DIMACS/IBM Workshop on Data Mining
in the Internet Age

Mining E-commerce Data: Challenges and Stories from the Trenches

Ronny Kohavi
Director, Data Mining
Blue Martini Software
ronnyk@bluemartini.com
http://robotics.Stanford.EDU/~ronnyk/

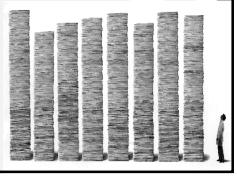
2 May 2000

Overview

- E-commerce: killer domain for data mining
- Blue Martini Software value proposition
- Architectural decisions
- Stories and Challenges
- Summary

E-Commerce: the Killer Domain for Data Mining

- E-commerce provides all the right ingredients for successful data mining, including:
  - Large amounts of data (many records)
  - Rich data with many attributes (wide)
  - Clean data
  - Actionable domain
  - Measurable return-on-investment



E-Commerce: the Killer Domain for Data Mining (II)

Why Data Mining and E-Commerce are a match made in heaven (or in the data warehouse):

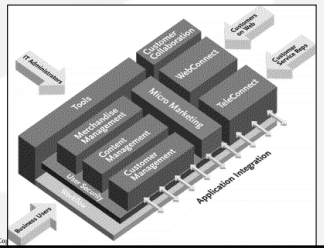
- Clickstreams at sites provide data that dwarfs large warehouses built in previous years
- Designed correctly, sites can assign many attributes to content on web pages, customers, products, purchases.
- Data is collected electronically at the webstore and it is clean (no painful legacy transformations)
- Insight derived from data mining analysis is easily turned into action, closing the loop with the transactional systems
- ROI is easy to measure at a webstore

Data Mining becomes Important

- Prior to 2000, horizontal data mining companies were bought for about $10M
  - Compression Sciences was bought by Gentia **($3M)**
  - HyperParallel was bought by Yahoo **($2.3M)**
  - Clementine was bought by SPSS **($7M)**
  - Thinking Machines was bought by Oracle (~ **$25M)**
- Around 2000, a phase shift occurred in valuations, which were $100M-$500M
  - NetPerceptions bought KD1 for 2.24 million shares worth **$116M** (now worth about $43 million)
  - Epiphany bought RightPoint (previously DataMind) for 3.6 million shares worth **$400M** (now worth about $240M)
  - Vignette bought DataSage for 3.16 million shares worth **$577** (now worth $456M after 3 to 1 split)
  - NeoVista bought by Accrue for 2.4 million shares worth $**140M** (now worth $70M)

Blue Martini Software

Blue Martini's solution is built on the vision of a complete E-commerce solution with integrated data mining, called Micro Marketing



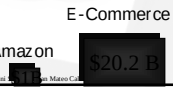
300+ employees
35+ customers including:
Levi's
Harley Davidson
Gymboree
The Men's Warehouse
gloss.com
gazelle.com
craft.com

Value Proposition

- A company's **brand** is a strategic asset
  - Avoid diluting it with a mediocre web store
  - Leverage the internet to build your brand and reinforce the same message from multiple touchpoints (web, call centers, wireless, bricks and mortar)
- Support personalized recommendations/cross-sells
  - Increase conversion rates (browse to buy)
  - Increase basket size (through cross sells, better navigation)
  - Increase customer retention and loyalty

The Webstore is an Experimental Laboratory

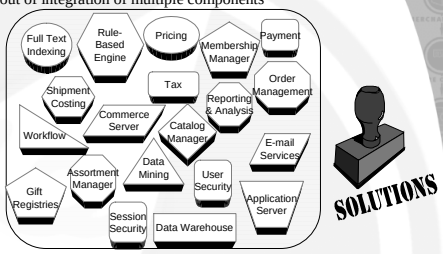
- Opening webstores will not dramatically impact revenues for established retailers, but lessons learned will affect other channels
- The webstore provides an experimental laboratory and a trend-discovery system
  - Which cross-sells work?
  - Which ads are effective?
  - What are people looking for (failed searches for pokédex)



Wal-Mart 1999 revenues: $162.8 B

Perils of Components

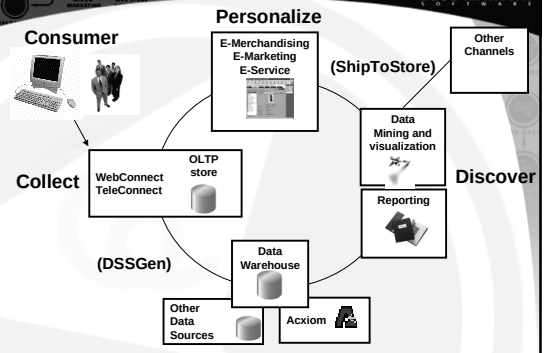
Blue Martini charges close to $1M for the software because we take the pain out of integration of multiple components



Average E-Commerce site costs $5.9M to assemble and $4.3M annually to maintain.

International Data Corporation, January 1999

Closing the Loop



Architectural Decisions

- Web pages are dynamic (jsp/jhtml) making API calls everywhere
  - Everything is in the database, the web pages access it.
  - Every object can have attributes (e.g., customers, products, assortments, orders, content pages) which are accessible in the webstore and later in mining
- Webstore's app server writes clickstreams
  - Knows about sessions - avoid sessionizing problems
  - Knows what was displayed - products, assortments
  - Can save additional information not in weblogs
  - Central DB repository - avoid need to join weblogs
  - Can investigate orders and clickstreams together

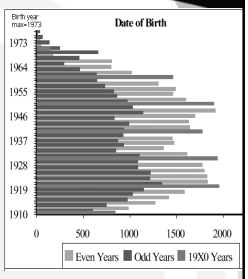
Architectural Decisions (II)

- Provide automatic ETL (Extract/Transform/Load) to build Data Warehouse, cutting the 70-90% data preparation step in KDD dramatically
- Provide access to syndicated data (Acxiom) About 60 customer attributes
- Personalization rules engine runs at the webstore, providing cross-sells, recommended images, assortments, etc.
- ShipToStore closes the loop, allowing model scores to be shipped back to store for personalization

Story I - Mysterious Birth Years

- **The KDD-98 data contained interesting anomalies for date of birth:**
  - Spikes on years ending in zero
  - No individuals were born prior to 1910
  - There were twice as many individuals who were born on even years as on odd years

Why?



Graph adopted from KDD'99 Competition by Jim Georges and Anne Milley, SAS institute SIGKDD Explorations, Jan 2000, Volume 1, Issue 2, page 79

Challenge I - Dates

- Dates/times are ***very*** important and appear very frequently in e-commerce, yet most data mining algorithms do not support them
- Common themes:
  - Provide well-used measurements in industry, such as Recency and Frequency (of RFM).
  - Provide strong support for date operations (days between dates, day-of-week, etc) and let the users define date operations

Is there something better we can do?

Story II - Gender Mystery

- A site has gender on the registration form
- Acxiom, a syndicated data provider, also provides gender
- There was a very large discrepancy between the percentage of males according to registration and by Acxiom

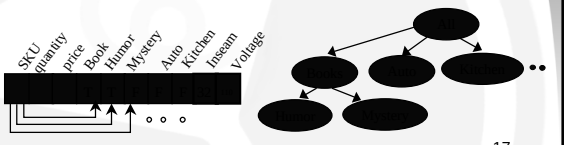
Why?

Challenge II - Unknown vs. N/A

- A *pregnant* attribute on a customer can have four values: yes, no (response), unknown (we didn't ask), and not applicable (e.g., for males or young children).
- E-commerce products have attributes that depend on their family:
  - Pants have length, inseam, fabric material
  - Microwaves have voltage, cubic inches, weight
- For data mining, we flatten all attributes and most values are N/A (e.g., voltage for pants)
- How should we handle N/A versus unknown?

Challenge II' - Product Hierarchies

- Products are typically arranged in a hierarchy. Most algorithms expect same-size records
- Common themes:
  - Flatten product attributes (lots of N/As).
  - Allow users to choose parts of hierarchy for pivots based on product id (SKU).
    Add Boolean columns from hierarchy



Story III - Low Conversion Rates

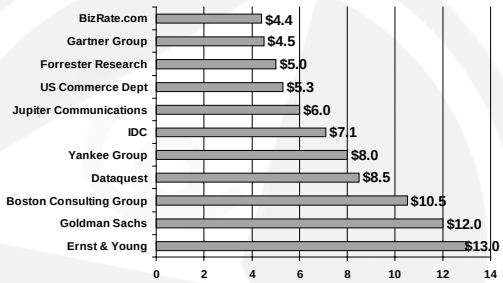
- Conversion rate is the ratio of buyers to browsers.
- High conversion rates are obviously desired
- Reports we generate can show highest and lowest conversion rates
- Why do some products have really low conversion rates?

Challenge III - Changing Sites

- One of the hardest problems is dealing with changing sites - all our I.I.D assumptions do not hold!
- It is very easy to change personalization rules (cross-sells/up-sells), images shown, messages, etc
- Sites changes their registration forms
- New products are introduced, old ones are deleted (especially with perishable products such as wines) or change attributes
- How do we deal with slowly changing dimensions? (or quickly changing?)

Story IV - Analyst numbers

4Q 1999 Online Holiday Shopping Revenues (Billions). Why don't the numbers agree?



Challenge IV - Scalability

- Yahoo had 465 million page views per day in December of 1999 (*)
- That's about 2-4GB of clickstream an hour, depending on the amount of clickstream information stored
- What can we do with such volumes?
- Are there useful aggregations of such data that can be done on the fly?

(*) CFO April 2000

More Challenges

- How can we prune the number of associations generated?
- What is the loss metric for displaying a cross-sell?
  - If you display the product most-likely to be purchased, would you display bread on all pages until bread was purchased?
  - If you display product with highest lift, would you still display it if the probability was 0.01% (up from 0.00001%)?
- Are there algorithms that could take a star-schema and mine it without flattening it (e.g., Query flocks)?
- Bots/Crawlers tend to skew statistics dramatically.
- How can marketing campaigns be taken into account?

Business Challenges

- Most companies do not have the expertise to build data mining transformations.
  Can we automate them?
  The more vertical, the easier this gets.
- How can we generate comprehensible models? Actionable models?

Summary

- E-commerce is the killer-domain for data mining
- Automatic generation of data warehouse and closing the loop back to store is key to making data mining usable and actionable
- Clickstreams need to be collected at the app-server level where meta information exists
- Challenges: Date handling, unknown vs. N/A, hierarchy support, constantly changing sites, scalability, and more