

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

Alan L. Montgomery and Brett Gordon  
**Carnegie Mellon University**

*Marketing Science Conference  
University of Alberta, Edmonton  
28 June 2002*



## Outline

- Clickstream Example
  - What topic is the user looking at on each page?
- Information Sources
  - Dmoz.org classification
  - Text classification
  - User browsing model
- Results
- Conclusions

# Clickstream Example

What topics is this user browsing on each of the following pages?

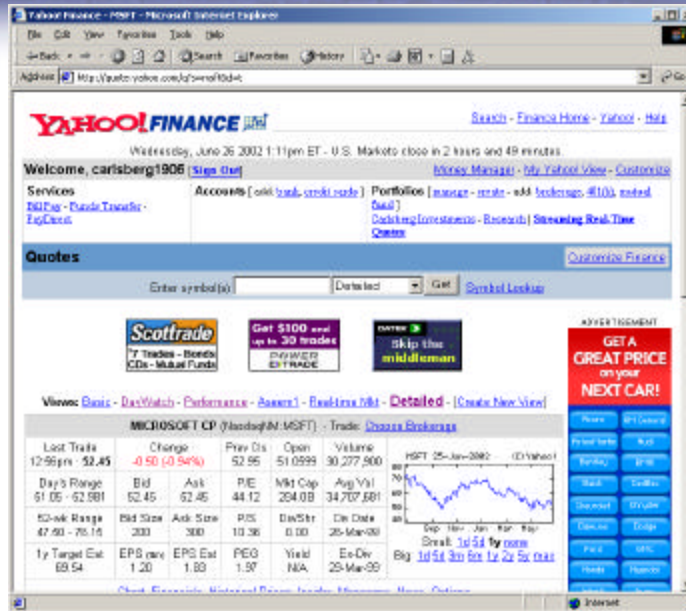
True Class: Business

The screenshot shows the Yahoo! Finance website in a Microsoft Internet Explorer browser window. The address bar shows the URL <http://www.yahoo.com>. The page content includes:

- Header:** "YAHOO! FINANCE" logo, navigation links (Search, Site Map, Yahoo!, Help), and a search bar.
- Personalization:** "Welcome, carlsberg1906" with a link to "See Opt".
- Services:** "Money Manager", "My Yahoo! View", "Customize".
- Accounts:** "add bank credit cards", "Portfolios", "manage", "create", "add", "brokerage", "401k", "mutual fund", "Certificates", "Investments", "Research", "Streaming Real-Time Quotes".
- Market Summary:** A line chart showing market performance with data for 04-Jun, 10am, 12pm, and 3pm. Key values include: "Dow: 8,599.55 (+27.24 (+.32%))", "S&P 500: 1,400.43 (-21.36 (-1.51%))", "NASDAQ: 960.09 (-14.83 (-1.44%))", "30-Day Bond: 4.727% (-0.184)", "NYSE Volume: 1,074,085,008", "NASDAQ Volume: 1,183,793,008".
- Top Stories:** "U.S. stocks punished by WorldCom scandal" with a sub-headline "WorldCom's shocking revelation of fraud has rocked Wednesday, but some buyers managed to secure the averages from their loss. Earnings, drug and retail news among the beneficiaries of investor bargain hunting." Below the story is a "Video from FinanceVision" section.
- Advertisements:** A "WATERHOUSE" advertisement for a new brokerage account offering "10 Free Trades".

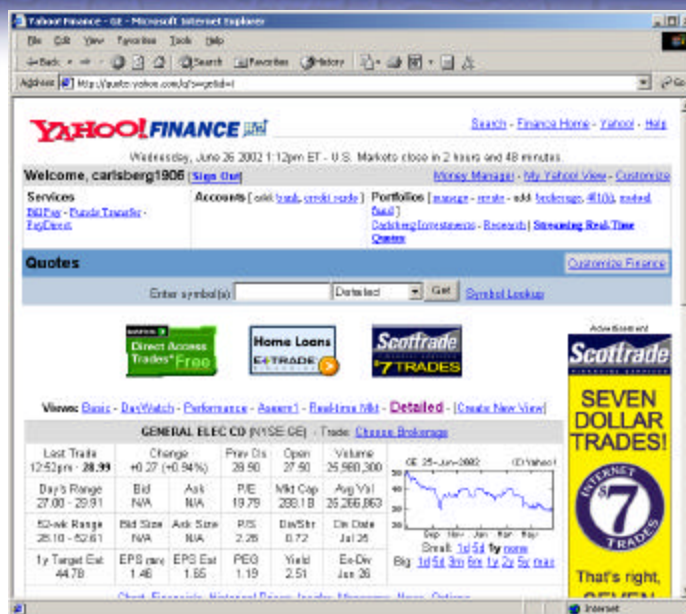
4

## True Class: Business



5

## True Class: Business



6

## True Class: Sports

ESPN.com - Microsoft Internet Explorer

Address: http://www.espn.com/tennis/tennis.html

Today in Chat  
8 p.m. ET  
Andy Katz talks NBA draft

Wednesday, June 25

### Petered Out

The struggles of Pete Sampras continue. George Bastl stunned the sixth-seeded Sampras 6-7, 6-2, 4-6, 5-6, 6-4 in a second-round match at Wimbledon on Wednesday. Sampras has won a record 13 Grand Slam singles titles but has not won a major in two years. No. 2 seed Maxie Safin was ousted by 54th-ranked Olivier Rochus.

[Bastl, Safin come up big](#) | [Wimbledon: So long, Sampras](#)

ESPN Insider  
Sign up for a free trial  
Get 30% more draft info including mock drafts, draft rumors and more.

All Systems Go  
The Rockets got the [draft picks they needed](#) today when they traded to make you into the No. 1 pick in the [draft](#) (ESPN Insider, 7:20 ET). The No. 1 pick is resolved, but there are still a lot of [trade rumors](#) [flaring](#) around New York, writes Andy Katz.

- [Draft Roundup: Update, 5 hours' worth](#)
- [Carmelo: Draft ready](#) | [Ask the coach](#) [ESPN.com](#)

ESPN.com's Most Timely  

- [Great catches that should start Oscar Pistorius](#)
- [Clayton Kershaw: Perfect, but he's not a superstar](#)
- [Carmelo Anthony: Ready to fly](#) | [Wade at 4](#)
- [Drew Brees: His second 1-year extension](#)
- [Quarterback ready to call for NFL career](#)
- [College basketball: For all the talk, it's still a long haul](#)

ESPN.com  
[Home](#) | [Baseball](#) | [ESPN Insider](#)  
 Register | [ESPN.com](#)  
 Profile of [draft](#)  
 No more 100% win or losses both  
 Super Bowl picks

7

## True Class: Sports

ESPN.com World Cup 2002 Home - Microsoft Internet Explorer

Address: http://worldcup.espn.com/tennis/tennis.html

FOOTBALLITIS IS SPREADING

Next News  
The Galizing Major, Kaiser Franz, "The Daisa Ponza" all you need to know about it's, sometimes - and "great Footballer" - among our top World Cup players. [...]

### Taking Care of Business

Ronaldo continues to prove himself in the 2002 World Cup.

Quick Stats  
 Leaders Goals  
 1. Ronaldo (BRA) 6  
 2. Klasev (GER) 5  
 3. Ronaldo (BRA) 5  
 4. Timonova (GER) 4  
 5. Pato (ITA) 4  
 More stats...

That's with more coverage  
espn.com's 24-hour video  
coverage of the World Cup final  
matchup and more Wednesday  
at 5 a.m. ET.

More Stats  

- [Germany's defensive tactics](#)
- [Brazil's first coach after poor showing](#)
- [U.S. Embassy breaks TV record](#)
- [FIFA scores 22 more all-time goals](#)

8

# True Class: Sports

The screenshot shows the ESPN.com website in a Microsoft Internet Explorer browser window. The page is titled "ESPN.com - Page 2 - TV Listings" and features a large "PAGE 2" banner. A "Summer Love" promotion is visible, including a "Trip to Hawaii". The main content is "Soccer Listings for Wednesday, Jun. 26", organized into three sections: ESPN, ESPN2, and ESPN Classic. Each section lists a soccer match with the time and location. A calendar for June 2002 is on the right, and a "ALSO SEE" section lists links to ESPN's TV listings.

ESPN	TIME (ET)
2002 WORLD CUP SEMI FINAL BRAZIL VS. TURKEY SAITAMA STADIUM, SAITAMA, JAPAN	7:25 am to 9:30 am

ESPN2	TIME (ET)
2002 WORLD CUP SEMI FINAL (RE-AR 2 HD) 3 BRAZIL VS. TURKEY SAITAMA STADIUM, SAITAMA, JAPAN	3:00 pm to 5:00 pm
WORLD CUP HIGHLIGHT CHARLOTTE, NC, USA	1:00 pm to 2:00 am

ESPN Classic	TIME (ET)
2002 WORLD CUP SEMI FINAL - (RE-AR 2 HD) 3 BRAZIL VS. TURKEY HD SAITAMA STADIUM, SOUTH KOREA	1:00 pm to 3:00 pm

9

# True Class: News

The screenshot shows the New York Times website in a Microsoft Internet Explorer browser window. The page is titled "The New York Times ON THE WEB" and is dated "UPDATED WEDNESDAY, JUNE 26, 2002 1:10 PM ET". The main headline is "WorldCom Says It Hid Expenses, Inflating Cash Flow \$3.8 Billion" by Steven Rosenberg and Alice Lipton. Other headlines include "Palestinians Confirm Their Plans to Hold Elections Next January" and "NATIONAL Public Defender Denied for Suspected American Taliban". A "MARKETS" section shows Dow Jones Industrial Average at 8,800.50, down 102.28 (-1.16%). A "WORLD'S LEADING CRUISE LINE" advertisement is also visible.

10

True Class: News

The screenshot shows the New York Times website's International section. The main article is titled "Security Tight for G-8 Talks at Idyllic Spot in Canada" by Clifford Kopp. The article discusses the G-8 summit in Kananaskis, Canada, and the security measures in place. A sidebar on the left contains navigation links for various news categories. A banner at the top of the article area reads "Buy Tickets Online for Select Broadway and Off-Broadway Shows".

11

True Class: News

The screenshot shows the New York Times website's Technology section. The main article is titled "WorldCom Says It Hid Expenses, Inflating Cash Flow \$5.8 Billion" by Steven Rosenberg and Alex Berenson. The article reports on WorldCom's admission of hidden expenses. Other articles on the page include "Master Piece of AOL Suffers First Setback in Cable Loss" and "Palm Reports a Narrower Loss as Hand-Held Sales Rebound". A sidebar on the left contains navigation links for various news categories. A banner at the top of the article area reads "Updated June 26, 2002 2:10 PM ET".

12

# True Class: News



13

User Demographics

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

14

## Information Sources

## Data

### 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

17

## 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

18

# 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)

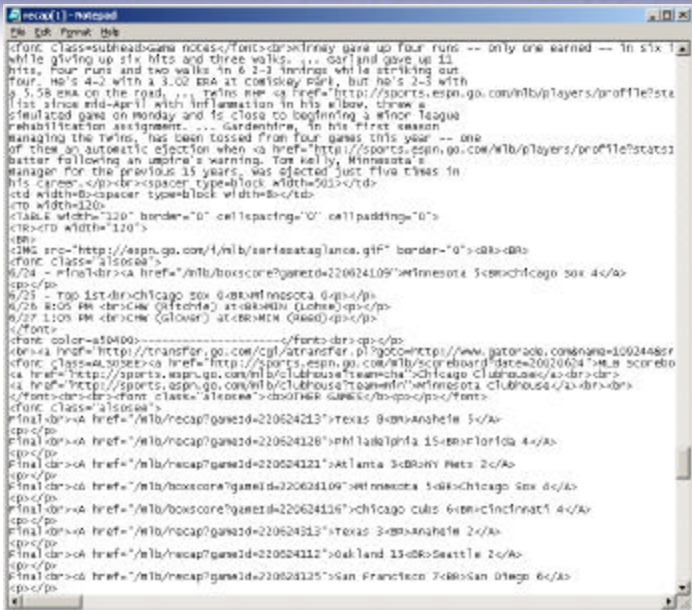
## True Class: Sports



The screenshot shows a web browser window displaying an ESPN.com article titled "Jones, Hunter cap late-inning comeback". The article reports on a game between the Minnesota Twins and the Chicago White Sox on Monday, June 24. The headline states that the Twins won 3-4. The article mentions that the Twins were down 4-2 in the eighth inning but scored two runs to win. Key players mentioned include Justin Jones, Matt Kraybill, and Bobby Howry. A sidebar on the left contains navigation links for various baseball topics. A Gatorade logo is visible in the bottom right corner of the article content.

21

## Page Contents = HTML Code + Regular Text



The screenshot shows a Notepad window displaying the raw HTML code of the ESPN.com page. The code includes various HTML tags such as `<h1>`, `<h2>`, `<p>`, `<a href="...">`, `<img alt="...">`, and `<table border="1">`. The code is a mix of regular text and HTML tags, illustrating the structure of the page's content.

22

## Tokenization & Lexical Parsing

- HTML code is removed
- Punctuation is removed
- All words are converted to lowercase
- Stopwords are removed
  - Common, non-informative words such as 'the', 'and', 'with', 'an', etc...

*Determine the term frequency (TF) of each remaining unique word*

23

## Result: Document Vector



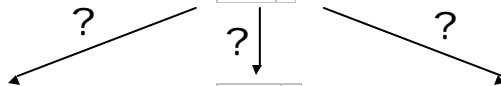
home	2
game	8
hit	4
runs	6
threw	2
ejected	1
baseball	5
major	2
league	2
bat	2

24

# 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

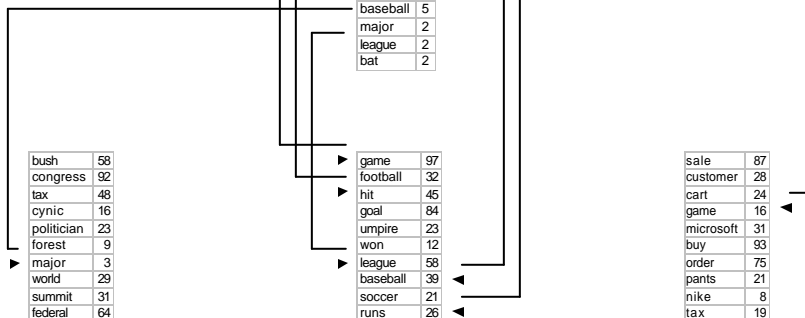
{Shopping Class}

25

# 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}

26

# 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

$P(\{\text{News}\} | \text{Test Doc}) = 0.02$

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

{News Class}

$P(\{\text{Sports}\} | \text{Test Doc}) = 0.91$

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

{Sports Class}

$P(\{\text{Shopping}\} | \text{Test Doc}) = 0.07$

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

{Shopping Class}

27

# 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

$P(\{\text{Sports}\} | \text{Test Doc}) = 0.91$

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

{Sports Class}

28

## 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:

$$d_c \sim M\{n, \{p_1^c, p_2^c, \dots, p_{|V|}^c\}\}$$

- Classification:  $\arg \max_{c \in C} P(c | d)$

$$\arg \max_{c \in C} P(c) \prod_{i=1}^{|V|} P(w_i | c)^{n_i^c}$$

$|V|$ : vocabulary size

$n_i^c$ : # of times word  $i$  appears in class  $c$

29

## 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

30

## 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:



32



# Markov Switching Model

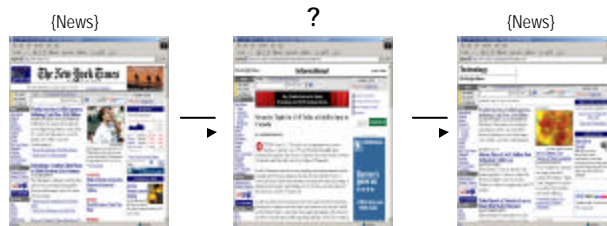
	arts	business	computers	games	health	home	news	recreation	reference	science	shopping	society	sports	adult
arts	83%	4%	5%	2%	1%	2%	6%	3%	2%	6%	2%	3%	4%	1%
business	3%	73%	5%	3%	2%	3%	6%	2%	3%	3%	3%	2%	3%	2%
computers	5%	11%	79%	3%	3%	7%	5%	3%	4%	4%	5%	5%	2%	2%
games	1%	3%	2%	90%	1%	1%	1%	1%	0%	1%	1%	1%	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

33

# Implications

- Suppose we have the following sequence:



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

34



## 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”


36



## Findings

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

37



## Conclusions

## Summary

- 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
- Future directions: larger datasets and newer text classification and user browsing models

39

## 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

40