Introduction to Computational Advertising

MS&E 239
Stanford University
Autumn 2011
Instructors: Dr. Andrei Broder and Dr. Vanja Josifovski
Yahoo! Research
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 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 Market Share

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

<table>
<thead>
<tr>
<th>Format</th>
<th>Q2 2009</th>
<th>Q2 2010</th>
<th>Q2 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td>47%</td>
<td>47%</td>
<td>49%</td>
</tr>
<tr>
<td>Display/banner</td>
<td>22%</td>
<td>23%</td>
<td>23%</td>
</tr>
<tr>
<td>Classifieds</td>
<td>10%</td>
<td>10%</td>
<td>8%</td>
</tr>
<tr>
<td>Digital video</td>
<td>4%</td>
<td>5%</td>
<td>6%</td>
</tr>
<tr>
<td>Lead generation</td>
<td>7%</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Rich media</td>
<td>7%</td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>Sponsorship</td>
<td>2%</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>Email</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
</tbody>
</table>

% of total

*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*
### US Online Ad Spending, by Format, 2009-2013

**Billions**

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

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

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

<table>
<thead>
<tr>
<th>% of 10,000 keywords in each category and CPC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>% share</strong></td>
</tr>
<tr>
<td>1. Insurance</td>
</tr>
<tr>
<td>2. Loans</td>
</tr>
<tr>
<td>3. Mortgage</td>
</tr>
<tr>
<td>4. Attorney</td>
</tr>
<tr>
<td>5. Credit</td>
</tr>
<tr>
<td>6. Lawyer</td>
</tr>
<tr>
<td>7. Donate</td>
</tr>
<tr>
<td>8. Degree</td>
</tr>
<tr>
<td>9. Hosting</td>
</tr>
<tr>
<td>10. Claim</td>
</tr>
<tr>
<td>11. Conference call</td>
</tr>
<tr>
<td>12. Trading</td>
</tr>
<tr>
<td>13. Software</td>
</tr>
<tr>
<td>14. Recovery</td>
</tr>
<tr>
<td>15. Transfer</td>
</tr>
<tr>
<td>16. Gas/electricity</td>
</tr>
<tr>
<td>17. Classes</td>
</tr>
<tr>
<td>18. Rehab</td>
</tr>
<tr>
<td>19. Treatment</td>
</tr>
<tr>
<td>20. Cord blood</td>
</tr>
</tbody>
</table>

*Note: English-language only; *US keyword price estimates

CPC per search engine

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

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

Ad North:
- Canon Camera at Circuit City
- Canon Camera

Ad East:
- Camera Cases and Bags
  To know Bogen Imaging Inc, just take a look at the premium brands. www.bogenimaging.us
- Canon Camera Battery
  Accessory
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.
User: useful ads

Also try: miele da3190, da3190 hood, miele da3190 hood, More...

Miele Appliances On Sale
Free Shipping on Miele Appliances Fast, Reliable Nationwide Delivery.
www.us-appliance.com/miele

DA3190 On Sale
www.AJMadison.com/MieleDA3190

Miele Range Hood DA3190 - krilion.com
Find great products like the Miele Range Hood in-stock and on-sale at a store near you. Krillion provides relevant local search results.
www.krilion.com/nop-Miele-Range_Hoods-DA3190

Da3190
A Giant Selection of da3190. Shop Here Now and Save.
www.become.com

Da3190
Create A Cooking Paradise. Save On DA3190
Rangelhoods.Shopzilla.com

See your message here...

Miele DA3190 36 in. Wall Mounted Range Hood - Shopzilla.com
Optimization

- Total utility of a Sponsored Search system is a balance of the individual utilities:
  \[ \text{Utility} = f(\text{UtilityAdvertiser}, \text{UtilityUser, UtilitySE}) \]
- 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:
  \[ \text{Utility} = \alpha \ast \text{UtilityAdvertiser} + \beta \ast \text{UtilityUser} + \gamma \ast \text{UtilitySE} \]
  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 > \( \alpha \)
2. Advertiser ROI per search > \( \beta \)

- Practically:
  1. Find all ads that have user utility above \( \alpha \)
  2. Optimize which ads to show based on an auction mechanism as discussed in the previous lecture (captures the \( \beta \))
Why do it this way?

(As opposed to first find all ads with utility > $\beta$, 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
## Yahoo data set statistics

<table>
<thead>
<tr>
<th>Property</th>
<th>One week</th>
<th>Six months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Queries</td>
<td>Hundreds of Millions</td>
<td>Tens of Billions</td>
</tr>
<tr>
<td>Number of Users</td>
<td>Tens of Millions</td>
<td>Hundreds of Millions</td>
</tr>
<tr>
<td>Average Query Length</td>
<td>3.0 Terms</td>
<td>3.0 Terms</td>
</tr>
<tr>
<td>Average Popular Query Length</td>
<td>1.6 Terms</td>
<td>1.7 Terms</td>
</tr>
<tr>
<td>Portion of first results page views</td>
<td>86.6%</td>
<td>90.6%</td>
</tr>
<tr>
<td>Portion of second results page views</td>
<td>7.4%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Portion of three or more pages views</td>
<td>6.0%</td>
<td>4.9%</td>
</tr>
</tbody>
</table>
Query Volume per Hour of the Day

% of Daily Traffic

0 6 12 18

Hour of Day

Distinct Queries
Total Queries

1.5% 2.5% 3.5% 4.5% 5.5% 6.5%
Query Volume: Day of Week

% of Weekly Traffic

Day of Week

Monday Tuesday Wednesday Thursday Friday Saturday Sunday

Distinct Queries
Total Queries
Topical Distribution of Web Queries

Figure 2.9. Breakdown of Categorized Queries
Textual Ads
Anatomy of a Textual Ad: the Visible and Beyond

Title

Creative

Display URL

Bid phrase: computational advertising
Bid: $0.5

ACL-08: HLT Tutorial
Computational Advertising Tutorial
Columbus, OH - June 15, 2008
research.yahoo.com

Landing URL: http://research.yahoo.com/tutorials/acl08_compadv/
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

Account 1
Campaign 1
Ad group 1
Creative 2

Account 2
Campaign 2
Ad group 2
Bid phrases

Account 3
Campaign 3
Ad group 3

Advertiser

Brand name appliances
Compare prices and save money
www.appliances-r-us.com

Can be just a single bid phrase, or thousands of bid phrases (which are not necessarily topically coherent)

{ Miele, KitchenAid, Cuisinart, …}
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

<table>
<thead>
<tr>
<th>Size of Pay-per-Click Keyword Inventory According to US Online Retailers, March 2009 (% of respondents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 words or less</td>
</tr>
<tr>
<td>101-200 words</td>
</tr>
<tr>
<td>201-500 words</td>
</tr>
<tr>
<td>501-750 words</td>
</tr>
<tr>
<td>751-1,000 words</td>
</tr>
<tr>
<td>1,001-5,000 words</td>
</tr>
<tr>
<td>5,001-10,000 words</td>
</tr>
<tr>
<td>10,000+ words</td>
</tr>
</tbody>
</table>

Source: Internet Retailer, “Search Engine Marketing” conducted by Knowledge Marketing, April 2009
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 captures 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 database 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 on slide 17. Some times the second phase bundled with the reordering
Reactive vs. predictive reordering

Example: Horse races

- **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**)

<table>
<thead>
<tr>
<th>Name</th>
<th>Starts</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>Earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well Armed</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$3,649,000</td>
</tr>
<tr>
<td>Rachel Alexandra</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>$2,746,914</td>
</tr>
<tr>
<td>Summer Bird</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>$2,023,040</td>
</tr>
<tr>
<td>Mine That Bird</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>$1,892,200</td>
</tr>
<tr>
<td>Regal Ransom</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>$1,650,000</td>
</tr>
<tr>
<td>Gio Ponti</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>$1,433,000</td>
</tr>
<tr>
<td>Einstein</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>$1,269,304</td>
</tr>
<tr>
<td>Gloria de Campasso</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$1,210,000</td>
</tr>
<tr>
<td>Pioneron de Nile</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>$1,090,000</td>
</tr>
<tr>
<td>Swift Temper</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>$1,079,497</td>
</tr>
</tbody>
</table>

Updated through 10/5/2009
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

[Diagram showing relationships between Query Sessions, Users, Queries, Ads, and Web pages with various connections labeled: issued, contains, clicks, co-occurrence, similarity, bid phrases, search result]
Query Rewriting for Sponsored Search
Typical query rewriting flow

- Typical of the DB approach to AM
- Rewrite the user query $q$ into $Q’ = (q_1, q_2, \ldots)$
- 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 \( q \rightarrow q' \)
- 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)

<table>
<thead>
<tr>
<th>Query rewriting technique</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Generating Query Substitutions: Jones et al, in Proc of WWW 2006</td>
<td>query logs (query sessions)</td>
</tr>
<tr>
<td>Using the Wisdom of the Crowds for Keyword Generation: Fuxman et al., In proc of WWW 2004</td>
<td>co-clicks on web search results</td>
</tr>
<tr>
<td>2. Simrank++: Query Rewriting through Link Analysis of the Click Graph: Atoanellis et al., In proc of VLDB 2008</td>
<td>co-clicks on ads</td>
</tr>
<tr>
<td>3. Learning Query Substitutions for Online Advertising: Broder et al. in Proc of ACM SIGIR 2008</td>
<td>query-to-ad similarity</td>
</tr>
<tr>
<td>5. Query Word Deletion Prediction: Jones at al., in Proc of ACM SIGIR 2003</td>
<td>query logs</td>
</tr>
</tbody>
</table>
Query Rewriting using Web Search Logs
Data Source

- Query Sessions
  - Query
  - Sessions
  - issued
  - contains

- Users
  - clicks

- Queries
  - search result
  - co-occurrence
  - similarity
  - bid phrases
  - clicks

- Ads
  - Web pages
  - contains

- Data Source
  - 65
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

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

[Jones & Fain SIGIR 2003]
Many substitutions are repeated

- Some substitutions are incidental
- Others repeat often with different users in different days
  - car insurance $\rightarrow$ auto insurance
    - 5086 times in a sample
  - car insurance $\rightarrow$ car insurance quotes
    - 4826 times
  - car insurance $\rightarrow$ geico  [brand of car insurance]
    - 2613 times
  - car insurance $\rightarrow$ progressive auto insurance
    - 1677 times
  - car insurance $\rightarrow$ carinsurance
    - 428 times
A principled way to determine when are we sure in the rewrites

- Determine if
  \[ p(rw | q) \gg p(rw) \]
- Since \( p(rw | q) = \frac{p(rw,q)}{p(q)} \), 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: \( P(rw | q) = P(rw | \neg q) \)
  - H2: \( P(rw | q) \neq P(rw | \neg q) \)
- The log likelihood ratio \(-2\log(L(H1)/L(H2))\) is asymptotically \( \chi^2 \) distributed
- Other statistical tests can be used – pick you favorite