Stages of Knowledge Discovery in Websites
Introduction

- **Physical store**
  - Layout and contents are the same for every customer
  - Customers leave very little useful trace

- **On-line store**
  - Layout and contents are easily modified and can be personalized to each customer
  - Every visit generates a trail of information on the customer’s experience
Overview: Three Stages

(1) Clickstream Analysis

(2) Advanced Data Mining

(3) Personalization
Stage 1: Clickstream Analysis

- **Data:** web server logs
- **How:** analytic reporting tools
Web Server Log

Example: Online Retail Store

LogFormat: "virtualname, host, date/time, method/URL, code, bytes, refererURL, useragent"

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Date/Time</th>
<th>Method/URL</th>
<th>Code</th>
<th>Bytes</th>
<th>RefererURL</th>
<th>User Agent</th>
</tr>
</thead>
</table>
| 121.32.143.220 | 05/19/2005 04:08:27 | POST /main.html | 200 | 0 | Google | Mozilla/5.0 (compatible; MSIE 5.0)
| 241.100.143.220 | 05/19/2005 04:08:30 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 24.201.105.203 | 05/19/2005 04:08:33 | POST /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:08:36 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 24.201.105.203 | 05/19/2005 04:08:39 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:08:42 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:08:45 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:08:48 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:08:51 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:08:54 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:08:57 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:08:59 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:09:02 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:09:04 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:09:07 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:09:09 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:09:12 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:09:15 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:09:18 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:09:20 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:09:23 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:09:25 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 82.231.127.34 | 05/19/2005 04:09:28 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:09:30 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
| 67.0.226.128 | 05/19/2005 04:09:33 | GET /main.html | 200 | 0 | | Mozilla/5.0 (compatible; MSIE 5.0)
Clickstream Analysis

Answers…

- Where do most visitors come from?
- Which search engines?
- What search terms are most often used?
- How long do visitors stay?
- How many pages do they visit on average?
- Which pages are most popular?
- When do they leave the site?
- Which pages do visitors commonly leave the website from?

Example report  (PDF)
Search Engine Optimization (SEO) Example

Number of Search Terms
(Referred by Search Engines)

Key Phrases
Key Words

Brigham Young University - Data Mining Lab (http://dml.cs.byu.edu)
Stage 2: Advanced Web Mining

- **Data:** stage 1 data, profile information, transactional data, outside demographic data, etc.

- **How:** design additional data collection mechanisms, mine additional data, add socio-economic or demographic data
Advanced Web Mining

Answers…

- What is the conversion rate?
- How many would-be customers begin shopping but drop out before check-out?
- How well did special offer X do?
- Who are the most profitable customers?
- What is being bought by whom?
- What interests do your customers have?
Simple Conversion Rate Example

- **Real Online E-Commerce Site**
  - Combine *transactional* data with *clickstream* data
  - During April, approximately *57%* of all visitors bought a product after logging in and viewing the final order details (43% abandoned checkout).

\[
\frac{\text{(# of customers that completed checkout)}}{\text{(# of visitors that viewed final order details)}} = \frac{799}{1414} = 57\%
\]
Stage 3: Personalization

**Data:**
- Everything from stages 1-2
- Pre-processed and real-time data

**How:** Data Mining Techniques
- Clustering/Segmentation
- Collaborative filtering
- Associations
- Rule-based
- State-based
Collaborative Personalization

Examples

- **Collaborative Filtering Based**
  - Recommendations/Predictions
    - Books (Amazon)
    - Music (Yahoo! Launchcast)
    - Movies (Yahoo! Movies)
    - Gifts (Yahoo! Shopping)
    - Websites (StumbleUpon), etc.

  "Collaborative filtering software is changing the way people choose music, books and other things, by helping them find things they like, but did not know about .... It helps people find things they might otherwise miss."

  -The Economist (“United we find,” March 2005)
Collaborative Filtering: Books

Example: Amazon.com (http://www.amazon.com/)

In this case, order information is leveraged to identify clusters of users with similar purchasing habits, the underlying assumption being that people with similar buying behavior are very likely to have similar interests.

Customers who bought **The Adventures of Huckleberry Finn (Penguin Classics)** by Mark Twain
- also bought **Treasure Island (Signet Classic)** by Robert Louis Stevenson
- **The Adventures of Huckleberry Finn (Bantam Classics)** by MARK TWAIN
- **Adventures of Huckleberry Finn (Modern Library Classics)** by MARK TWAIN
- **20,000 Leagues Under the Sea** by Jules Verne
- **The Swiss Family Robinson** by JOHANN WYSS
Collaborative Filtering: Music

Example: Yahoo’s Launchcast (http://lauch.yahoo.com/)
Collaborative Filtering: Movies

Still needs work, but is improving.

To improve your recommendations, rate more movies.
Personalized Search

Example: Google (http://www.google.com/psearch)

“Personalized Search uses the information from your search history or other information you provide us to improve your Google search results. Personalized Search algorithms use the information to improve your Google search results by boosting results that are more relevant to you.”
Personalization Possibilities

- Serve the *right product* to the *right person* at the *right time* (1-to-1 marketing).
  - Save time for users/customers
  - Boost revenue for the businesses
  - Build trust with users/customer
    - Unsuccessful (or extreme) personalization can lose trust and loyalty.

- Many Undiscovered Possibilities
Review: Stages

- (1) Clickstream Analysis
- (2) Advanced Data Mining
- (3) Personalization