Introduction to Computational Advertising

MS&E 239

Stanford University

Autumn 2011

Instructors: Dr. Andrei Broder and Dr. Vanja Josifovski

Yahoo! Research

General course info

- Course Website: http://www.stanford.edu/class/msande239/
- Instructors
  - **Dr. Andrei Broder**, Yahoo! Research, broder@yahoo-inc.com
  - **Dr. Vanja Josifovski**, Yahoo! Research, vanjaj@yahoo-inc.com
- TA: **Krishnamurthy Iyer**
  - Office hours: Tuesdays 6:00pm-7:30pm, Huang
- Course email lists
  - Staff: msande239-aut1112-staff
  - All: msande239-aut1112-students
  - Please use the staff list to communicate with the staff
- Lectures: 10am ~ 12:30pm Fridays in HP
- Office Hours:
  - After class and by appointment
  - Andrei and Vanja will be on campus for 2 times each to meet and discuss with students. Feel free to come and chat about even issues that go beyond the class.

Course Overview (subject to change)

1. 09/30 Overview and Introduction
2. 10/07 Marketplace and Economics
3. 10/14 Textual Advertising 1: Sponsored Search
4. 10/21 Textual Advertising 2: Contextual Advertising
5. 10/28 Display Advertising 1
6. 11/04 Display Advertising 2
7. 11/11 Targeting
8. 11/18 Recommender Systems
9. 12/02 Mobile, Video and other Emerging Formats
10. 12/09 Project Presentations

Lecture 3:
Textual Advertising - Sponsored Search

Disclaimers

- This talk presents the opinions of the authors. It does not necessarily reflect the views of Yahoo! inc or any other entity.
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Yahoo! or any other company.
- These lectures benefitted from the contributions of many colleagues and co-authors at Yahoo! and elsewhere. Their help is gratefully acknowledged.

Lecture 3 plan

- Review of Sponsored Search interactions
- Textual Ads
- Web queries
- Ad Selection
  - Overview of ad selection methods
  - Exact Match
  - Advanced Match
- Advanced Match
  - Query rewriting for advanced match
  - Use of click graphs for advanced match
- In class presentation – Advertising on Facebook

Sponsored Search Review

Some statistics

Sponsored Search Market Share

US Online Ad Revenue Share, by Format, Q2 2009-Q2 2011

% of total

	Q2 2009	Q2 2010	Q2 2011
Search	47%	47%	49%
Display/banner	22%	23%	23%
Classifieds	10%	10%	8%
Digital video	4%	5%	6%
Lead generation	7%	5%	5%
Rich media	7%	6%	5%
Sponsorship	2%	2%	3%
Email	1%	1%	1%

Note: numbers may not add up to 100% due to rounding
Source: Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers (PwC), "IAB Internet Advertising Revenue Report: 2011 First Six Months Results," Sep 28, 2011

132940 www.eMarketer.com

Spending per format

US Online Ad Spending, by Format, 2009-2013
billions

	2009	2010	2011	2012	2013
Paid search	$9.49	$11.01	$12.11	$13.93	$16.02
Display	$4.21	$4.51	$4.96	$5.50	$6.22
Classifieds	$3.22	$3.22	$3.45	$3.65	$3.84
Internet video/rich media	$2.21	$2.74	$3.34	$4.21	$5.34
Social media	$0.56	$0.74	$0.98	$1.33	$1.79
Mobile	$0.39	$0.58	$0.83	$1.23	$1.85
Internet radio	$0.23	$0.26	$0.29	$0.32	$0.35
Podcast	$0.03	$0.03	$0.04	$0.04	$0.04
Total	**$20.34**	**$23.08**	**$25.98**	**$30.20**	**$35.44**

Note: at current prices; numbers may not add up to total due to rounding
Source: ZenithOptimedia, "Advertising Expenditure Forecasts," Oct 3, 2011

133120 www.eMarketer.com

The Key Words

Top 20 Search Keyword Categories and Cost per Click* in Google AdWords, Q2 2011

% of 10,000 keywords in each category and CPC

	% share	CPC*
1. Insurance	24.0%	$54.91
2. Loans	12.8%	$44.28
3. Mortgage	9.0%	$47.12
4. Attorney	3.6%	$47.07
5. Credit	3.2%	$36.06
6. Lawyer	3.0%	$42.51
7. Donate	2.5%	$42.02
8. Degree	2.2%	$40.61
9. Hosting	2.2%	$31.91
10. Claim	1.4%	$45.51
11. Conference call	0.9%	$42.05
12. Trading	0.8%	$33.19
13. Software	0.8%	$35.29
14. Recovery	0.7%	$42.03
15. Transfer	0.6%	$29.86
16. Gas/electricity	0.6%	$54.62
17. Classes	0.5%	$35.04
18. Rehab	0.5%	$33.59
19. Treatment	0.3%	$37.18
20. Cord blood	0.2%	$27.80

*Note: English-language only; *US keyword price estimates*
Source: WordStream as cited in press release, July 18, 2011

130217 www.eMarketer.com

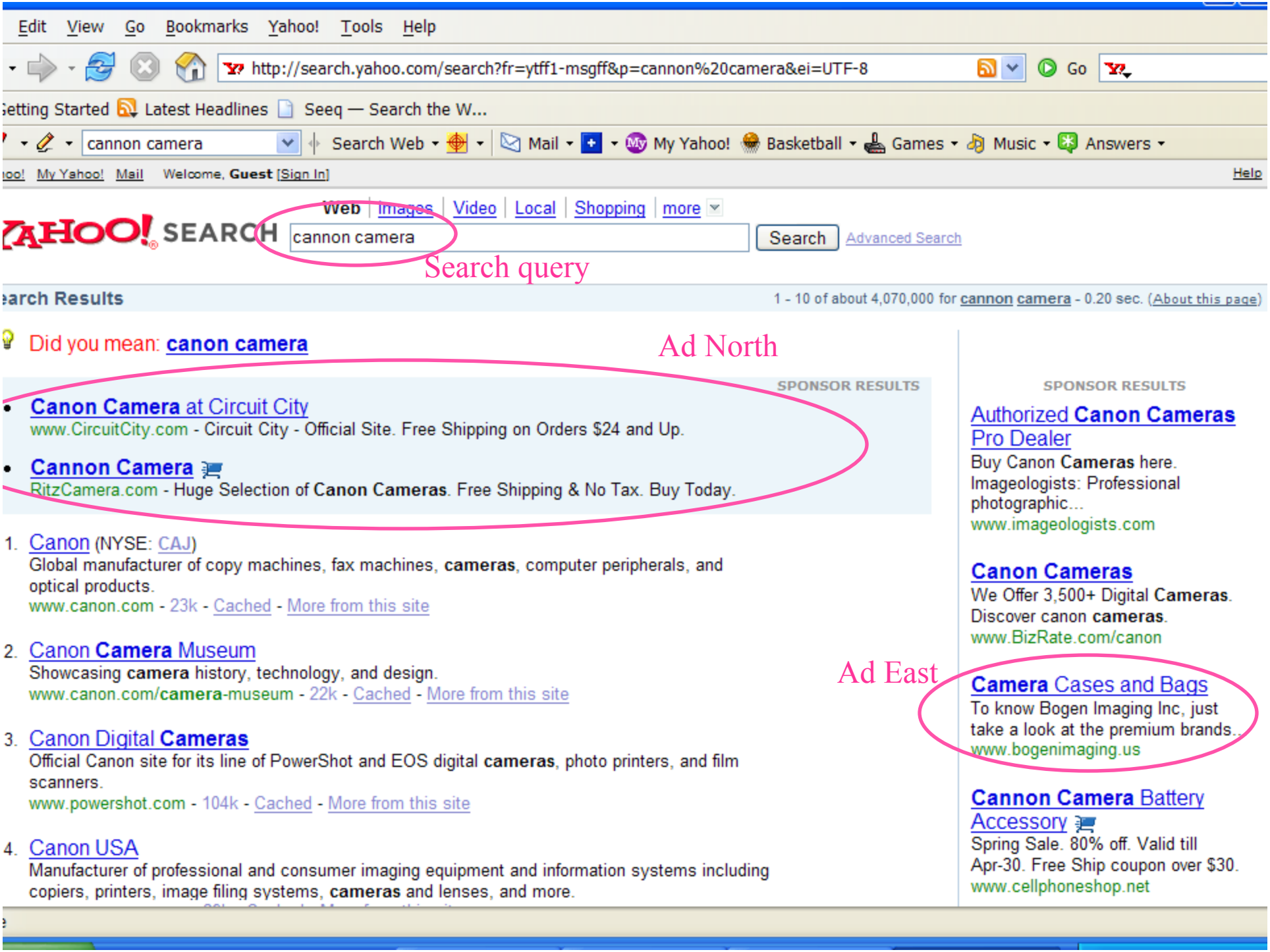
CPC per search engine

Paid Search Cost per Click (CPC) Worldwide, by Search Engine, 2007-Q3 2011

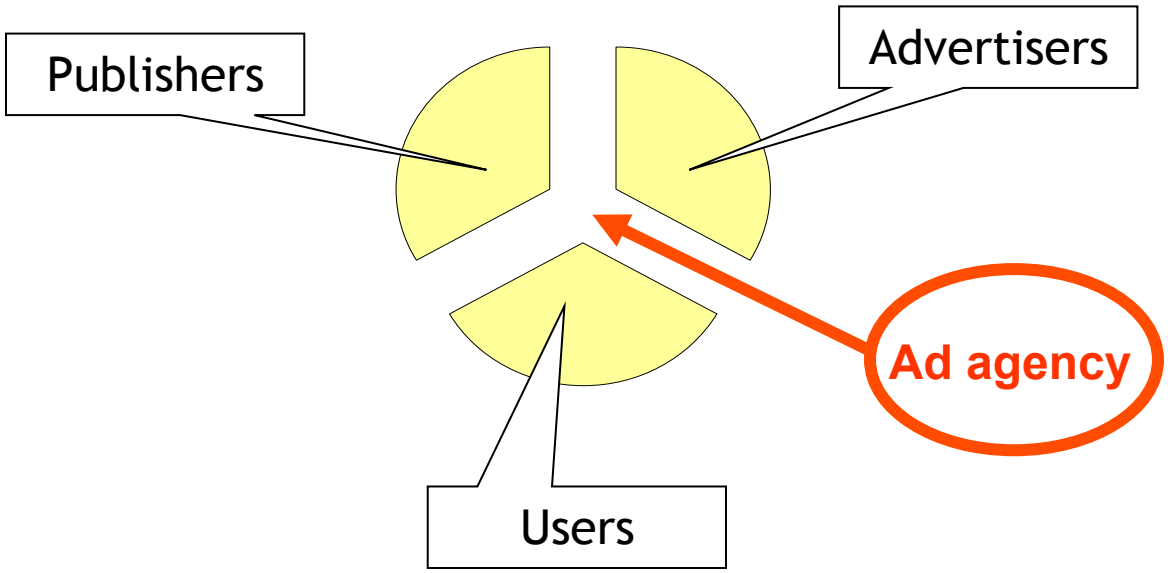
	2007	2008	2009	2010	Q1 2011	Q2 2011	Q3 2011
Ask.com	-	$0.36	$0.38	$0.71	$0.69	$0.68	$0.67
Baidu	$0.13	$0.14	$0.16	$0.25	$0.35	$0.42	$0.45
Google	$1.22	$1.06	$0.72	$0.74	$0.79	$0.81	$0.83
Bing	$1.28	$1.60	$0.64	$0.92	$1.02	$1.08	$1.05
Rambler	-	$0.06	$0.80	$0.00	$0.13	$0.13	$0.13
Yahoo!	$1.59	$1.16	$0.74	$0.79	-	-	-
Yandex	$1.16	$1.09	$0.56	$0.61	$0.61	$0.62	$0.65
Total	**$1.19**	**$1.04**	**$0.71**	**$0.79**	**$0.79**	**$0.83**	**$0.85**

Source: Covario, "Third Quarter Paid Search Rebounds from Second Quarter Lull," Oct 12, 2011

133478 www.eMarketer.com


Edit View Go Bookmarks Yahoo! Tools Help
http://search.yahoo.com/search?fr=ytff1-msgff&p=cannon%20camera&ei=UTF-8
Go
Getting Started Latest Headlines Seeq — Search the W...
cannon camera
Search Web Mail My Yahoo! Basketball Games Music Answers
My Yahoo! Mail Welcome, Guest [Sign In] Help
Web Images Video Local Shopping more
YAHOO! SEARCH
cannon camera
Search
Advanced Search
Search query
Search Results
1 - 10 of about 4,070,000 for cannon camera - 0.20 sec. (About this page)
Did you mean: canon camera
Ad North
SPONSOR RESULTS
Canon Camera at Circuit City
www.CircuitCity.com - Circuit City - Official Site. Free Shipping on Orders $24 and Up.
Cannon Camera
RitzCamera.com - Huge Selection of Canon Cameras. Free Shipping & No Tax. Buy Today.
1. Canon (NYSE: CAJ)
Global manufacturer of copy machines, fax machines, cameras, computer peripherals, and optical products.
www.canon.com - 23k - Cached - More from this site
2. Canon Camera Museum
Showcasing camera history, technology, and design.
www.canon.com/camera-museum - 22k - Cached - More from this site
3. Canon Digital Cameras
Official Canon site for its line of PowerShot and EOS digital cameras, photo printers, and film scanners.
www.powershot.com - 104k - Cached - More from this site
4. Canon USA
Manufacturer of professional and consumer imaging equipment and information systems including copiers, printers, image filing systems, cameras and lenses, and more.
Ad East
SPONSOR RESULTS
Authorized Canon Cameras Pro Dealer
Buy Canon Cameras here. Imageologists: Professional photographic...
www.imageologists.com
Canon Cameras
We Offer 3,500+ Digital Cameras. Discover canon cameras.
www.BizRate.com/canon
Camera Cases and Bags
To know Bogen Imaging Inc, just take a look at the premium brands...
www.bogenimaging.us
Cannon Camera Battery Accessory
Spring Sale. 80% off. Valid till Apr-30. Free Ship coupon over $30.
www.cellphoneshop.net

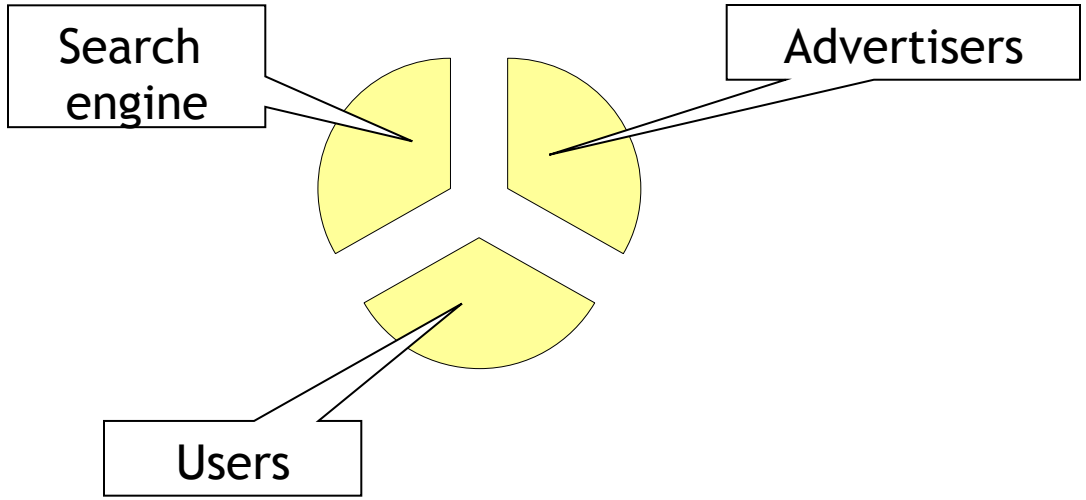
The general interaction picture: Publishers, Advertisers, Users, & “Ad agency”



- Each actor has its own goal (more later)

The simplified picture for sponsored search

- All major search engines (Google, MSN, Yahoo!) are simultaneously
  1. **search results provider**
  2. **ad agency**



- Sometimes full picture: SE provides ad results to a different search engine: e.g. Google to Ask.

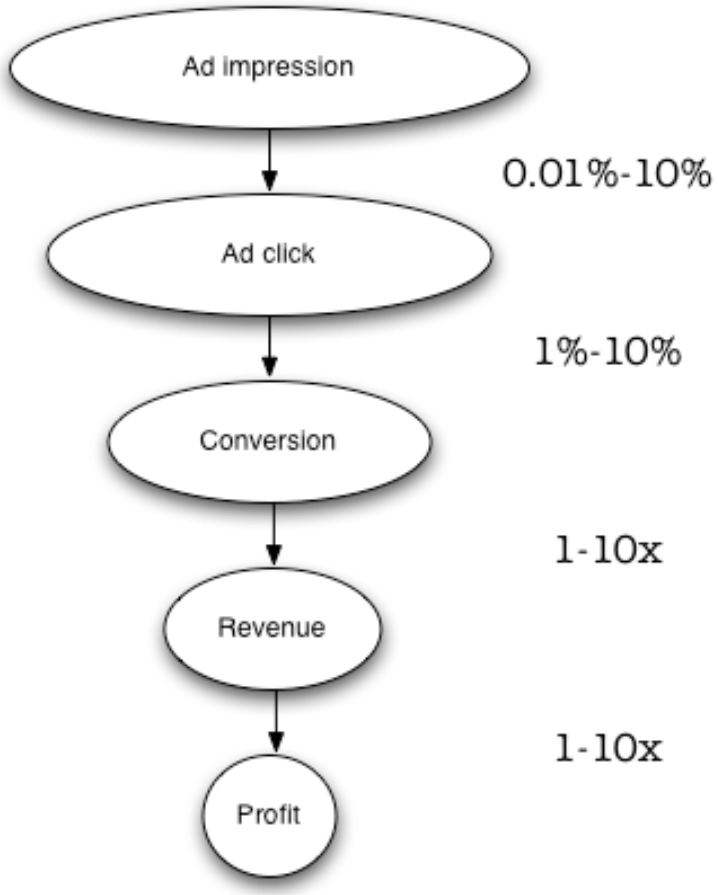
Interactions in Sponsored Search

- **Advertisers**:
  - Submit ads associated to certain bid phrases
  - Bid for position
  - Pay CPC

Lecture 2

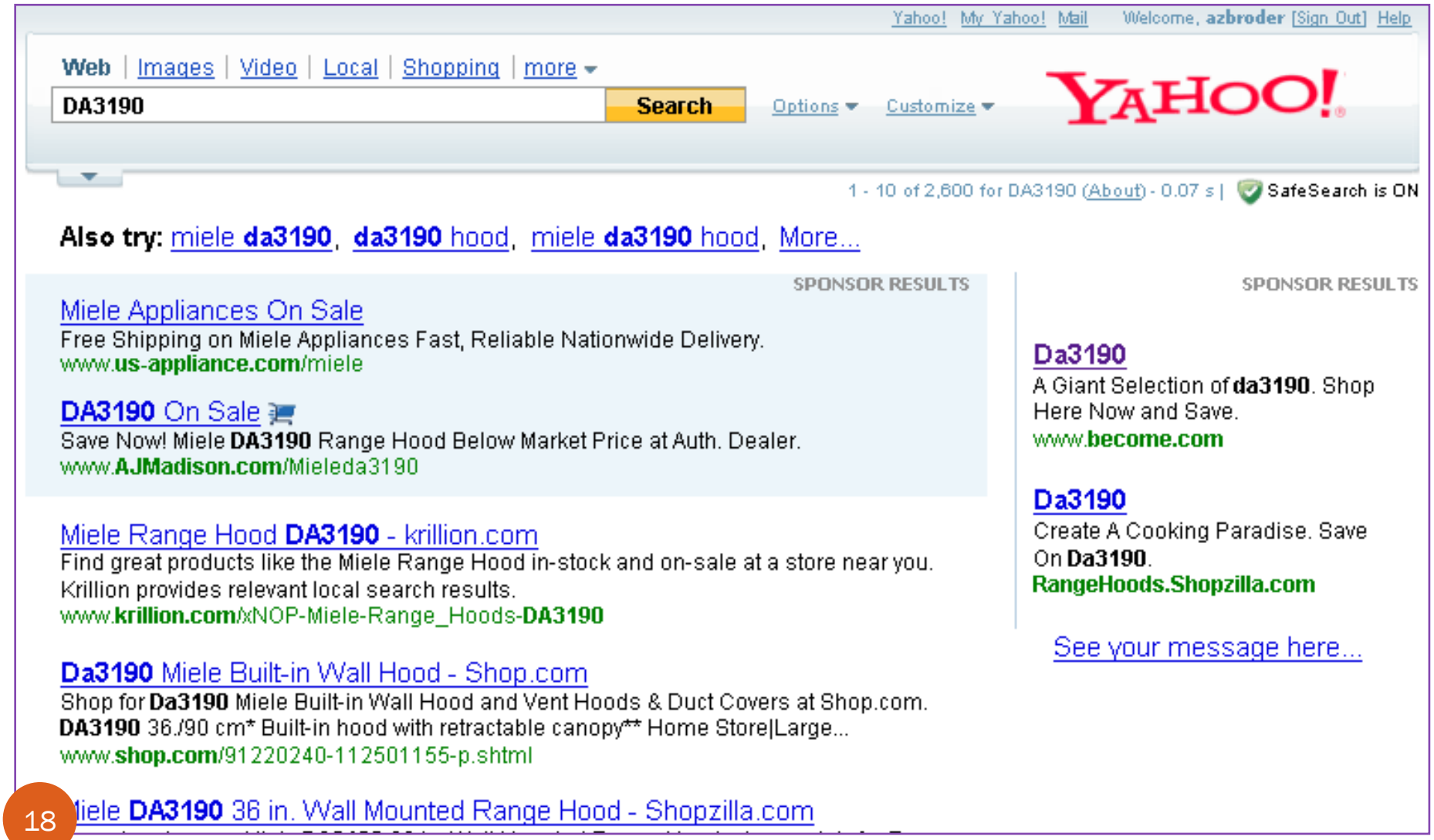
- **Users**
  - Make queries to search engine, expressing some intent
- **Search engine**
  - Executes query against web corpus + other data sources
  - Executes query against the ad corpus
  - Displays a Search Results Page (SERP) = integration of web results, other data, and ads
- **Each of the SE, Advertisers, and Users has its own utility**

Advertiser Utility: The Value Funnel



- Value = Long Term Profit
- What is the value of each event?
- **Immediate value –** profit of the current event (conversion and below)
- **Future value** – increase of the future profit due to the user action:
  - Ad impression might bring future conversions
  - Revenue events (upon user satisfaction) bring repeat customers
- Approximation: value declines by a linear coefficient as we move upwards

User: useful ads



Conflicts and synergies

- Some utilities are aligned, some are in conflict
- **Aligned:**
  - SE and advertisers want more clicks – revenue for SE + volume for advertisers
  - SE and users want good ads – ads offer information + users click and SE makes money
- **Conflicting utilities:**
  - Higher cost per click better for SE but worse for advertiser
  - Irrelevant ads with high PPC annoy most users but still get some clicks → Revenue for SE, ROI for advertiser
- **How to balance?**

Optimization

- Total utility of a Sponsored Search system is a balance of the individual utilities:

  $$\text{Utility} = f(\text{UtilityAdvertiser}, \text{UtilityUser}, \text{UtilitySE})$$

- Function f() combines the individual utilities
- How to choose an appropriate combination function?
  - Model the long-term goal of the system
  - Parameterized to allow changes in the business priorities
  - Simple – so that business decisions can be done by the business owners!
- Example: convex linear combination:

  $$\text{Utility} = \alpha * \text{UtilityAdvertiser} + \beta * \text{UtilityUser} + \gamma * \text{UtilitySE}$$

  where $\alpha + \beta + \gamma = 1$

Utility – more pragmatic view

- Long term utilities are hard to capture/quantify
- Instead

  Maximize per search revenue subject to
  1. User utility per search > α
  2. Advertiser ROI per search > β

- Practically:
  1. Find all ads that have user utility above α
  2. Optimize which ads to show based on an auction mechanism as discussed in the previous lecture (captures the β)

Why do it this way?

(As opposed to first find all ads with utility > β, etc)

- **Ad relevance**: is a simple proxy for total utility:
  - Users – better experience
  - Advertisers – better (more qualified) traffic but possible volume reduction
  - SE gets revenue gain through more clicks but possible revenue loss through lower coverage
- However, ad relevance does not solve all problems
  - When to advertise: certain queries are more suitable for advertising than others
  - Interaction with the algorithmic side of the search

Web Queries

Queries are the driver of sponsored search

- Queries are a (very) succinct representation of the user's intent
  - The ultimate driver of the ad selection
  - Describe the need of the user
- Intent entropy is low in sponsored search!
- Before any grand design, let's look at the queries and their characteristics

Web User Needs – Types of Queries

- Type of queries [Brod02, RL04]
  - **Informational** – want to learn about something (~40% / 65%)

    `Swine Flu prevention`
  - **Navigational** – want to go to that page (~25% / 15%)

    `Alitalia US`
  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service

      `New York weather`
    - Downloads

      `Mars surface images`
    - Shop

      `Nokia mp3 phone`
  - Gray areas
    - Find a good hub

      `Rome hotels`
    - Exploratory search "see what's there"

Web Queries Statistics

- **Act 1**: 2003/2006 (Steve Beitzels thesis):
  - AOL queries
  - Dataset 1: a week of queries from Dec. 2003
  - Dataset 2: six months of queries Sept. 2005-Feb. 2005
- **Act 2**: 2010/2011 Yahoo! search queries (thanks John Xia and Zhaohui Zheng!)
  - Dataset 1: a week of queries from Sept. 2010
  - Dataset 2: six months of queries Oct. 2010 - Mar. 2011

Yahoo data set statistics

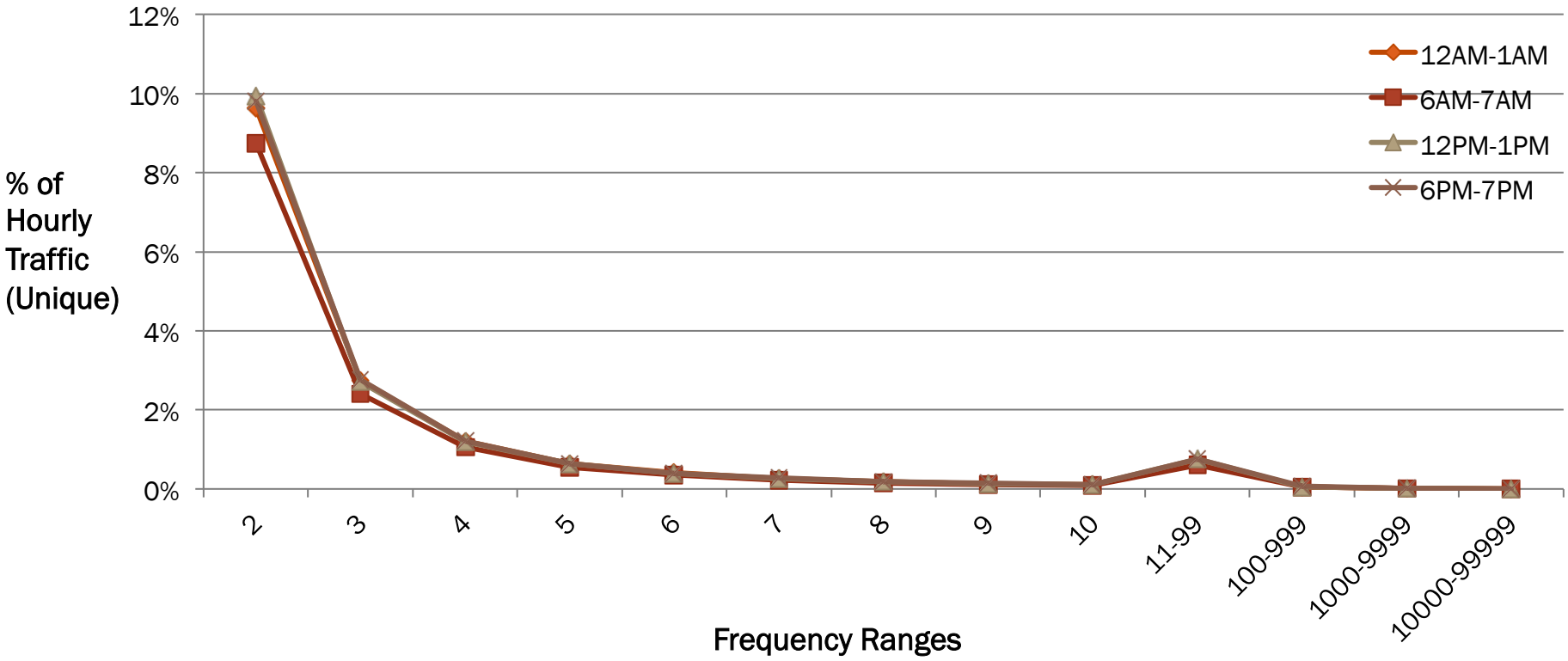
Property	One week	Six months
Number of Queries	Hundreds of Millions	Tens of Billions
Number of Users	Tens of Millions	Hundreds of Millions
Average Query Length	3.0 Terms	3.0 Terms
Average Popular Query Length	1.6 Terms	1.7 Terms
Portion of first results page views	86.6%	90.6%
Portion of second results page views	7.4%	4.5%
Portion of three or more pages views	6.0%	4.9%

AOL data set statistics

Table 2.1. Aggregate Query Log Statistics

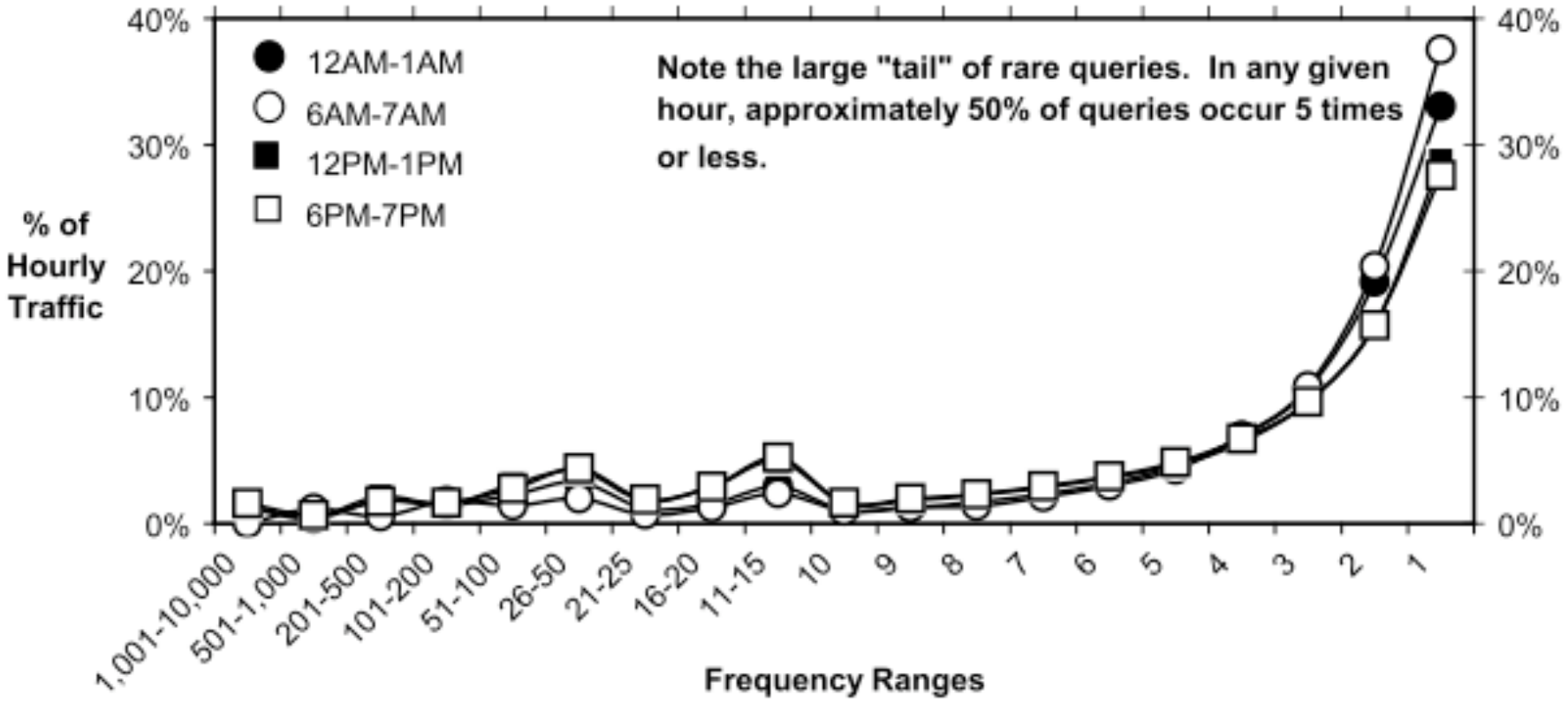
Property	One week	Six months
Number of Queries	Hundreds of Millions	Billions
Number of Users	Tens of Millions	Tens of Millions
Average Query Length	2.2 Terms	2.7 Terms
Average Popular Query Length	1.7 Terms	1.7 Terms
Portion of users viewing first results page	81%	79%
Portion of users viewing second results page	18%	15%
Portion of users viewing three or more pages	1%	6%

Frequency breakdown
The long tail

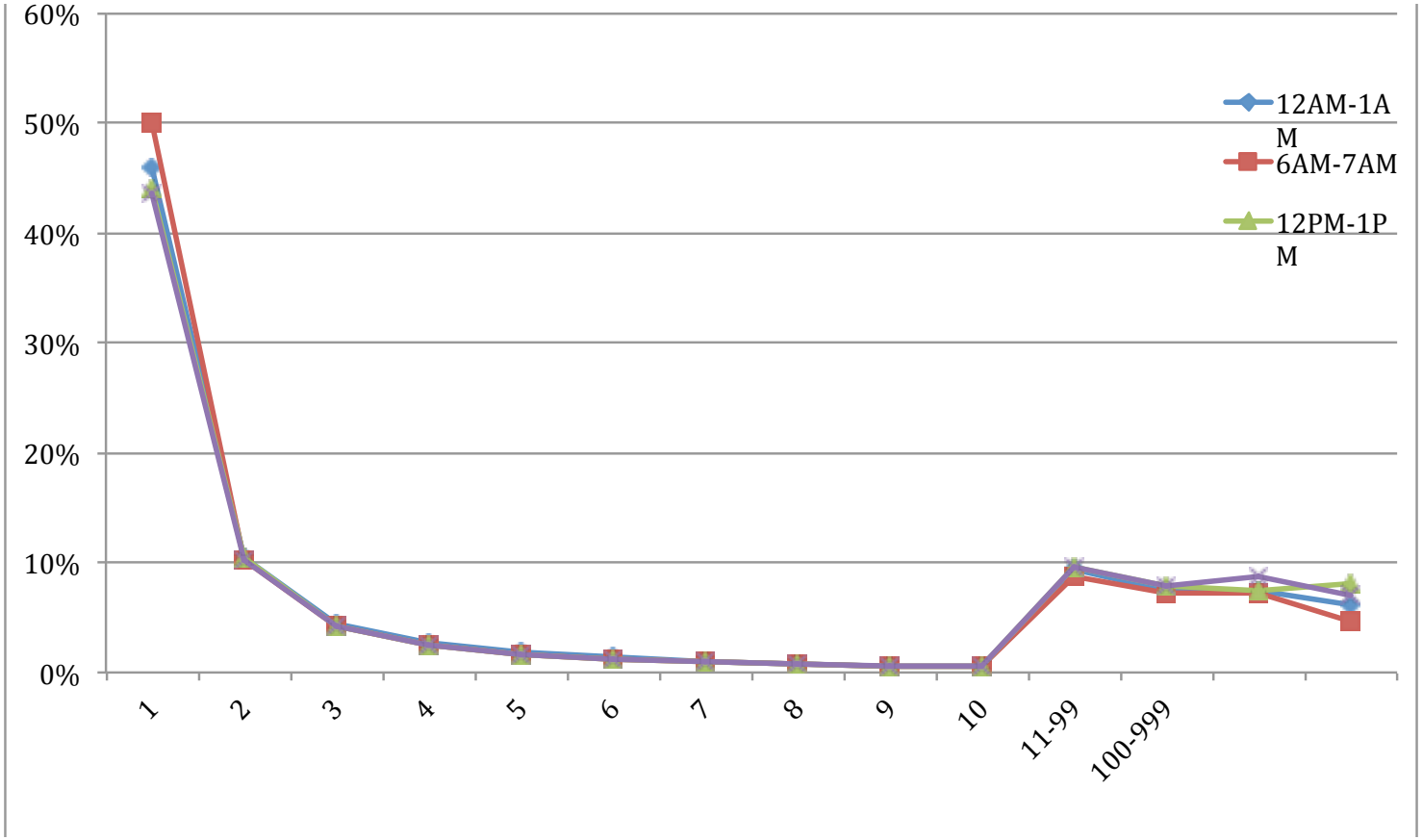


- Note the large "tail" of rare queries. In any given hour, more than 80% of queries occur one time.

Frequency breakdown
The long tail

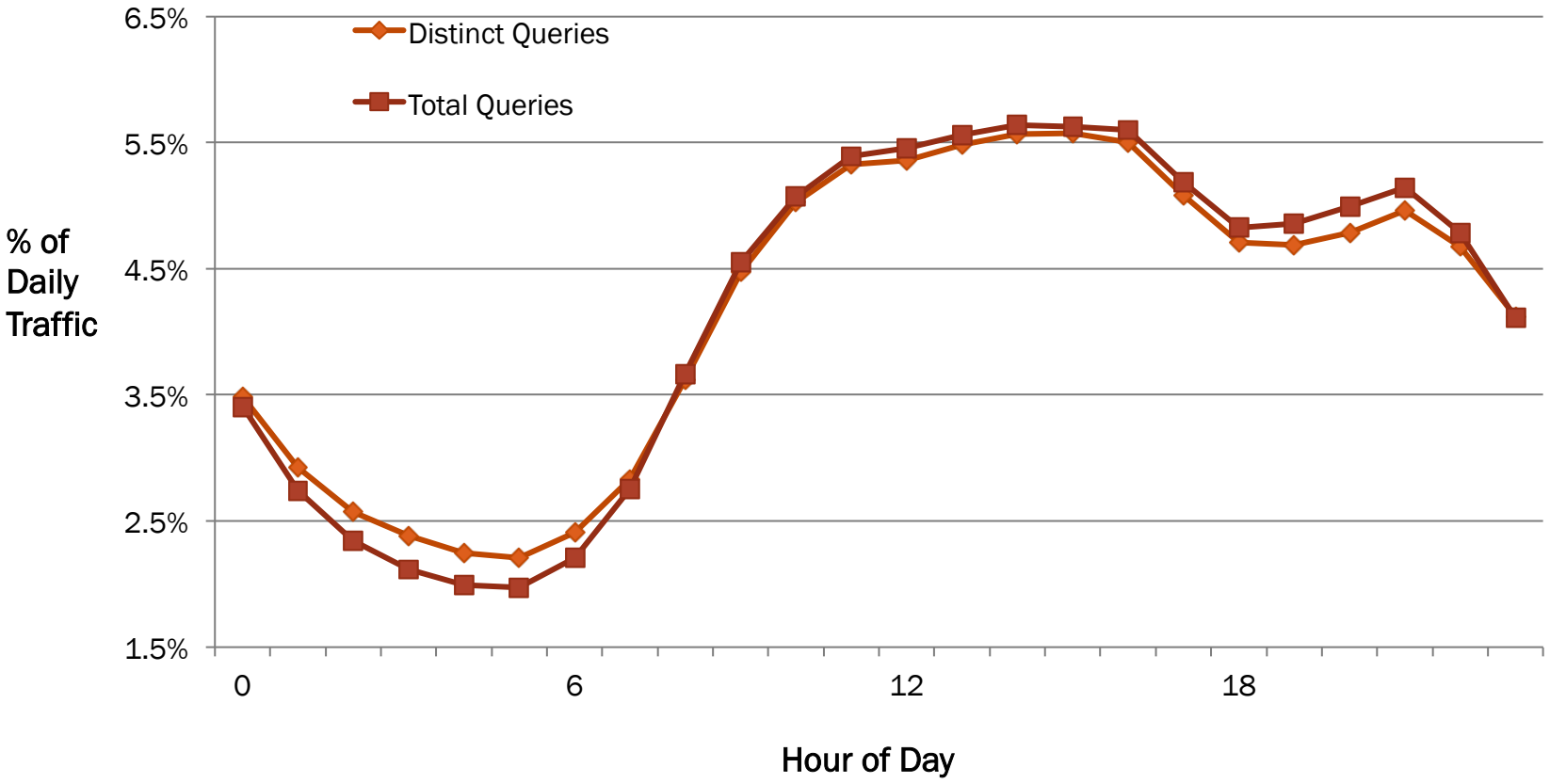


Volume per frequency



- Volume is mostly in the tail and in the head

Query Volume per Hour of the Day



Query Volume per Hour of the Day

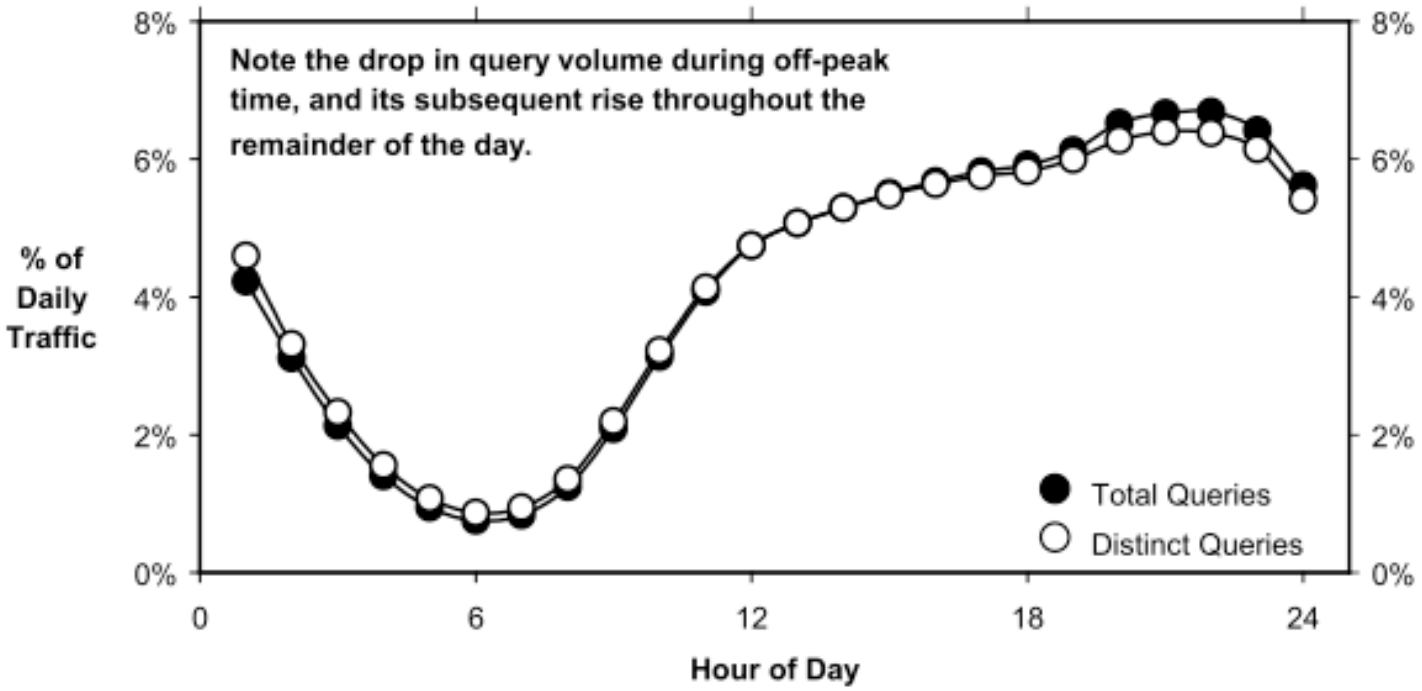
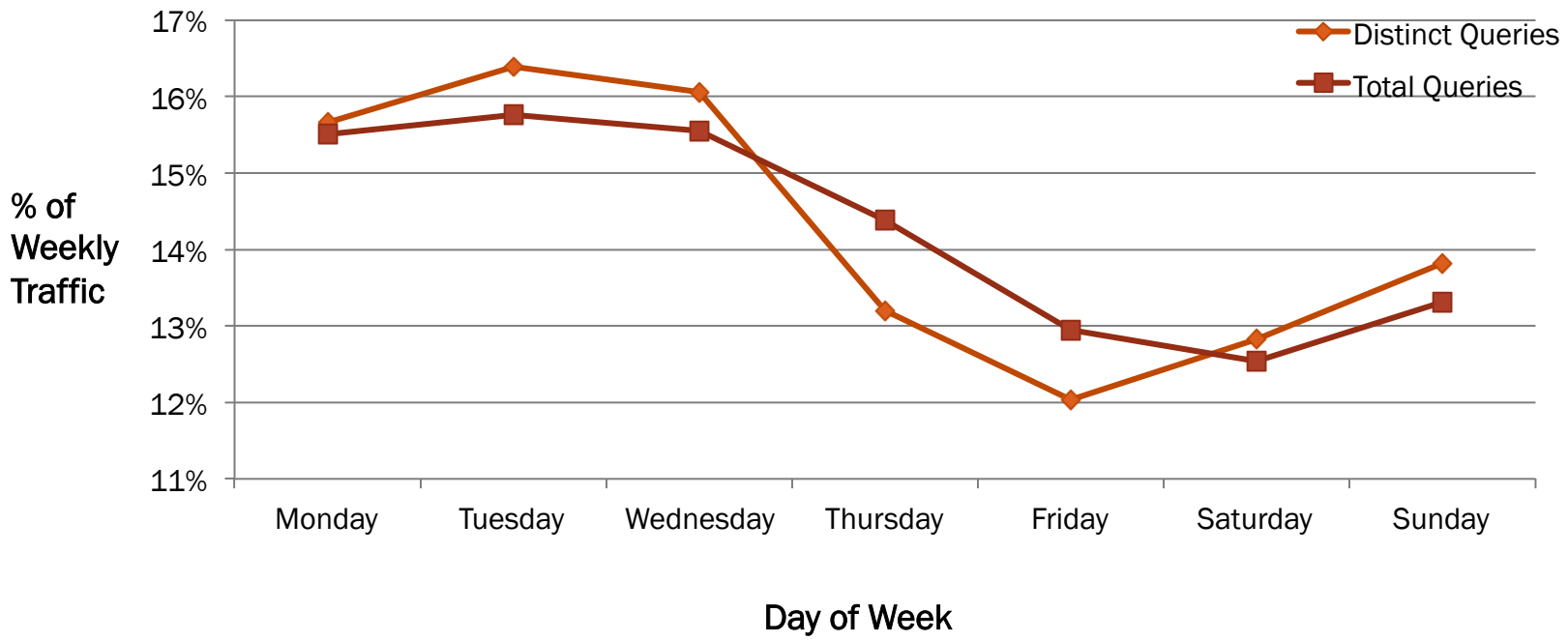


Figure 2.1. Query Volume Over a Day

Query Volume: Day of Week



Query Volume: Day of Week

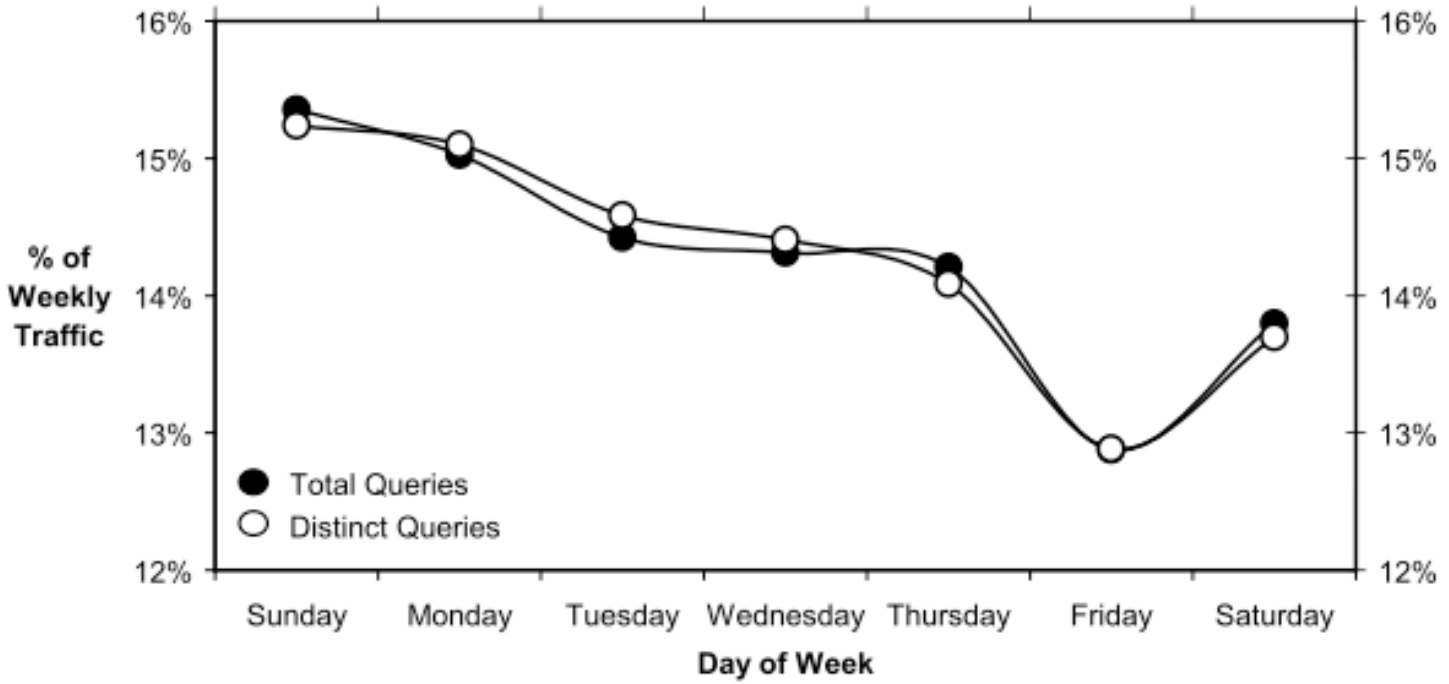
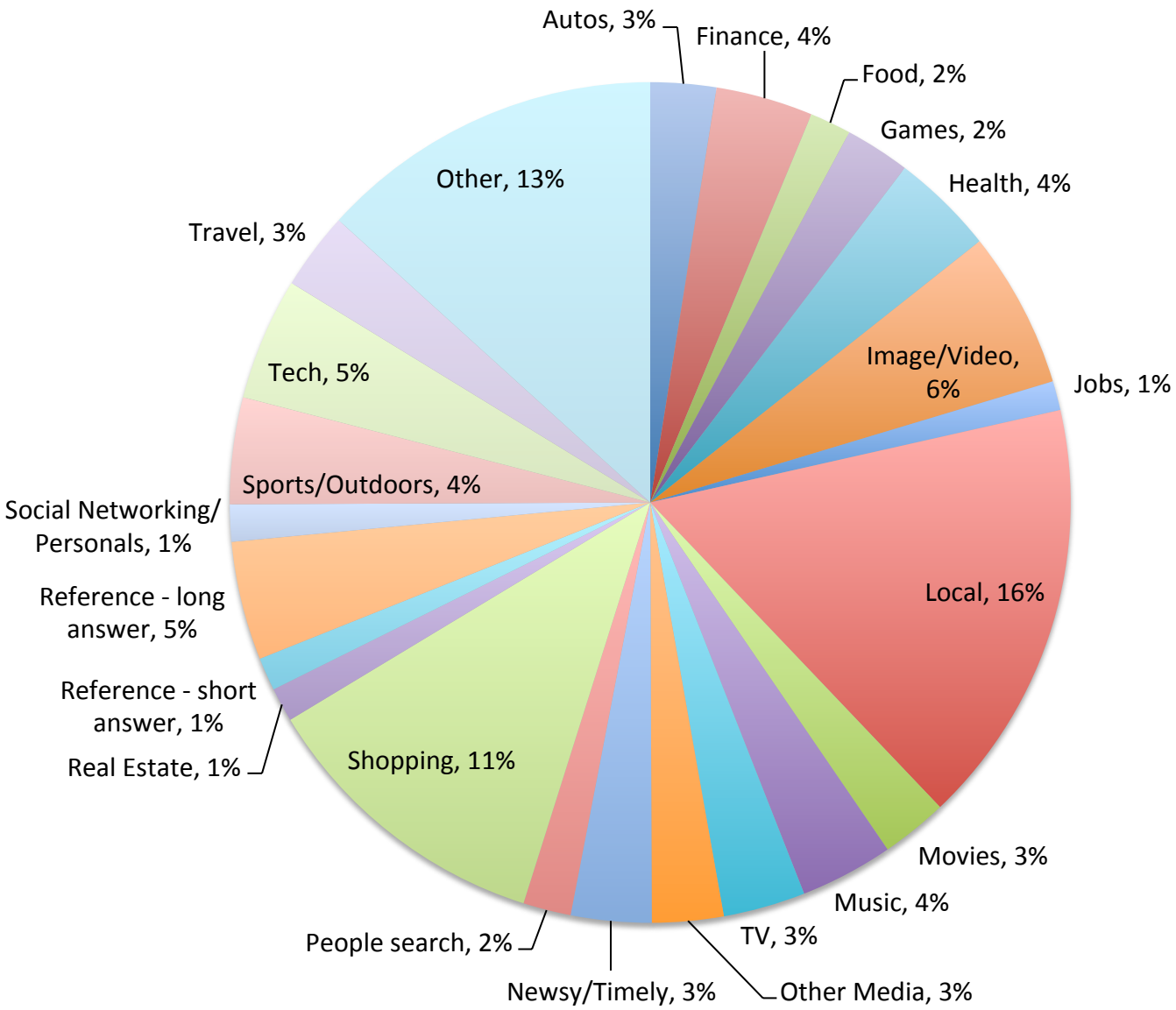


Figure 2.3. Average Volume of Days in the Week

Topical Distribution of Web Queries



Topical Distribution of Web Queries

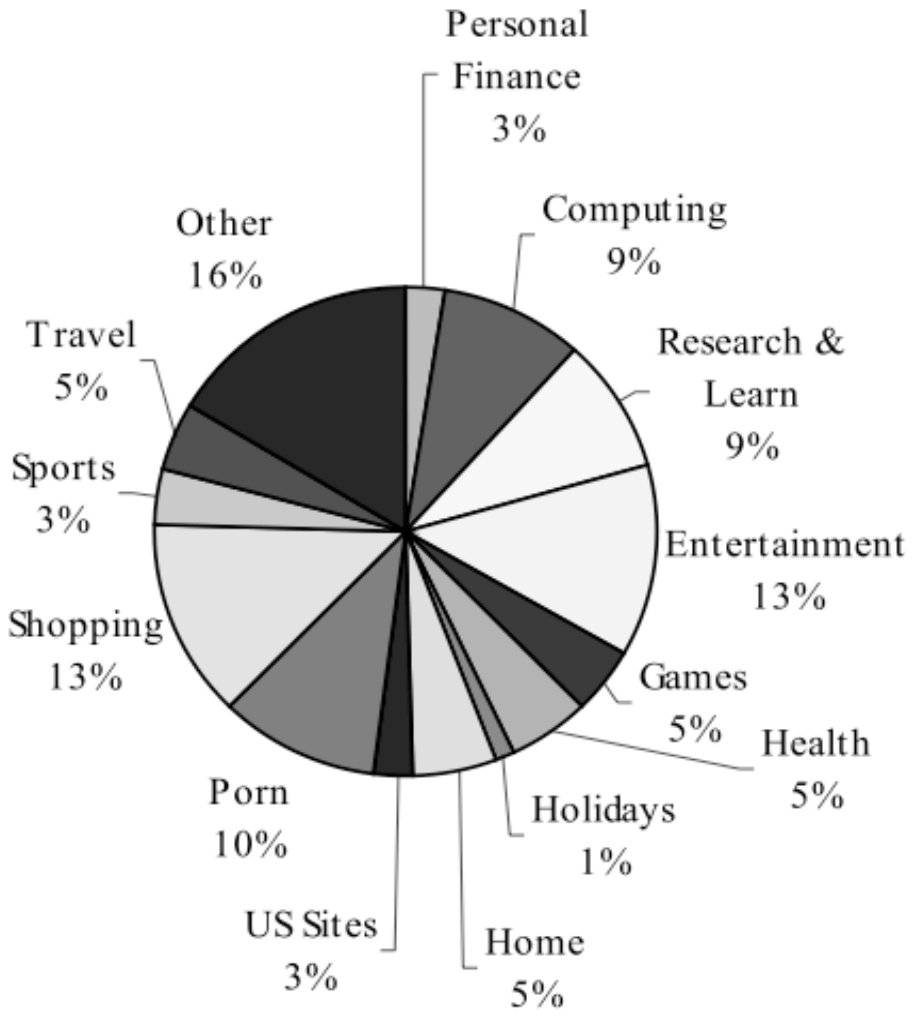


Figure 2.9. Breakdown of Categorized Queries

Long tail: why is it there

- How to explain the long tail?
- Web query example. Two options:
  - Most people query the 'usual' queries; a few do the 'unusual' ones
  - Large number people query the 'usual' queries; Most people also do a few unusual queries
- Study with online retailers supports the second hypothesis [Goel et al CIKM 2009]
  - Everybody is a bit eccentric, consuming both popular and niche products
  - However, consumers exhibit varying degrees of eccentricity
  - Availability of tail supply boosts even sales of popular items - one stop shop. (How does this map to search engines?)

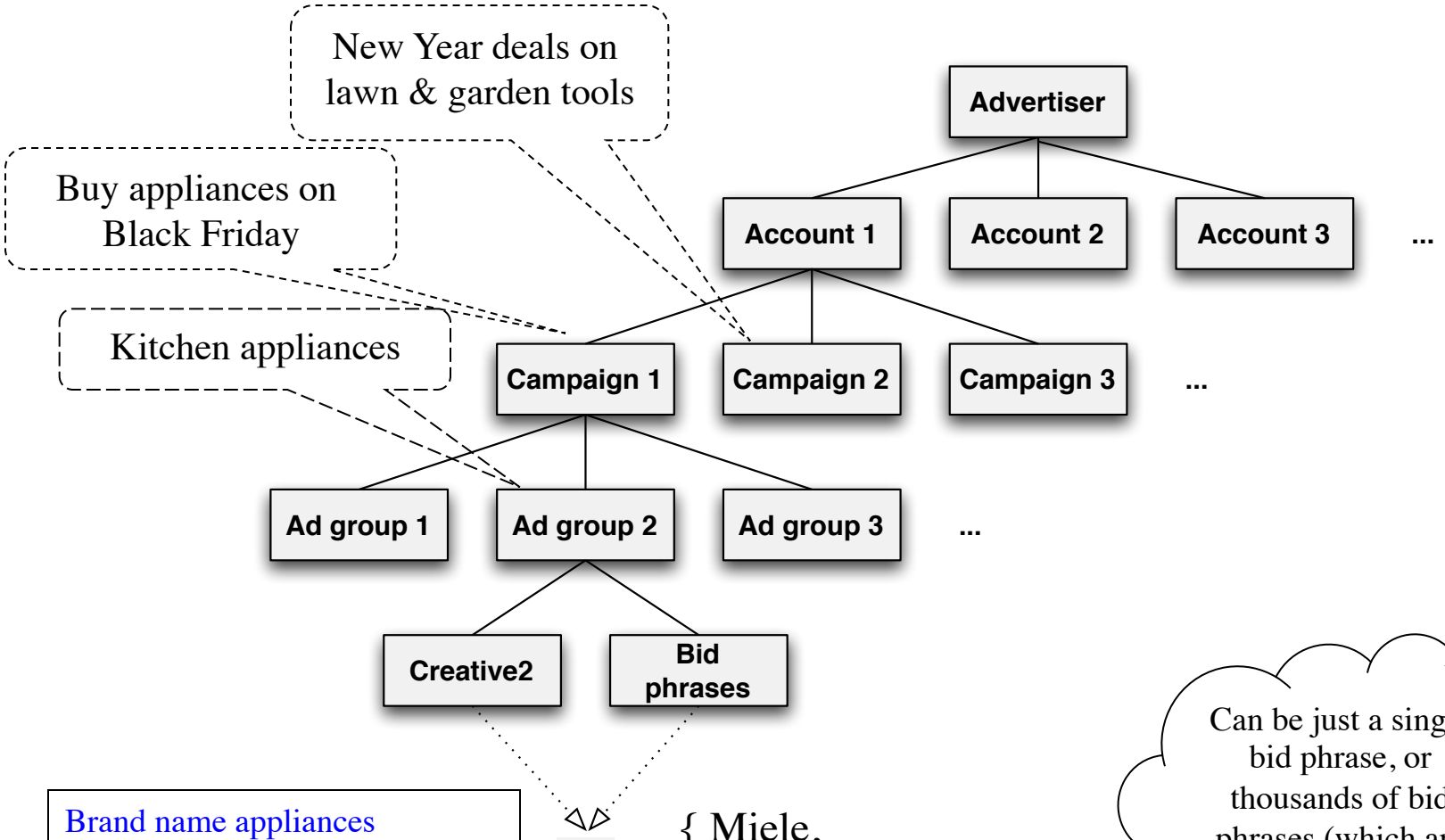
Textual Ads

Anatomy of a Textual Ad: the Visible and Beyond

Beyond a Single Ad

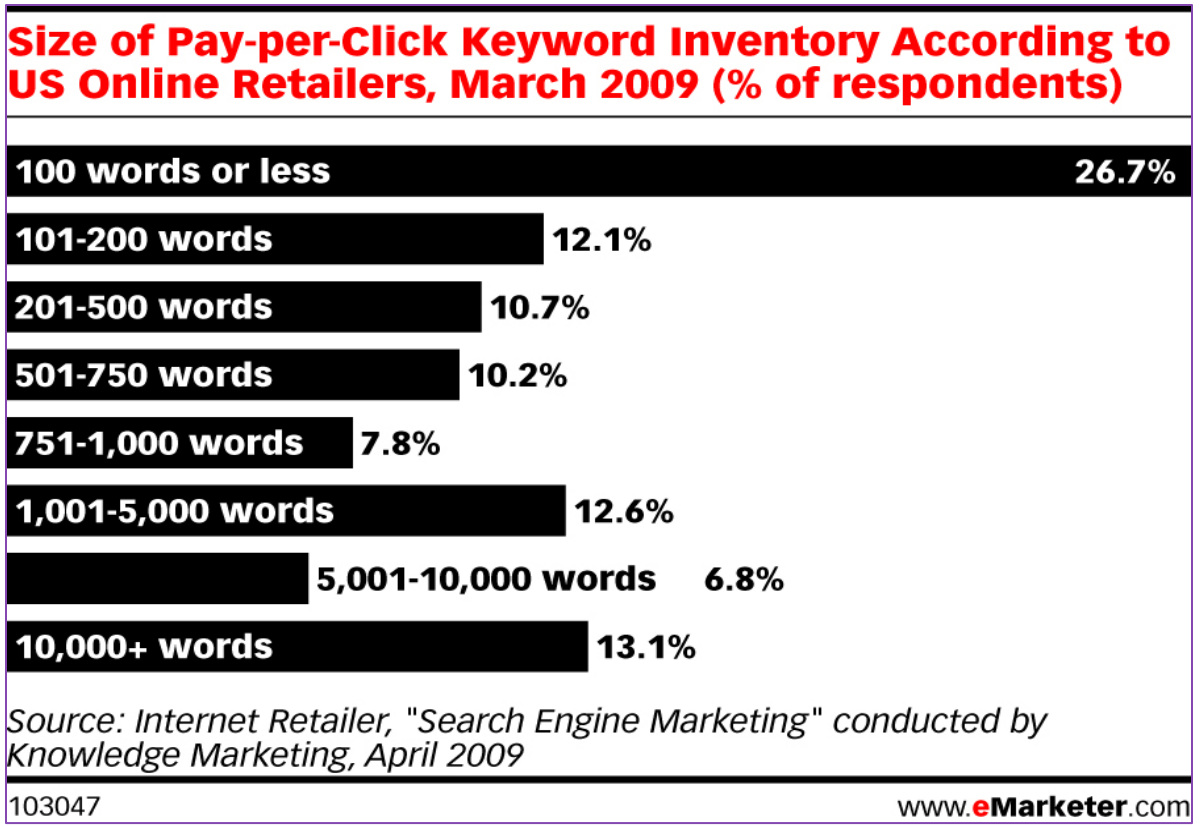
- Advertisers can sell multiple products
- Might have budgets for each product line and/or type of advertising (AM/EM) or bunch of keywords
- Traditionally a focused advertising effort is named a campaign
- Within a campaign there could be multiple ad creatives
- Financial reporting based on this hierarchy

Ad schema


New Year deals on lawn & garden tools
Buy appliances on Black Friday
Kitchen appliances
Advertiser
Account 1
Account 2
Account 3
...
Campaign 1
Campaign 2
Campaign 3
...
Ad group 1
Ad group 2
Ad group 3
...
Creative2
Bid phrases
Ad
Brand name appliances
Compare prices and save money
www.appliances-r-us.com
{ Miele, KitchenAid, Cuisinart, …}
Can be just a single bid phrase, or thousands of bid phrases (which are not necessarily topically coherent)

Taxonomy of sponsored search ads

- Advertiser type
  - Ubiquitous: bid on query logs. Yahoo Shopping, Amazon, Ebay, . . .
  - Mom-and-pop's shop
  - Everything in the middle



Ad-query relationship

- **Responsive**: satisfy directly the intent of the query
  - query: Realgood golf clubs
  - **ad**: Buy Realgood golf clubs cheap!
- **Incidental**: a user need not directly specified in the query
  - **Related**: Local golf course special
  - **Competitive**: Sureshot golf clubs
  - **Associated**: Rolex watches for golfers
  - **Spam**: Vitamins

Types of Landing Pages

[H. Becker, AB, E. Gabrilovich, VJ, B. Pang, SIGIR 2009]

- Classify landing page types for all the ads for 200 queries from the 2005 KDD Cup labeled query set. Four prevalent types:

I. **Category (37.5%):** Landing page **c**aptures the broad category of the query

II. **Search Transfer (26%):** Land on dynamically generated search results (same q) on the advertiser's web page

a) Product List – search within advertiser's web site

b) Search Aggregation – search over other web sites

III. **Home page (25%)**: Land on advertiser's home page

IV. **Other (11.5%):** Land on promotions and forms

Ad Selection

Dichotomy of sponsored search ad selection methods

- **Match types**
  - Exact – the ad's bid phrase matches the query
  - Advanced - the ad platform finds good ads for a given query
- **Implementation**
  - Database lookup
  - Similarity search
- **Phased selection**
- **Reactive vs predictive**
  - Reactive: try and see using click data
  - Predictive: generalize from previous ad placement to predict performance
- **Data used (for predictive mostly)**
  - Unsupervised
  - Click data
  - Relevance judgments

Match types

- For a given query the engine can display two types of ads:
- **Exact match** (EM)
  - The advertiser bid on that specific query a certain amount
- **Advanced match** (AM) or "Broad match"
  - The advertiser did not bid on that specific keyword, but the query is deemed of interest to the advertiser.
  - Advertisers usually opt-in to subscribe to AM

Exact Match Challenges

- **What is an exact match**?
  - Is "Miele dishwashers" the same as
    - Miele dishwasher (singular)
    - Meile dishwashers (misspelling)
    - Dishwashers by Miele (re-order, noise word)
  - Query normalization
- **Which exact match to select among many?**
  - Varying quality
    - Spam vs. Ham
    - Quality of landing page
  - Suitable location
  - More suitable ads (E.g. specific model vs. generic "Buy appliances here")
  - Budget drain
    - Cannot show the same ad all the time
  - Economic considerations (bidding, etc)

Advanced match

- Significant portion of the traffic has no bids
  - Advertisers need volume
  - Search engine needs revenue
  - Users need relevance!
- Advertisers do not care about bid phrases – they care about conversions = selling products
- How to cover all the relevant traffic?
- From the SE point of view AM is much more challenging

Advertisers' dilemma: example

- Advertiser can bid on "broad queries" and/or "concept queries"
  - Suppose your ad is:
    - "**Good prices on Seattle hotels**"
  - Can bid on any query that contains the word **Seattle**
- Problems
  - What about query "**Alaska cruises start point**"?
  - What about "**Seattle's Best Coffee Chicago**"
- Ideally
  - Bid on any query related to Seattle as a travel destination
  - We are not there yet …
- Market Question: Should these "broad matches" be priced the same?
  - Whole separate field of research
- In the remaining of the lecture we will discuss several mechanisms for advanced match

Implementation approaches

1. **The data base approach** (original Overture approach)
   - Ads are records in a data base
   - The bid phrase (BP) is an **attribute**
   - On query q
     - For EM consider all ads with BP=q
2. **The IR approach** (modern view)
   - Ads are documents in an **ad corpus**
   - The bid phrase is a meta-datum
   - On query q run q against the ad corpus
     - Have a suitable ranking function (more later)
     - BP = q (exact match) has high weight
     - No distinction between AM and EM

The two phases of ad selection

- **Ad Retrieval**: Consider the whole ad corpus and select a set of most viable candidates (e.g. 100)
- **Ad Reordering**: Re-score the candidates using a more elaborate scoring function to produce the final ordering
- Why do we need 2 phases:
  - Ad Retrieval:
    - considers a larger set of ads, using only a subset of available information
    - might have a different objective function (e.g. relevance) than the final function
  - Ad Reordering
    - Limited set of ads with more data and more complex calculations
    - Must use the bid in addition to the retrieval score (e.g. revenue as criteria for the ordering, implement the marketplace design()
- Note that this is all part of the α on slide 17. Some times the second phase bundled with the reordering

Reactive vs. predictive reordering

Example: Horse races

Name	Starts	1st	2nd	3rd	Earnings
Well Armed	4	1	1	0	$3,649,000
Rachel Alexandra	8	8	0	0	$2,746,914
Summer Bird	8	4	1	1	$2,023,040
Mine That Bird	6	1	2	2	$1,892,200
Regal Ransom	3	2	0	0	$1,650,000
Gio Ponti	6	4	1	0	$1,433,000
Einstein	6	2	1	2	$1,269,304
Gloria de Campeao	2	0	1	0	$1,210,000
Pioneerof the Nile	5	3	1	0	$1,090,000
Swift Temper	8	3	3	1	$1,079,497

Updated through 10/6/2009

- Reactive:
  - Follow Summer Bird
  - See how it did in races
  - Predict the performance
- Predictive
  - Make a model of a horse: weight, jockey weight, leg length
  - Find the importance of each feature in predicting a win/position
  - Predict performance of unseen (and seen) horses based on the importance of these features
- **When we have enough information for a given horse use it (reactive), otherwise use model (predictive)**

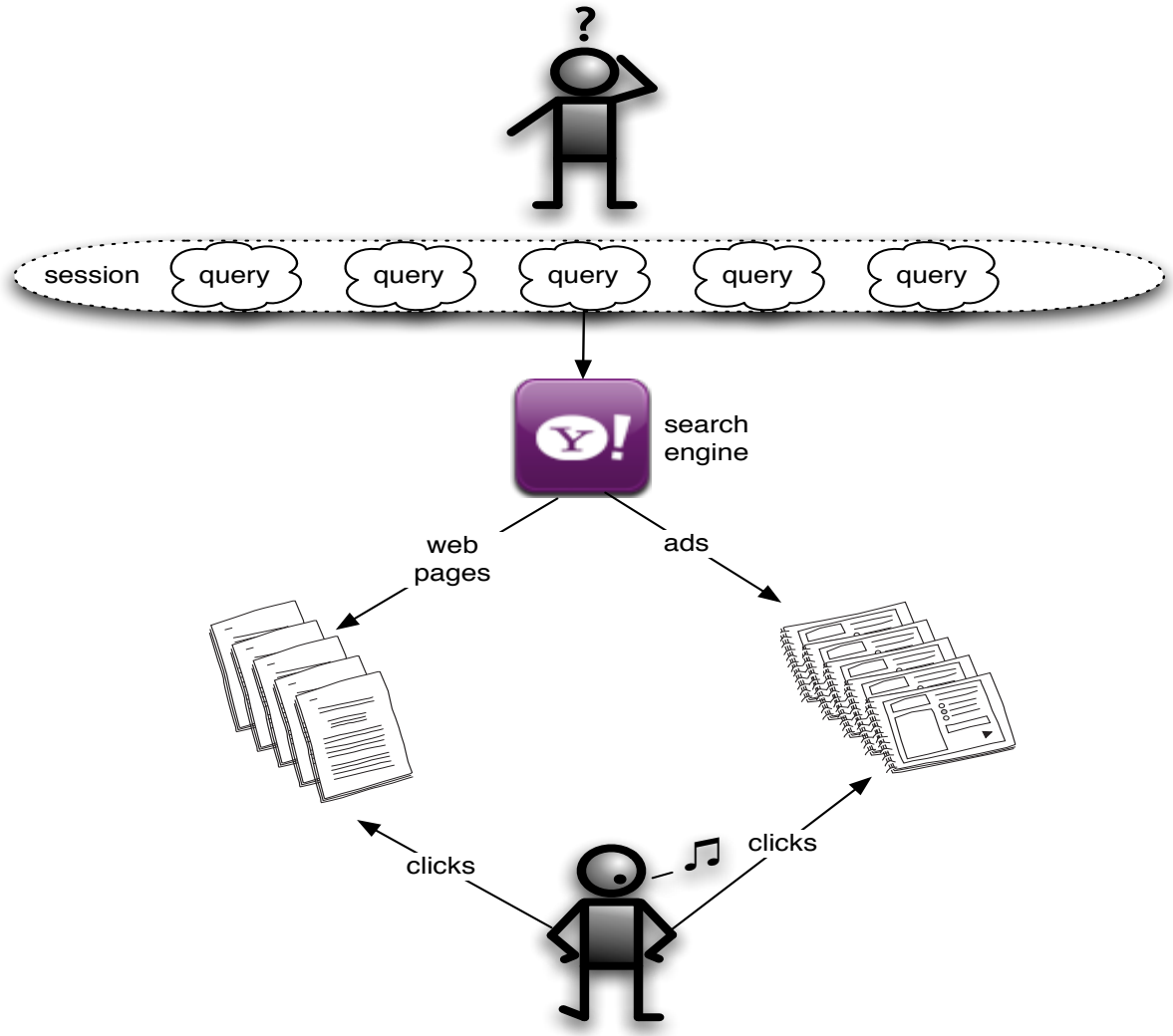
Reactive vs predictive methods in sponsored search

- All advanced match methods aim to maximize some objective
  - Ad-query match
  - query-rewrite similarity
- What is the unit of reasoning?
- Individual queries/ads
  - Can we try all the possible combinations enough times and conclude? We might for common queries and ads
  - Recommender system type of reasoning (query q is similar to query q')
- Features of the queries and ads: words, classes, etc
  - Generalize from the ads to another space
  - Predict performance of unseen ads and queries
- Hybrid approaches:
  - What if we aggregate CTR at campaign level?
  - Get two predictions, how to combine?

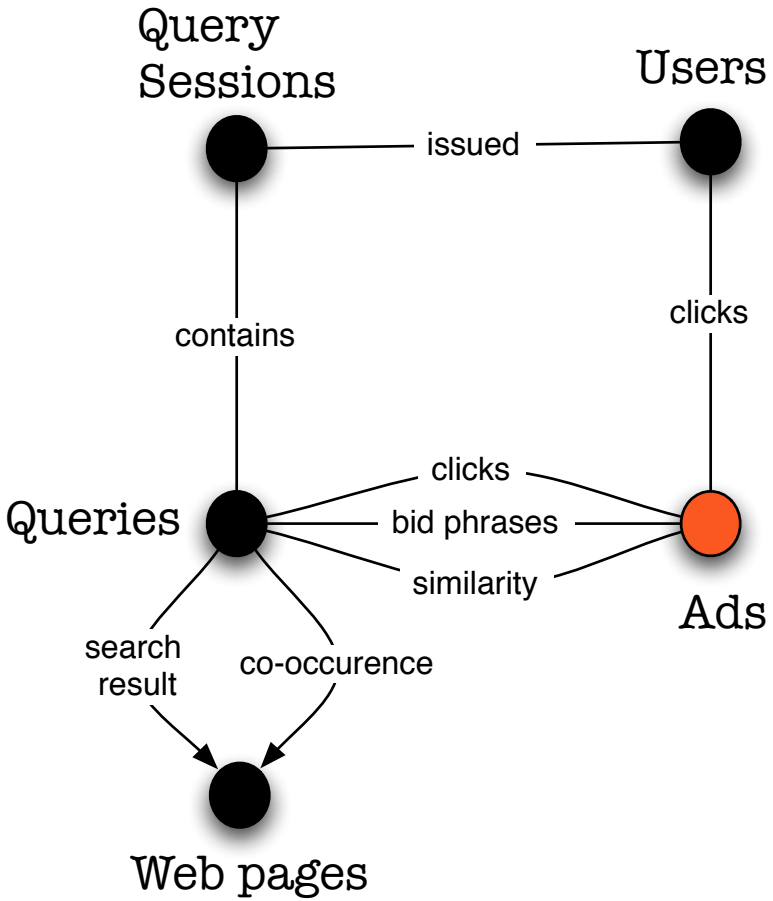
Indication of success: relevance and click data

- **Relevance data**
  - Limited editorial resources
  - Editors require precise instruction of relevance
  - How to deal with multiple dimensions?
  - Editors cannot understand every domain and every user need
- **Click data**
  - Higher volume – might need sampling
  - Binary (click/no click)
  - Click-through-rate (CTR) usually very low (a few percent)
  - People do not click on ads even when they are relevant
  - Much more noise

Sponsored search ad selection is data driven. It is computational!



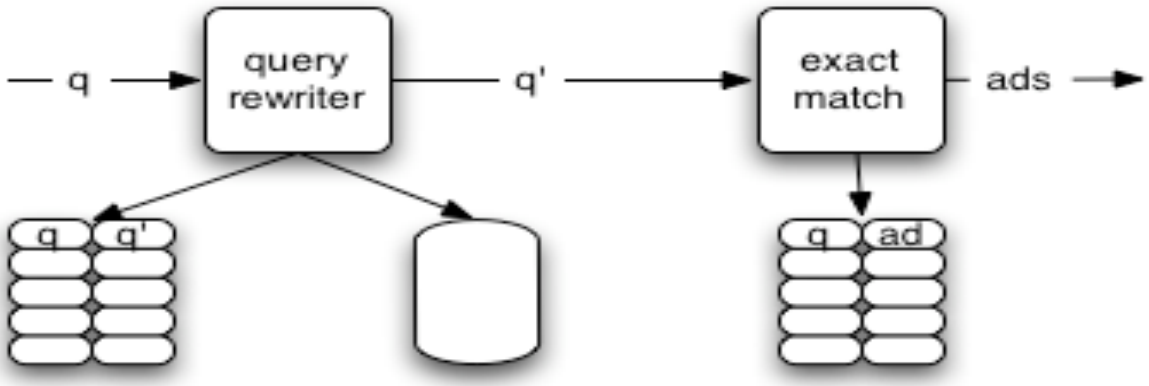
Data Source


Query Sessions
Users
issued
contains
clicks
clicks
Queries
bid phrases
similarity
Ads
search result
co-occurence
Web pages

Query Rewriting for Sponsored Search

Typical query rewriting flow

- Typical of the DB approach to AM
- Rewrite the user query q into Q' = (q1, q2, …)
- Use EM to select ads for Q'
- Fits well in the current system architectures



Keyword suggestion – related problem

- Guessing the keyword for the advertiser has some risks
  - Tolerance/value of precision vs. volume differs among advertisers
  - Additional issue: what to charge the advertiser in advanced match
- Semi-automatic approach:
  - Propose rewrites to advertisers
  - Let them chose which ones are acceptable
  - Advertiser determines the bid
- Keyword suggestion tools draw upon similar data and technologies as advanced match

Online vs. offline rewriting

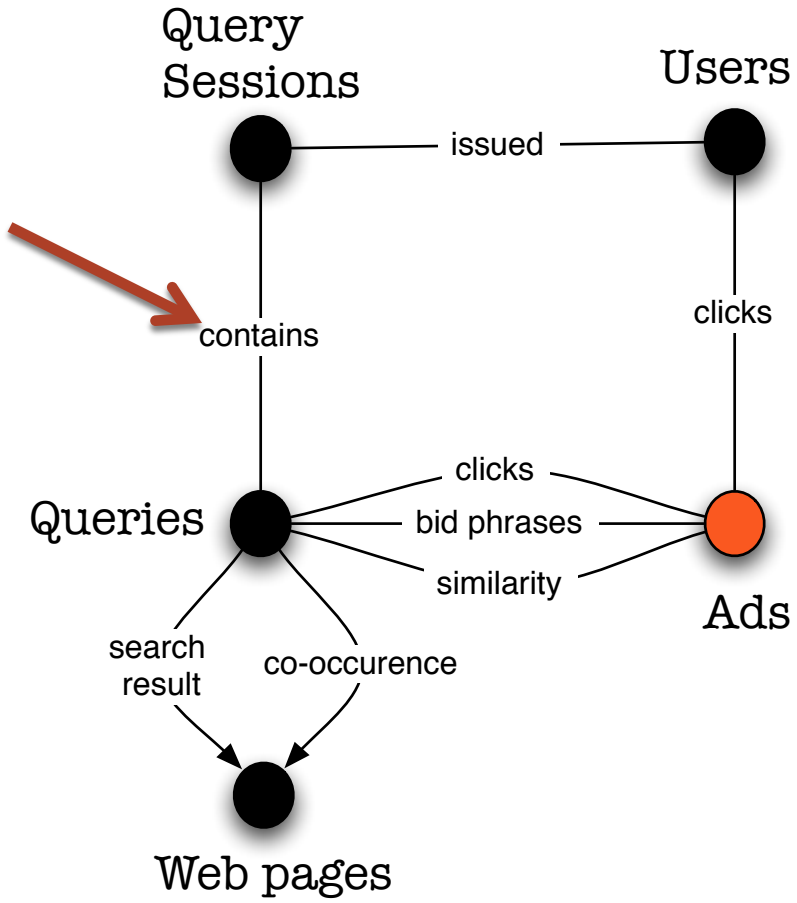
- **Offline**
  - Process queries offline
  - Result is a table of mappings q→q'
  - Can be done only for queries that repeat often
  - More resources can be used
  - Question: what common queries we should be rewriting: where we need depth of market
  - What queries do we rewrite into?
- **Online**
  - For rare queries offline not practical or simply does not work
  - Lot less time to do analysis (a few ms)
  - Limited amount of data (memory bound, time bound)

Sponsored Search: query rewriting reading list (part 1)

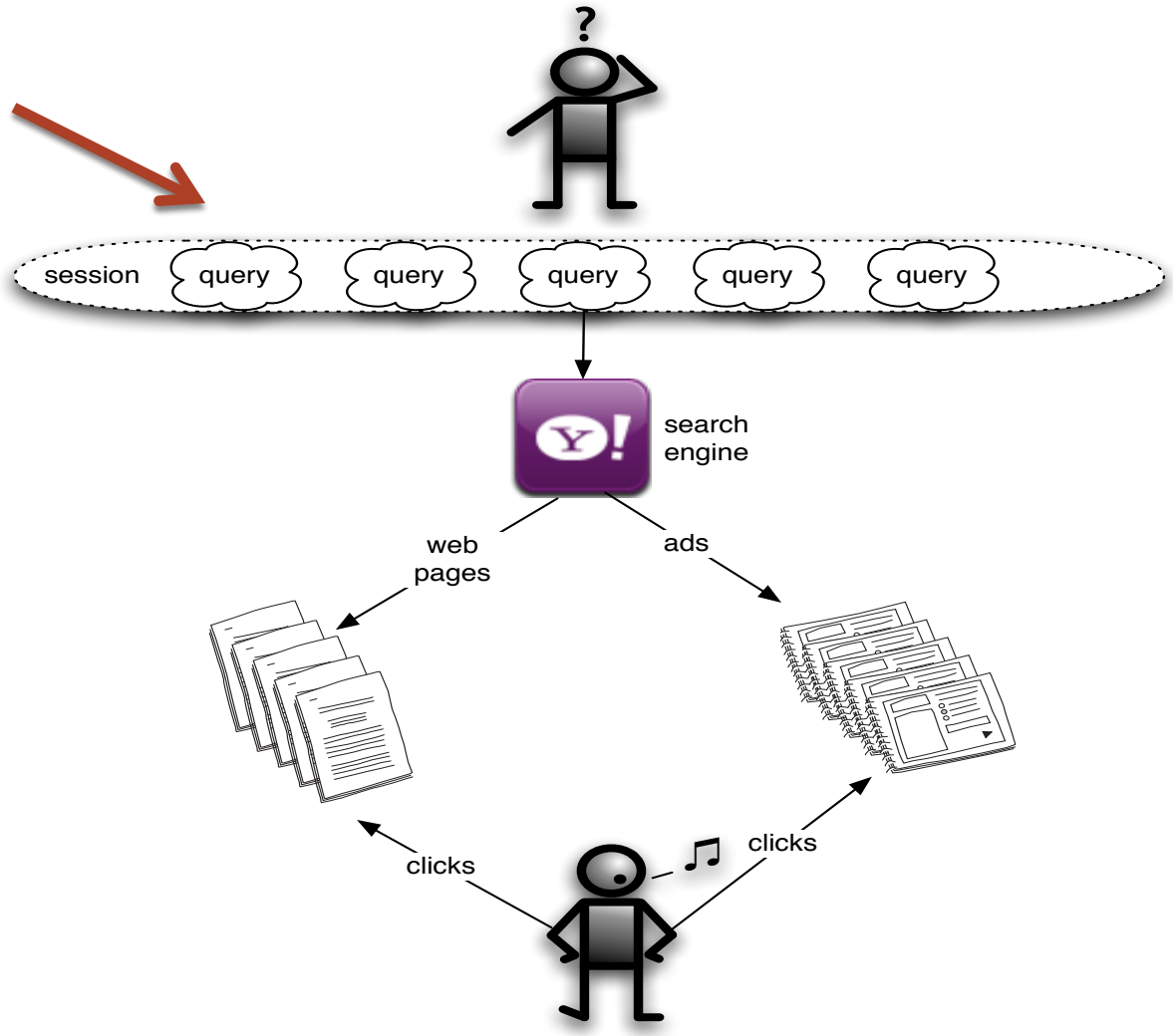
	Query rewriting technique	Data source
1.	Generating Query Substitutions: Jones et al, in Proc of WWW 2006	query logs (query sessions)
	Using the Wisdom of the Crowds for Keyword Generation: Fuxman et al., In proc of WWW 2004	co-cliks on web search results
2.	Simrank++: Query Rewriting through Link Analysis of the Click Graph: Atoanellis et al., In proc of VLDB 2008	co-clicks on ads
3.	Learning Query Substitutions for Online Advertising: Broder et al. in Proc of ACM SIGIR 2008	query-to-ad similarity
4.	Online Expansion of Rare Queries for Sponsored Search: Broder et al, In Proc. of WWW 2009	query-to-query similarity
5.	Query Word Deletion Prediction: Jones at al., in Proc of ACM SIGIR 2003	query logs

Query Rewriting using Web Search Logs

Data Source


Query Sessions
Users
issued
contains
clicks
clicks
Queries
bid phrases
similarity
Ads
search result
co-occurence
Web pages

Data source: relationship between queries and sessions



User sessions

- A user uses the search engine to complete a task
- Task completion will usually take several steps:
  - Queries
  - Browsing
- For query rewriting we can focus on the query stream
- Finding the session boundaries – research problem
  - Time period (all queries within 24hrs)
  - Machine learned approach based on query similarity or labeled set
- How to identify queries that are suitable for rewriting?
  - Examine the different types of rewrites that the users do
  - Get enough instances of the rewrite to be able to determine its value

Example session: trying to find the web page of this course

1. Computation in Advertising class Stanford ← first try
2. Computation in Advertising ← generalization, try find more general info on CA
3. Computational Advertising class Stanford ← got terminology right, back to task
4. VTA timetables Palo Alto ← another sessions (interleaved)
5. Computational Advertising Andrey Brodski Stanford ← back to work: specialization
6. Computational Advertising Andrei Broder Stanford ← spelling correction
7. Raghavan Manning Stanford class ← give up, start another session

Half of the Query Pairs are Reformulations

Type	Example	%
switch tasks	mic amps -> create taxi	53.2%
insertions	game codes -> video game codes	9.1%
substitutions	john wayne bust -> john wayne statue	8.7%
deletions	skateboarding pics -> skateboarding	5.0%
spell correction	real eastate -> real estate	7.0%
mixture	huston's restaurant -> houston's	6.2%
specialization	jobs -> marine employment	4.6%
generalization	gm reabtes -> show me all the current auto rebates	3.2%
other	thansgiving -> dia de acconde gracias	2.4%

[Jones & Fain SIGIR 2003]

Many substitutions are repeated

- Some substitutions are incidental
- Others repeat often with different users in different days
  - car insurance → auto insurance
    - 5086 times in a sample
  - car insurance → car insurance quotes
    - 4826 times
  - car insurance → geico [brand of car insurance]
    - 2613 times
  - car insurance → progressive auto insurance
    - 1677 times
  - car insurance → carinsurance
    - 428 times

A principled way to determine when are we sure in the rewrites

- Determine if

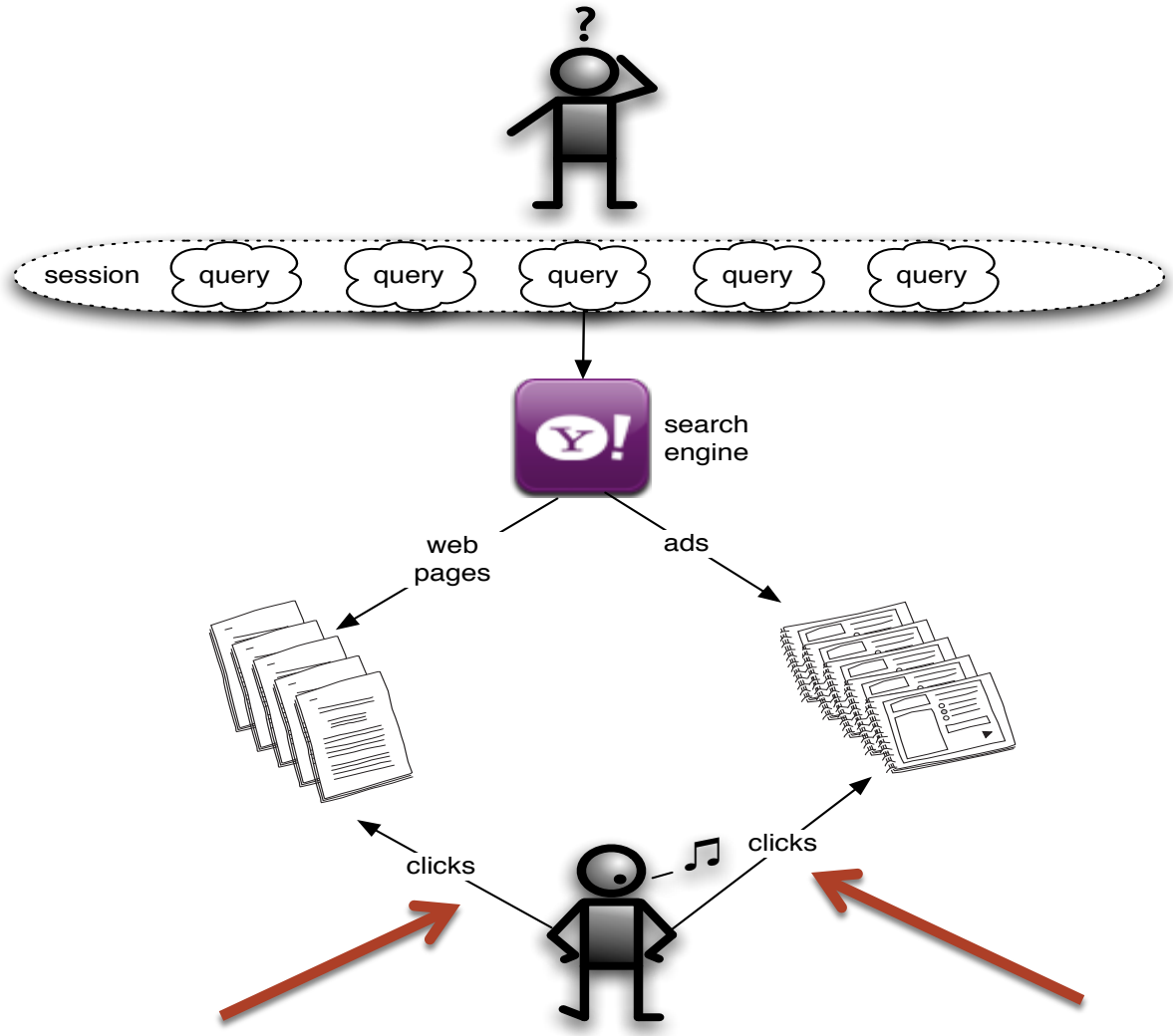
  $p(rw \mid q) >> p(rw)$
- Since $p(rw \mid q) = p(rw,q)/p(q)$, this depends on the relative magnitude of $p(rw,q)$ and $p(q)$, $p(rw)$
- How do we estimate $p(rw,q)$ and $p(q)$?
- Maximum likelihood: frequencies in the training data
- Assume an underlying distribution – binomial
- Test two hypothesis:
  - H1: $P(rw \mid q) = P(rw \mid \neg q)$
  - H2: $P(rw \mid q) \neq P(rw \mid \neg q)$
- The the log likelihood rato $-2\log(L(H1)/L(H2)$ is asymptotically χ^2 distributed
- Other statistical tests can be used – pick you favorite

Query logs as rewrite source – summary

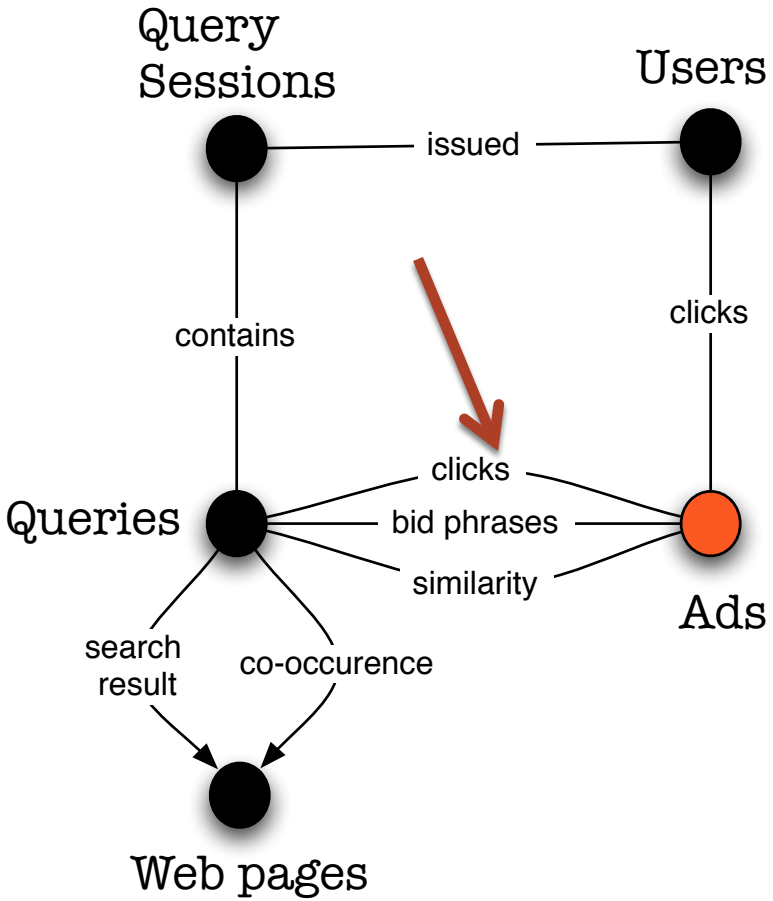
- Use the knowledge of the users to generate rewrites
- Practical and useful approach, however a few tough challenges:
  - Sessions boundaries
  - Type of the rewrites
  - Requires relatively high frequency of rewrites to be detected

Clicks graphs and random walks for query rewrite generation

Data source: clicks



Data source


Query Sessions
Users
issued
contains
clicks
clicks
Queries
bid phrases
similarity
Ads
search result
co-occurence
Web pages

Problem definition

- Given a bipartite graph G:

  $G = (V, E)$

  $V = V_q \bigcup V_a$

  $E = \{e_1 \ldots e_k\} \quad e_i = (q, a, w) \quad q \in V_q \quad a \in V_a \quad w \in R$

  - Vq – nodes representing queries
  - Va – nodes representing ads
  - Problem virtually the same if Va represents URLs of organic search
  - Edges connect queries with ads.
  - Each edge has one or more weights
- For each pair of queries (q_1, q_2) determine the similarity $sim(q_1, q_2)$

Weighting the edges

- **Un-weighted**: there is an edge for each ad query pair where there is at least on click
  - Issue – some ads get a lot more clicks than others for the same query
- **Clicks**: weight the edges with the number of clicks on the (q,a) combination
  - Pairs with higher number of impressions get more clicks even if the relationship is not as strong
- **CTR**: keep the ratio between the clicks and impressions
  - CTR of 0.5 differs in confidence when we have 1 or 10K impressions
- We will examine the use of CTR as weights

Interpreting clicks: positional bias

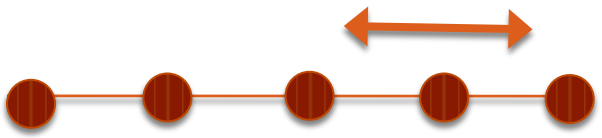
- Ads shown on position 1 are more likely to get clicks even if they are less relevant
- How does this impact the training in our click-based weighting system?
- If the clicks of an ad are all at position 1
  - Are those clicks because the ad was relevant?
  - Or are those clicks caused by the inherent bias of the user to click the top ad?
- A study has shown that even if you swap the ads on position 1 and 2, position 1 still gets more clicks

De-biasing click data - click models

- To deal with this bias we need a model of user behavior
- Model #1: p(click)=p(seen)p(relevant)
  - Ads at position 1 are more likely to be seen than other positions
  - Ads at position 1 are more likely to be relevant: ranked retrieval
- We need to separate the positional and relevancy effect
- Use normalized CTR by the expected CTR at a position:
  - "The ad a is twice more likely to be clicked than an average ad at the same position"
- Count an impression only if the ad has been seen – if there is a click on a lower position – "Cascade model"
  - [Craswell et al, WSDM 2008]
- Active research area

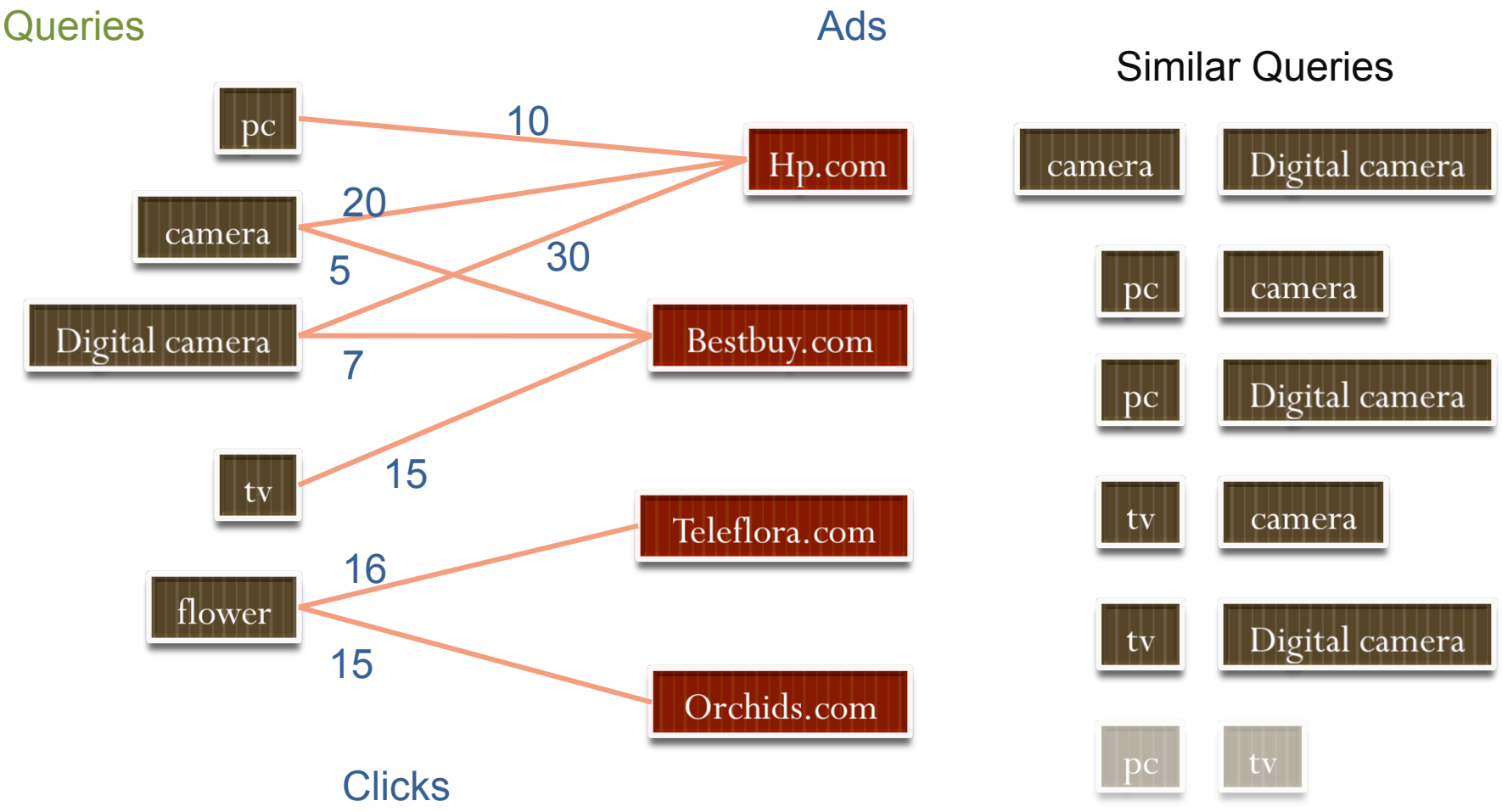
Random Walks in Graphs

- Imagine a 'drunkard' walking in a city
- At intersection she takes a random decision which way to go
  - Including back from where he came
  - Imagine this in 1 dimension



- What is the probability that he will reach a certain place (node)?
- **Static distribution** – node probabilities converge as the number of steps goes toward infinity
- **Markov process**: discrete steps, next position depends only on the current (no memory)

Click Graph from sponsored search



Similar Queries

camera	Digital camera
pc	camera
pc	Digital camera
tv	camera
tv	Digital camera
pc	tv

Ioannis Antonellis et al PVLDB08

Random walk in bi-partite graph

- Two types of nodes
- Can move from one type to the other only
- The walker goes back-and forth between the two sides
- Many different applications in Computational Advertising
  - Query-clicks on ads
  - Query- clicks on URLS
  - ….

Simrank – Similarity ranking algorithm for bipartite graphs

- Intuition:
  - "Two queries are similar if they are connected to similar ads"
  - "Two ads are similar if they are connected to similar queries"
- Assume similarity is a measure between 1 and 0 (like probability)
- A query is "very" similar to itself: sim(q,q) = 1
- Initially, we know nothing about the similarity with other queries: sim (q,q') = 0 iff q ≠ q'
- Establish similarity of two queries based on the ads they connect to
- Then the same on the ad side
- Random walk starting at q and q' simultaneously – end up in the same node
- Iterative procedure: at each iteration similarity propagates through the the graph

83

Simrank algorithm

- E(q): set of ads connected to q
- N(q): # of ads connected to q
- $sim_k(q,q')$: q-q' similarity at k-th iteration
- Initially sim(q,q) = 1, sim(q,q') = 0, sim(a,a) = 1, sim(a,a') = 0

$$sim_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} sim_{k-1}(i,j)$$

$$sim_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} sim_{k-1}(i,j)$$

- C – constant between 0 and 1, ensures diminishing impact with increased number of steps (small k, impact of the further nodes goes to 0)
- Theorem: Solution always exist and is unique

84

Simrank in matrix notation

Input:

transition matrix P,

scalar decay factor C,

scalar number of iterations k

Output: similarity matrix S

For i = 1:k, do

S = C PT S P

Set diagonal entries of S to 1

end

- Millions of queries and millions of ads: rely on already existing large scale matrix manipulation algorithms

Simrank

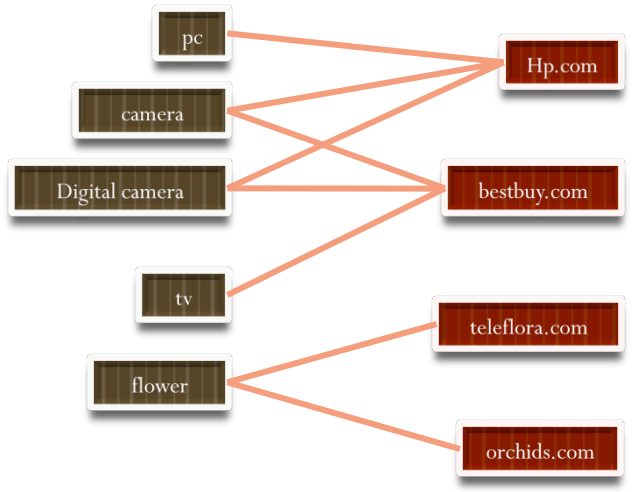
0th Iteration

	pc	camera	digital camera	tv	flower
pc	1				
camera	0	1			
digital camera	0	0	1		
tv	0	0	0	1	
flower	0	0	0	0	1

C = 0.8

$$s_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s_{k-1}(i,j)$$

$$s_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} s_{k-1}(i,j)$$



Simrank

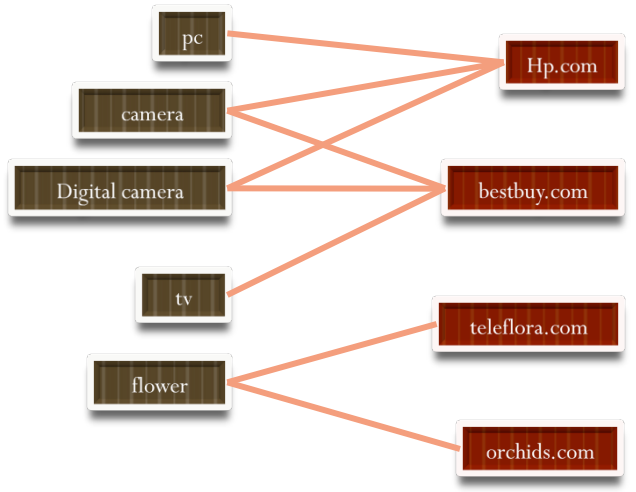
1st Iteration

	pc	camera	digital camera	tv	flower
pc	1				
camera	0.0889	1			
digital camera	0.0889	0.1778	1		
tv	0	0.0889	0.0889	1	
flower	0	0	0	0	1

C = 0.8

$$s_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s_{k-1}(i,j)$$

$$s_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} s_{k-1}(i,j)$$



Simrank

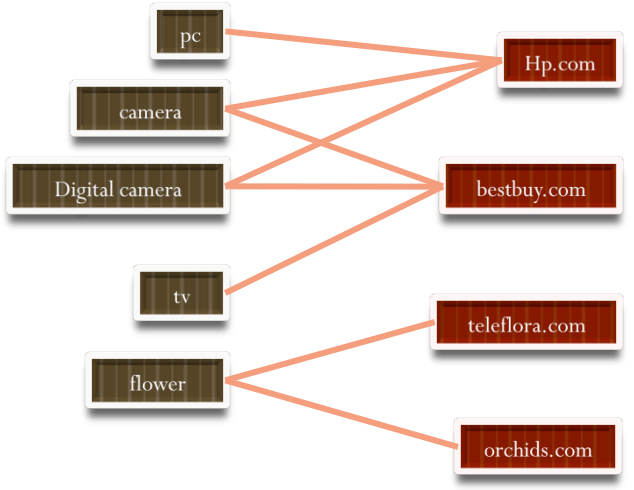
2nd Iteration

	pc	camera	digital camera	tv	flower
pc	1				
camera	0.1244	1			
digital camera	0.1244	0.2489	1		
tv	0.0356	0.1244	0.1244	1	
flower	0	0	0	0	1

C = 0.8

$$s_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s_{k-1}(i,j)$$

$$s_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} s_{k-1}(i,j)$$



Simrank

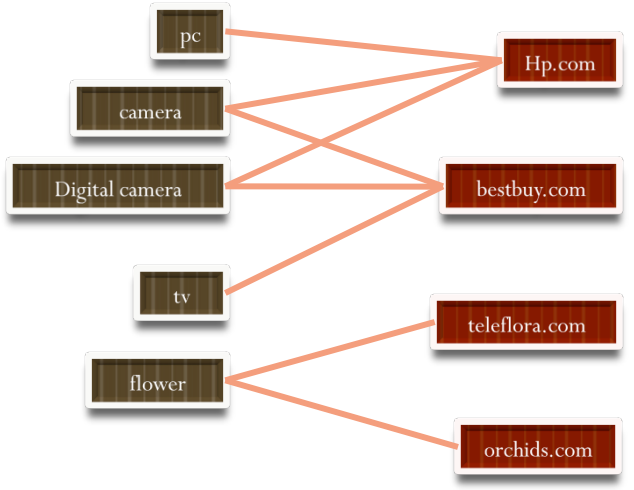
12th Iteration

	pc	camera	digital camera	tv	flower
pc	1				
camera	0.1650	1			
digital camera	0.1650	0.33	1		
tv	0.0761	0.1650	0.1650	1	
flower	0	0	0	0	1

C = 0.8

$$s_k(q,q') = \frac{C}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s_{k-1}(i,j)$$

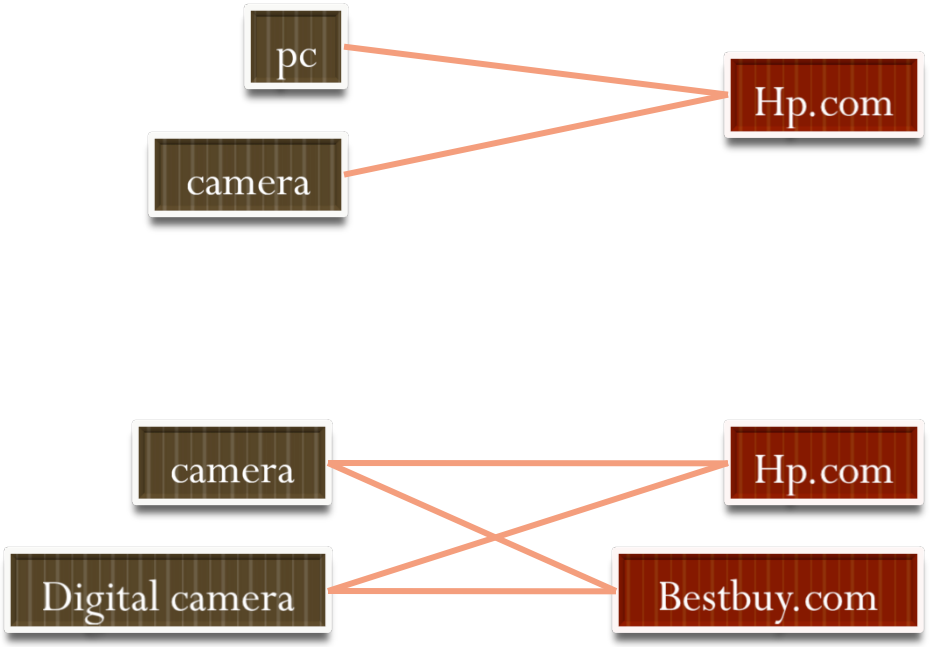
$$s_k(a,a') = \frac{C}{N(a)N(a')} \sum_{i \in E(a)} \sum_{j \in E(a')} s_{k-1}(i,j)$$



Some issues with Simrank

- Complete bipartite graphs: every node on the left links to every node on the right
- Why complete graphs?
- Allows for simplified setting to examine the algorithm performance
- Finding: Simrank scores in complete bipartite graphs are sometimes counter-intuitive

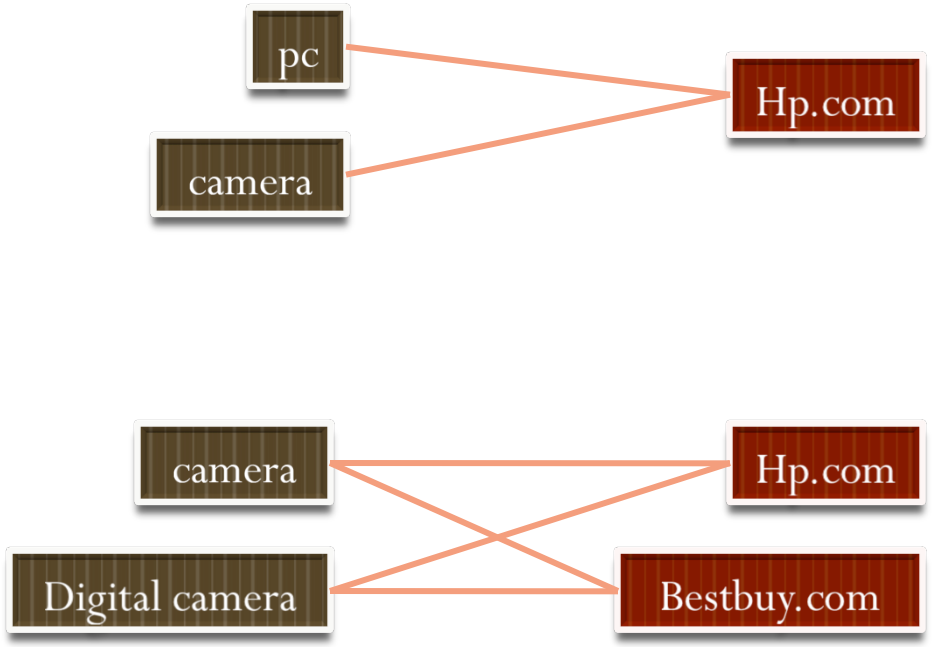
Example: Similarity depends on the size of one of the graph partition



iteration	Camera – digital camera	Pc - camera
1	0.4	0.8
2	0.62	0.8
3	0.65	0.8
4	0.66	0.8
5	0.664	0.8
6	0.665	0.8

C = 0.8

Solution: give the immediate neighborhood more weight



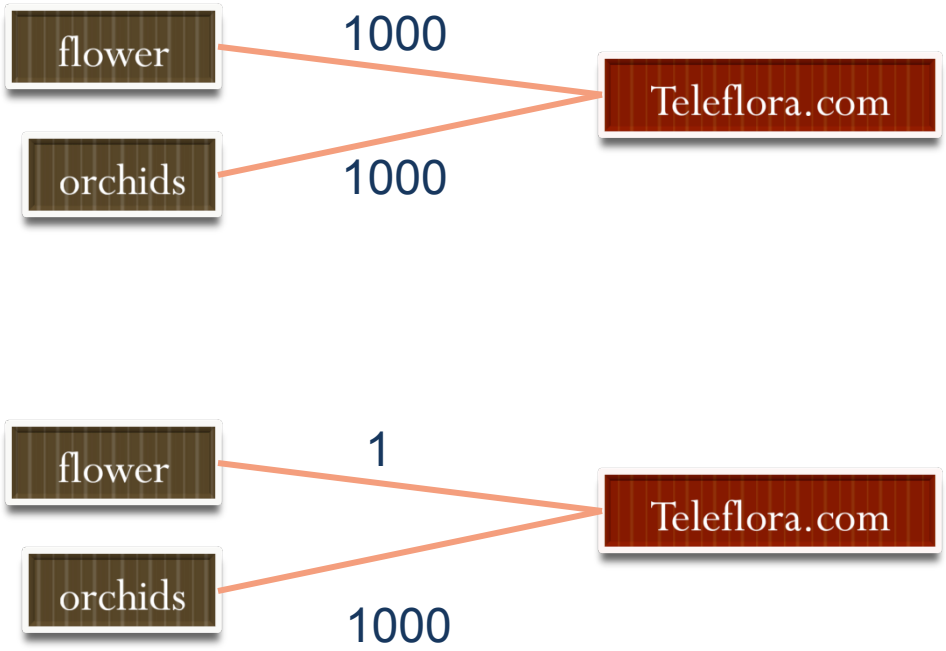
iteration	Camera – digital camera	Pc - camera
1	0.3	0.4
2	0.42	0.4
3	0.468	0.4
4	0.4872	0.4
5	0.49488	0.4
6	0.497952	0.4

$$evidence(q,q') = \sum_{i=1}^{E(q)\cap E(q')} \frac{1}{2^i}$$

C = 0.8

$$sim^{k}_{evidence}(q,q') = evidence(q,q') \cdot sim_{k}(q,q')$$

Weighted Simrank



- Absolute value and variance of weights matters
- The algorithm so far would output the same result in both cases

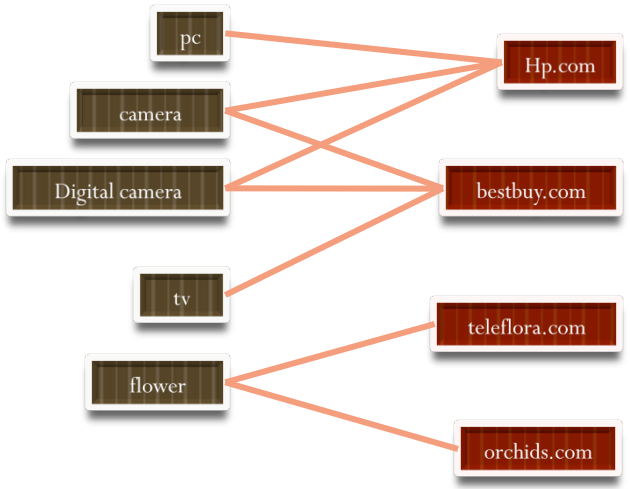
Weighted Simrank

$$p(a,i) = spread(i) \cdot normalized\_weight(a,i), \forall i \in E(a)$$

$$p(a,a) = 1 - \sum_{i \in E(a)} p(a,i)$$

$$spread(i) = \frac{1}{\text{var}iance(i)}$$

$$normalized\_weight(a,i) = \frac{w(a,i)}{\sum_{j \in E(a)} w(a,j)}$$



Another way to define the random walk

- Walk starting at a query node q
- Move by the edges of the graph
- Probability of move depending on the relative weight of the edge with regard to other edges originating at the same node
- A special "sink" node ω connected to all nodes in the graph
  - Move to ω with probability α
  - Walk ends in ω
- The stationary probability (probability after infinite number of steps) of the walk ending at node q' reflects the similarity between queries

The formalization

- Given a query rewrite target

$$P(q \sim q') = (1-\alpha) \sum_{a \in E(q)} p_{qa} P(a \sim q')$$

$$P(a \sim q') = (1-\alpha) \sum_{q \in E(a)} p_{aq} P(q \sim q')$$

$$p_{qa} = \frac{w_{qa}}{\sum_{b \in E(q)} w_{qb}}$$

Random walks - summary

- Both approaches have a static distribution
- Related to flow of electricity in a electrical network
  - voltage at q'
  - ground at the sink
  - probabilities are voltages on q
- Random walks use all the information in the graph
- To guarantee convergence and limit the processing somewhat simple use of the graph weights
- Is most of the information in the first hop?
- Avoid multiple iterations, but use more elaborate metrics in hop 1

Two alternatives for non-iterative similarity metrics

- Pearson correlation coefficient: linear relationship between the variables

$$sim(q,q') = \frac{\sum_a (p_{qa} - \overline{p_q})(p_{q'a} - \overline{p_{q'}})}{\sqrt{\sum_a (p_{qa} - \overline{p_q})^2 (p_{q'a} - \overline{p_{q'}})^2}}$$

- Potential issue: skewed by a single weight. Balance the weights and the number of features matched (circumstantial evidence):

$$Jacard(q,q') = \frac{|E(q) \cap E(q')|}{|E(q) \cup E(q')|}$$

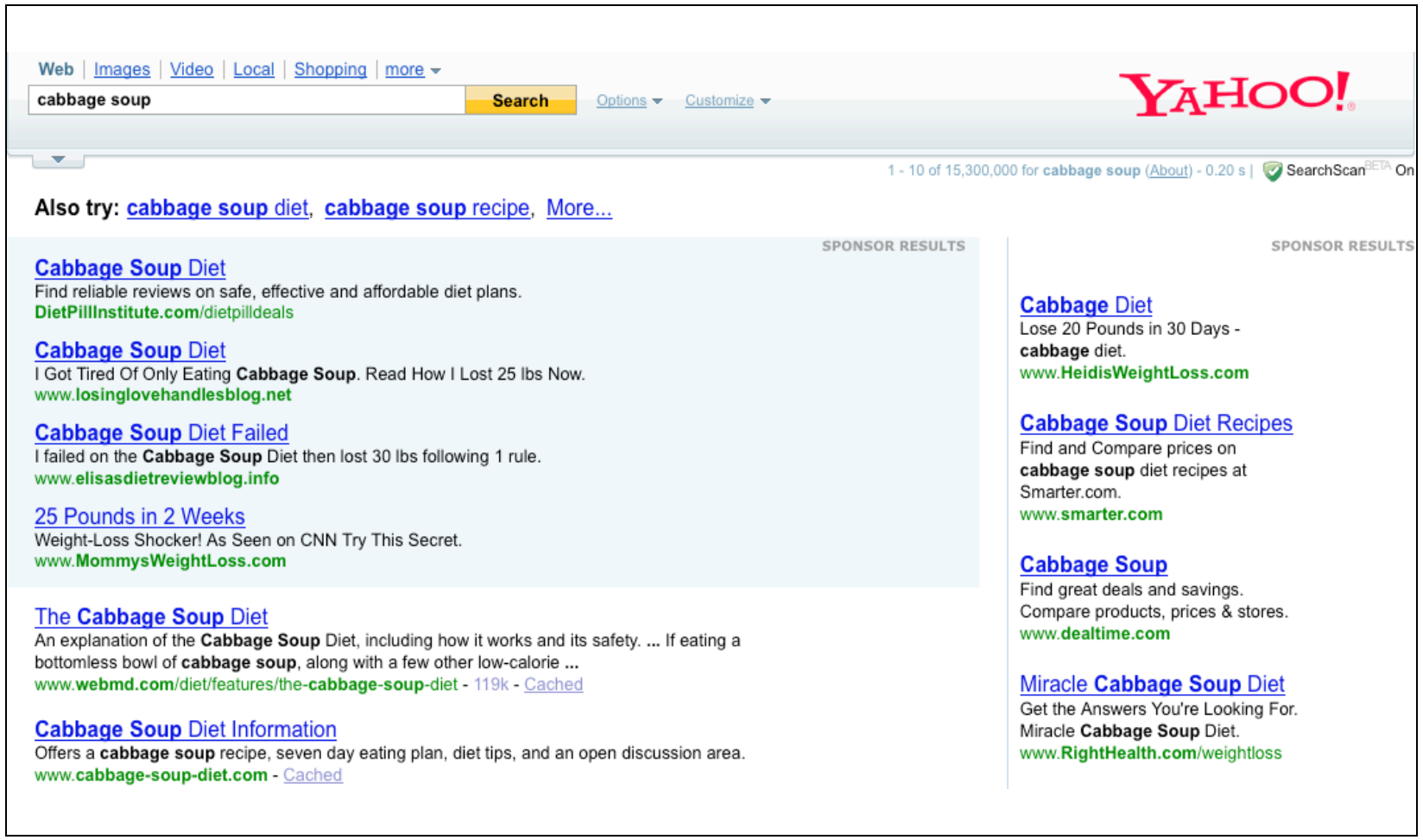
$$sim_j(q,q') = sim(q,q') + \gamma\, Jacard(E(q),E(q'))$$

Summary: Generating rewrites from click graphs

- Use the knowledge of the advertisers: co-bidding
- Validate by users: clicks
- Applicable to both query-ad and query-web search url graphs
- Very powerful approach
- How much information is in the immediate neighborhood vs. the whole graph is completely data dependent
  - Are there good rewrite pairs that are not likely to have co-bid from advertisers? (unlikely)
  - Not enough clicks in immediate links (more probable) – frequent queries function like hubs
- Start with a simple approach and see if we get enough recall

Online query rewriting

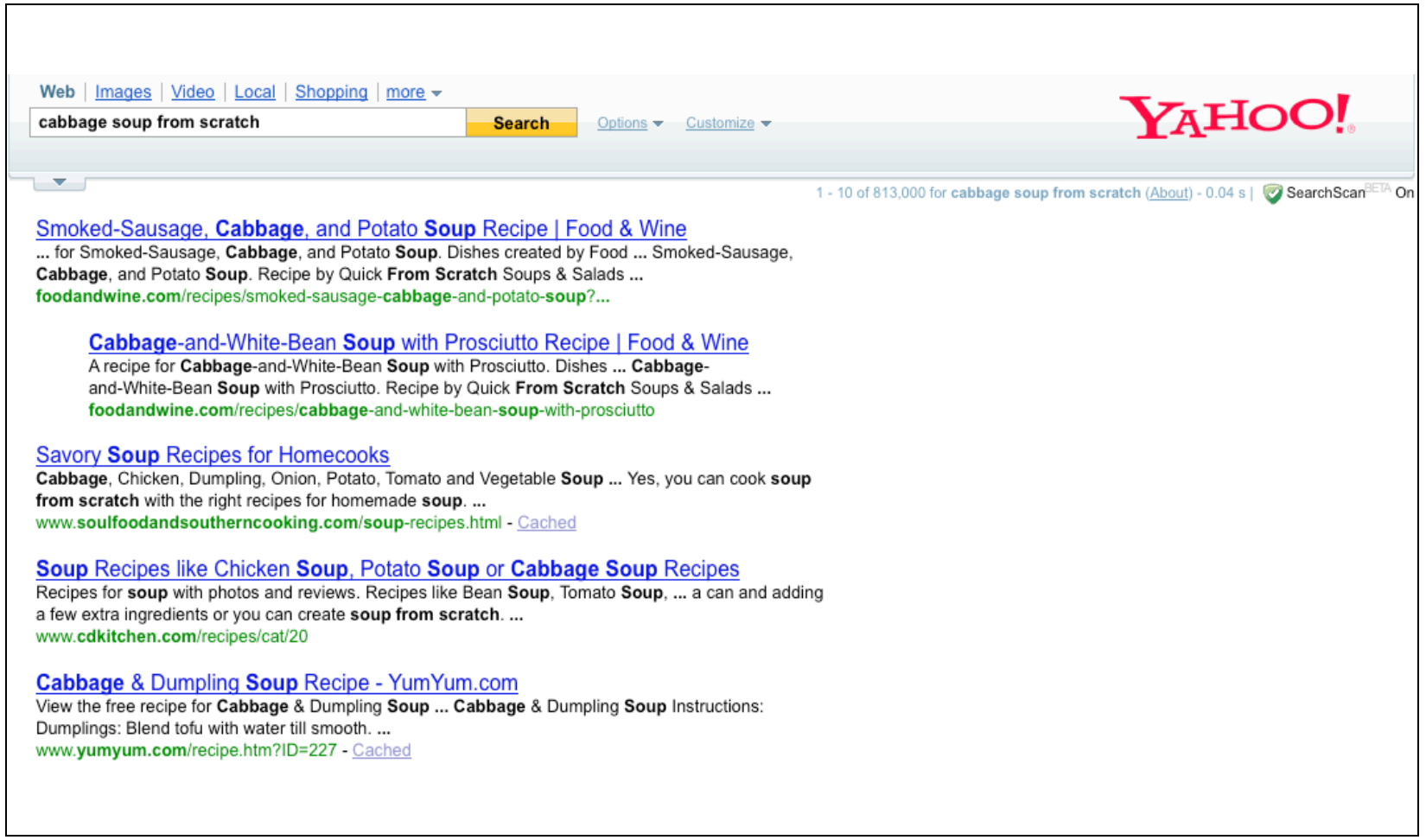
Sponsored Search: Decent Performance in the Head



Sponsored Search in the Tail

- Many tail queries display no sponsored search results
- Advertising in the tail is challenging
  - Longer / rare queries are more difficult to interpret
  - Exact match and phrase match are much less likely
  - Click-based relevance predictors are poor, due to data sparseness
- Considerable monetization potential in tail, if done correctly
- This us what advanced match is mostly about!

Sponsored Search in the Tail



Query fragment rewriting

Query fragment rewriting

- Particularly of interest for tail queries
  - too rare to process offline and store rewrites in a table
- Tail queries often contain repeating sub-phrases = *query segments*
  - "Cheap car insurance" vs. "Inexpensive car insurance"
  - "Cheap ski trips" vs. "Inexpensive ski trips"
- Individually, the queries do not repeat enough to learn a substitution
- However, the substitution

  `cheap → inexpensive`

  can be learned from many queries that contain this substitution

Using fragment rewriting online

A query comes into the system:

1. Fragment the query
2. Look up into the fragment rewrite table
3. Produce all possible variants

   k fragments, each having s substitutions →

   total number of rewrites is of order s^k
4. Reduce the space: consider only rewrites that result into an existing bid phrase

   How would you do this algorithmically?
5. Retrieve the ads

Example: “Fishing in Santa Cruz”

- Segment the query:
  - “Fishing| in | Santa Cruz”
- Substitution:
  1. Fishing → Surfcasting
  2. in → near
  3. Santa Cruz → Capitola
- Rewrites
  1. Fishing in Capitola
  2. Fishing near Santa Cruz
  3. Surfcasting in Santa Cruz
  4. Fishing near Capitola
  5. …

Off-line sub-query rewriting

- Some of the rewriting methods we presented suffer from *data sparsity*:
  - Query logs: we need significant occurrences of the rewrite-query pairs
  - Click graphs: we need many clicks/impressions
- **Solution**: Modify these methods to work with segments instead of queries
  - Query logs: count segment rewrites
  - Click graphs: nodes can be segments (a query induces multiple edges)

Problem: Correct query fragmentation

- Studied in NLP literature but still open research problem,
- How to determine that "`new york`" should be treated as a single entity as opposed to "`new countertop`"
- Simple strategy: pair-wise mutual information:

  $$p(a,b) > k * p(a)p(b)$$

  (the compound a.b is much more frequent than expected if a and b are independent)
  - maximum likelihood estimate from a corpus of queries
  - can use larger corpus: web pages
- More sophisticated approaches
  - Use dictionaries of entities: people names, places, companies
  - Conditional Random Fields or Markov models to capture the sequential structure of the query

Lecture Summary

Key messages

- Sponsored search is the main channel for textual advertising on the web
- Web queries are a (very) succinct representation of the user's intent
- Query volumes follow a power law with a long tail. There are billions of unique queries
- Ads are selected in sponsored search using an exact match to the bid phrase or advanced match to the whole ad
- Main ad selection approaches are the database approach (lookup for exact match) and the IR approach where we look up using multiple features
- Advanced match approaches use click and relevance judgments to learn how to match ads to queries
- Query rewrite is a common advanced match technique where the query is rewritten into another query that is used to retrieve the ads by exact match
- Users often rewrite queries in a single session – a good source for learning rewrites for advanced match
- Random walks in query-ad click graphs are another mechanism to establish query similarity

Questions?

We welcome suggestions about all aspects of the course: msande239-aut1112-staff

Thank you!

broder@yahoo-inc.com
vanjaj@yahoo-inc.com

http://research.yahoo.com