Introduction to Computational Advertising

MS&E 239
Stanford University
Autumn 2010
Instructors: Andrei Broder and Vanja Josifovski
Lecture 6: Using click data for ad selection
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.
Course Overview

1. 09/24 Introduction
2. 10/01 Textual Advertising and Sponsored Search part 1
3. 10/08 Marketplace
4. 10/15 Sponsored Search part 2 and IR intro
5. 10/22 Content Match and IR ad retrieval
6. 10/29 Using click data for textual ad selection
7. 11/05 Display advertising
8. 11/12 Targeting
9. 11/19 Mobile/Social/Rich formats
10. 12/03 Project presentations
Lecture overview: CTR estimation

- Checkpoint: relevant concepts in ad selection
- Bid phrase clustering
- Beyond the bid phrase
- Topical models
- CTR aggregation
- Explore-Exploit for CTR aggregation
- What happens after the click
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)

- **Today’s lecture applies to Ad Reordering mostly**
Reactive vs. predictive methods

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)**
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
  - Order of magnitude of judgments: 10Ks – 100Ks

- **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
  - Order of magnitude: millions to tens of billions
Why use the click data at all?

- If an ad has been shown enough times for a query we would know what the reaction of the users is
  - Can we use the click data from similar ads?
- **First goal**: Use click history to improve ad selection
- **Second goal**: Establish a probability of click $p(\text{click} | \text{ad query})$
  - Ad retrieval establishes a relative ordering of the ads: Ad at position 1 better than ad at position 2
  - How much better is ad1 from ad2?
    - Important in establishing revenue estimates: $R=p(\text{click}) \times \text{cost}$
    - The only way to get the $p(\text{click})$ is to use click history
Click data properties

- Volume curves: frequency in given period of time
- Power law: \( y = ax^k \)
- Linear function when we log both sides: \( \log(y) = k \log(x) + \log(a) \)
- Seen in many web scenarios: query volumes, impressions, clicks
- Content match example:
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 boots even sales of popular items - one stop shop. (How does this map to search engines?)
Interaction of two power law distributions

- Sponsored search: both ad impressions and query frequency power law
  - Correlated – frequent ads are intended for frequent queries
  - However, not all ads bidding on frequent queries have high number of impressions
- Less data: break query occurrences per given ad
- Enough data for a given query-ad pair only in the head of the curves
Short-term vs. longer term loop

- The very common ads and queries appear rapidly
- Small number of queries/ads have disproportionately high impact on the users
- Need quick reaction in case things go wrong
- Short term loop:
  - Monitors the impression/click data stream online
  - Adjusts for deviations of expected performance in the head of the volume curve
  - Can also be used for other purposes: personalization based on short term history
Clicks in the tail

- In total, a large fraction of the impressions and clicks are in the tail
  - Area under the tail part of curve significant (e.g. 40%)
- We cannot reason for individual events:
  - Too many unique events
  - Cost per invested time: cost/number of uses too high
  - Too little data to achieve statistical significance
- Aggregat queries, pages, ads into larger groups with enough events
Case studies
Clustering of Bid Phrases for CTR Prediction in Sponsored Search

Predicting CTR Using Keyword Clusters: M. Regelson, D. Fain, EC 2006
Scope and motivation

• Assumption 1: CTR depend only on the ad bid phrase
  • Exact Match

• Assumption 2: future can be predicted from the past:
  • 90% of queries show no periodic changes in search volume at a level of one week

Maximum likelihood estimate: CTR = clicks/impressions

• Variance depends on the number of impressions:
  • 1 click of two impressions: CTR=0.5; std error = 0.35

• Reliable CTR estimates for
  • Low impression terms
  • New terms
First cut: CTR estimates from volume and position (rank)

- Volume and rank correlated with the CTR
  - Rank cannot be used to select ads
  - Trust Bias: swap results 1 and 2, 2 gets consistently more clicks (Joachims et al. SIGIR 2008)

- Per decile break up of CTR:
  - Sort queries by volume
  - decile 1- queries up to 10% of the total volume

- Why do we see a bump in decile 2?

Figure 2: Relative CTR by search volume decile and rank for the first week of December 2005.

Regelson, Fain EC2006
Predicting CTR based on keyword clustering

- Related keywords exhibit similar CTR
  - Relatedness is a measure of the semantic similarity
  - CTR related to the query semantics
- Hierarchical clustering
  - Allows multiple levels of resolution – back-off up the hierarchy when no data at certain level
Keyword clustering: top-down

- Bipartite graph of advertisers and bid phrases
- Top-down method: a series of recursive graph partitioning
- Partition into two subgraphs A and B to minimize conductance:
  - $\phi = c / \min(D_A, D_B)$
  - $D_A$ is the number of edges in partition A
  - NP-hard, some efficient approximation algorithms
Bottom-up clustering algorithm

- Build a keyword only graph
  - Single step or iterative techniques that we discussed in Lect. 4
  - E.g. Pearson correlation coefficient
- Single-link clustering method:
  - Sort links by weight
  - Examine each edge from a singleton cluster
  - Merge singleton the node into the other cluster node
Top down or bottom up clustering?

- Top-down better when small number of clusters needed
- Bottom up better when larger number of clusters needed
- Error accumulates
- Minimize the number of steps needed to come to the right number of clusters
- Precision: clusters cannot be too big
- Bottom-up likely to perform better
Use the clusters to predict CTR

- Combine estimates at different level of hierarchy:
  - Use the parent data when there is no data at the lower level
- How far up the tree to go?
  - Top of the tree – overall average CTR
- Where do we start?
  - Too much variance at the leafs
- Given CTR estimate for each term and its error, how to measure the performance of the prediction?
- Do we need different solutions for different frequency ranges?
  - B1 (1-10), B2 (10-50), B3 (50, ...)
Results

Table 3: Mean error across terms and weeks on entire 100K sample, on set B1 (0-10 impressions), on set B2 (11-50 impressions), and on set B3 (more than 50 impressions) for each prediction method. Predictions for each week were computed using historical data from the previous week.

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Mean Error</th>
<th>Mean Error (B1)</th>
<th>Mean Error (B2)</th>
<th>Mean Error (B3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Rank One Rate</td>
<td>0.775</td>
<td>0.304</td>
<td>0.735</td>
<td>2.423</td>
</tr>
<tr>
<td>2 Volume Decile Rate</td>
<td>0.824</td>
<td>0.257</td>
<td>0.797</td>
<td>2.442</td>
</tr>
<tr>
<td>3 Cluster and Parent Cluster Rate (omitting term)</td>
<td>0.468</td>
<td>0.171</td>
<td>0.525</td>
<td>1.177</td>
</tr>
<tr>
<td>4 Cluster, Parent and Grandparent Cluster Rate (omitting term)</td>
<td>0.485</td>
<td>0.169</td>
<td>0.539</td>
<td>1.253</td>
</tr>
<tr>
<td>5 Cluster through Great-Grandparent Cluster Rate (omitting term)</td>
<td>0.499</td>
<td>0.169</td>
<td>0.550</td>
<td>1.311</td>
</tr>
<tr>
<td>6 Term Rate</td>
<td>0.398</td>
<td>0.268</td>
<td>0.539</td>
<td>0.493</td>
</tr>
<tr>
<td>7 Term and Cluster Rate</td>
<td>0.409</td>
<td>0.249</td>
<td>0.537</td>
<td>0.610</td>
</tr>
<tr>
<td>8 Term, Cluster, and Parent Cluster Rate</td>
<td>0.418</td>
<td>0.230</td>
<td>0.528</td>
<td>0.728</td>
</tr>
<tr>
<td>9 Term, Cluster, Parent and Grandparent Cluster Rate</td>
<td>0.430</td>
<td>0.221</td>
<td>0.532</td>
<td>0.816</td>
</tr>
<tr>
<td>10 Term and Cluster through Great-Grandparent Cluster Rate</td>
<td>0.440</td>
<td>0.214</td>
<td>0.536</td>
<td>0.886</td>
</tr>
</tbody>
</table>
Conclusion: Clustering of keywords

- Clustering helps for cases when there is no (reliable) historical information.
- Many ways to cluster
- Related to query rewriting
  - Different objective might require different methods
- Cluster size should be small: 10-20 keywords
Predicting Clicks: beyond the bid phrase

Predicting Clicks: Estimating the Clickthrough Rate for New Ads:
M. Richardson, E. Dominowska, R. Ragno, WWW2008
Using Ad Features to Predict CTR

- Still assuming **exact match**: the CTR depends solely on the information in the ad
  - the query is equal to the ad bid phrase
- Predict CTR by extracting features from the ad and building a model based on a training set
  \[ z = \sum_i w_i \cdot f_i(ad) \] \[ \text{ctr} = \frac{1}{1 + e^{-z}} \]
- Logistic regression as a model:
- Cross entropy loss function \( H(P,Q) = \sum_i p(i) \log(q(i)) \)
- For each feature
  - Normalize to zero mean, unit standard dev, crop outliers to 5S
  - Add derived features \( \log(f+1) \) and \( f^2 \)
Features: bid phrase based

- Ads with the same bid phrase
- Ads with similar bid phrases – logit of CTR and counts:

\[
f_0(\text{ad}) = \frac{\alpha \overline{\text{CTR}} + N(\text{ad}_\text{term}) \text{CTR}(\text{ad}_\text{term})}{\alpha + N(\text{ad}_\text{term})}
\]

\[
R_{mn}(t) = \left\{ \begin{array}{l}
\{ \text{ad} : |t - \text{ad}_\text{term}| = m \text{ and } |	ext{ad}_\text{term} - t| = n \\
\{ |\text{ad}_\text{term} \cap t| > 0 \text{ and } 
\end{array} \right\}
\]

\[
\text{CTR}_{mn}(\text{term}) = \frac{1}{\sum_{R_{mn}(\text{term})} \text{CTR}_x}
\]

Table 1: Term and Related Term Results

<table>
<thead>
<tr>
<th>Features</th>
<th>MSE (x 1e-3)</th>
<th>KL Divrg. (x 1e-2)</th>
<th>% Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (CTR)</td>
<td>4.79</td>
<td>4.03</td>
<td>-</td>
</tr>
<tr>
<td>Term CTR</td>
<td>4.37</td>
<td>3.50</td>
<td>13.28%</td>
</tr>
<tr>
<td>Related term CTRs</td>
<td>4.12</td>
<td>3.24</td>
<td>19.67%</td>
</tr>
</tbody>
</table>

Richardson et al. WWW2008
Beyond bid phrases: ad quality

- CTR varies considerably for ads with the same bid phrase
  - Digital camera – 3x; surgery 5x
  - This is the lower bound on the error even if we have perfect bid phrase clustering!
- Does the CTR depend on the ad quality?
  - CTR of organic search depends on snippet
  - What ad features to use to predict the click response of the users?
Ad features

- Five categories considered (~80 features in total):
  - **Appearance**: number of words in each part; word length; capitalization; punctuation (!#$****)
  - **Attention capture**: action words (“buy”, “join”,…), numbers (prices, discounts)
  - **Landing page**: complexity of the HTML, etc.
  - **Relevance**: bid term in the title, body; subset of the term,…
  - **Reputation**: short clean urls are expensive – more reputable domain

- One feature for the 10K most common words in title/body
More Features

• Ad group specificity:
  • Entropy of the results of the bid phrase classification
  • Number of bid phrases in the ad group

• Web search features:
  • Query frequency
  • Web page frequency
Results for ad quality, ad group and search data features

Table 2: Ad Quality Results

<table>
<thead>
<tr>
<th>Features</th>
<th>MSE (x 1e-3)</th>
<th>KL Diverg. (x 1e-2)</th>
<th>% Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (CTR)</td>
<td>4.79</td>
<td>4.03</td>
<td>-</td>
</tr>
<tr>
<td>Related term CTRs</td>
<td>4.12</td>
<td>3.24</td>
<td>19.67%</td>
</tr>
<tr>
<td>+ Ad Quality</td>
<td>4.00</td>
<td>3.09</td>
<td>23.45%</td>
</tr>
<tr>
<td>+ Ad Quality without uni</td>
<td>4.10</td>
<td>3.20</td>
<td>20.72%</td>
</tr>
</tbody>
</table>

Table 3: Order Specificity results

<table>
<thead>
<tr>
<th>Features</th>
<th>MSE (x 1e-3)</th>
<th>KL Diverg. (x 1e-2)</th>
<th>% Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (CTR)</td>
<td>4.79</td>
<td>4.03</td>
<td>-</td>
</tr>
<tr>
<td>CTRs &amp; Ad Quality</td>
<td>4.00</td>
<td>3.09</td>
<td>23.45%</td>
</tr>
<tr>
<td>+ Order Specificity</td>
<td>3.75</td>
<td>2.86</td>
<td>28.97%</td>
</tr>
</tbody>
</table>

Table 4: Search Engine Data results. AQ means the Ad Quality feature set, and OB means the Order Specificity.

<table>
<thead>
<tr>
<th>Features</th>
<th>MSE (x 1e-3)</th>
<th>KL Diverg. (x 1e-2)</th>
<th>% Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (CTR)</td>
<td>4.79</td>
<td>4.03</td>
<td>-</td>
</tr>
<tr>
<td>+ Search Data</td>
<td>4.68</td>
<td>3.91</td>
<td>3.11%</td>
</tr>
<tr>
<td>CTRs &amp; AQ &amp; OS</td>
<td>3.75</td>
<td>2.86</td>
<td>28.97%</td>
</tr>
<tr>
<td>+ Search Data</td>
<td>3.73</td>
<td>2.84</td>
<td>29.47%</td>
</tr>
</tbody>
</table>

Figure 4. Frequency of advertisement word unigrams, sorted by overall frequency. The light and dark gray lines give the relative frequency of unigrams in low and high CTR ads.

Richardson et al. WWW2008
Empirical vs. estimated CTR with regard to ad views

Figure 6: Expected mean absolute error in CTR as a function of the number of times an ad is viewed.
Conclusion

- Evidence that the ad contains more information than the bid phrase that can be used in ad selection
  - Can be used to improve both for EM and AM
- Strong signal from ad and ad group features
- Consistent with the search based approach
  - Can be expanded to include features based on the query and the matched ad
  - Use click data instead of relevance judgments
- Estimate CTR for new ads
Using topics (clusters) for ad selection

A hidden class page-ad probability model for Contextual Advertising. A. Ratnaparkhi In Proc. of TROA 2008
Bridging the gap between the ads and the pages

- Generative model:
  - There is a set of predefined topics
  - Given a topic, we can generate an ad
  - Ads generated by generating a length $l$, and then drawing from a language model distribution of words: $ad = \{a_1, \ldots a_l\}$
  - Same for pages $page = \{b_1, \ldots b_m\}$
  - View the clicks as a result of such generative process
  - All clicks explained by the topics
“Generate” the click using a hidden topic
Inducing topics from click data

- Collect clicks of (page, ad) pairs
- Decide on the number of topics
- Estimate topic-based probability model on this data
- Premise: (page, ad) click share a single topic
- Optimization objective: maximize the likelihood of the given training set

\[ P(T) = \sum_{(page, ad) \in T} p(page, ad) \]
Probability model

\[ p(c, \text{ad}, \text{page}) = p(c)p(\text{ad}|c)p(\text{page}|c) = p(c) \cdot \prod_{i=1}^{\text{size(ad)}} q_{\text{ad}}(a_i|c) \cdot \prod_{i=1}^{\text{size(page)}} q_{\text{page}}(b_i|c) \]

\[ p(\text{ad}, \text{page}) = \sum_c p(c, \text{ad}, \text{page}) \]

- Term and length probabilities are dependent on class
- Ad and page have separate vocabularies
- Topics are not observed – hidden variables
  - Parameter estimation
- Model \( \theta \) - the red boxes – use Expectation Maximization algorithm
EM algorithm for parameter updates

- Each iteration has E-step and M-step
- E-step
  - Imagine that (page, ad) training instance occurs with topic $c$
  - Weight instance by $p(c | ad, page)$ using current parameter estimates: $\theta_n$

$$p(c|ad, page, \theta_n) = \frac{p(c, ad, page | \theta_n)}{\sum_{c'} p(c', ad, page | \theta_n)}$$

- Collect counts
M-step

- Use counts collected in E-step for new estimates
- Maximize the conditional expectation of the complete data log-likelihood:

\[
Q(\theta|\theta_n) = E_{C|T,\theta_n} \left[ \log \prod_{ad, page \in T} p(c, ad, page|\theta) \right] \\
= E_{C|T,\theta_n} \left[ \sum_{ad, page \in T} \log p(c, ad, page|\theta) \right] \\
= \sum_{ad, page \in T} E_{c|ad, page, \theta_n} \left[ \log p(c, ad, page|\theta) \right] \\
= \sum_{ad, page \in T} \sum_{ad, page \in T} p(c|ad, page, \theta_n) \log p(c, ad, page|\theta)
\]

- Example solution:

\[
q_{ad}(a|c) = \frac{\sum_{ad, page \in T} p(c|ad, page, \theta_n) \sum_{i=1}^{size(ad)} \delta(a, a_i)}{\sum_{ad, page \in T} p(c|ad, page, \theta_n) size(ad)}
\]
The mechanics of EM in this application

- Two types of parameters:
  a) Per training instance:
  b) Cut across different instances: feature-class

- Parameters depend on each other according to the optimization objective
  - maximize log likelihood of the parameters

- Repeat until conversion:
  1. estimate a) based on b)
  2. estimate b) based on a)

- One outcome: “Unifying effect”
  - Instances with similar composition start getting similar probabilities for the same classes
  - Features that appear in instances with similar probability will get similar class probabilities
Synthetic example

- 2 underlying classes: *flavors* and *sports*
- Each line represents a click:

<table>
<thead>
<tr>
<th>Ad text</th>
<th>Page text</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanilla</td>
<td>chocolate mint</td>
</tr>
<tr>
<td>vanilla</td>
<td>strawberry banana</td>
</tr>
<tr>
<td>football</td>
<td>golf shoes</td>
</tr>
<tr>
<td>soccer</td>
<td>tennis shoes</td>
</tr>
</tbody>
</table>
## Synthetic example (cont’d)

| Ad words  | $q_{\text{ad}}(w|c1)$ | $q_{\text{ad}}(w|c2)$ |
|-----------|------------------------|------------------------|
| vanilla   | 0                      | 1                      |
| soccer    | 0.5                    | 0                      |
| football  | 0.5                    | 0                      |

| Page words | $q_{\text{page}}(w|c1)$ | $q_{\text{page}}(w|c2)$ |
|------------|--------------------------|--------------------------|
| golf       | 0.25                     | 0                        |
| strawberry | 0                        | 0.25                     |
| banana     | 0                        | 0.25                     |
| tennis     | 0.25                     | 0                        |
| shoes      | 0.5                      | 0                        |
| mint       | 0                        | 0.25                     |
| chocolate  | 0                        | 0.25                     |

<table>
<thead>
<tr>
<th>Ad text</th>
<th>Page text</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanilla</td>
<td>chocolate mint</td>
</tr>
<tr>
<td>vanilla</td>
<td>strawberry banana</td>
</tr>
<tr>
<td>football</td>
<td>golf shoes</td>
</tr>
<tr>
<td>soccer</td>
<td>tennis shoes</td>
</tr>
</tbody>
</table>
Examples from real click data

| Ad words                    | $q_{ad}(w|c1)$ | $q_{ad}(w|c2)$ |
|----------------------------|----------------|----------------|
| currency trading           | 0.0820891          |                |
| home loans                 | 0.067927            |                |
| countrywide home loans     | 0.067313            |                |
| free demo                  | 0.0417377           |                |
| forex currency             | 0.0365207           |                |
| forex currency trading     | 0.0364378           |                |
| invoice price              | 0.00946243          |                |
| used autos                 | 0.00848382          |                |
| for sale                   | 0.00793812          |                |
| real estate                | 0.00680954          |                |
| birthday party             | 0.0501497           |                |
| birthday invitations       | 0.0338498           |                |
| free birthday              | 0.0336109           |                |
| free birthday invitations  | 0.033425            |                |
| party idea                 | 0.024849            |                |
| party invitation           | 0.0225438           |                |
| birthday party invitation  | 0.0203071           |                |
| party ideas                | 0.0202639           |                |
| ice cream                  | 0.012521            |                |
| alabama job                | 0.0121976           |                |

| Page words                 | $q_{page}(w|c1)$  | $q_{page}(w|c2)$ |
|----------------------------|-----------------|-----------------|
| yahoo finance              | 0.0599609        |                |
| exchange rates             | 0.0369237        |                |
| view exchange              | 0.0366989        |                |
| exchange rate              | 0.0339454        |                |
| financial news             | 0.0209973        |                |
| news yahoo                 | 0.0209596        |                |
| dollar exchange            | 0.0162197        |                |
| dollar exchange rate       | 0.0162182        |                |
| canadian dollar            | 0.0115819        |                |
| new york                   | 0.00555826       |                |
| birthday party             | 0.0751149        |                |
| party ideas                | 0.072431         |                |
| birthday party ideas       | 0.0662678        |                |
| party planning             | 0.0360878        |                |
| party planning ideas       | 0.0325496        |                |
| party party                | 0.020292         |                |
| for sale                   | 0.0113448        |                |
| nick jr                    | 0.0110672        |                |
| classifieds ads            | 0.00916032       |                |
| classified ads             | 0.00912854       |                |

Ratnaparkhi TROA2008
Conclusion

• The vocabulary gap between the ads and the pages can be bridged by using topics
• Topics as hidden variables
• Topics can be discovered using click data
• Might be a better fit for the data
• Easier to bootstrap if you already have a running system
• Sometimes we can guess ‘labels’ for the topics based on the feature weights
CTR aggregation

Using click data for CTR estimation

• Can we estimate the CTR from the click data that we already have?

• If an ad-page pair has been seen significant (thousands) significant amount of times, we could estimate precisely

• Issues:
  • Infrequent pairs
  • New pages/ads

Figure 1: Distribution of (page, ad) impressions: Plots are on log-log scale but ticks are on the original scale.
Estimation in the “tail”

- Use an existing, well-understood hierarchy
  - Categorize ads and webpages to leaves of the hierarchy
  - CTR estimates of siblings are correlated

  ➔ The hierarchy allows us to aggregate data

- Coarser resolutions
  - provide *reliable* estimates for rare events
  - which then influences estimation at finer resolutions
Units of estimation: region defined by the class hierarchy
System overview

Retrospective data
[URL, ad, isClicked] → Crawl
a sample of URLs → Classify pages and ads

Rare event estimation using hierarchy

Impute impressions, fix sampling bias
Sampling of webpages

- Naïve strategy: sample at random from the set of URLs
  - Sampling errors in impression volume AND click volume
- Instead, we propose:
  - Crawling all URLs with at least one click, and
  - a sample of the remaining URLs
  - Variability is only in impression volume
Imputation of impression volume

#impressions =

\[ n_{ij} + m_{ij} + x_{ij} \]

- Clicked pool
- Sampled Non-clicked pool
- Excess impressions (to be imputed)

- Column constraint:
  \[ \sum n_{ij} + K \cdot \sum m_{ij} \]

- Row constraint:
  \[ \sum \text{impressions on ads of this ad class} \]

- Total impressions (known)
Imputation of impression volume

- **Region**
  \[ = \text{(page node, ad node)} \]

- **Region Hierarchy**
  \[ \rightarrow \text{A cross-product of the page hierarchy and the ad hierarchy} \]
Imputation of impression volume

[level constraint]

sумs to

[block constraint]

Level i

Level i +1
Imputing $x_{ij}$

Iterative Proportional Fitting [Darroch +/1972]

Initialize $x_{ij} = n_{ij} + m_{ij}$

**Top-down:**

- Scale all $x_{ij}$ in every block in $Z^{(i+1)}$ to sum to its parent in $Z^{(i)}$
- Scale all $x_{ij}$ in $Z^{(i+1)}$ to sum to the row totals
- Scale all $x_{ij}$ in $Z^{(i+1)}$ to sum to the column totals

*Repeat for every level $Z(i)$*

**Bottom-up:** Similar
Imputation: Summary

• Given
  • $n_{ij}$ (impressions in clicked pool)
  • $m_{ij}$ (impressions in sampled non-clicked pool)
  • # impressions on ads of each ad class in the ad hierarchy

• We get
  • Estimated impression volume
    $$\tilde{N}_{ij} = n_{ij} + m_{ij} + x_{ij}$$
    in each region $ij$ of every level
System overview

Retrospective data
[page, ad, isclicked]

Crawl
a sample of pages

Classify pages and ads

Rare event estimation using hierarchy

Impute impressions, fix sampling bias
Rare rate modeling

1. Freeman-Tukey transform

\[
y_r = \frac{1}{2} \left( \sqrt{\frac{c_r}{\tilde{N}_r}} + \sqrt{\frac{c_r + 1}{\tilde{N}_r}} \right)
\]

- Distinguishes between regions with zero clicks based on the number of impressions
- **Variance stabilizing transformation:** \( \text{Var}(y) \) is independent of \( \text{E}[y] \) \( \Rightarrow \) needed in further modeling
- \( y \mid S_r, \beta^{(d(r))} \approx N(\beta^{(d(r))} + S_r, V_r) \)
- Simplify: assume that \( V_r = V/N_r \)
- Main part of the model: \( S_r = \text{parent}(S_r) + w_r, \quad w_r \approx N(0, W_d^{(r)}) \)

\# clicks in region r
\# impressions in region r
Rare rate modeling

Generative Model (Tree-structured Markov Model)

Unobserved “state”
Rare rate modeling

- Fitting using a Kalman filtering algorithm
  - **Filtering**: Recursively aggregate data from leaves to root
  - **Smoothing**: Propagates information from root to leaves
Experimental evaluation

- Compare 4 models
- **TS:** tree-structured model
- **LM (level-mean):** each level smoothed independently
- **NS (no smoothing):** CTR proportional to $1/\bar{N}$
- Random
Experiments
Experiments

- Few impressions → Estimates depend more on siblings
- Enough impressions → little “borrowing” from siblings
Conclusions

• A method to estimate
  • rates of extremely rare events
  • at multiple resolutions
  • under severe sparsity constraints

• Two parts
  • Imputation of impressions ➔ incorporates hierarchy, fixes sampling bias
  • Tree-structured generative model ➔ extremely fast parameter fitting
Issue with click aggregation

- How to bootstrap the process?
- Learn only based on what the system show
- Are there any other ads that could yield good performance?
- Try all possible page-ad pairs enough times to decide
  - It has been done for display advertising!
  - Trying out costs money!
- Not feasible for even modest ad/page counts
- How to explore the space of ad placements in efficient way?
Explore-Exploit for CTR aggregation

Bandits for Taxonomies: A Model-based Approach S. Panday et al. SDM 2007
Exploration

- Click based systems learn from the users’ responses
- Users respond to what the system shows
- Self-feeding loop
- How to explore other options, not currently used by the system
  - Limited amount of options – explore every possibility
  - Not feasible in textual advertising
- Need a mechanism to guide the exploration
- Option: matching dimensions (e.g. ad and query words)
  - Too many combinations (apple-ipod)
Taxonomies for dimensionality reduction

- Used for CTR aggregation
- Dimensionality reduction
- Can we use the taxonomies to guide the exploration?
Bandits

Pull arms sequentially so as to maximize the total expected revenue
• Estimate payoff probabilities $p_i$
• Bias the estimation process towards better arms
• Learn and preserve revenue at the same time
Bandits for content match

- Content match: query known (row chosen)
- We need to chose among the ads
- Every cell has its own CTR
- Too many cells to exhaustively search the space
Explore-Exploit

- Balance between the exploitation and exploration
- Exploitation: pull the arm that we know will return most
- Exploration: pull the arm that could get lead us to a better return in the future
- Exploration costs: each impression that we explore we are likely to lose revenue
- Design a policy that will optimize the long term revenue
- Bandit flow:
  - Assign priority. Based on the currently available information
  - Pull the arm with the highest priority
  - Update the priorities
Ad/page taxonomy: multi-level policy

Assumption: CTRs in a block are homogeneous
Multi-level Policy (Allocation)

- Classify webpage ➞ page class, parent page class
- Run bandit on ad parent classes ➞ pick one ad parent class
- Run bandit among cells ➞ pick one ad class
- In general, continue from root to leaf ➞ final ad
Why allocation in multiple levels

- Efficiency: lower level has many cells (6k x 6k) in this study
- Improved revenue: find the relevant areas quicker
  - Higher levels guide the exploration (homogeneity assumption)
  - Less to explore – smaller budget for exploration
- Both levels use the same criteria for selection:
  \[
  \arg \max_a (\overline{CTR}_{qa} + \sqrt{\frac{2\log N_q}{N_{aq}}})
  \]
- \(N_q\) – number of times the query has been seen
  - Higher \(N_q\): more important is the query
- \(Naq\) – number of times the query-ad has been seen
  - Higher \(Naq\): more certainty in \(\overline{CTR}_{qa}\)
Conclusion

• When having a CTR guided system, exploration is a key component
• Short term penalty for the exploration needs to be limited (exploration budget)
• Most exploration mechanisms use a weighted combination of the predicted CTR rate (average) and the CTR uncertainty (variance)
• Exploration in a reduced dimensional space: class hierarchy
• Top down traversal of the hierarchy to determine the class of the ad to show
Multi-level Policy (Estimation)

- CTRs in a block are homogeneous
  - Observations from one cell also give information about others in the block
- How can we model this dependence?
Multi-level Policy (Estimation)

- Shrinkage Model

\[ S_{\text{cell}} \mid \text{CTR}_{\text{cell}} \sim \text{Bin} \left( N_{\text{cell}}, \text{CTR}_{\text{cell}} \right) \]
\[ \text{CTR}_{\text{cell}} \sim \text{Beta} \left( \text{Params}_{\text{block}} \right) \]

All cells in a block come from the same distribution
Intuitively, this leads to shrinkage of cell CTRs towards block CTRs

\[ \omega = \gamma / (\gamma + N_{\text{cell}}) \]

\[ E[\text{CTR}] = \omega \cdot \text{Prior}_{\text{block}} + (1 - \omega) \cdot \frac{S_{\text{cell}}}{N_{\text{cell}}} \]
What happens after the ad click

Predict user behavior after the ad click

- Desired action by the user on the advertisers site is called \textit{conversion}. Examples:
  - A sale of a camera on a digital camera sales site
  - Enter information for follow up on a carpet buying site
  - Petition signing on a petition site
- Conversion is a better proxy for the return that the advertisers gets for the investment in sponsored search
- Ultimate goal: optimize ad selection for better conversion
How to obtain conversion data?

- Conversion tracking
  - 30%-50% of advertisers sign into conversion tracking
  - Instrument their site
  - Search engine use this information to improve conversion
  - Reports to the advertisers – help the advertiser improve conversion

- Conversion data is not always available. Proxies for conversion:
  - Number of clicks on the site
  - Time spent on the site (dwell time)

- This study: using Yahoo! toolbar data
Activity after click

Attenberg et al, KDD 2008
Does CTR relate to site activity?

![Graph showing expected number of clicks by click-through rate](chart)

Attenberg KDD 2008
Summary – use of click data

- Click data is one of the major sources for improving ad selection
- If every ad appeared enough times on every query/page all we would need to do is count
- Most of the methods cope with the data sparsity
  - Product of two power law volumes
- Aggregations – key for obtaining more data by generalization
  - Introduce noise
- Project the click data over bid phrases. Clustering for aggregation
- Hidden variables ("topics") to explain the clicks
- Hierarchical aggregation with supervised classification
- Data depends on the current system settings – need exploration strategy
Thank you!

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

http://research.yahoo.com
This talk is Copyright 2009. Authors retain all rights, including copyrights and distribution rights. No publication or further distribution in full or in part permitted without explicit written permission.