



Matthew Smith

smitty@byu.edu

Brigham Young University, Data Mining Lab

July 2005

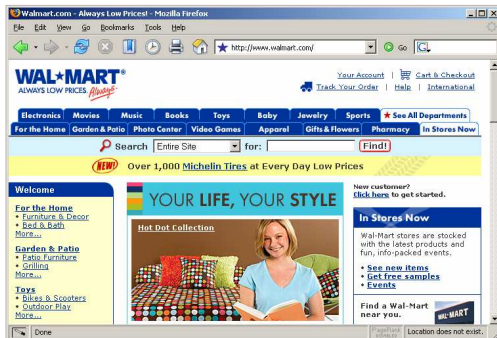
Stages of Knowledge Discovery in Websites

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



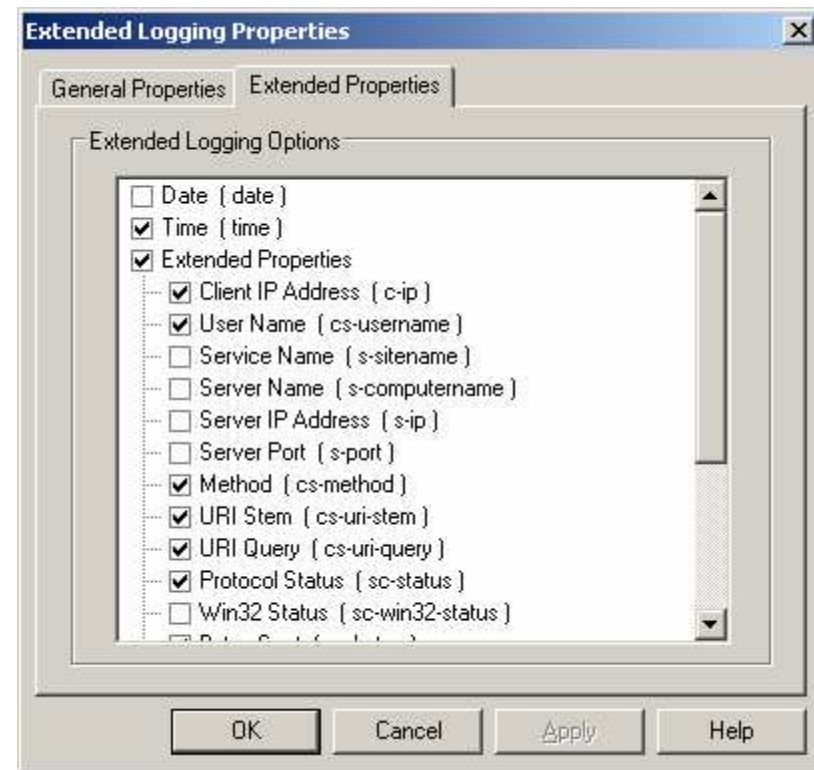
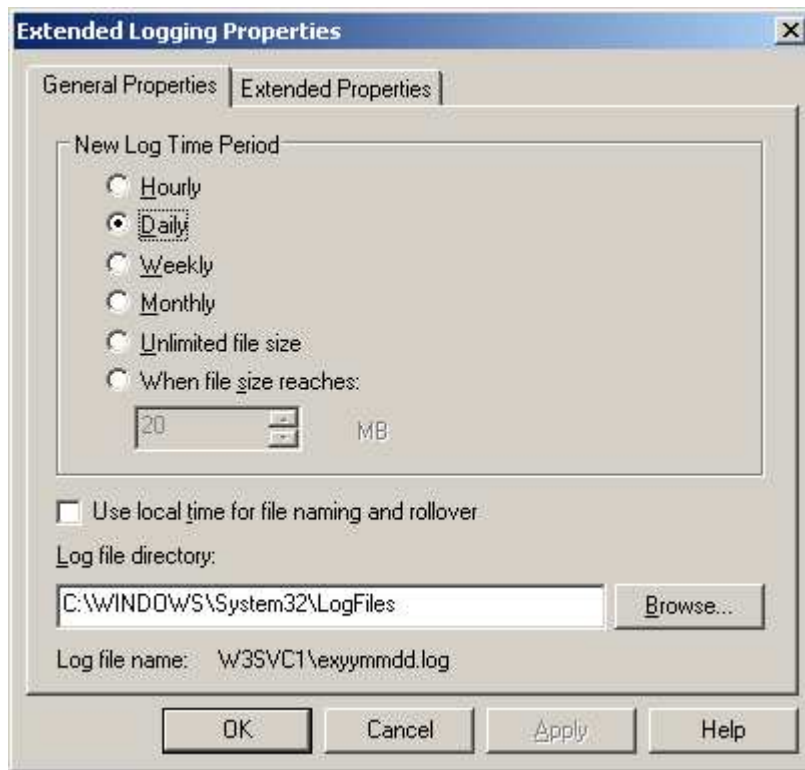
(3) Personalization

(2) Advanced Data Mining

(1) Clickstream Analysis

Stage 1: Clickstream Analysis

- *Data:* web server logs
- *How:* analytic reporting tools

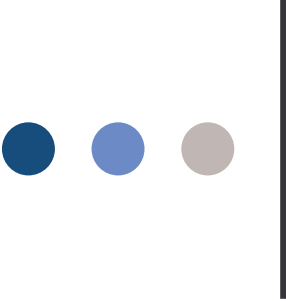


Web Server Log

Example: Online Retail Store

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

```
webusage.log - Notepad
File Edit Format View Help
213.100.143.220 - - [29/May/2005:04:06:26 -0700] "GET http://www.google.se/search?hl=sv&q=lds+art&meta= HTTP/1.1" 200 82517 "-" "Mozilla/
213.100.143.220 - - [29/May/2005:04:06:26 -0700] "GET /ePOS/this_category=17&store=1&form=shared3/gm/main.html&design=1 HTTP/1.1" 200 825
213.100.143.220 - - [29/May/2005:04:07:15 -0700] "GET /ePOS/this_category=17&store=439&form=shared3/gm/browse.html&design=1 HTTP/1.1" 200
12.218.82.122 - - [29/May/2005:04:07:25 -0700] "GET http://www.google.com/search?hl=en&q=king+benjamin HTTP/1.1" 200 85209 "-" "Mozilla/4
12.218.82.122 - - [29/May/2005:04:07:25 -0700] "GET /ePOS/this_category=124&store=1&item_number=1330687&form=shared3/gm/detail.html&desig
213.100.143.220 - - [29/May/2005:04:08:54 -0700] "GET /ePOS/this_category=17&store=439&form=shared3/gm/main.html&design=1 HTTP/1.1" 200 8
195.229.241.184 - - [29/May/2005:04:10:49 -0700] "GET http://www.google.com/search?hl=en&r=&rls=GGLD%2CGGLD%3A2004-33%2CGGLD%3Aen&q=char
195.229.241.184 - - [29/May/2005:04:10:49 -0700] "GET /ePOS/this_category=76&store=1&item_number=8807654&form=shared3/gm/detail.html&desi
24.202.106.203 - - [29/May/2005:04:11:38 -0700] "GET http://www.google.ca/search?hl=en&q=grey+olsen&meta= HTTP/1.1" 200 85352 "-" "Mozi
24.202.106.203 - - [29/May/2005:04:11:38 -0700] "GET /ePOS/this_category=102&store=1&item_number=1-57734-880-x&form=shared3/gm/detail.htm
213.100.143.220 - - [29/May/2005:04:18:08 -0700] "GET /ePOS/this_category=17&store=439&form=shared3/gm/browse.html&design=1 HTTP/1.1" 200
82.231.127.34 - - [29/May/2005:04:20:30 -0700] "GET http://www.google.fr/search?hl=fr&q=lds+music+by+Lex+de+Azevedo&meta= HTTP/1.1" 200 2
82.231.127.34 - - [29/May/2005:04:20:30 -0700] "GET /ePOS/width=615&vlink=%230028CA&valign=top&topmargin=0&this_category=76&text=%23000000
82.231.127.34 - - [29/May/2005:04:20:59 -0700] "GET /ePOS/this_category=76&store=439&item_number=1242458&form=shared3/gm/detail.html&desi
81.64.156.214 - - [29/May/2005:04:21:39 -0700] "GET http://www.google.fr/search?q=emma+smith+photo&sourceid=mozilla-search&start=0&start=
81.64.156.214 - - [29/May/2005:04:21:39 -0700] "GET /ePOS/this_category=70&store=1&item_number=3266834&form=shared3/gm/detail.html&design
69.67.226.128 - - [29/May/2005:04:22:09 -0700] "GET http://www.google.com/search?hl=en&q=As+a+Man+Thinketh+by+James+Allen&btnG=Google+sea
69.67.226.128 - - [29/May/2005:04:22:09 -0700] "GET /ePOS/this_category=91&store=1&item_number=1-57008-968-x&form=shared3/gm/detail.html&
82.231.127.34 - - [29/May/2005:04:22:32 -0700] "GET /ePOS/this_category=76&store=439&item_number=78302754132&form=shared3/gm/detail.html&
69.67.226.128 - - [29/May/2005:04:23:39 -0700] "GET /ePOS/store=439&item_number=1-57008-968-x&form=shared3/catalogs/common/large_image_po
81.192.29.230 - - [29/May/2005:04:26:56 -0700] "GET http://www.google.fr/search?hl=fr&q=bookstore+shipping+dh&meta= HTTP/1.1" 200 91519
81.192.29.230 - - [29/May/2005:04:26:56 -0700] "GET /ePOS/this_category=202&store=439&form=shared3/gm/main.html&design=1 HTTP/1.1" 200 91
64.124.85.76 - - [29/May/2005:04:27:34 -0700] "GET - HTTP/1.1" 200 27 "-" "Mozilla/5.0 (compatible; BecomeBot/2.3; MSIE 6.0 compatible; +
64.124.85.76 - - [29/May/2005:04:27:34 -0700] "GET /robots.txt HTTP/1.1" 200 27 "-" "Mozilla/5.0 (compatible; BecomeBot/2.3; MSIE 6.0 com
141.154.118.110 - - [29/May/2005:04:27:58 -0700] "GET http://www.google.com/search?sourceid=navclient&ie=UTF-8&rls=GGLD,GGLD:2004-20,GGLD
141.154.118.110 - - [29/May/2005:04:27:58 -0700] "GET /ePOS/this_category=94&store=1&item_number=1-59038-322-2&form=shared3/gm/detail.htm
141.154.118.110 - - [29/May/2005:04:28:39 -0700] "GET /ePOS/this_category=66&store=1&item_number=1-57345-267-x&form=shared3/gm/detail.htm
210.50.218.113 - - [29/May/2005:04:29:28 -0700] "GET http://www.google.com.au/search?hl=en&q=isaiah-prophet+&meta= HTTP/1.1" 200 85557 "-"
210.50.218.113 - - [29/May/2005:04:29:28 -0700] "GET /ePOS/this_category=89&store=1&item_number=1-57345-942-9&form=shared3/gm/detail.html
210.50.218.113 - - [29/May/2005:04:29:50 -0700] "GET /ePOS/this_category=309&store=439&form=shared3/gm/browse.html&design=1 HTTP/1.1" 200
210.50.218.113 - - [29/May/2005:04:30:13 -0700] "GET /ePOS/this_category=83&store=439&listtype=begin&form=shared3/gm/browse.html&design=1
62.45.82.98 - - [29/May/2005:04:30:50 -0700] "GET http://www.google.nl/search?hl=nl&q=books+mary+ross&meta= HTTP/1.1" 200 85145 "-" "Mozi
62.45.82.98 - - [29/May/2005:04:30:50 -0700] "GET /ePOS/this_category=94&store=1&item_number=1591560608&form=shared3/gm/detail.html&desig
64.124.85.76 - - [29/May/2005:04:31:02 -0700] "GET /ePOS?form=local/gm/detail.html&item_number=1570088772&store=1&associateID=meridian HT
62.45.82.98 - - [29/May/2005:04:31:14 -0700] "GET /ePOS/this_category=242&store=439&form=shared3/gm/main.html&design=1 HTTP/1.1" 200 9564
62.45.82.98 - - [29/May/2005:04:31:25 -0700] "GET /ePOS/width=615&vlink=%230028CA&valign=top&topmargin=0&this_category=242&text=%23000000.
62.45.82.98 - - [29/May/2005:04:31:36 -0700] "GET /ePOS/width=615&vlink=%230028CA&valign=top&topmargin=0&this_category=242&text=%23000000.
62.45.82.98 - - [29/May/2005:04:31:48 -0700] "GET /ePOS/width=615&vlink=%230028CA&valign=top&topmargin=0&this_category=242&text=%23000000.
62.45.82.98 - - [29/May/2005:04:31:57 -0700] "GET /ePOS/width=615&vlink=%230028CA&valign=top&topmargin=0&this_category=242&text=%23000000.
62.45.82.98 - - [29/May/2005:04:32:03 -0700] "GET /ePOS/width=615&vlink=%230028CA&valign=top&topmargin=0&this_category=242&text=%23000000.
62.45.82.98 - - [29/May/2005:04:32:09 -0700] "GET /ePOS/width=615&vlink=%230028CA&valign=top&topmargin=0&this_category=242&text=%23000000.
```

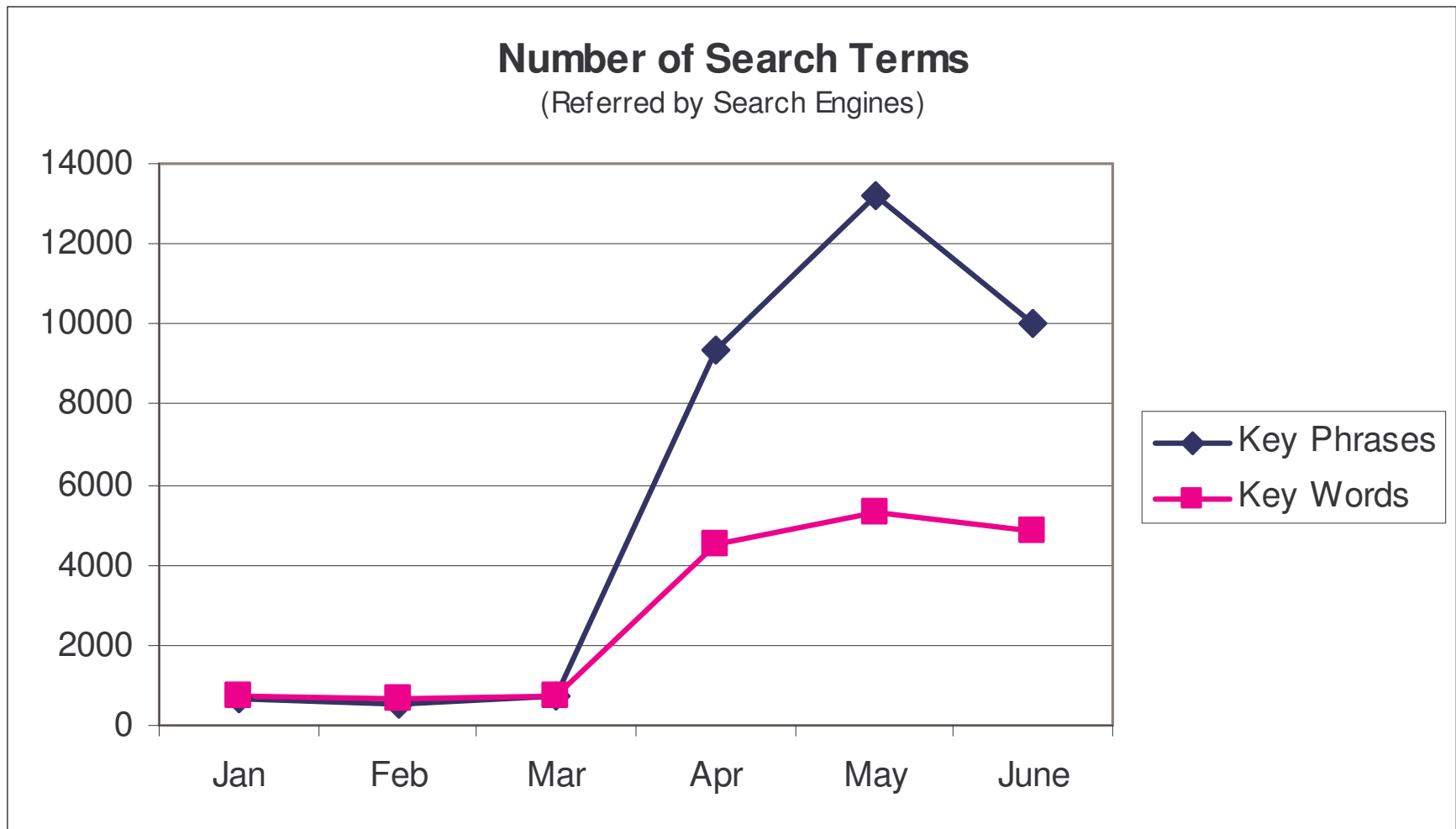


Clickstream Analysis Answers...

- 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





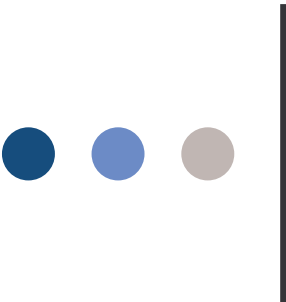
Stage 2: Advanced Web Mining

- ***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

o 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).

$$\frac{\text{(# of customers that completed checkout) } \mathbf{799}}{\text{(# of visitors that viewed final order details) } \mathbf{1414}} = \mathbf{57\%}$$



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

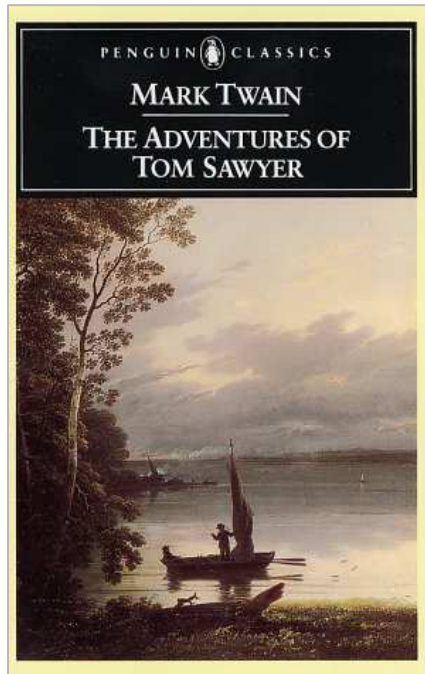
- Collaborative Filtering Based
 - Recommendations/Predictions
 - Books (Amazon)
 - Music (Yahoo! Launchcast)
 - Movies (Yahoo! Movies)
 - Gifts (Yahoo! Shopping)
 - Websites (StumbleUpon), etc.

“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/>)



Customers who bought **?** also bought

- [The Adventures of Huckleberry Finn \(Penguin Classics\)](#) by [Mark Twain](#)
- [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](#)

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

Collaborative Filtering: Music


Example: Yahoo's Launchcast (<http://lauch.yahoo.com/>)





Collaborative Filtering: Movies

Personalized Recommendations for Idmissions
You have rated 34 movies. Rating movies helps us understand your movie taste. Rate more movies to improve your recommendations.

Movies in Theaters
Location: 84604 [[Change Location](#)]

 **Batman Begins** (PG-13)
[Showtimes & Tickets](#) | [Add to My Lists](#)
Yahoo! Users: **B+** 42888 ratings
The Critics: **B+** 15 reviews
 Don't Recommend Again Seen It? Rate It!

 **Robots** (PG)
[Showtimes & Tickets](#) | [Add to My Lists](#)
Yahoo! Users: **B** 17519 ratings
The Critics: **B+** 11 reviews
 Don't Recommend Again Seen It? Rate It!

 **The Hitchhiker's Guide to the Ga...** (PG)
[Showtimes & Tickets](#) | [Add to My Lists](#)
Yahoo! Users: **B-** 17487 ratings
The Critics: **B-** 13 reviews
 Don't Recommend Again Seen It? Rate It!

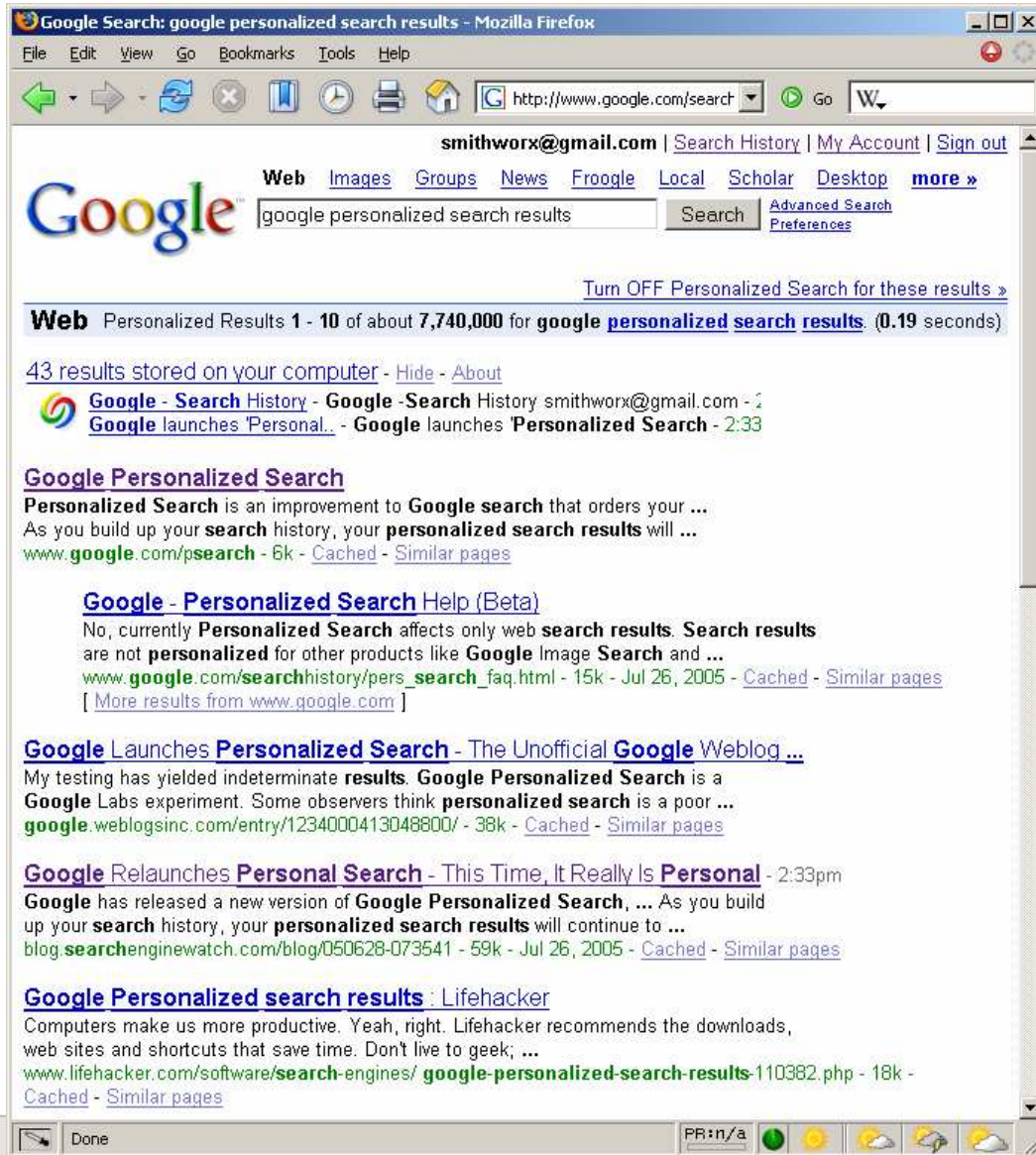
More Recommended Movies in Theaters...

To improve your recommendations, [rate more movies](#)

Still needs work, but is improving.

Personalized Search

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



The screenshot shows a Mozilla Firefox browser window with the address bar at <http://www.google.com/search>. The search bar contains the text "google personalized search results" and the search button is labeled "Search". The page title is "Google Search: google personalized search results - Mozilla Firefox". The search results are for "Personalized Results 1 - 10 of about 7,740,000 for google personalized search results (0.19 seconds)". The first result is "Google - Search History - Google - Search History smithworx@gmail.com - 2" with a snippet "Google launches 'Personal...' - Google launches 'Personalized Search' - 2:33". The second result is "Google Personalized Search" with a snippet "Personalized Search is an improvement to Google search that orders your ... As you build up your search history, your personalized search results will ... www.google.com/psearch - 6k - Cached - Similar pages". The third result is "Google - Personalized Search Help (Beta)" with a snippet "No, currently Personalized Search affects only web search results. Search results are not personalized for other products like Google Image Search and ... www.google.com/searchhistory/pers_search_faq.html - 15k - Jul 26, 2005 - Cached - Similar pages [More results from www.google.com]". The fourth result is "Google Launches Personalized Search - The Unofficial Google Weblog ..." with a snippet "My testing has yielded indeterminate results. Google Personalized Search is a Google Labs experiment. Some observers think personalized search is a poor ... google.weblogsinc.com/entry/1234000413048800/ - 38k - Cached - Similar pages". The fifth result is "Google Relaunches Personal Search - This Time, It Really Is Personal - 2:33pm" with a snippet "Google has released a new version of Google Personalized Search, ... As you build up your search history, your personalized search results will continue to ... blog.searchenginewatch.com/blog/050628-073541 - 59k - Jul 26, 2005 - Cached - Similar pages". The sixth result is "Google Personalized search results : Lifehacker" with a snippet "Computers make us more productive. Yeah, right. Lifehacker recommends the downloads, web sites and shortcuts that save time. Don't live to geek; ... www.lifehacker.com/software/search-engines/google-personalized-search-results-110382.php - 18k - Cached - Similar pages". The browser's status bar at the bottom shows "Done" and "PR:n/a".

“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

