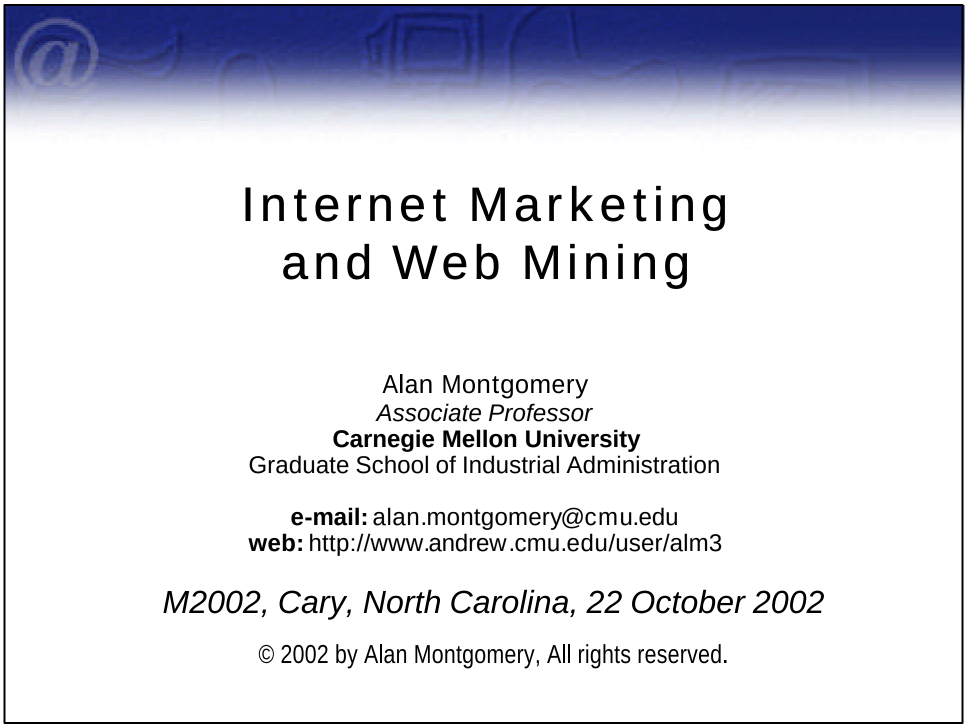
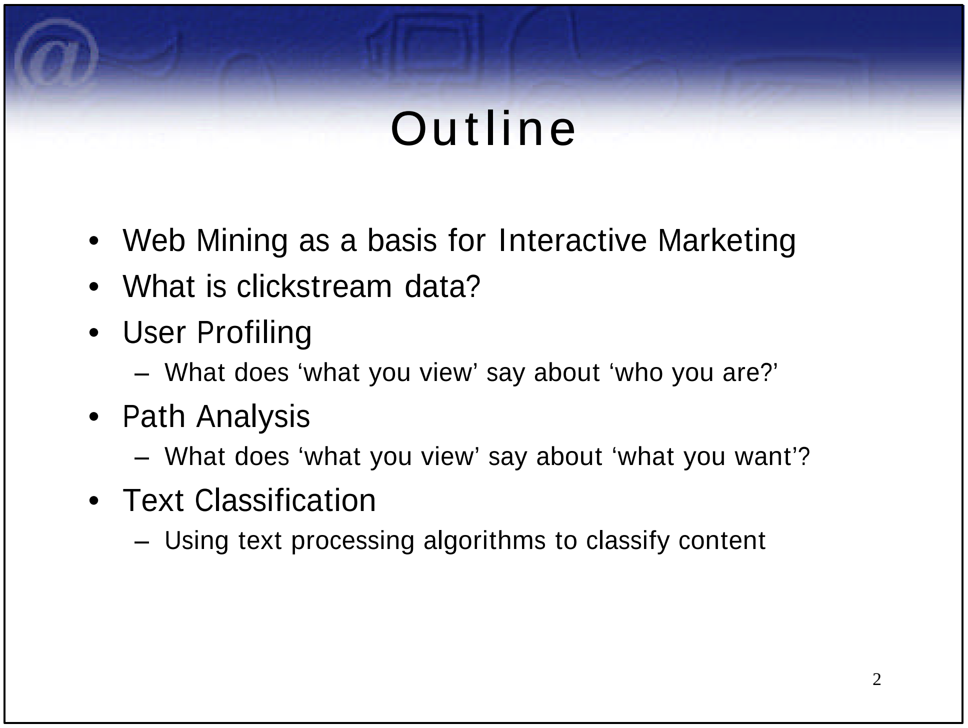


Internet Marketing and Web Mining

Alan Montgomery
Associate Professor
Carnegie Mellon University
Graduate School of Industrial Administration

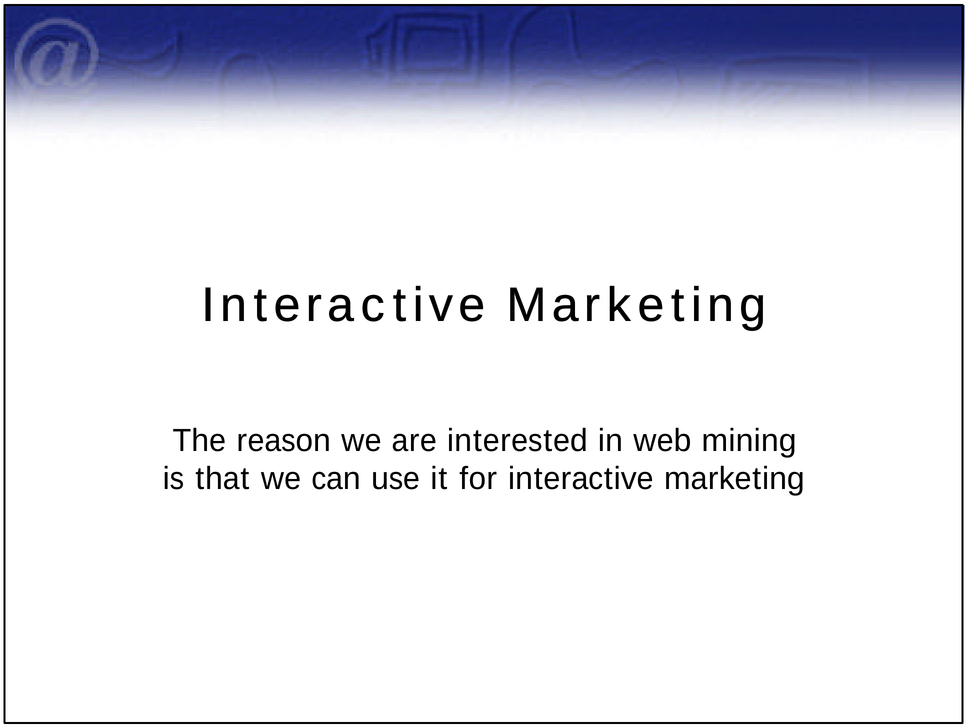
e-mail: alan.montgomery@cmu.edu
web: http://www.andrew.cmu.edu/user/alm3

M2002, Cary, North Carolina, 22 October 2002



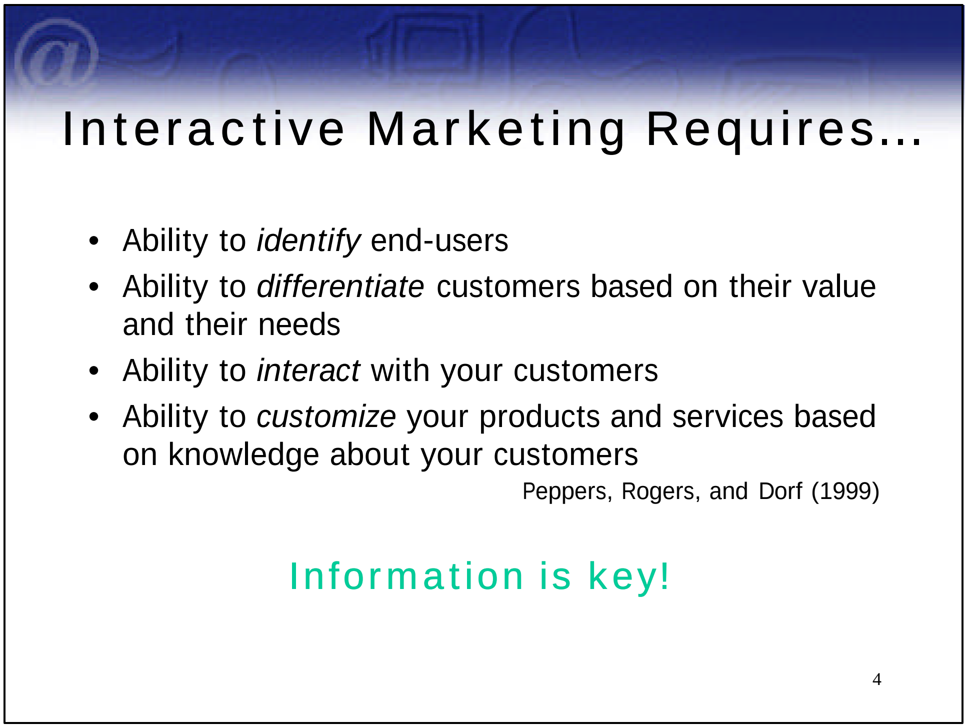
Outline

- Web Mining as a basis for Interactive Marketing
- What is clickstream data?
- User Profiling
  - What does ‘what you view’ say about ‘who you are?’
- Path Analysis
  - What does ‘what you view’ say about ‘what you want’?
- Text Classification
  - Using text processing algorithms to classify content



Interactive Marketing

The reason we are interested in web mining
is that we can use it for interactive marketing



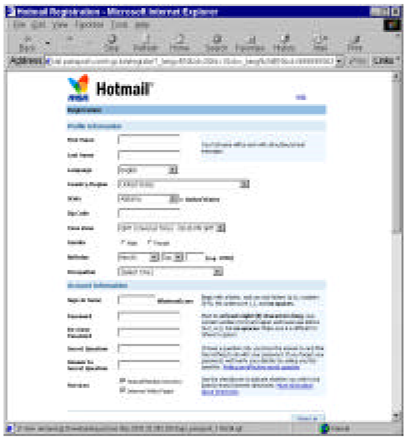
Interactive Marketing Requires...

- Ability to *identify* end-users
- Ability to *differentiate* customers based on their value and their needs
- Ability to *interact* with your customers
- Ability to *customize* your products and services based on knowledge about your customers

Peppers, Rogers, and Dorf (1999)

Information is key!

Active Learning



- Making a user subscribe to your web site
- Linking browsing behavior to offline purchasing (Frequent Flier Accounts, Mailing Lists, ...)

Passive Learning

- How can we learn about a user without asking questions?
- Observe their actions and choices
- Marketing research tells us that the most predictive information is what consumers have done in the past

<u>User Demographics</u>

Sex:	Male
Age:	55
Occupation:	Service Worker
State:	Washington

Learning

- The web is a rich environment for both active and passive
- Most overlook passive because it requires higher degree of sophistication, generally data mining tools
- But can be much more powerful and help fulfill all the promises of interactive marketing

Defining Clickstream Data

The raw input for web mining

What is clickstream data?

- A record of an individual’s movement through time at a web site
- Contains information about:
  - Time
  - URL content
  - User’s machine
  - Previous URL viewed
  - Browser type

Sources of clickstream data

- Web Servers
  - Each hit is recorded in the web server log
- Media Service Providers
  - DoubleClick, Flycast
- ISP/Hosting Services
  - AOL, Juno, Bluelight.com
- Marketing Research Companies
  - ComScore Media Metrix and NetRatings

User Profiling

What does ‘where you go’ say
about ‘who you are’?

"On the Internet, nobody knows you're a dog."

New Yorker, 5 July 1993, p. 61

Is this user male or female?

User visits the following five sites in the Doubleclick network

95% probability that user is female

Bayesian updating formula

Test the hypothesis that a user is female by updating the current guess using new information

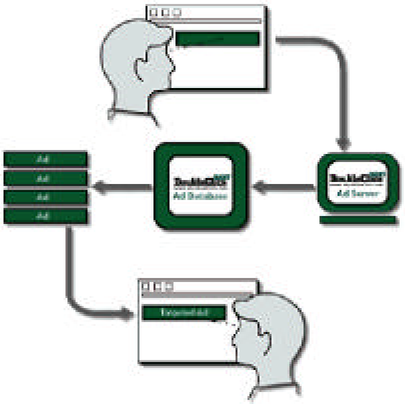
$$\overset{\text{New probability}}{\bar{\bar{p}}} = \frac{\overset{\text{New information}}{p} \cdot \overset{\text{Old probability}}{\bar{p}}}{\underbrace{p \cdot \bar{p}}_{\text{Female}} + \underbrace{(1-p)(1-\bar{p})}_{\text{Male}}}$$

Probability user is female

	Probability a Female Visits the site	*Probability visitor is Female given visits to*
Overall Internet	45%	45.0%
cbs.com	54%	49.0%
ivillage.com	66%	65.1%
libertynet.org	63%	76.0%
nick.com	57%	80.8%
onlinepsych.com	83%	95.4%

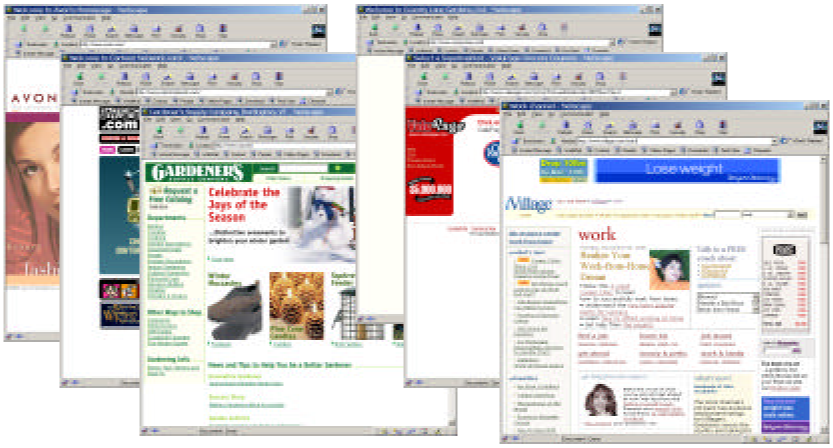
Best Guess

Banner Ad Generation by DoubleClick



Source: http://www.doubleclick.com/publishers/service/how_it_works.htm

What can we learn?



17

A Full Month of Browsing

Example

% of female visitors during one month (Media Metrix):

48%	aol.com	63%	libertynet.org
64%	astronet.com	39%	lycos.com
75%	avon.com	27%	netradio.net
52%	blue-planet.com	57%	nick.com
56%	cartoonnetwork.com	59%	onhealth.com
54%	cbs.com	83%	onlinepsych.com
76%	country-lane.com	44%	simplenet.com
47%	eplay.com	76%	thriveonline.com
41%	halcyon.com	59%	valupage.com
70%	homearts.com	71%	virtualgarden.com
66%	ivillage.com	66%	womenswire.com

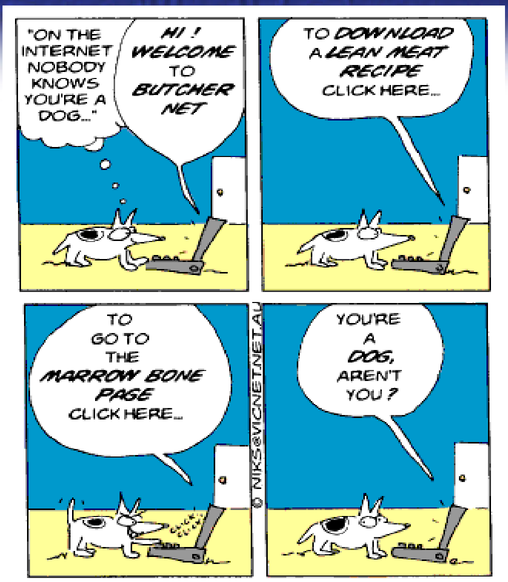
99.97% probability that user is female

18

Key Points of User Profiling

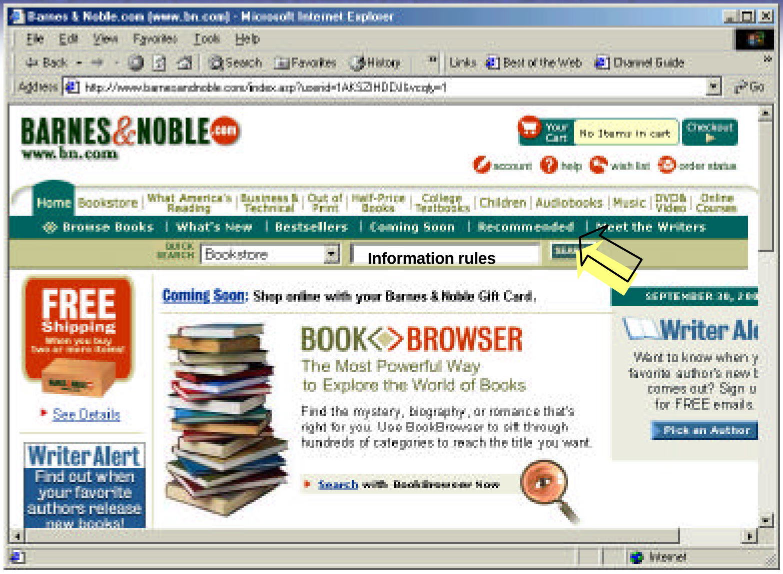
We can identify ‘who you are’ from ‘where you go’

- What the user views on the web reveals their interests and preferences
  - We can personalize the web experience without explicitly requiring customers to login and identify themselves
- Browsing and product choices can reveal key information about interest and price sensitivity
- Requires marketers to be smarter in designing their websites and analyzing their information. Big profitability gains if this is done correctly.


"ON THE INTERNET NOBODY KNOWS YOU'RE A DOG..."
HI ! WELCOME TO BUTCHER NET
TO DOWNLOAD A LEAN MEAT RECIPE CLICK HERE...
TO GO TO THE MARROW BONE PAGE CLICK HERE...
YOU'RE A DOG, AREN'T YOU ?
© NIKS@VICNETNET.AU

Http://www.moreinfo.com/au.cranlerma/fol2.htm

Clickstream Example #1



Barnes & Noble.com (www.bn.com) - Microsoft Internet Explorer

Address: http://www.barnesandnoble.com/index.asp?userid=1AKSZHDDJ&vcqty=1

QUICK SEARCH: Bookstore — Information rules

Coming Soon: Shop online with your Barnes & Noble Gift Card.

BOOK BROWSER

The Most Powerful Way to Explore the World of Books

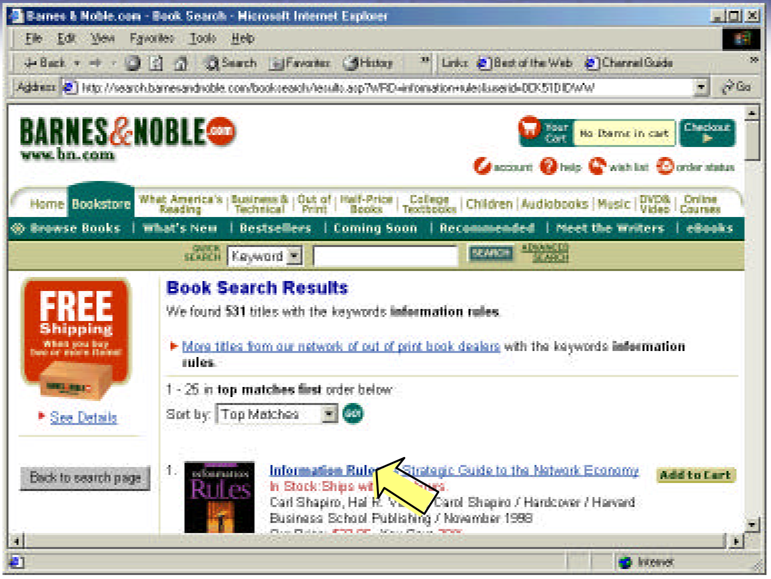
Find the mystery, biography, or romance that's right for you. Use BookBrowser to sift through hundreds of categories to reach the title you want.

Search with BookBrowser Now

Writer Alert: Find out when your favorite authors release new books!

Want to know when y favorite author's new b comes out? Sign u for FREE emails.

File Edit View Favorites Tools Help
Back Search Favorites History Links Best of the Web Channel Guide Go
BARNES&NOBLE.com
www.bn.com
Your Cart
No Items in cart
Checkout
account help wish list order status
Home Bookstore What America's Reading Business & Technical Out of Print Half-Price Books College Textbooks Children Audiobooks Music DVD & Video Online Courses
Browse Books What's New Bestsellers Coming Soon Recommended Meet the Writers
FREE Shipping
When you buy two or more items!
See Details
SEPTEMBER 30, 20
Pick an Author
Internet



Barnes & Noble.com - Book Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links Best of the Web Channel Guide
Address http://search.barnesandnoble.com/booksearch/results.asp?WRD=information+rules&userid=0CK51DIDWW
Go
BARNES&NOBLE.com
www.bn.com
Your Cart No Items in cart Checkout
account help wish list order status
Home Bookstore What America's Reading Business & Technical Out of Print Half-Price Books College Textbooks Children Audiobooks Music DVD& Video Online Courses
Browse Books | What's New | Bestsellers | Coming Soon | Recommended | Meet the Writers | eBooks
QUICK SEARCH Keyword SEARCH ADVANCED SEARCH
FREE Shipping
See Details
Back to search page
Add to Cart
Internet

Book Search Results

We found **531** titles with the keywords **information rules**.

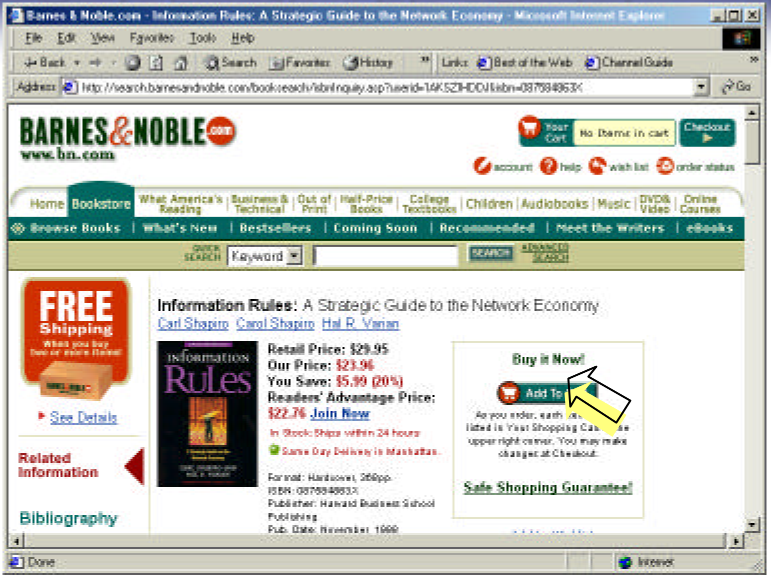
More titles from our network of out of print book dealers with the keywords **information rules**.

1 - 25 in **top matches first** order below.

Sort by: Top Matches GO!

1. Information Rules: A Strategic Guide to the Network Economy
In Stock: Ships wi... ...hours.
Carl Shapiro, Hal R. V... Carol Shapiro / Hardcover / Harvard Business School Publishing / November 1998

23



Barnes & Noble.com - Information Rules: A Strategic Guide to the Network Economy - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links Best of the Web Channel Guide
Address http://search.barnesandnoble.com/booksearch/isbninquiry.asp?userid=14K5ZIHDDJ&isbn=087584863X
Go
BARNES&NOBLE.com
www.bn.com
Your Cart No Items in cart Checkout
account help wish list order status
Home Bookstore What America's Reading Business & Technical Out of Print Half-Price Books College Textbooks Children Audiobooks Music DVD& Video Online Courses
Browse Books | What's New | Bestsellers | Coming Soon | Recommended | Meet the Writers | eBooks
QUICK SEARCH Keyword SEARCH ADVANCED SEARCH
FREE Shipping
See Details
Related Information
Bibliography
Done
Internet

Information Rules: A Strategic Guide to the Network Economy

Carl Shapiro Carol Shapiro Hal R. Varian

Retail Price: $29.95
Our Price: $23.96
You Save: $5.99 (20%)
Readers' Advantage Price: $22.76 Join Now

In Stock: Ships within 24 hours.
Same Day Delivery in Manhattan.

Format: Hardcover, 368pp.
ISBN: 087584863X
Publisher: Harvard Business School Publishing
Pub. Date: November 1998

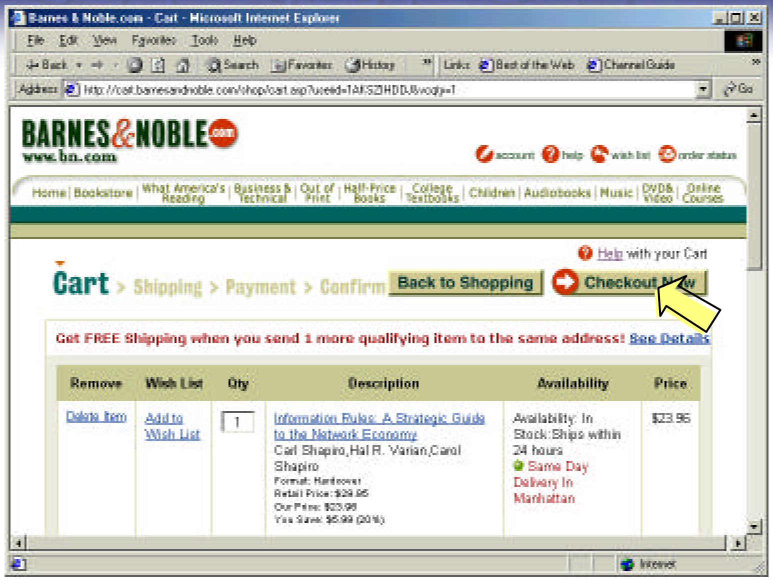
Buy it Now!

Add To...

As you order, each ... listed in Your Shopping Ca... ...e upper right corner. You may make changes at Checkout.

Safe Shopping Guarantee!

24

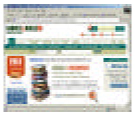

Barnes & Noble.com - Cart - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links Best of the Web Channel Guide
Address http://cart.barnesandnoble.com/shop/cart.asp?uceid=1AI!SZIHDDJ&vcqly=1
Go
BARNES&NOBLE.com
www.bn.com
account help wish list order status
Home | Bookstore | What America's Reading | Business & Technical | Out of Print | Half-Price Books | College Textbooks | Children | Audiobooks | Music | DVD& Video | Online Courses
Help with your Cart
Cart > Shipping > Payment > Confirm
Back to Shopping
Checkout Now
Internet

Get FREE Shipping when you send 1 more qualifying item to the same address! See Details

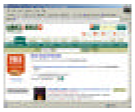
Remove	Wish List	Qty	Description	Availability	Price
Delete Item	Add to Wish List	1	Information Rules: A Strategic Guide to the Network Economy Carl Shapiro,Hal R. Varian,Carol Shapiro Format: Hardcover Retail Price: $29.95 Our Price: $23.96 You Save: $5.99 (20%)	Availability: In Stock:Ships within 24 hours Same Day Delivery In Manhattan	$23.96

Predicting Purchase Conversion

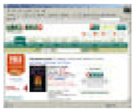
Home



Category



Product



Shopping Cart

What is the chance of this user making a purchase during this session?

1st viewing = 7%

2nd viewing = 14%

3rd viewing = 20%

4th viewing = 60%

Clickstream Example #2

Will this user buy?

{Home}

{Category}

{Category}

{Category}

{Shop Cart}

{Account}

.

.

.

User 1 Demographics

Sex:	Male
Age:	55
Occupation:	Service Worker
State:	Washington

Clickstream Example #3

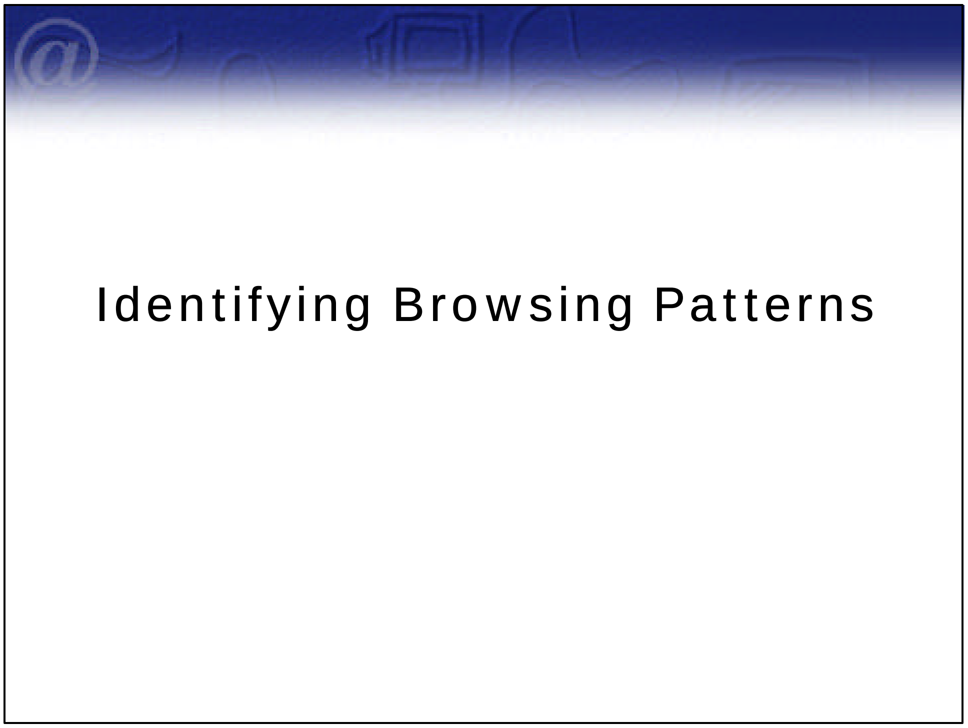
Will this user buy?

No Purchase

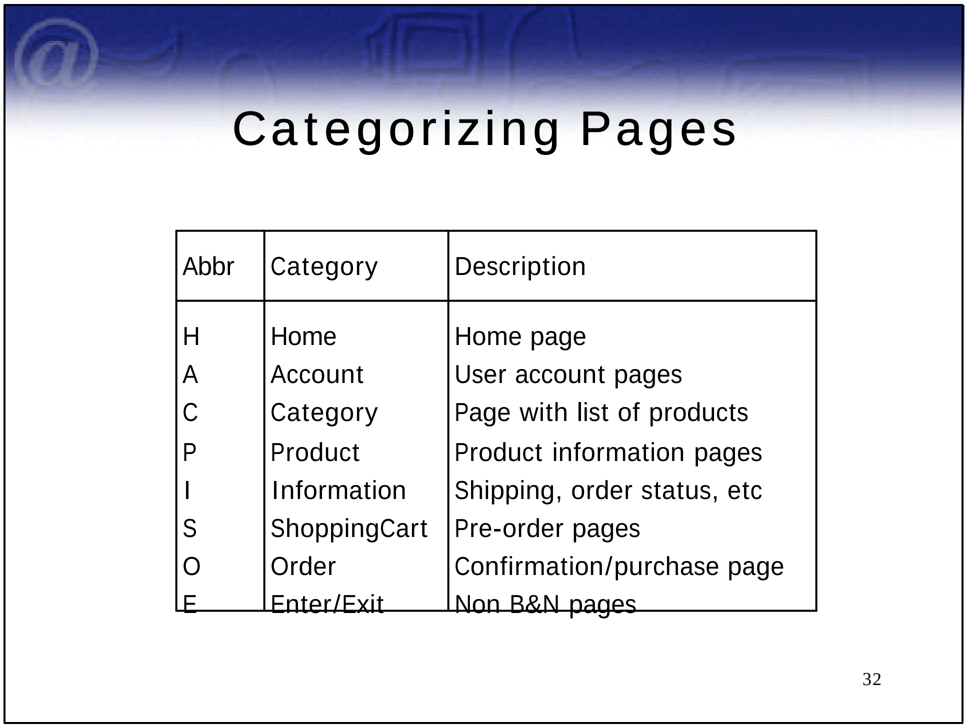
{Home}
{Information}
{Home}
{Information}
{Category}
{Category}
.
.
.

User 2 Demographics

Sex:	Female
Age:	17
Occupation:	Student
State:	Virginia



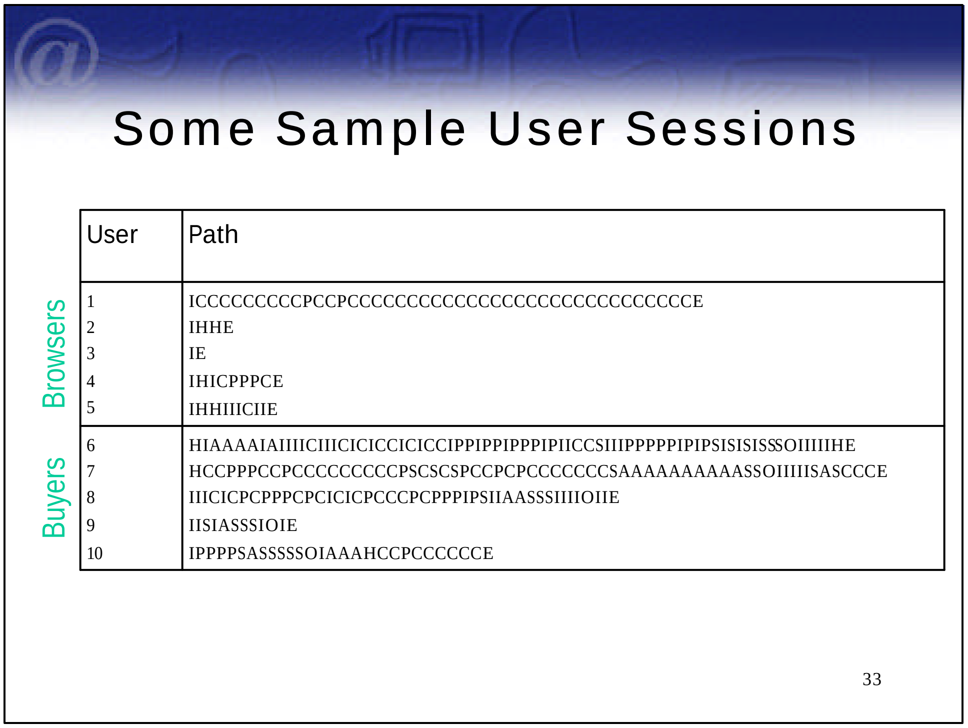
Identifying Browsing Patterns



Categorizing Pages

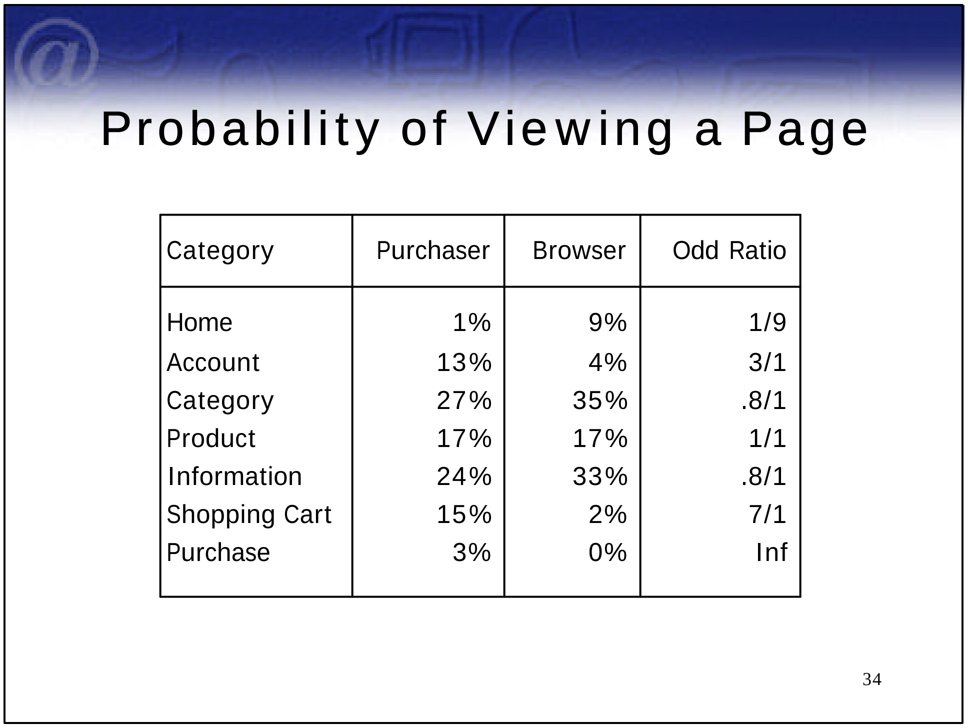
Abbr	Category	Description
H	Home	Home page
A	Account	User account pages
C	Category	Page with list of products
P	Product	Product information pages
I	Information	Shipping, order status, etc
S	ShoppingCart	Pre-order pages
O	Order	Confirmation/purchase page
E	Enter/Exit	Non B&N pages

32



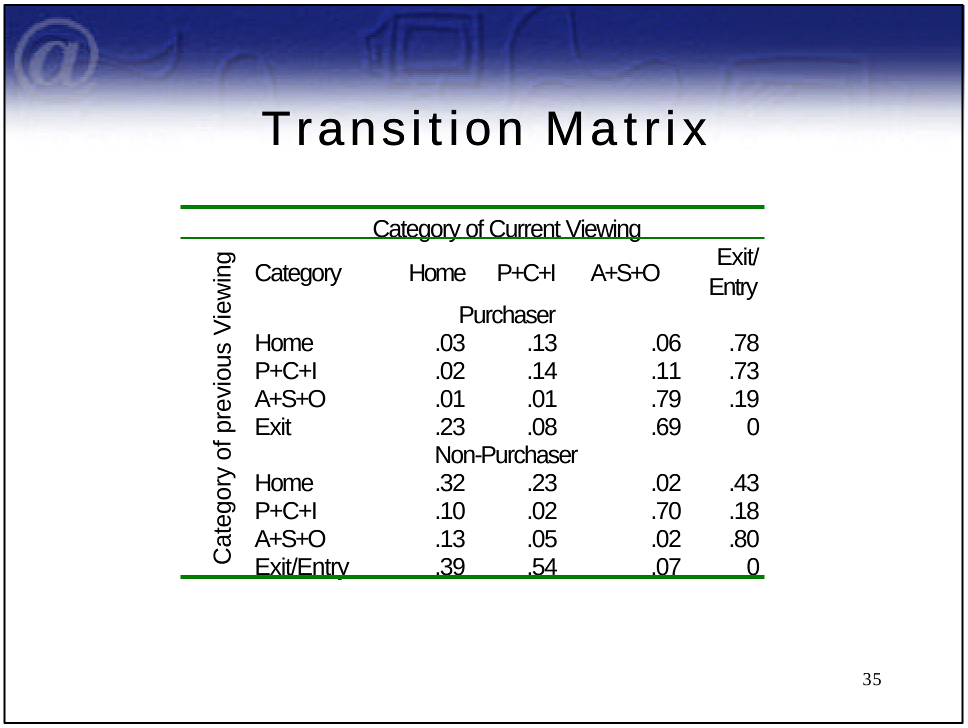
Some Sample User Sessions

	User	Path
Browsers	1	ICCCCCCCCCPCCPCCCCCCCCCCCCCCCCCCCCCCCCCCCCCE
	2	IHHE
	3	IE
	4	IHICPPPCE
	5	IHHIIICIIE
Buyers	6	HIAAAAIAIIIICIIICICICCICICCIPPIPPIPPPIPIICCSIIIPPPPPIPIPSISISISSSOIIIIIHE
	7	HCCPPPCCPCCCCCCCCCPSCSCSPCCPCPCCCCCCCSAAAAAAAAAASSOIIIIISASCCCE
	8	IIICICPCPPPCPCICICPCCCPCPPPIPSIIAASSSIIIIOIIE
	9	IISIASSSIOIE
	10	IPPPPSASSSSSOIAAAHCCPCCCCCCE



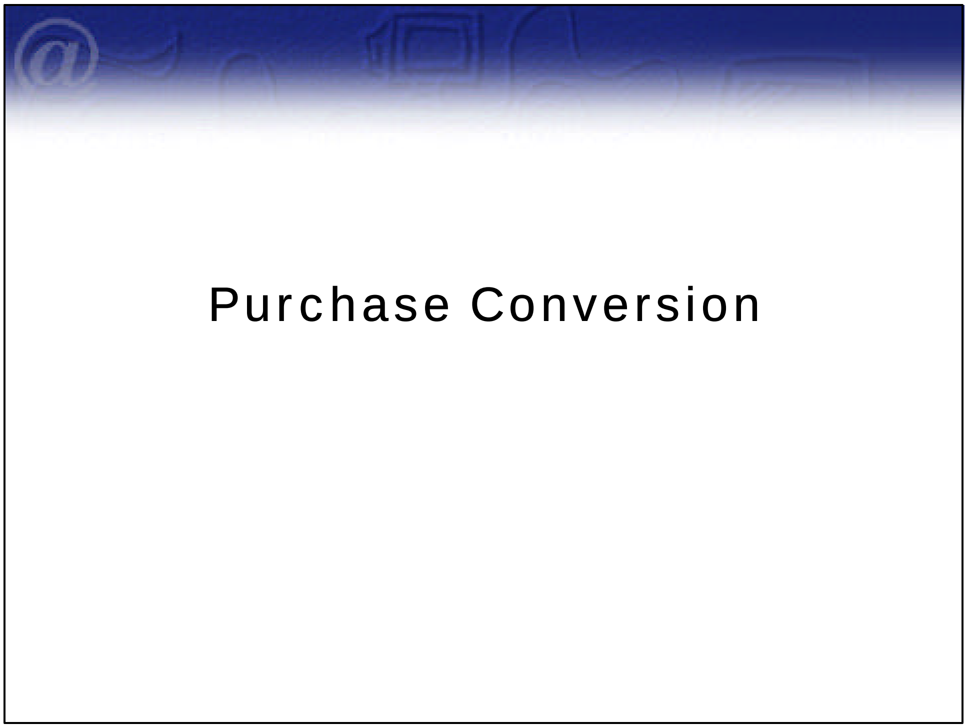
Probability of Viewing a Page

Category	Purchaser	Browser	Odd Ratio
Home	1%	9%	1/9
Account	13%	4%	3/1
Category	27%	35%	.8/1
Product	17%	17%	1/1
Information	24%	33%	.8/1
Shopping Cart	15%	2%	7/1
Purchase	3%	0%	Inf



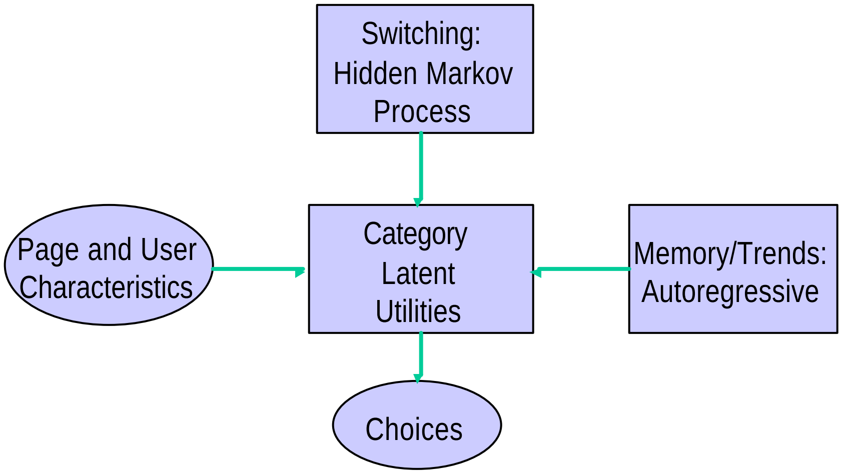
Transition Matrix

		Category of Current Viewing			
Category of previous Viewing	Category	Home	P+C+I	A+S+O	Exit/ Entry
		Purchaser			
	Home	.03	.13	.06	.78
	P+C+I	.02	.14	.11	.73
	A+S+O	.01	.01	.79	.19
	Exit	.23	.08	.69	0
		Non-Purchaser			
	Home	.32	.23	.02	.43
	P+C+I	.10	.02	.70	.18
	A+S+O	.13	.05	.02	.80
	Exit/Entry	.39	.54	.07	0



Purchase Conversion

Describing the Model



Switching: Hidden Markov Process

Page and User Characteristics

Category Latent Utilities

Memory/Trends: Autoregressive

Choices

Will this user buy?

13.8%

12.3%

13.2%

14.3%

35.3%

52.4%

.
.
.

User 1 Demographics

Sex:	Male
Age:	55
Occupation:	Service Worker
State:	Washington

Will this user buy?

0.24%

0.26%

0.06%

0.05%

0.04%

0.03%

.

.

.

User 2 Demographics

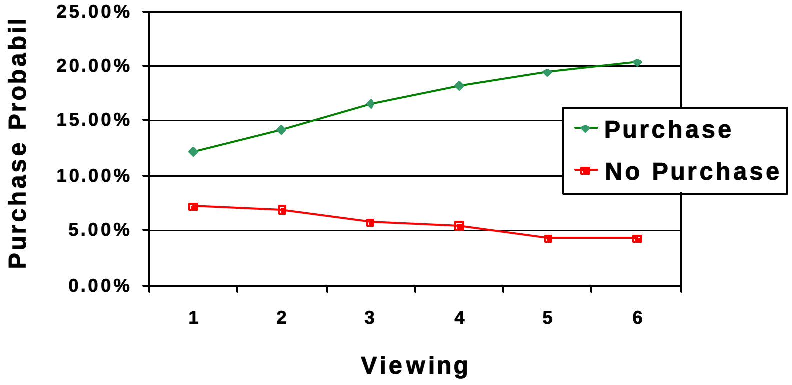
Sex: Female

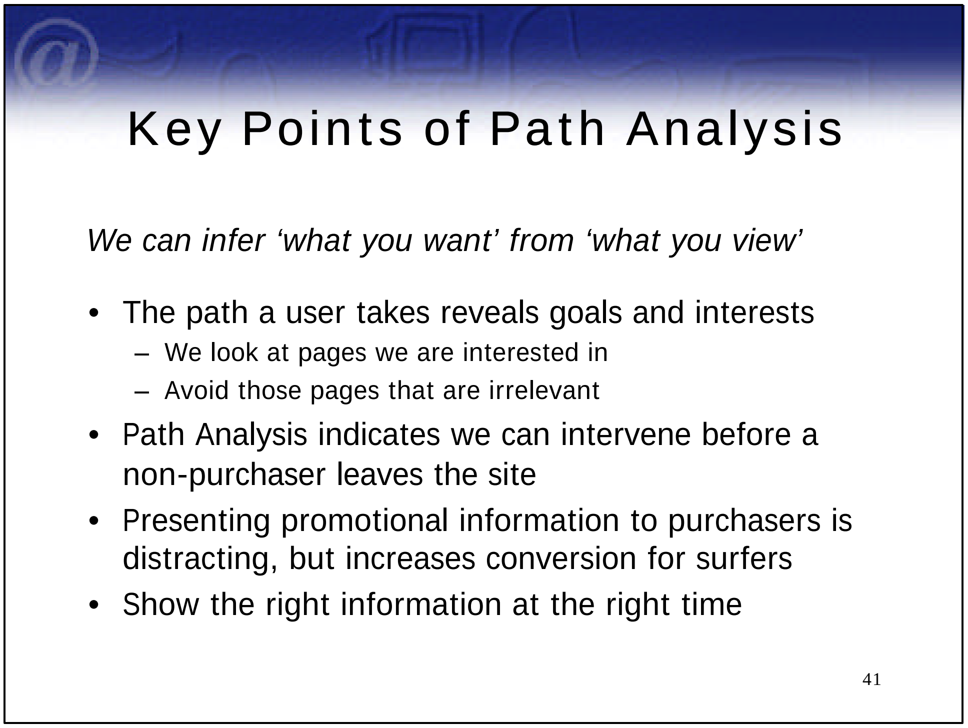
Age: 17

Occupation: Student

State: Virginia

Predicting Purchase Conversion

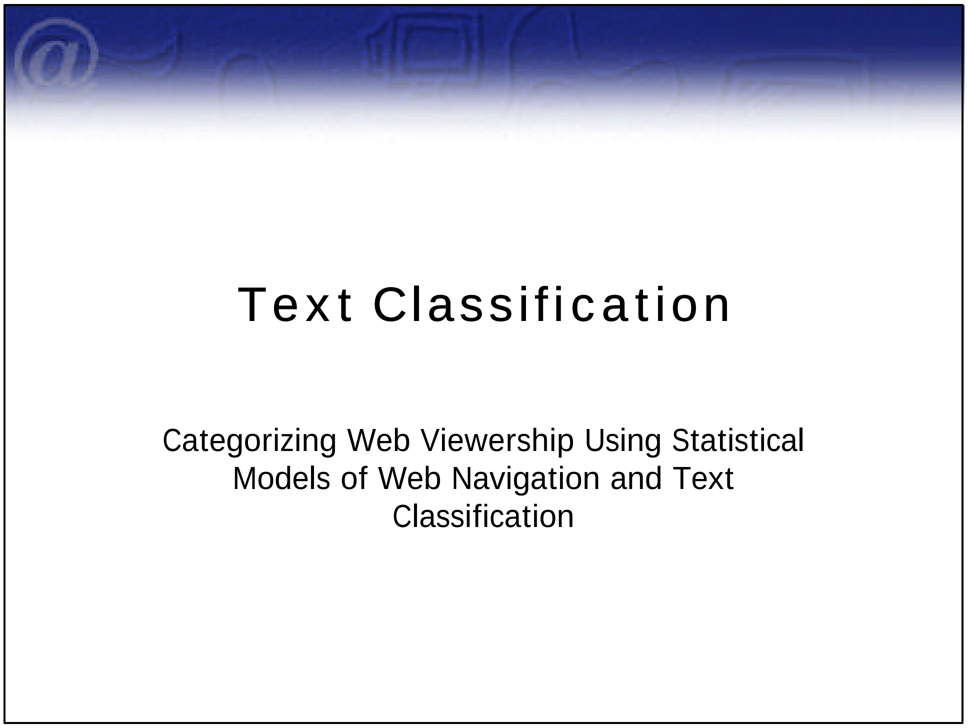




Key Points of Path Analysis

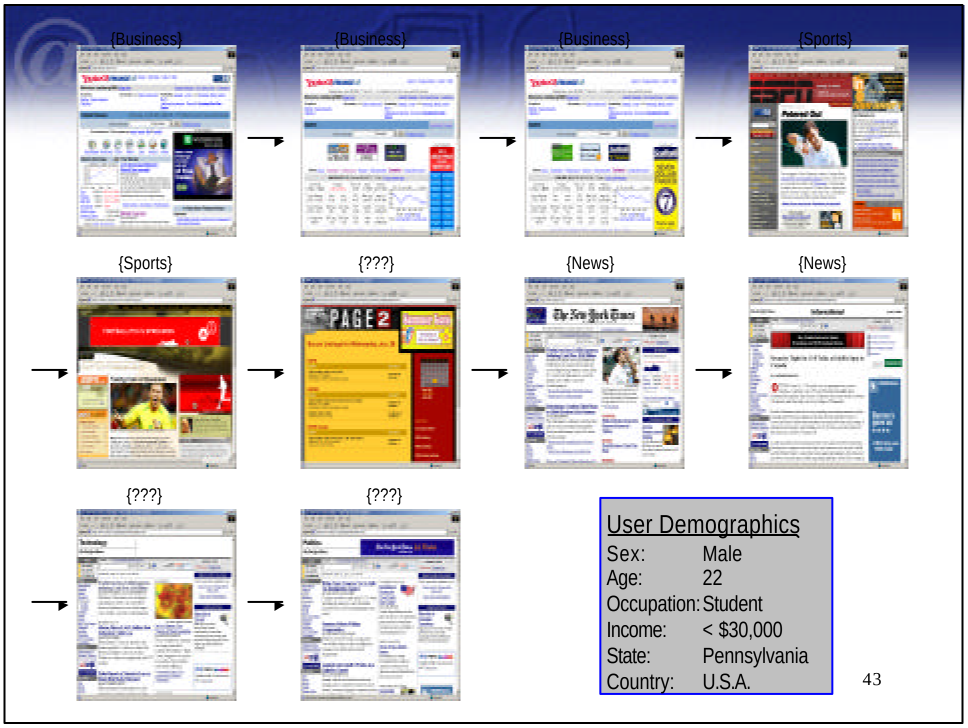
We can infer ‘what you want’ from ‘what you view’

- The path a user takes reveals goals and interests
  - We look at pages we are interested in
  - Avoid those pages that are irrelevant
- Path Analysis indicates we can intervene before a non-purchaser leaves the site
- Presenting promotional information to purchasers is distracting, but increases conversion for surfers
- Show the right information at the right time



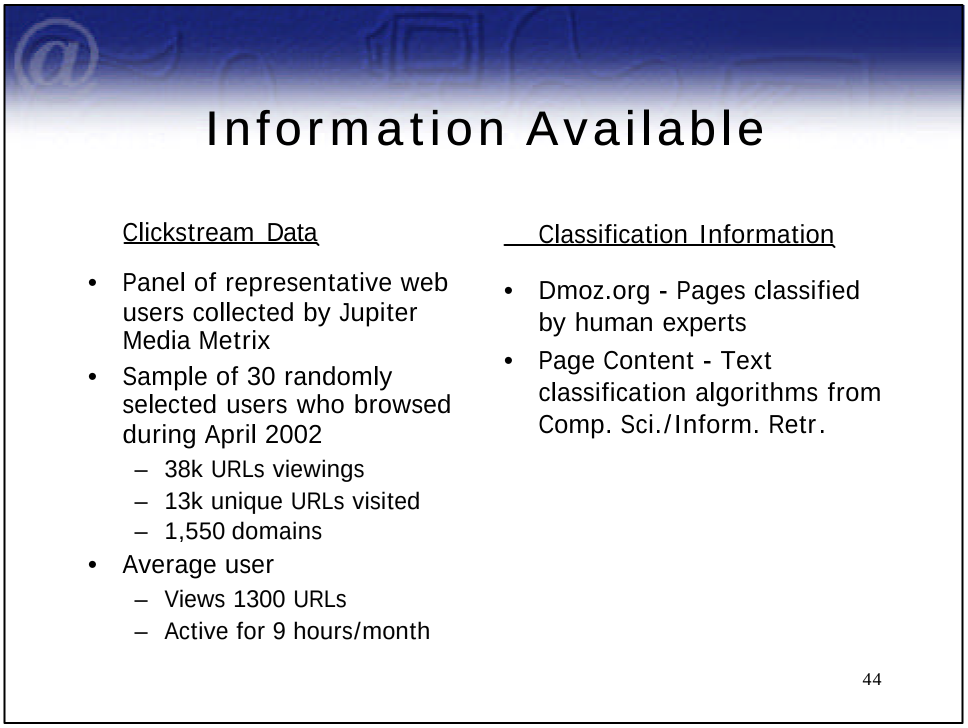
Text Classification

Categorizing Web Viewership Using Statistical
Models of Web Navigation and Text
Classification



User Demographics

Sex:	Male
Age:	22
Occupation:	Student
Income:	< $30,000
State:	Pennsylvania
Country:	U.S.A.



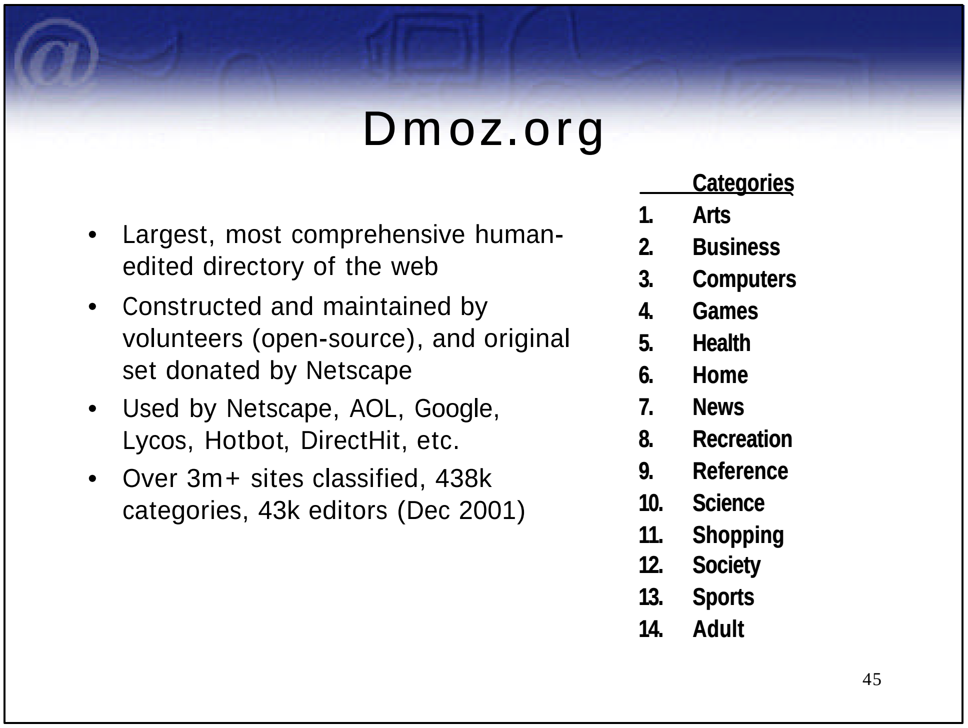
Information Available

Clickstream Data

- Panel of representative web users collected by Jupiter Media Metrix
- Sample of 30 randomly selected users who browsed during April 2002
  - 38k URLs viewings
  - 13k unique URLs visited
  - 1,550 domains
- Average user
  - Views 1300 URLs
  - Active for 9 hours/month

Classification Information

- Dmoz.org - Pages classified by human experts
- Page Content - Text classification algorithms from Comp. Sci./Inform. Retr.

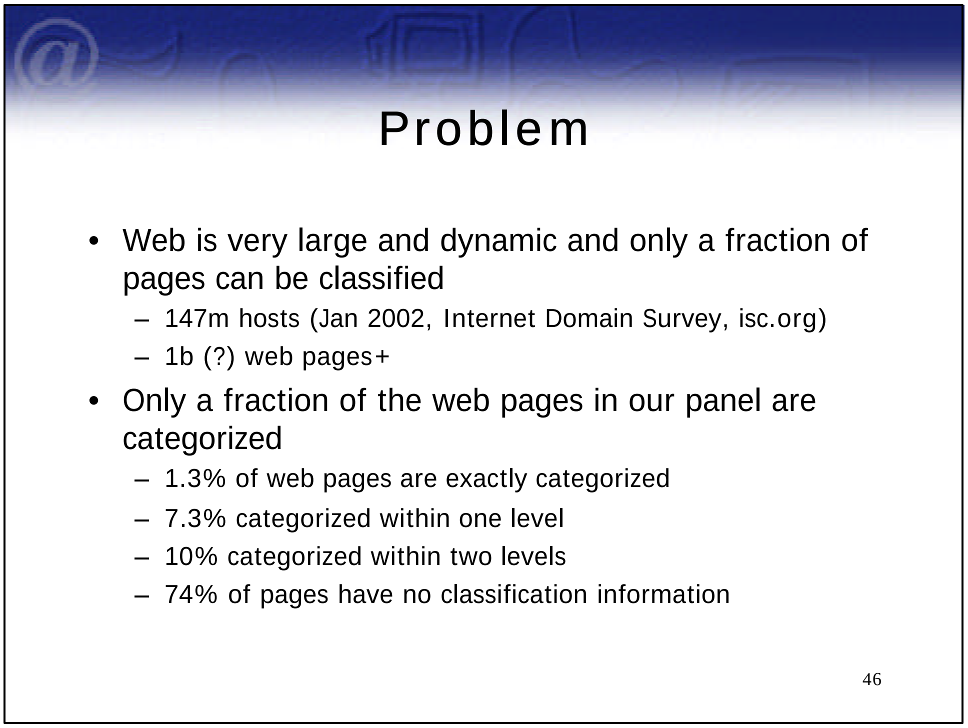


Dmoz.org

- Largest, most comprehensive human-edited directory of the web
- Constructed and maintained by volunteers (open-source), and original set donated by Netscape
- Used by Netscape, AOL, Google, Lycos, Hotbot, DirectHit, etc.
- Over 3m+ sites classified, 438k categories, 43k editors (Dec 2001)

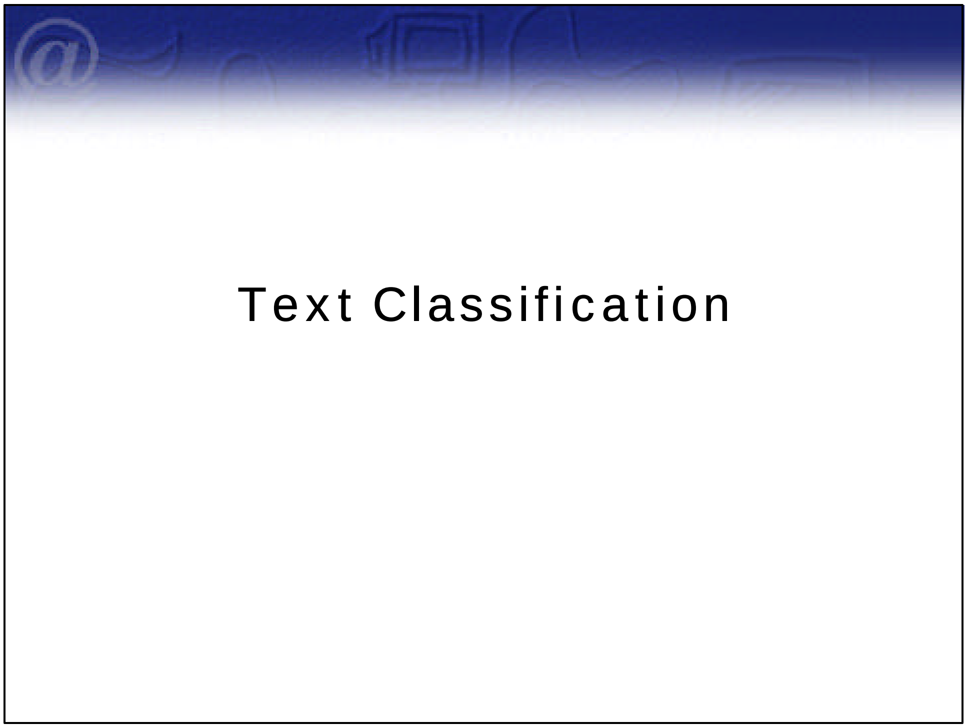
Categories

1. **Arts**
2. **Business**
3. **Computers**
4. **Games**
5. **Health**
6. **Home**
7. **News**
8. **Recreation**
9. **Reference**
10. **Science**
11. **Shopping**
12. **Society**
13. **Sports**
14. **Adult**

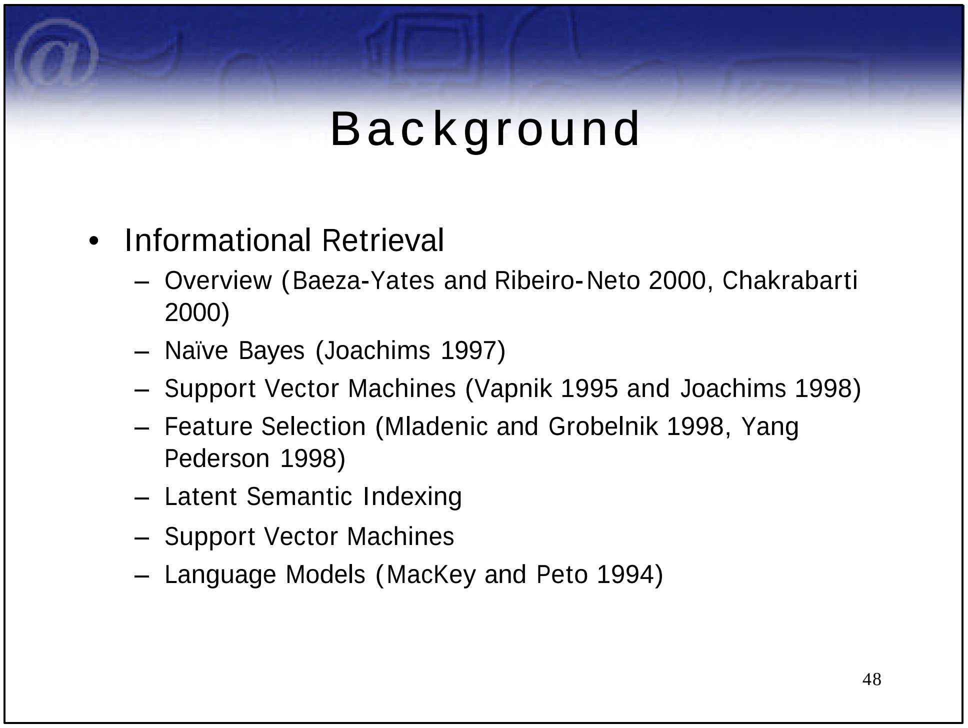


Problem

- Web is very large and dynamic and only a fraction of pages can be classified
  - 147m hosts (Jan 2002, Internet Domain Survey, isc.org)
  - 1b (?) web pages+
- Only a fraction of the web pages in our panel are categorized
  - 1.3% of web pages are exactly categorized
  - 7.3% categorized within one level
  - 10% categorized within two levels
  - 74% of pages have no classification information



Text Classification



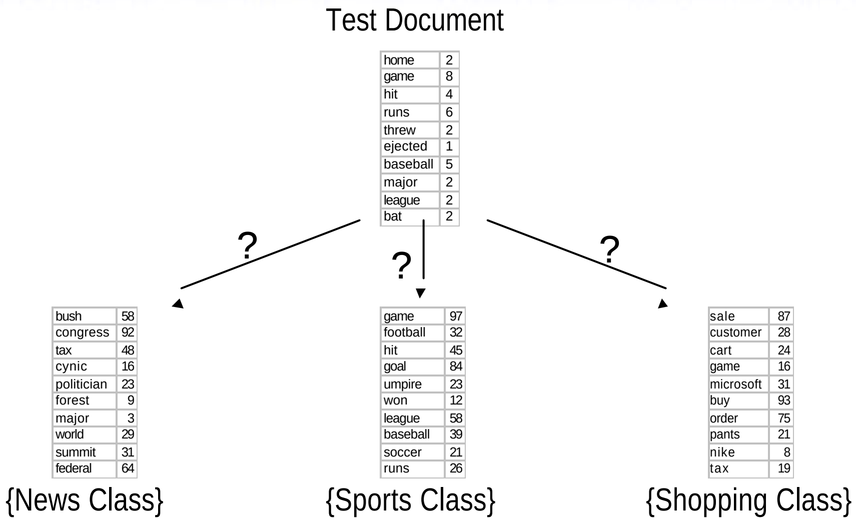
Background

- Informational Retrieval
  - Overview (Baeza-Yates and Ribeiro-Neto 2000, Chakrabarti 2000)
  - Naïve Bayes (Joachims 1997)
  - Support Vector Machines (Vapnik 1995 and Joachims 1998)
  - Feature Selection (Mladenic and Grobelnik 1998, Yang Pederson 1998)
  - Latent Semantic Indexing
  - Support Vector Machines
  - Language Models (MacKey and Peto 1994)

Result: Document Vector

home	2
game	8
hit	4
runs	6
threw	2
ejected	1
baseball	5
major	2
league	2
bat	2

Classifying Document Vectors



Test Document

home	2
game	8
hit	4
runs	6
threw	2
ejected	1
baseball	5
major	2
league	2
bat	2

bush	58
congress	92
tax	48
cynic	16
politician	23
forest	9
major	3
world	29
summit	31
federal	64

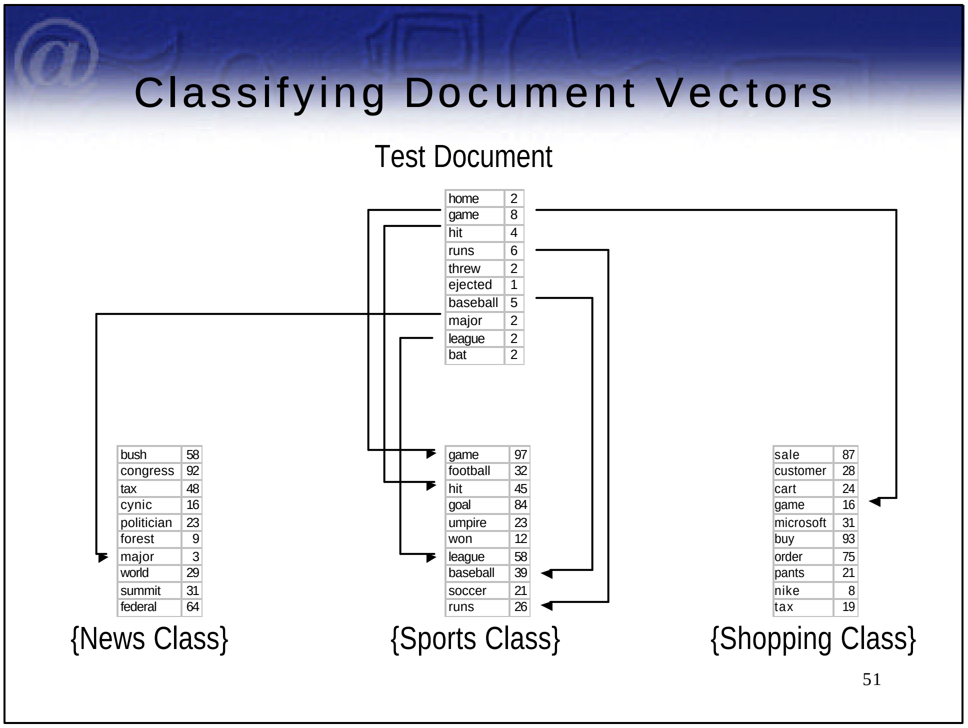
{News Class}

game	97
football	32
hit	45
goal	84
umpire	23
won	12
league	58
baseball	39
soccer	21
runs	26

{Sports Class}

sale	87
customer	28
cart	24
game	16
microsoft	31
buy	93
order	75
pants	21
nike	8
tax	19

{Shopping Class}



Classifying Document Vectors

Test Document

home	2
game	8
hit	4
runs	6
threw	2
ejected	1
baseball	5
major	2
league	2
bat	2

bush	58
congress	92
tax	48
cynic	16
politician	23
forest	9
major	3
world	29
summit	31
federal	64

{News Class}

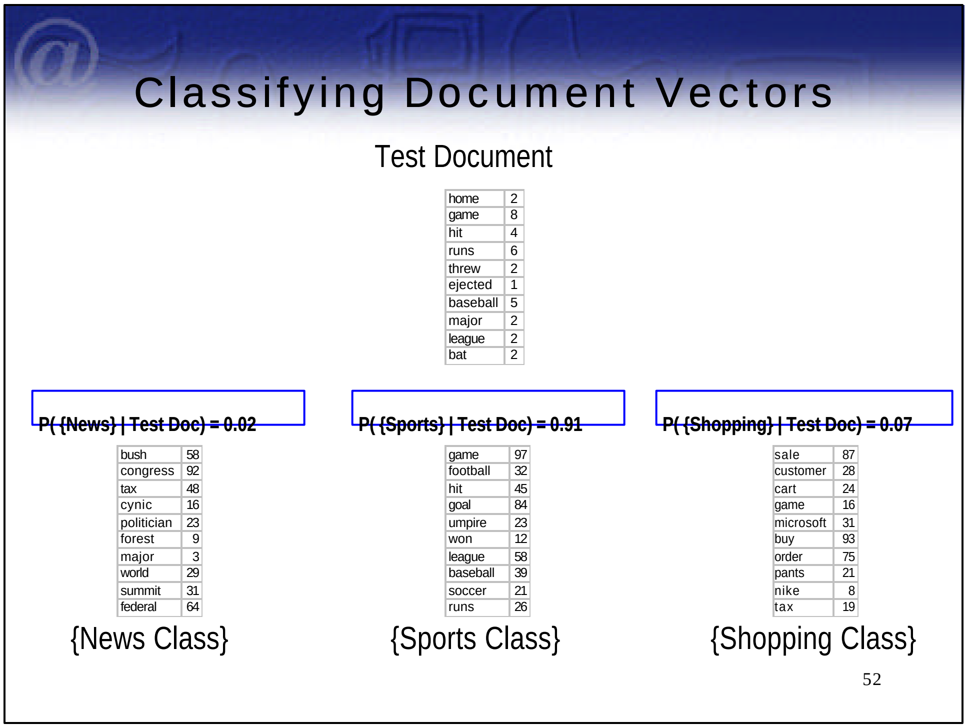
game	97
football	32
hit	45
goal	84
umpire	23
won	12
league	58
baseball	39
soccer	21
runs	26

{Sports Class}

sale	87
customer	28
cart	24
game	16
microsoft	31
buy	93
order	75
pants	21
nike	8
tax	19

{Shopping Class}

51



Classifying Document Vectors

Test Document

home	2
game	8
hit	4
runs	6
threw	2
ejected	1
baseball	5
major	2
league	2
bat	2

P({News} | Test Doc) = 0.02

bush	58
congress	92
tax	48
cynic	16
politician	23
forest	9
major	3
world	29
summit	31
federal	64

{News Class}

P({Sports} | Test Doc) = 0.91

game	97
football	32
hit	45
goal	84
umpire	23
won	12
league	58
baseball	39
soccer	21
runs	26

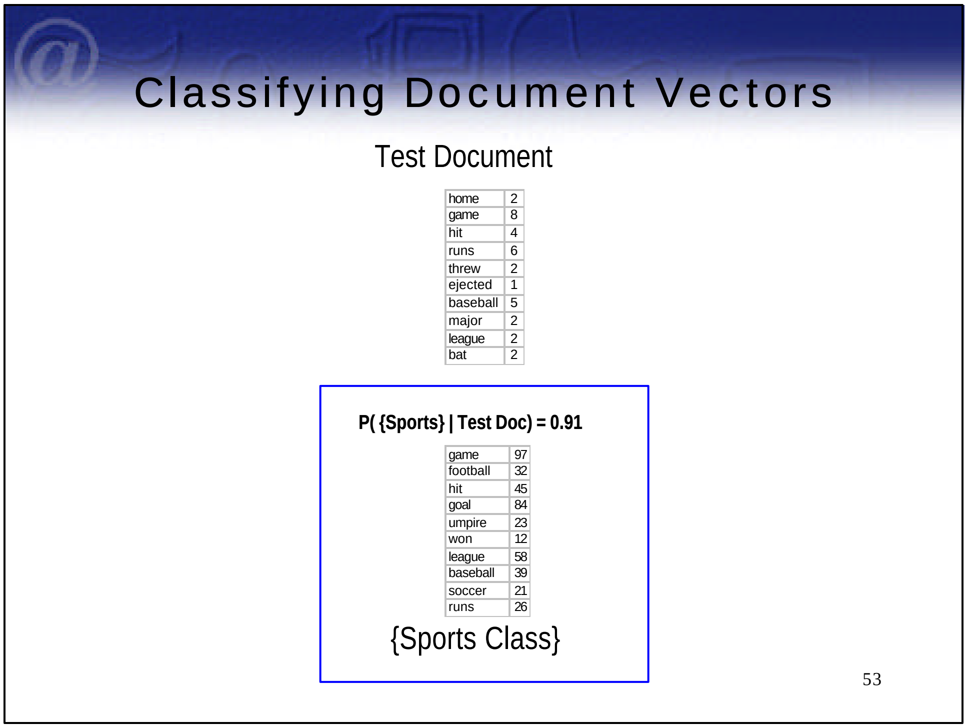
{Sports Class}

P({Shopping} | Test Doc) = 0.07

sale	87
customer	28
cart	24
game	16
microsoft	31
buy	93
order	75
pants	21
nike	8
tax	19

{Shopping Class}

52



Classifying Document Vectors

Test Document

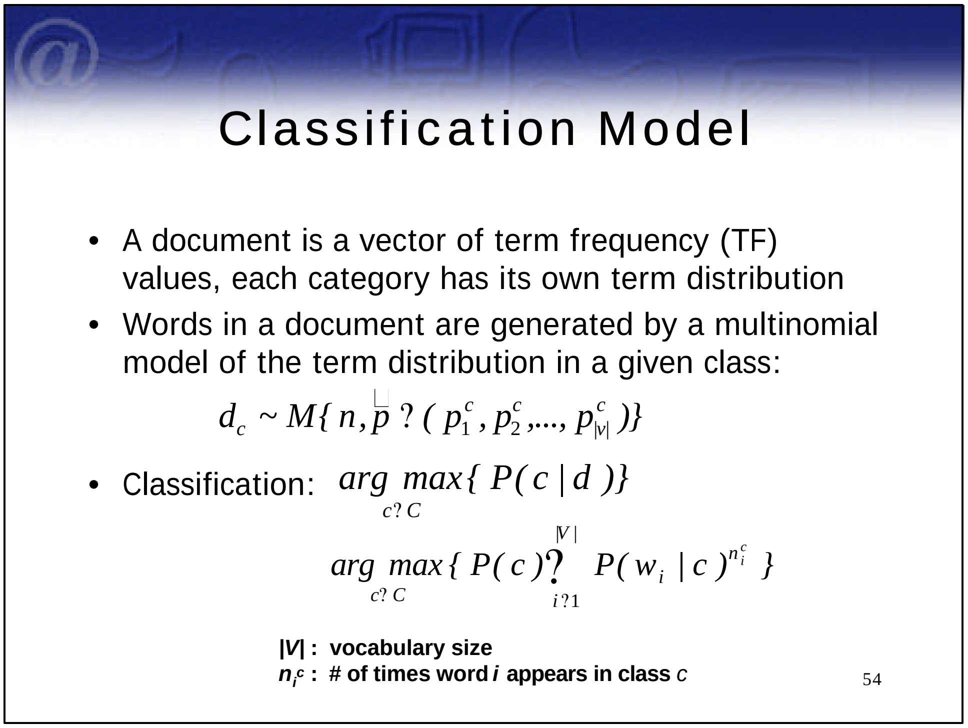
home	2
game	8
hit	4
runs	6
threw	2
ejected	1
baseball	5
major	2
league	2
bat	2

P({Sports} | Test Doc) = 0.91

game	97
football	32
hit	45
goal	84
umpire	23
won	12
league	58
baseball	39
soccer	21
runs	26

{Sports Class}

53



Classification Model

- A document is a vector of term frequency (TF) values, each category has its own term distribution
- Words in a document are generated by a multinomial model of the term distribution in a given class:

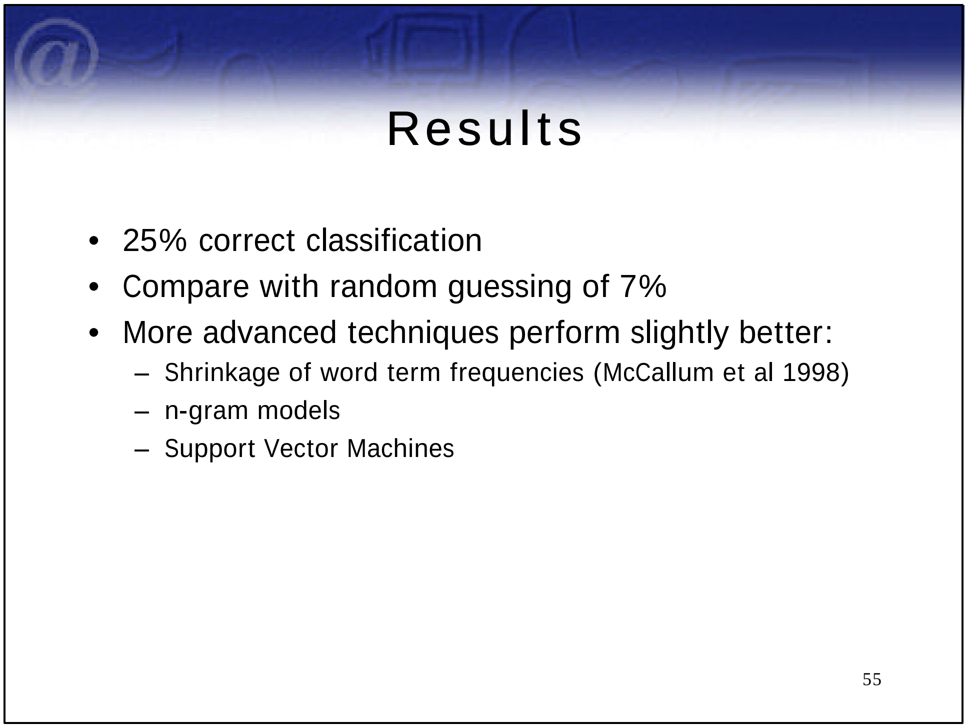
$$d_c \sim M\{ n, \vec{p} \; ? \; (p_1^c , p_2^c ,..., p_{|V|}^c)\}$$

- Classification: $\underset{c ? C}{arg\ max}\{ P(c \,|\, d)\}$

$$\underset{c ? C}{arg\ max}\{ P(c) \underset{i ? 1}{\overset{|V|}{?}} P(w_i \,|\, c)^{n_i^c} \}$$

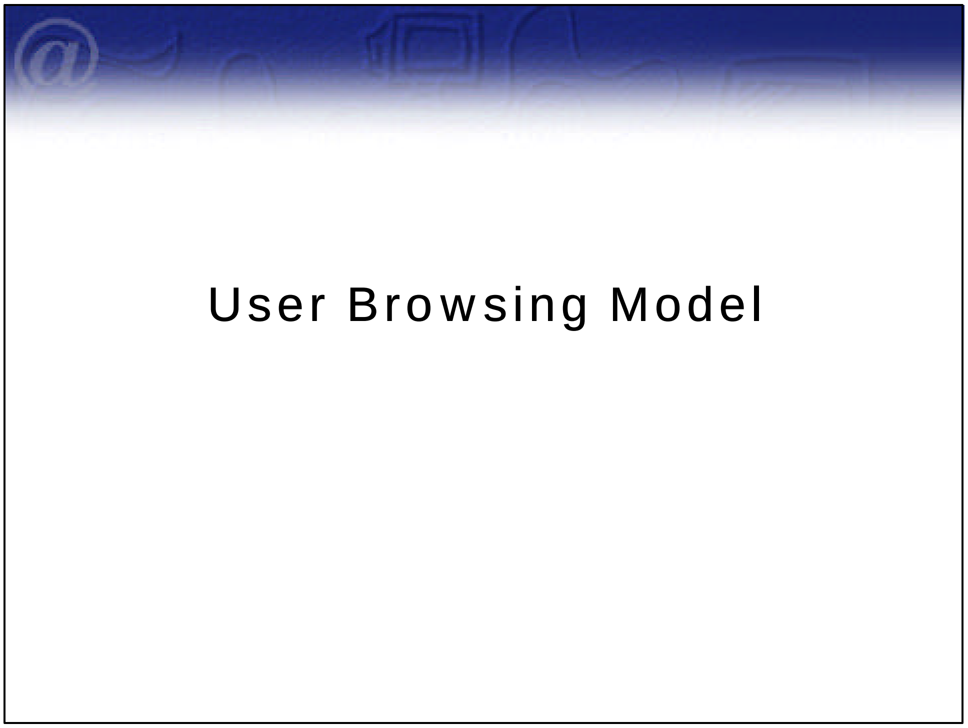
|*V*| : vocabulary size
n_i^c* : # of times word *i* appears in class *c

54



Results

- 25% correct classification
- Compare with random guessing of 7%
- More advanced techniques perform slightly better:
  - Shrinkage of word term frequencies (McCallum et al 1998)
  - n-gram models
  - Support Vector Machines



User Browsing Model

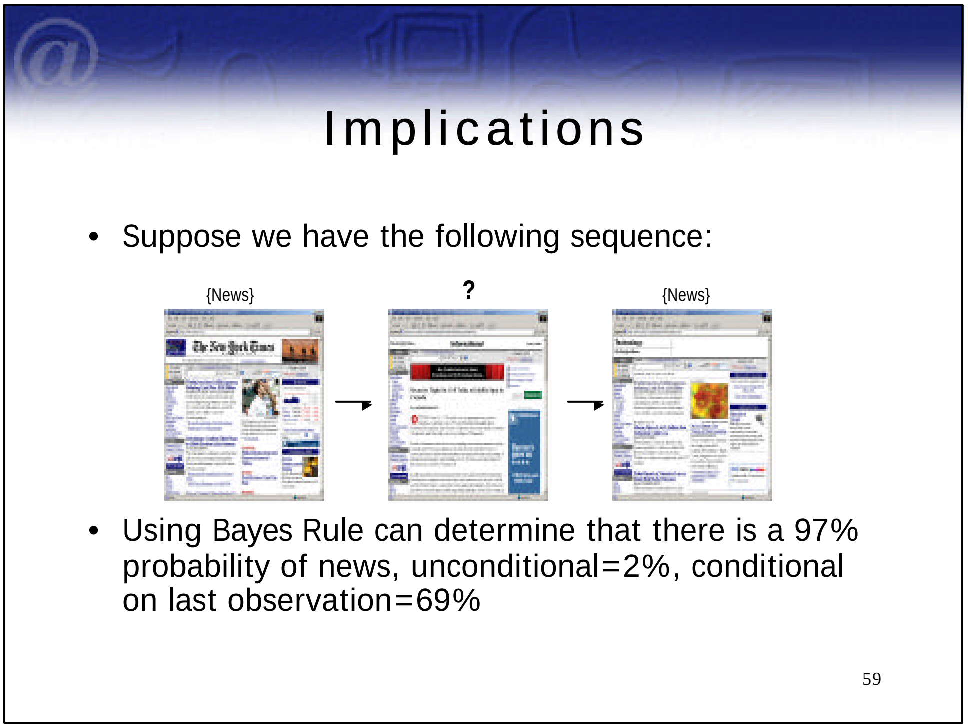
User Browsing Model

- Web browsing is “sticky” or persistent: users tend to view a series of pages within the same category and then switch to another topic
- Example:

Markov Switching Model

	arts	usiness	mputers	games	health	home	news	creation	ference	science	hopping	society	sports	adult
arts	83%	4%	5%	2%	1%	2%	6%	3%	2%	6%	2%	3%	4%	1%
business	3%	73%	5%	3%	2%	3%	6%	2%	3%	3%	3%	2%	3%	2%
computers	5%	11%	79%	3%	3%	7%	5%	3%	4%	4%	5%	5%	2%	2%
games	1%	3%	2%	90%	1%	1%	1%	1%	0%	1%	1%	1%	1%	0%
health	0%	0%	0%	0%	84%	1%	1%	0%	0%	1%	0%	1%	0%	0%
home	0%	1%	1%	0%	1%	80%	1%	1%	0%	1%	1%	1%	0%	0%
news	1%	1%	1%	0%	1%	0%	69%	0%	0%	1%	0%	1%	1%	0%
recreation	1%	1%	1%	0%	1%	1%	1%	86%	1%	1%	1%	1%	1%	0%
reference	0%	1%	1%	0%	1%	0%	1%	0%	85%	2%	0%	1%	1%	0%
science	1%	0%	0%	0%	1%	1%	1%	0%	1%	75%	0%	1%	0%	0%
shopping	1%	3%	2%	1%	1%	2%	1%	1%	0%	1%	86%	1%	1%	0%
society	1%	1%	2%	0%	2%	1%	3%	1%	2%	2%	0%	82%	1%	1%
sports	2%	1%	1%	0%	0%	0%	3%	1%	1%	0%	0%	1%	85%	0%
adult	1%	1%	1%	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	93%
	16%	10%	19%	11%	2%	3%	2%	6%	3%	2%	7%	6%	5%	7%

Pooled transition matrix, heterogeneity across users



Implications

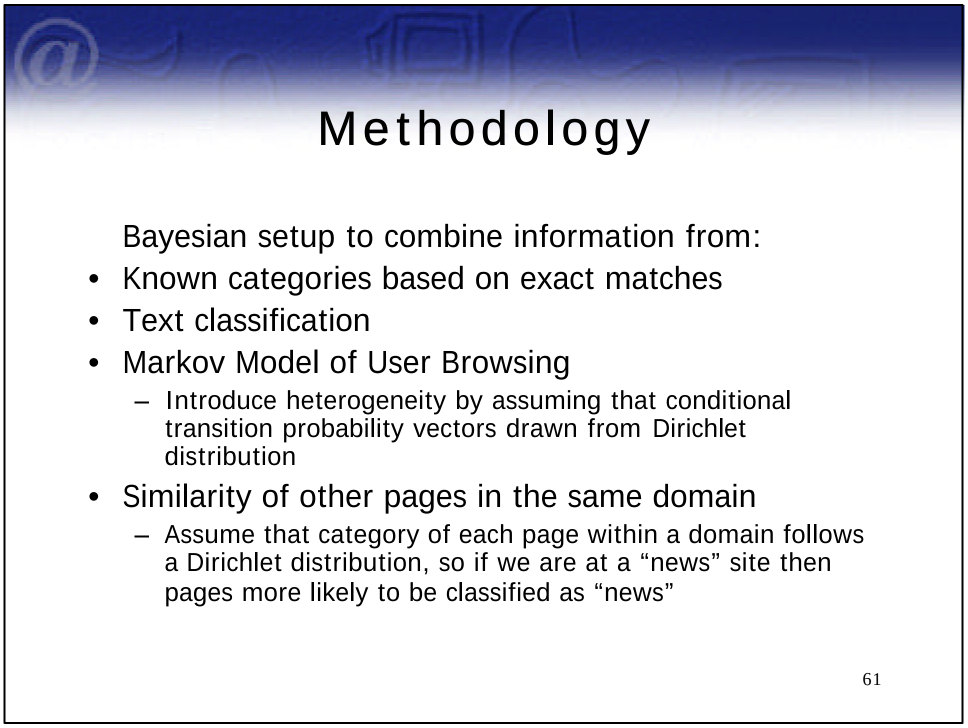
- Suppose we have the following sequence:

{News} ? {News}

- Using Bayes Rule can determine that there is a 97% probability of news, unconditional=2%, conditional on last observation=69%

59

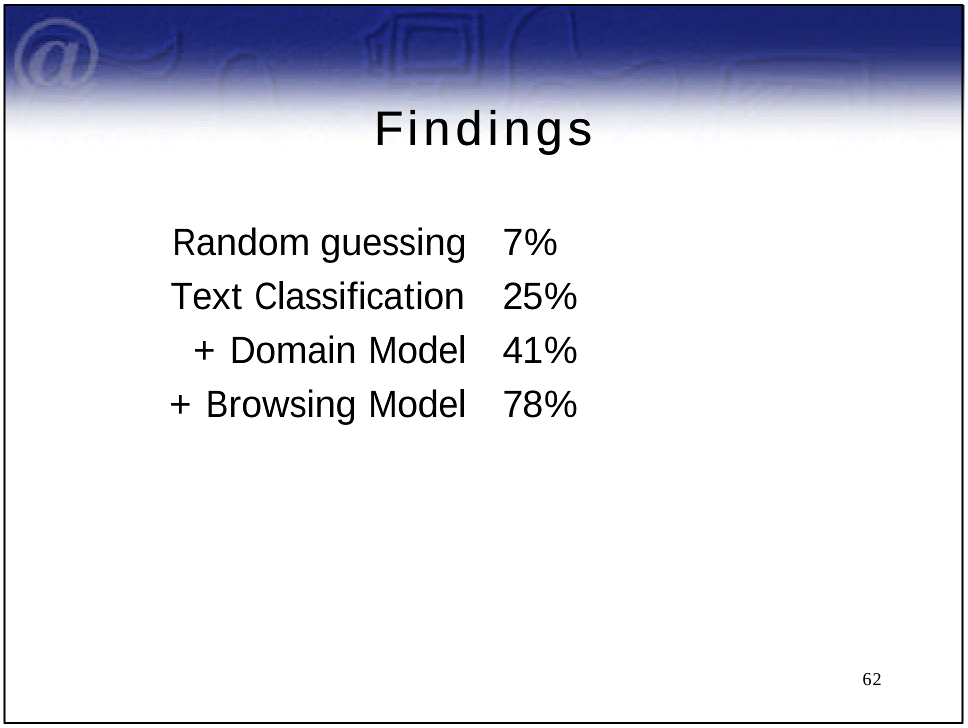
Results



Methodology

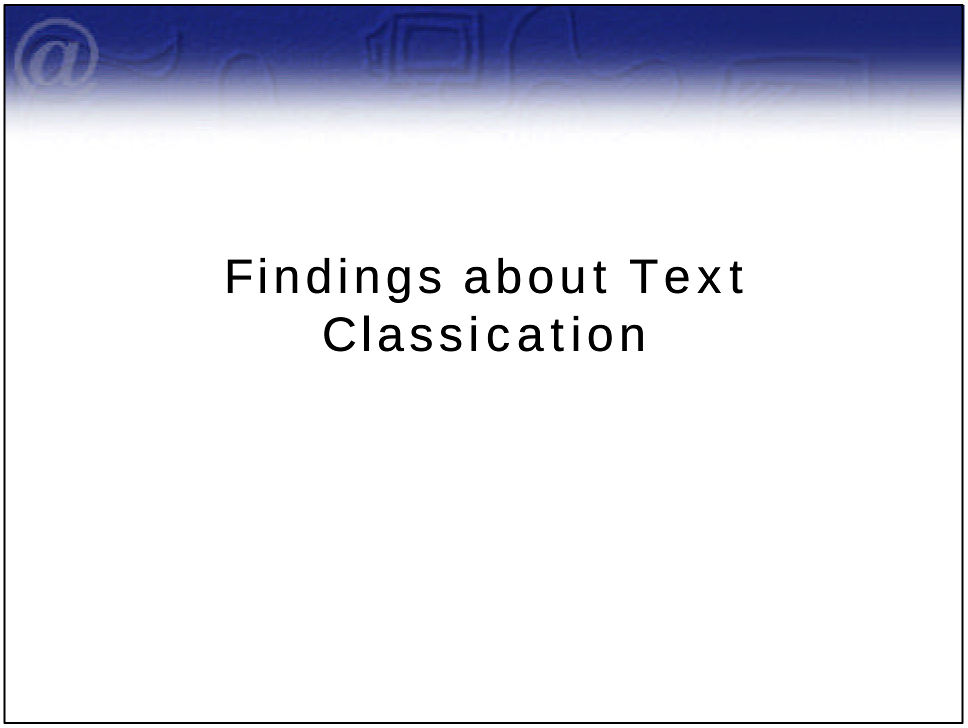
Bayesian setup to combine information from:

- Known categories based on exact matches
- Text classification
- Markov Model of User Browsing
  - Introduce heterogeneity by assuming that conditional transition probability vectors drawn from Dirichlet distribution
- Similarity of other pages in the same domain
  - Assume that category of each page within a domain follows a Dirichlet distribution, so if we are at a “news” site then pages more likely to be classified as “news”

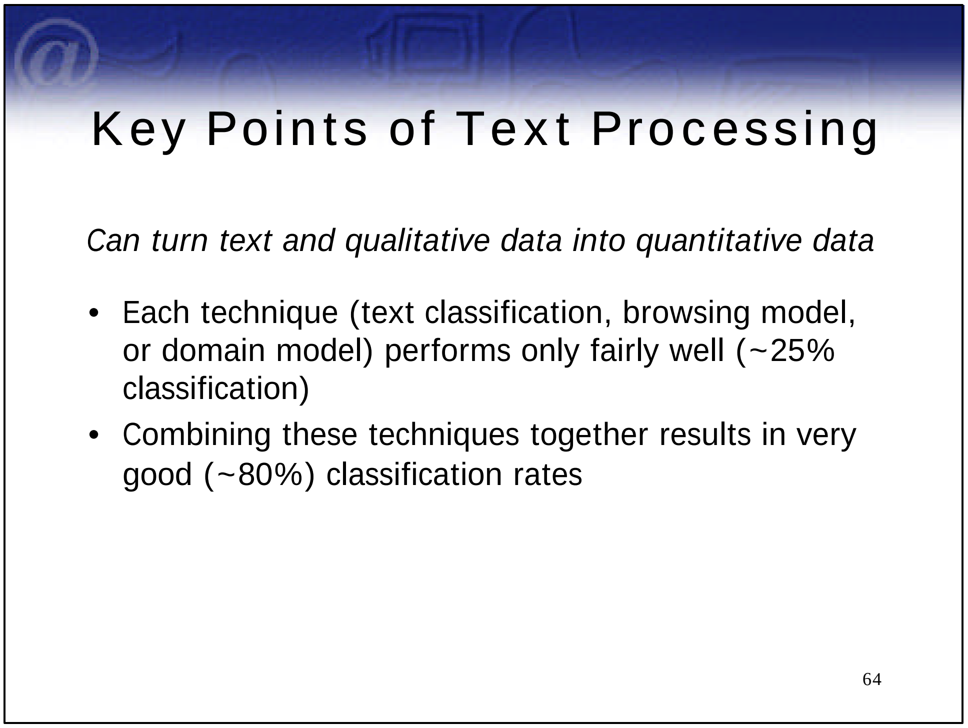


Findings

Random guessing	7%
Text Classification	25%
+ Domain Model	41%
+ Browsing Model	78%



Findings about Text Classication



Key Points of Text Processing

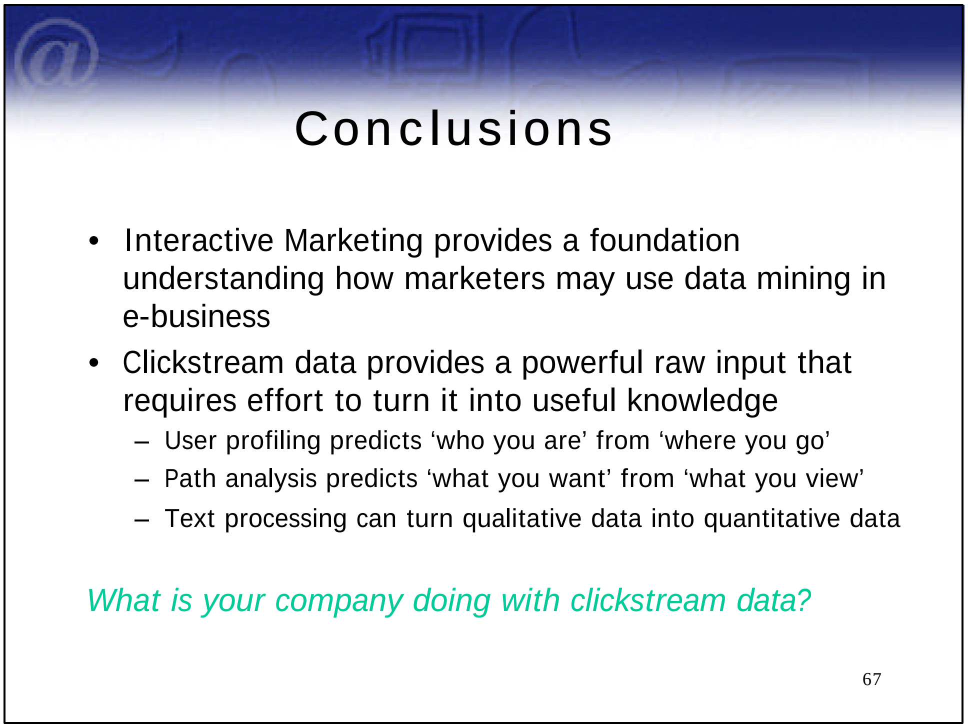
Can turn text and qualitative data into quantitative data

- Each technique (text classification, browsing model, or domain model) performs only fairly well (~25% classification)
- Combining these techniques together results in very good (~80%) classification rates

Applications

- Newsgroups
  - Gather information from newsgroups and determine whether consumers are responding positively or negatively
- E-mail
  - Scan e-mail text for similarities to known problems/topics
- Better Search engines
  - Instead of experts classifying pages we can mine the information collected by ISPs and classify it automatically
- Adult filters
  - US Appeals Court struck down Children’s Internet Protection Act on the grounds that technology was inadequate

Session Conclusions



Conclusions

- Interactive Marketing provides a foundation understanding how marketers may use data mining in e-business
- Clickstream data provides a powerful raw input that requires effort to turn it into useful knowledge
  - User profiling predicts ‘who you are’ from ‘where you go’
  - Path analysis predicts ‘what you want’ from ‘what you view’
  - Text processing can turn qualitative data into quantitative data

What is your company doing with clickstream data?

67