Software  
ronnyk@bluemartini.com

Mehran Sahami  
Systems Scientist  
E.piphany, Inc.  
sahami@epiphany.com
Panel Participants

- Jim Bozik from Acxiom Corp (data provider)
- Dorian Pyle from Data Miners (consulting)
- Rob Gerritsen from Exclusive Ore (consulting)
- Steve Belcher from Unica (horizontal to vertical)
- Ken Ono from Angoss (horizontal)
Panel Format

Panel is 90 minutes

- Introduction - 10 minutes
- Panelists: 1 minute intro by Mehran/Ronny
  5 minutes opening statement
- Discussion: 30 minutes
- Panelists: 3 minutes closing statement
- Concluding remarks

10 40 70 85 90
Panelist Questions

- Eight questions were sent to panelists for opinions and interest rating
- Waterfall model based on responses:
  - Each panelist was asked to address two different questions
  - Each question is being answered by two consecutive panelists
  - Questions were chosen so that consecutive panelists do not agree on answer
Questions (I of II)

Q1: Solutions versus Tools
What should companies sell?

Q2: Who are the users of the data mining?
Business users or analysts?

Q3: Will data mining functionality be successfully integrated into databases?

Q4: Do models need to be interpretable?
Questions (II of II)

Q5: Is there a future for horizontal data mining tool providers?

Q6: Will industry-standard APIs be adopted? Will they help horizontal data mining companies?
Ronny Kohavi (Blue Martini Software)

- Joined Blue Martini Software in Sept 1998
  - Director of Data Mining
- Previous experience
  - MineSet manager, SGI
  - MLC++ project, Stanford University
  - Co-chair (with Jim Gray) of KDD-99's industrial track
  - Co-editor (with Foster Provost) of upcoming issue of the Data Mining and Knowledge Discovery journal special issue on: E-commerce and Data Mining
- Ph.D. in Computer Science from Stanford
Mehran Sahami (E.piphany)

- Joined E.piphany in 1998
  - Systems Scientist leading data mining R & D
  - Manager of Real-Time Products development

- Previous experience
  - DM research at Xerox PARC, SRI, and Microsoft
  - Consultant in text mining/classification/clustering
  - Lecturer at Stanford University

- Ph.D. in Computer Science from Stanford
Jim Bozik (Acxiom)

- Joined Acxiom in 1997
  - Works directly with data mining clients
  - Leads effort researching analytical software

- Previous experience
  - Retail Marketing and Analysis at Signet Bank
  - Business Research Division at Hallmark Cards
  - Statistical Research Division at the U.S. Census

- BA in Mathematics and CS, MA in Statistics
Q1: Solutions versus Tools: what should companies sell?

- We are interested in SOLVING PROBLEMS, not in BUYING SOFTWARE.

- SDS has an ongoing process of evaluating software that...
  - Enhances the ‘Analyst Toolkit’
  - Offers ways to create a more visually dynamic product

- We offer the following advice. It sounds like common sense, but you’d be surprised...
  - LISTEN! Ask about the areas of application, users, objectives (e.g., Don’t talk about NN if NN have limited use)
  - Provide explicit guidance on the proper configuration, and file size constraints for evaluation software
Q2: Who are the users of the data mining? Business users or analysts?

- In our environment, the users are ANALYSTS.

- We believe the issue is not the SCIENCE of analysis, but the ART of analysis.
  - Is the data received what you expected?
  - How do you spot problems in data? Are they really problems?
  - When do you create variables to enhance a model? Which ones?
  - How do you create a model that is intuitively appealing to a client?
Dorian Pyle (Data Miners)

- Joined Data Miners in 1998
  - Consultancy company with Michael Berry and Gordon Linoff

- Previous experience
  - 25 years of modeling experience, including at Naviant and Thinking Machines Corporation
  - Author of Data Preparation for Data Mining
  - Upcoming textbook Mining for Models
The Questions

- Who uses data mining?
- Will data mining functionality be successfully integrated into databases?
The Problem

I really hate this damn computer,
I think I ought to sell it.
It never does just what I want,
But only what I tell it!

Sign in computer room. Circa 1975.
The ideal!

Listen to what we ask ....

.... do what we want, NOT what we said!
The Perception

Hmmmm. I wonder what I can do with that? I can’t really imagine anything to do with it!

Now this is magic! It’s going to solve all my problems!
The Way Today

- Marketing
- Telemarketing
- Churn
- Development
- Manufacturing
Tomorrow - A core technology

- Task comprehension
- Voice recognition
- Problem framing
- Automatic alerts
- Intelligent agents
- Situation monitoring
- Results presentation
  - Mobile contact
  - Web alert
- Data access
- Web searching
- Knowledge acquisition
Q: Who uses Data Mining?

- Business Managers
- Business Analysts
- Architects
- Planners
- Financial Analysts
- Plant Managers
- Marketers
- Investors

A: Anyone who needs answers to questions based on available data (No imagination about how needed!)
Q: Incorporate in Databases?

- Data access
- Web searching
- Knowledge acquisition

A: No - for business reasons …..

- First, the range of data to be accessed is not just in a database
- Second, the questions asked require multiple methods of inquiry - no “one-size-fits-all”
- Third, performance and currency (for now)
Q: Incorporate in Databases?

A: No - for technical reasons.

Not just NN & DT. No common primitives for new techniques (evolution programming, algebra evolvers, swarm clusters, semantic nets, Baysian nets, thematic association, …..)
Rob Gerritsen (Exclusive Ore)

- Founded Exclusive Ore in 1997
  - Focus on data mining consulting and technology
  - Research in integrating data mining and RDBMS

- Previous experience
  - 31+ years experience in data management/mining
  - Co-founder and VP Technology at Two Crows
  - Associate Professor at The Wharton School

- Ph.D. in System Science from CMU
The Questions

Q3: Will data mining functionality be successfully integrated into databases?

Q4: Do models need to be interpretable?
Q3: DM into Databases?

YES!

- It’s natural
  - Models are no more than abstracted/reformatted data
  - Data mining can benefit from database integration

- It’s inevitable
  - Competitive pressure
A model is an abstraction of the data and belongs with the data.

There is nothing more in a model than what is already in the data.
DM Will Benefit from DBMS - I

- Model management
  - Version control, model comparisons

- Model deployment
  - Predictions right in the database

- Understanding the model
  - Browse, query, compare rules

- Incremental modeling
  - Revise models when new data arrives
DM Will Benefit from DBMS -II

- Model monitoring
  - Continuous validation of models on new data
- Security services
  - Extraction opens big security hole!
- Better performance
DM into Database - Inevitable

- Expand the database as an enterprise platform
- Happening now
  - Informix/Red Brick SQL Extensions
  - Compaq SQL/MX
  - Oracle acquires Darwin
Q4: Models be Interpretable?

YES!

- For the model builder
  - Avoid costly/stupid mistakes
- For the business user
  - “Trust me it works”
Business Risks are Too Great

- **Direct mail**
  - Would you eliminate 25% of your list without knowing why? You risk reducing revenue by 25%!

- **Medical**
  - Patient complains of recurrent headaches, but model says no brain cancer risk. Do you want to know why?

- **Lender**
  - Would you deny lending me $50K without telling me why?
Steve Belcher (Unica Technologies)

- Consultant at Unica Technologies
- Previous experience
  - Worked in IT and Data Mining for 16 years
  - Taught in graduate and undergraduate programs at several colleges
- Dissertation on application of neural networks in financial forecasting
Q4: Do models need to be interpretable?

- Models need to work. This does imply validation

- Interpretability is subject to customer needs

- Required in some applications - Fair Lending practices
Q5: Is there a future for horizontal data mining tool providers?

- Unique perspective
- A very limited future
- Vendor consolidation.
- Vertical apps are easier to use
- Models must be able to be used in a business environment
- DM Futures - embedded systems
Ken Ono (Angoss)

- VP of Technology at Angoss
  - Head of development for data mining solutions
  - Chief architect for the data mining product suite

- Other responsibilities at Angoss
  - Embedding technologies
  - OEMing technology
  - Other licensing transactions with partners
ANGOSS Products

- Provider of KnowledgeSEEKER & KnowledgeSTUDIO
- STUDIO designed from ground up to achieve:
  - Programmability and embedability (DCOM/ActiveX)
  - Tight integration with database (In-Place Mining)
  - Visualization and exploration for visual data mining and knowledge discovery
  - Ease of use
- Price points that make it much easier to start data mining
Q5: Future for horizontal DM tool providers?

- **OEM** - One of many approaches
  - Data Mining is a complex technology that can apply to many different industries
  - State of software industry makes it easy to encapsulate DM components
  - Why should solution providers have to learn intricacies of DM algorithms?
  - Can hide & automate complexities by leveraging domain knowledge thus widen market

- **Analytic Market** - small but important
  - Will continue to grow
  - Expert individual can created better models than an application that hides & automates process
Q5: Future for horizontal DM tool providers?

- Creation of predictive models (algorithms) will be incorporated into databases and will become commodities quickly dropping in price.
  - Microsoft OLE DB for DM & Oracle’s purchase of Darwin are the beginning of this.

- DM Vendors must leverage & enhance functionality of database.

- Client side tools are still required for data exploration and discovery of new and interesting insights.
Q6: Will standard APIs be adopted, will they help horizontal DM corp.?

- Standards are already starting to emerge
- OLE DB for DM from Microsoft
  - Provides an easy way to create and deploy predictive models
  - Legions of developers can integrate DM with much less risk than writing to “one company API”.
  - Paves way for wide deployment of low risk PM’s.
    - “What banner do I display?” = low risk.
    - “Should I give this person a loan?” = high risk
  - Creates infrastructure for deployment of models
Q6: Will standard APIs be adopted, will they help horizontal DM corp.?

PMML - another piece of the puzzle

- Predictive Model Mark Up Language
- XML extension for describing the contents of a predictive model
- Defines a way for a PM to be
  - transferred between environments
  - persisted in a repository
  - searched and queried (find me a model that …)
Q6: Will standard APIs be adopted, will they help horizontal DM corp.?

Will it help DM vendors?
- Will reduce cost of ownership of adopting and providing solutions that contain DM
- Will increase level of awareness about data mining (especially OLE DB for DM)
- Will increase demand for data mining
- Will increase competition