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 4:
Contextual Advertising

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 4 plan

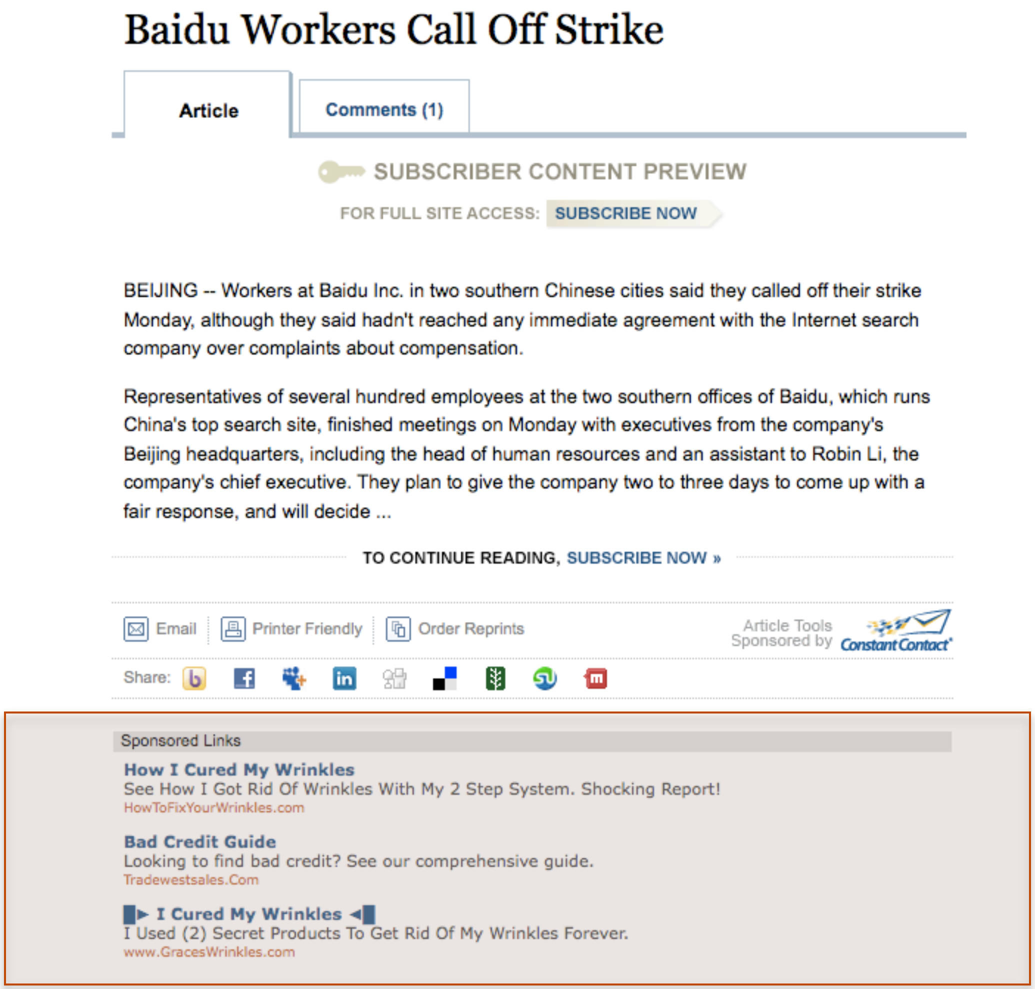
- Contextual advertising basics
- Ad selection in contextual advertising
- Class evaluation at 12:30

Content Match Basics

Contextual Advertising (Content Match)

- Textual advertising on third party web pages
- Complement the content of the web page with paid content
- Ubiquitous on the web
- **Supports the diversity of the web**
  - **Sites small and big rely on CM revenue to cover for the cost of existence**
- Players
  - Google: Adsense
  - Microsoft: ContentAds

Contextual Ads - Example



Baidu Workers Call Off Strike

SUBSCRIBER CONTENT PREVIEW

FOR FULL SITE ACCESS: SUBSCRIBE NOW

BEIJING -- Workers at Baidu Inc. in two southern Chinese cities said they called off their strike Monday, although they said hadn't reached any immediate agreement with the Internet search company over complaints about compensation.

Representatives of several hundred employees at the two southern offices of Baidu, which runs China's top search site, finished meetings on Monday with executives from the company's Beijing headquarters, including the head of human resources and an assistant to Robin Li, the company's chief executive. They plan to give the company two to three days to come up with a fair response, and will decide ...

TO CONTINUE READING, SUBSCRIBE NOW »

Sponsored Links

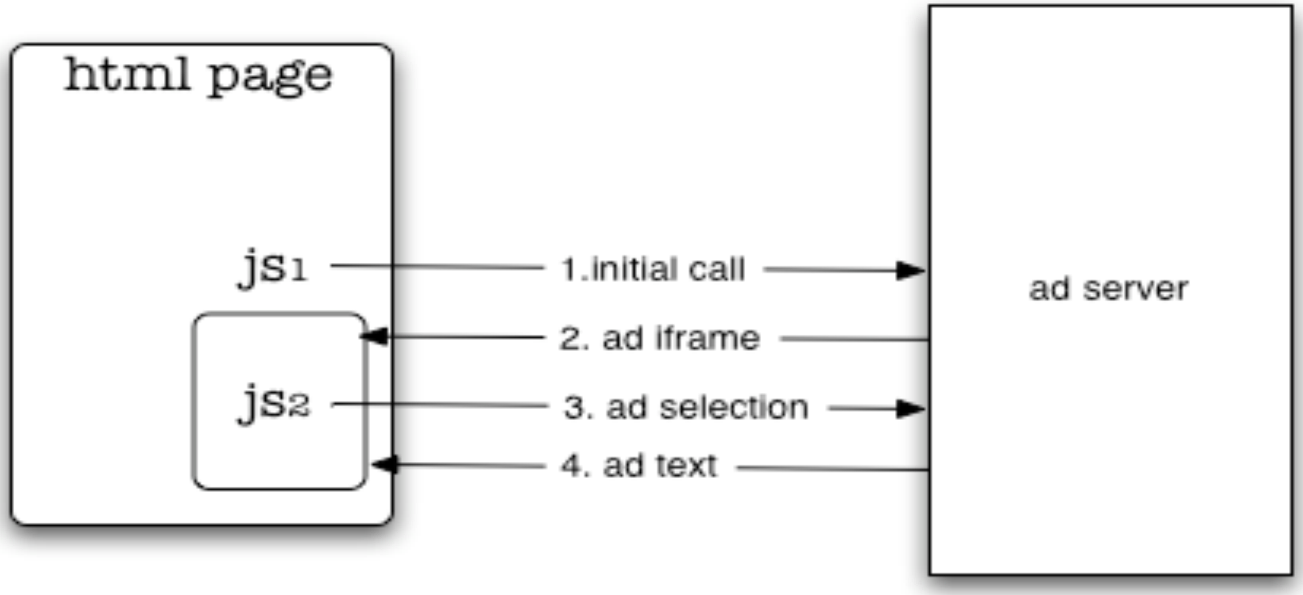
How I Cured My Wrinkles
See How I Got Rid Of Wrinkles With My 2 Step System. Shocking Report!
HowToFixYourWrinkles.com

Bad Credit Guide
Looking to find bad credit? See our comprehensive guide.
Tradewestsales.Com

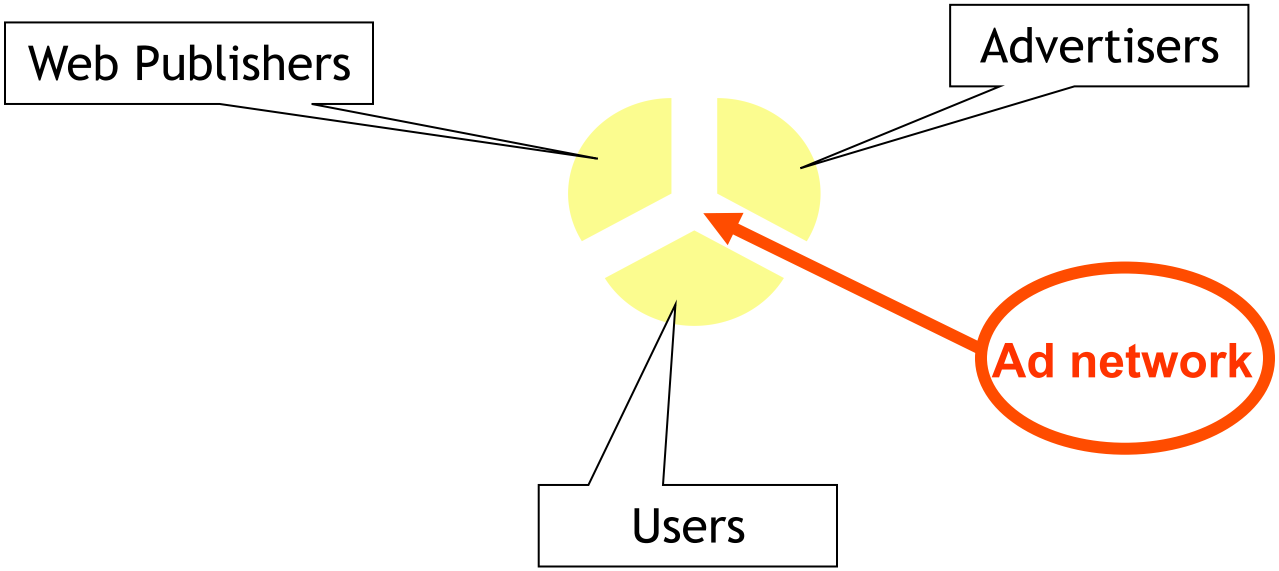
▶ I Cured My Wrinkles ◀
I Used (2) Secret Products To Get Rid Of My Wrinkles Forever.
www.GracesWrinkles.com

How does it all work: the front end

- Two main approaches:
  1. Page fully built by publisher using ads supplied by the ad network.
     - E.g.: XML feed (Usually done with large partners.)
  2. Dynamic loading of ads:



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



Web Publishers

Advertisers

Users

Ad network

Relationship to sponsored search

- Main goal is to increase volume for textual campaigns in sponsored search
- Same type of ads
- Virtually always companion campaign for sponsored search
    - Advertiser opts into contextual advertising

Some differences with Sponsored Search

- Coverage at 100% usually – do not want to leave empty slots on the page
  - Trade-off with display advertising
- Lesser role of the ad network, increased role of the publisher
  - Ad Network: which ads
  - Publisher: how many/where/how
- Ad selection using the content of a web page
  - Much more text
  - Less focused
  - Less intentional

Content Match: The Challenges

- Very thin margin business
- CTR very low – orders of magnitude, ranges in ranges 0.001-0.1%.
  - Higher CTR variance
- Lower conversions – less of a clear intent
- High volume - many page views per day
- More difficult ad placement – not as intentional as search and more difficult for the advertisers to help
- Lower earnings: 1) lower bids 2) share revenue with the publisher
- Other benefits:
  - User tracking

Content match ad selection

Ad selection methods: what information is provided from the page

- Publisher can supply different information to the ad network
- **Page Content**
  - Process the content of the page
  - Cannot be done on-line: crawl
  - Most flexible from the ad selection perspective
- **Page Snippet**
  - Part of the page
  - How much can we process online?
  - How much is enough?
- **Custom Keywords**
  - Sponsored Search – like mechanism
  - Least flexibility in ad selection
  - More control for the publisher

Two main implementation strategies

- **Phrase extraction (from the publisher page)**
  - Map CM to Sponsored Search
  - Extract phrases from the page
  - Use these phrases to select ads (exact match or advanced match in Sponsored Search)
  - Ads selected on a single feature (phrase) from the page and the ad
  - Historically first approach
- **IR approach**
  - Treat CM as a *document similarity* problem
  - Pages are compared to the ads in corpus in a common feature space
  - Bid phrase one of the features used in matching
  - Ads selected based on multiple (overlapping) features of the page and threads

Contextual Advertising Ad Selection: Case studies

	Paper	Method
1.	**Finding Advertising Keywords on Web Pages. Wen-tau Yih et al. In Proc. of WWW 2006**	phrase extraction ad selection
2.	**Impedance coupling in content-targeted advertising. Ribero-Neto et al. In Proc of SIGIR 2005**	IR ad selection
3.	**A Semantic Approach to Contextual Advertising..Broder et al, In Proc. of SIGIR 2007**	IR ad selection
4.	**Contextual Advertising by Combining Relevance with Click Feedback. D. Chakrabarti et al. In Proc of WWW 2008**	IR ad selection using clicks
5.	**To Swing or not to Swing: Predicting when (not) to Advertise. Broder et al, CIKM 2008**	various

Key Information Retrieval Concepts

Finding the "best ad" as an Information Retrieval (IR) problem

- Representation: Treat the ads as documents in IR

  [Ribeiro-Neto et al. SIGIR 2005] [Broder et al. SIGIR2007] [Broder et al. CIKM2008]
- Optimization/solution: Retrieve the ads by evaluating the query over the ad corpus

Details

- Analyze the "query" and extract query-features

  Query = full context (content, user profile, environment, etc)
- Analyze the documents (= ads) and extract doc-features
- Devise a scoring function = predicates on q-features and d-features + weights
- Build a **search engine** that produces quickly the ads that maximize the scoring function
- In the following **documents → ads**

IR from 100,000 feet

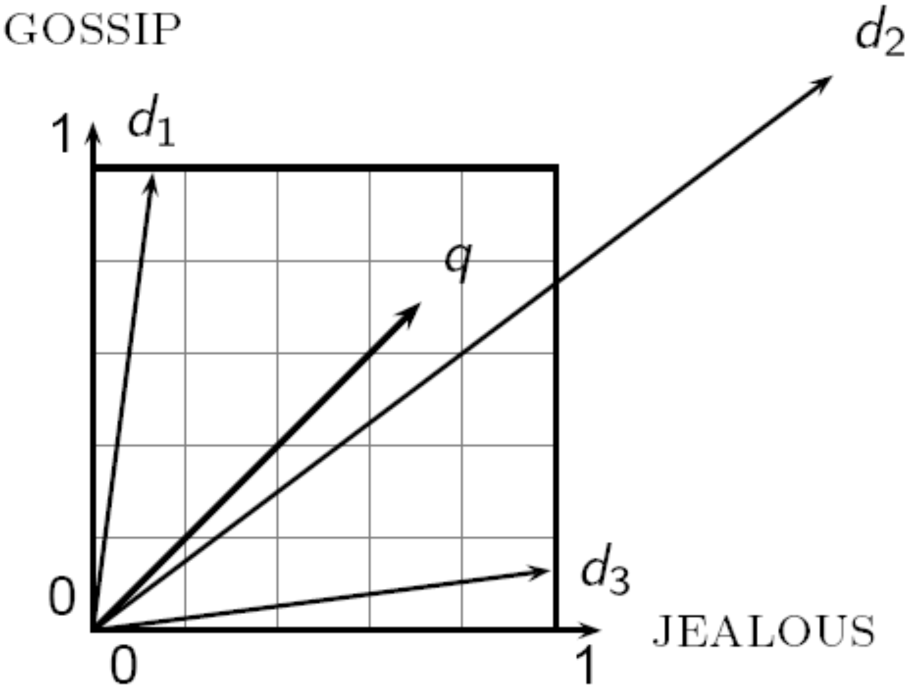
- Collection: Fixed set of **documents**
- Query: Description of the user's information need
- Goal: Retrieve documents with information that is relevant to user's information need and helps him complete a task
  - How would you formulate the task in the ad retrieval case?

Basics of similarity search

- Sim(a,p) is a function of a set of **features** of **a** and **p**
  - **a** and **p** are vectors
- Usually calculate similarity in a **high dimensional space** of features. Orders of magnitude numbers for textual ads:
  - unique words ~ 1M-2M
  - sequences (phrases) ~ 10M

One similarity measure: vector space proximity

- Common similarity measure – the cosine of the angle between the vectors **cosine similarity** or **dot product**
  - Product of the weights of the common dimensions



Cosine(query, document)

Dot product

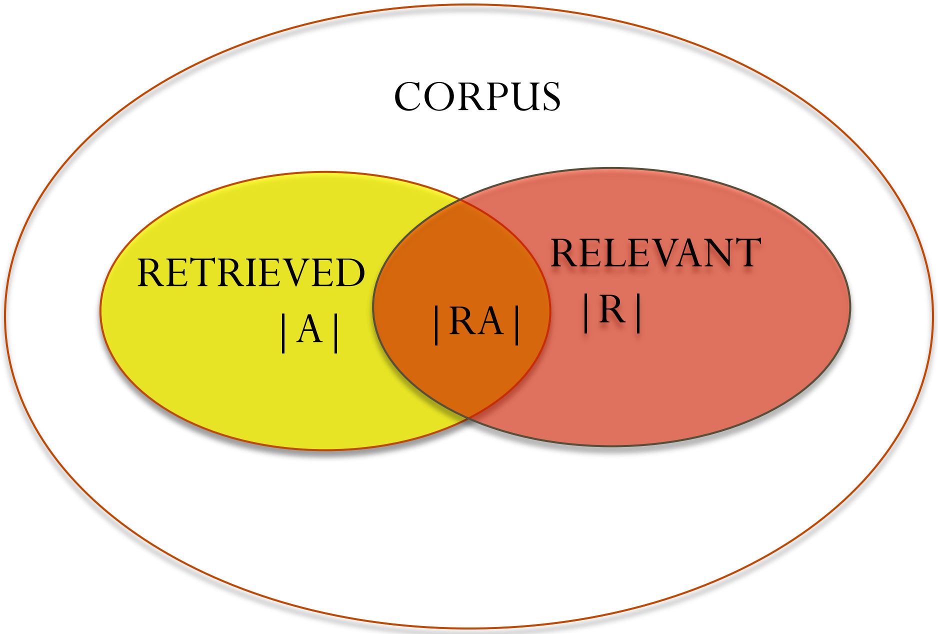
Unit vectors

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

q_i is the weight of term *i* in the query
d_i is the weight of term *i* in the document
cos(*q*,*d*) is the cosine similarity of *q* and *d* … or,
equivalently, the cosine of the angle between *q* and *d*.

Metrics: Precision-Recall

- **Precision**: fraction of retrieved documents that are relevant
  - P=|RA|/|A|
- **Recall**: proportion of relevant documents (ads) in the retrieved documents
  - P = |RA|/|R|



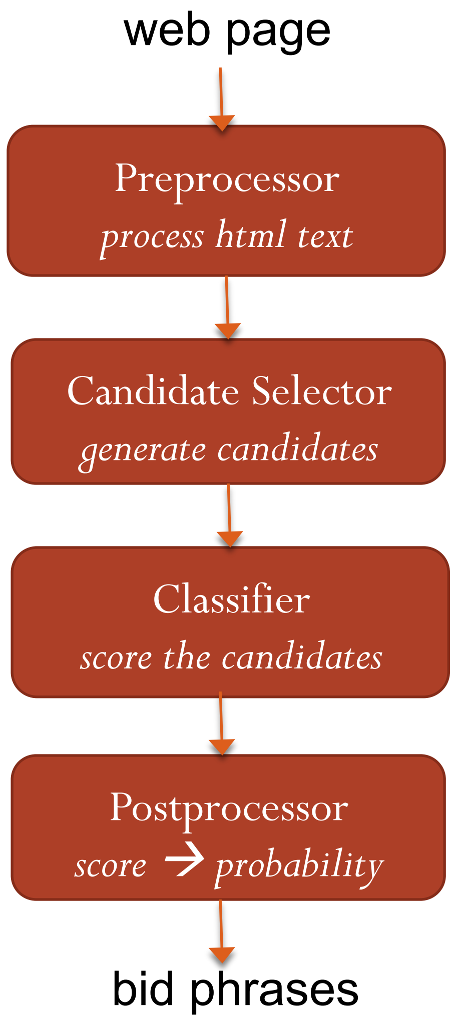
Phrase Extraction for Contextual Advertising

Finding Advertising Keywords on Web Pages. Wen-tau Yih et al. In Proc. of WWW 2006

Contextual Advertising by single feature

- Goal: given a page find phrases that are good for placing ads
- **Reverse search problem**: given a page, find the queries that would match (summarize) the content of this page
- Select ads based on a single selected keyword:
  - Contextual Advertising translated into database approach of Sponsored Search
  - Reuse of the Sponsored Search infrastructure – lower cost

System architecture



web page

Preprocessor

process html text

Candidate Selector

generate candidates

Classifier

score the candidates

Postprocessor

score → probability

bid phrases

Candidate Selection

- All phrases of length up to 5 (including single words)
  - Within a single page block (sentence)
- Two dimensions of candidate selection:
  - Individual occurrences extracted separately vs. combining all occurrences into entry per page (*separate vs. combined*)
  - Consider the phrase as a whole
  - Label individual words with their relationship with a phrase:
    - **B**eginning of a phrase
    - **I**nside a phrase
    - **L**ast word of a phrase
    - …

Classifier

- Given a phrase predict if it is "keyword" (usable for selecting ads)
- Binary classifier:
  - Logistic regression model $P(Y=1 \mid x=\overline{x}) = \frac{1}{1+e^{-xw}}$
  - x is vector of features of a given phrase
  - w is a vector of importance weights learned from the training set

Features

- **Linguistic features**: is a noun; is a proper name; is a noun phrase; are all words in the phrase of the same type
- **Capitalization**: any/all/first word capitalization
- **Section based features**:
  - Hypertext – is the feature extracted from anchor text
  - Title
  - Meta tags
  - URL
- **IR features**: tf, idf, log(tf), log(idf), sentence length, phrase length, relative location in the document
- **Query log features**: log(phrase frequency), log(first/second/interior word frequency)

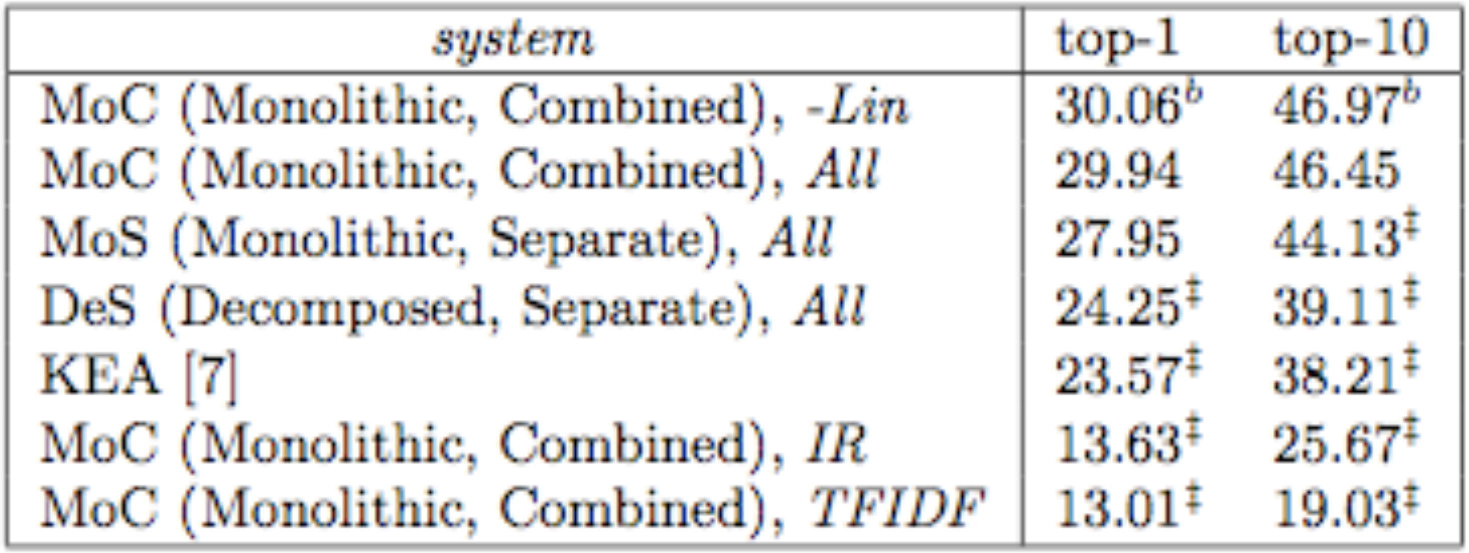
Experiments: Data

- 828 pages
- Indexed by MSN
- Have ads
- In the Internet Archive
- One page per domain
- Eliminate foreign and adult pages
- Editors (8) instructed to seek highly prominent keywords with advertising potential

Measuring the extraction quality

- Editorial judgments
- Precision-recall – might be too difficult
  - Too long for the judges to find all the relevant phrases
  - Given a phrase – influence the judges
- A proxy for P-R
  - top-1 = top-1 results is in the list selected by the editor, count across the set of pages
  - top-10 = % of top-10 results in the editor set, averaged over the set of pages

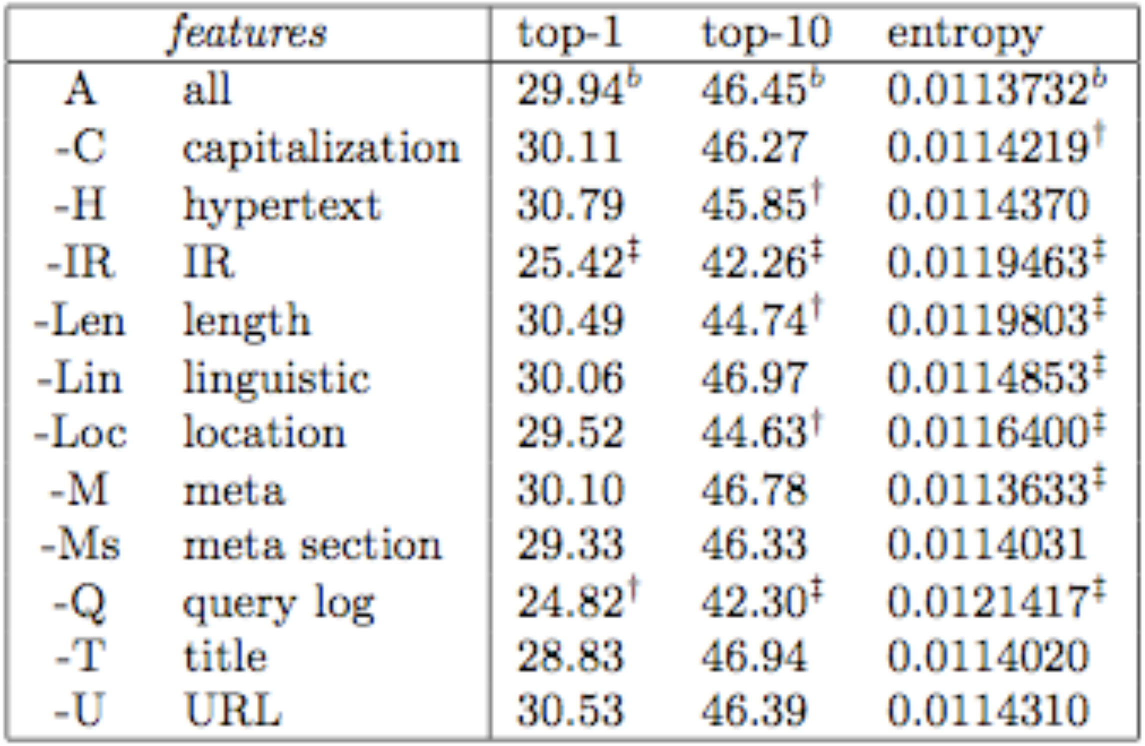
Main result



system	top-1	top-10
MoC (Monolithic, Combined), *-Lin*	30.06$^{\flat}$	46.97$^{\flat}$
MoC (Monolithic, Combined), *All*	29.94	46.45
MoS (Monolithic, Separate), *All*	27.95	44.13‡
DeS (Decomposed, Separate), *All*	24.25‡	39.11‡
KEA [7]	23.57‡	38.21‡
MoC (Monolithic, Combined), *IR*	13.63‡	25.67‡
MoC (Monolithic, Combined), *TFIDF*	13.01‡	19.03‡

Table 1: Performance of different systems

Feature importance



	features	top-1	top-10	entropy
A	all	29.94$^\flat$	46.45$^\flat$	0.0113732$^\flat$
-C	capitalization	30.11	46.27	0.0114219†
-H	hypertext	30.79	45.85†	0.0114370
-IR	IR	25.42‡	42.26‡	0.0119463‡
-Len	length	30.49	44.74†	0.0119803‡
-Lin	linguistic	30.06	46.97	0.0114853‡
-Loc	location	29.52	44.63†	0.0116400‡
-M	meta	30.10	46.78	0.0113633‡
-Ms	meta section	29.33	46.33	0.0114031
-Q	query log	24.82†	42.30‡	0.0121417‡
-T	title	28.83	46.94	0.0114020
-U	URL	30.53	46.39	0.0114310

Table 3: The system performance by removing one set of features in the MoC framework

Conclusion

- Mapping Contextual Advertising to Sponsored Search
  - Extract phrases from the publisher’s web page
  - Select ads using exact or advanced match on this phrase
- Ad selection using a single feature
- Approach based on logistic regression trained on editorial judgments
  - Editors extracting salient terms from pages
- Combining the information from multiple occurrences and treating the phrases as single units yields best results
- IR and query log features account for almost all of the signal
- Low precision – difficult problem

IR methods for content match ad retrieval

Impedance coupling in content-targeted advertising. Ribeiro-Neto et al. SIGIR 2005

Using more than one feature in ad matching

- The phrase extraction approach uses one feature of the page (phrase) to select the ads
- Risk with ambiguous phrases: ‘Tahoe’ is a destination as well as a truck model.
- Can we select ads based on multiple features from the page?
  - What are the features of the ad?
  - How to weight the features?
  - What metrics to use to relate the ads to the pages?

Formalism for comparing ads and pages: Vector Space Model

- Represent each ad **a** as a vector: $\mathbf{a} = \{\mathbf{w}_{1a}, \mathbf{w}_{2a}, \ldots, \mathbf{w}_{na}\}$
  - In this study: **a** is the visible part of the ad (title and abstract)
- Represent the page **p** as a vector in the same space $\mathbf{p} = \{\mathbf{w}_{1p}, \mathbf{w}_{2p}, \ldots, \mathbf{w}_{np}\}$
- Weights using tf-idf method (last lecture)
- Use cosine of the angle between the vectors to rank the ads for a given page – denoted by **sim()**

Basic set of measures

- **AD(p, a) = sim(p,a)** – based on the visible parts of the ad
- **KW(p,a) = sim(p, kw(a))** – based on the keyword of a
- **AD_KW(p, a) = sim(p, a ∪ kw(a))** - using both the visible parts and the keyword
- Assuming that **kw(a)** summarizes well the essence of **a**, assure the presence of **kw(a)** in **p**

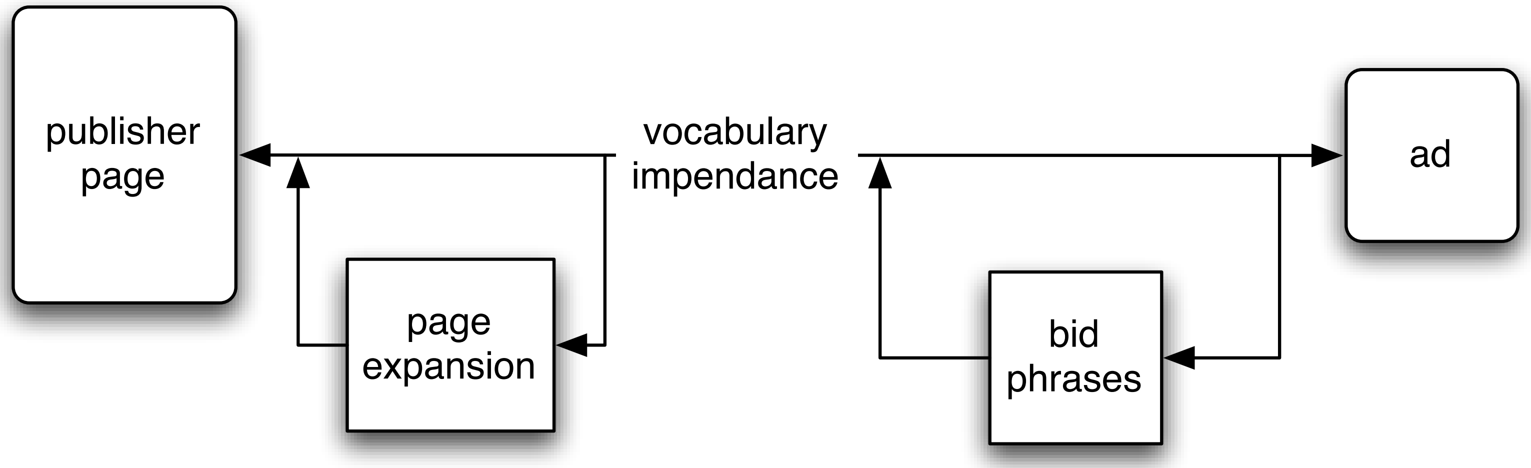
$$ANDKW(p,a) = \begin{cases} sim(p,a) & \text{if } kw(a) \subseteq p \\ 0 & otherwise \end{cases}$$

$$AAK(p,a) = \begin{cases} sim(p,a \cup kw(a)) & \text{if } kw(a) \subseteq p \\ 0 & otherwise \end{cases}$$

The Vocabulary impedance Problem

- Language and the topic of the page and the ad can differ substantially:
  - Publisher page belongs to a broader/narrower contextual scope
  - Ads concise in nature
  - ‘Hidden topic’ – not mentioned in the ad and/or the page
- Intersection of the vocabularies of related pages and ads can be low: *vocabulary impedance problem*

Solution: Impedance Coupling



Bayesian network model for page expansion using similar pages

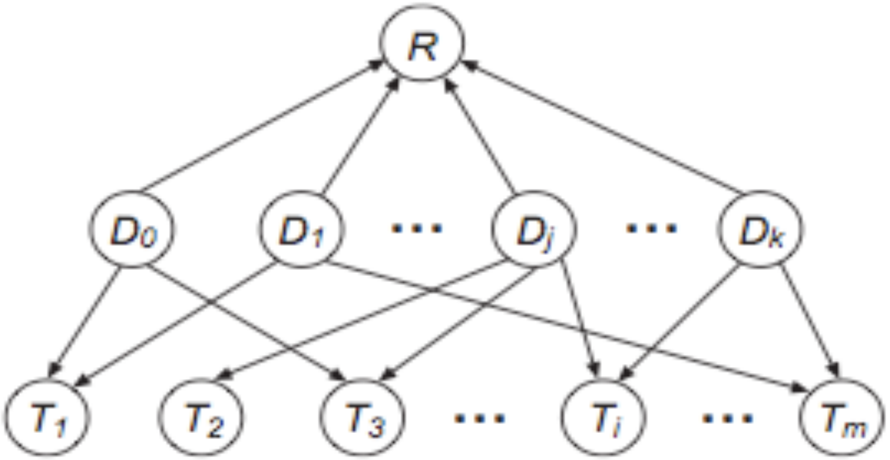


Figure 3: Bayesian network model for our impedance coupling technique.

$$P(T_i|R) = \frac{1}{P(R)} \sum_{\mathbf{d}} P(T_i|\mathbf{d})P(R|\mathbf{d})P(\mathbf{d}) \qquad (2)$$

$$P(T_i|R) = \frac{\nu}{P(R)} \sum_{j=0}^{k} P(T_i|\mathbf{d_j})P(R|\mathbf{d_j})$$

Page expansion, continued

$$P(T_i|\mathbf{d_j}) = \eta \; w_{ij}$$

$$P(R|\mathbf{d_j}) \;\; = \;\; \begin{cases} (1-\alpha) & j = 0 \\ \alpha \; sim(r, d_j) & 1 \leq j \leq k \end{cases}$$

$$P(T_i|R) = \rho \; ((1-\alpha) \; w_{i0} + \alpha \sum_{j=1}^{k} w_{ij} \;\; sim(r, d_j))$$

AAK_T(p, a) = sim(r, a)
AAK_EXP(p,a) = AAK(p ∪ r, a)

feature selection for r: $P(T_i|R)/P(T_{top}|R) > 0.05$

Page expansion results: it works

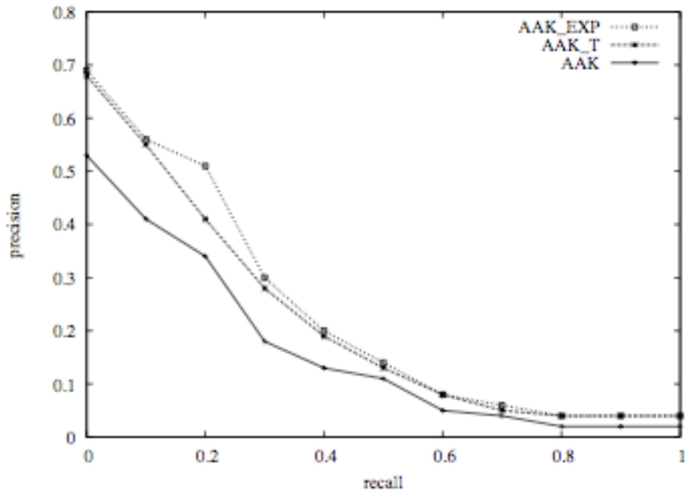


Figure 6: Impact of using a new representation for the triggering page, one that includes expansion terms.

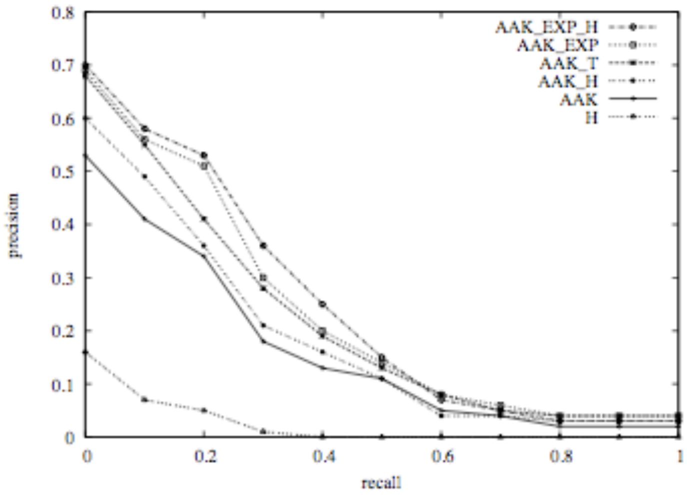


Figure 8: Comparison among our ad placement strategies.

Ad expansion: landing page content is useful

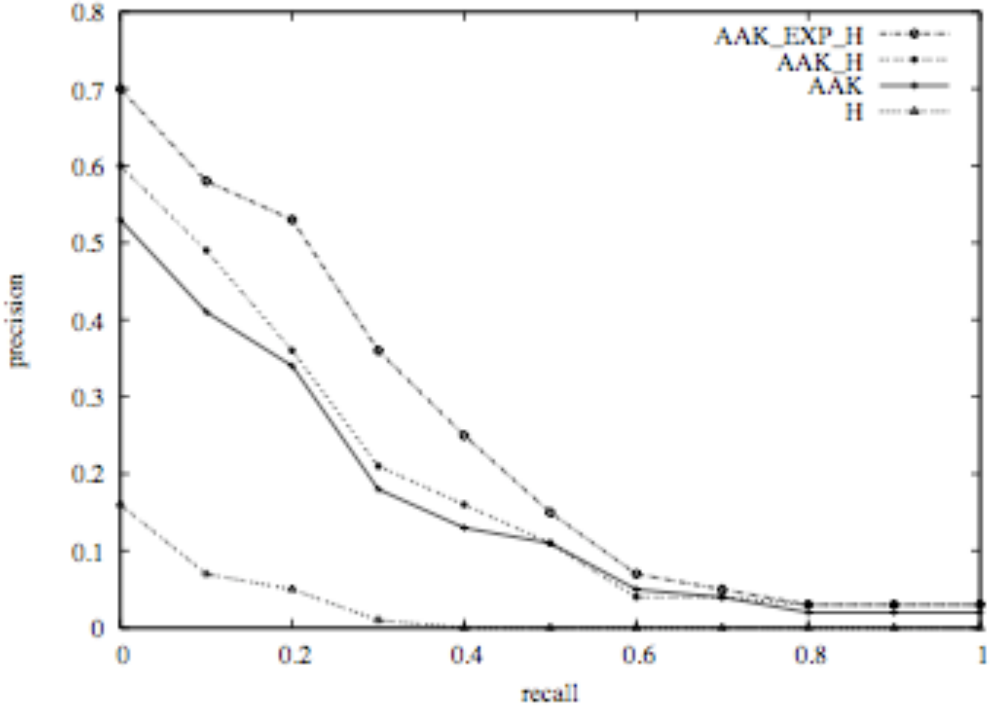


Figure 7: Impact of using the contents of the page pointed by the ad (the hyperlink).

Summary

- Using IR techniques to match ads and pages
- Both the ad and the page are mapped to a common vector space
- Cosine of the angle between the ad and the page as the basic similarity measure
  - Bid phrase as a required feature – projection of the space
- Expanding pages using terms from similar pages improves results
- Landing page contains useful data for ad selection
- Some practical considerations:
  - How long are the queries?
  - How much is the cost of this method?

Holistic view at the page in Contextual Advertising

Semantic-Syntactic Approach to Contextual Advertising. AB, M. Fontoura, L. Riedel, VJ. SIGIR 2007

Motivation

- Even with using multiple features there is still a risk that the subset used in matching does not represent the semantics of the page
- Can we somehow summarize the content of the whole page into a small number of features?
  - This work: supervised approach based on classification
- Use external knowledge: taxonomies
  - This work: a topical taxonomy
- What is a better signal: page class or page words? Or both?

Semantic-syntactic match

- Figure out the topic of the page
  - Classification of the page into a commercial oriented taxonomy
- Pre-classify all the ads into the same taxonomy
- Restrict the matching to ads that are in related categories
- Use word similarity to improve the match

Page and ad classification

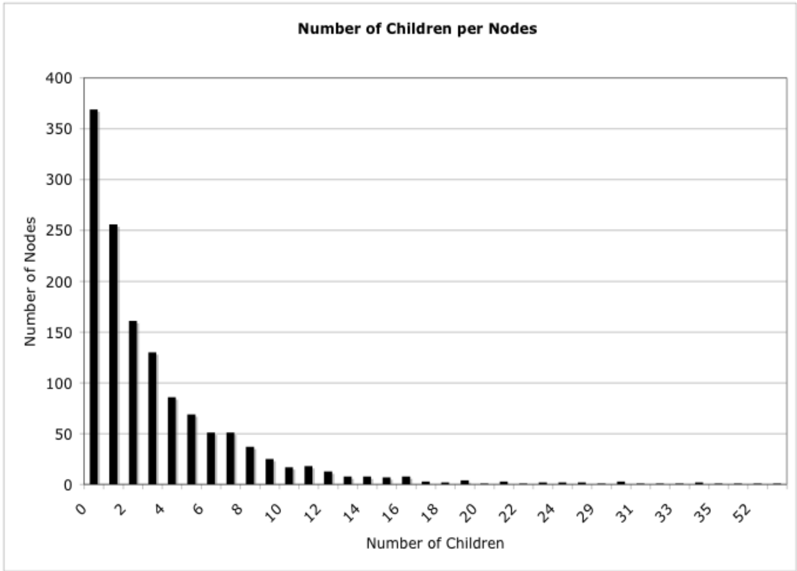
- Use a large scale classification to relate pages and ads
  - Need a taxonomy with sufficient resolution
- We used a taxonomy of 6,000+ nodes, primarily built for classifying commercial interest queries
  - Each node is a collection of query terms
- Rocchio-style nearest neighbor classifier
  - Meta-document produced of the queries at each node

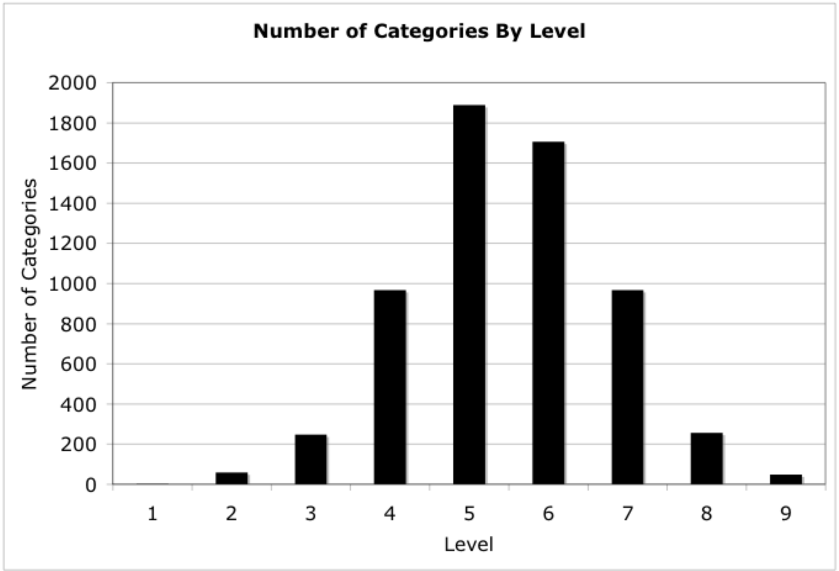
$$C = \alpha \frac{1}{|D_r|} \sum_{\vec{d} \in Dr} \vec{d} + (1 - \alpha) \frac{1}{|D_{nr}|} \sum_{\vec{d} \in Dnr} \vec{d}$$

Taxonomy requirements: intuition

- Enough resolution to be useful
- Not too specific to make maintenance too costly:
  - Electronics - too broad
  - Electronics/Digital Camera/Canon - feasible
  - Electronics/Digital Camera/Canon/XT10i - hard to maintain

Taxonomy statistics



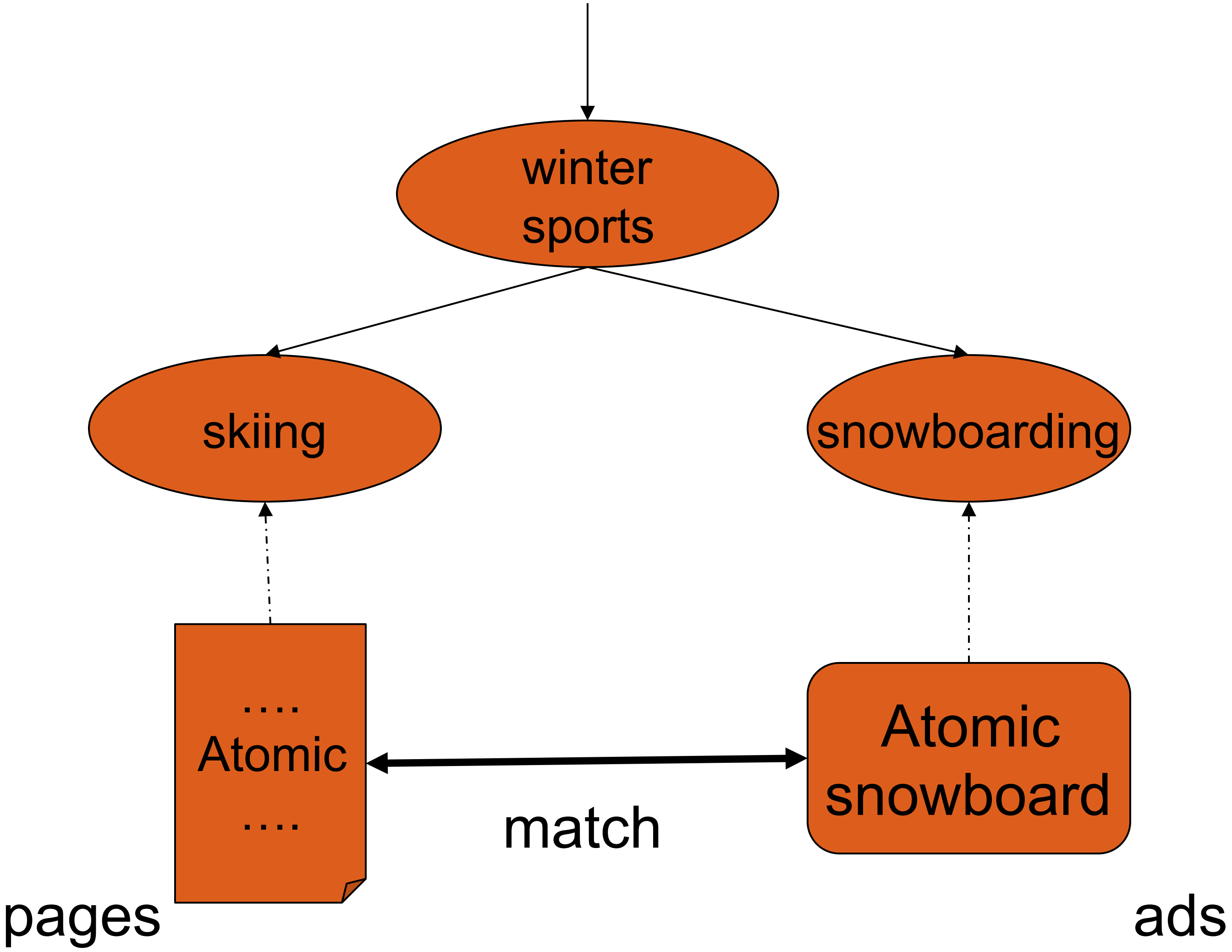


Scoring

- For a given page score every ad, select the top-k ads
- Linear combination of 2 scores:
  - Taxonomy score (semantic distance)
  - Word and phrase score (syntactic distance)
- Allow generalization in the taxonomy

$$Score(p_i, a_i) = \alpha \cdot TaxScore(Tax(p_i), Tax(a_i)) + (1 - \alpha) \cdot KeyordScore(p_i, a_i)$$

Generalization paths



Semantic and syntactic scores

- Semantic component - class based

$$\sum_{d \in Tax(x_i)} cWeight(d) = 1 \qquad idist(c,p) = \frac{n_c}{n_p}$$

$$TaxScore(PC, AC) =$$

$$\sum_{pc \in PC} \sum_{ac \in AC} idist(LCA(pc, ac), ac) \cdot cWeight(pc) \cdot cWeight(ac)$$

- Syntactic component - term vector cosine

$$tWeight(kw^{si}) = weightSection(S_i) \cdot tf\_idf(kw)$$

$$KeyordScore(p_i, a_i) = \frac{\sum_{i \in |K|} tWeight(pw_i) \cdot tWeight(kw_i)}{\sqrt{\sum_{i \in |K|} (tWeight(pw_i))^2} \sqrt{\sum_{i \in |K|} (tWeight(aw_i))^2}}$$

Searching the ad space

- Ad search done in real time - how to make it fast enough?
- Index the ads using a inverted index
  - Use the page features as the query
- Find top-k ads with the highest score
- Monotonic scoring function that has the two sub-scores
- Evaluate the query using a variant of the WAND doc-at-a-time algorithm [Broder et al.]

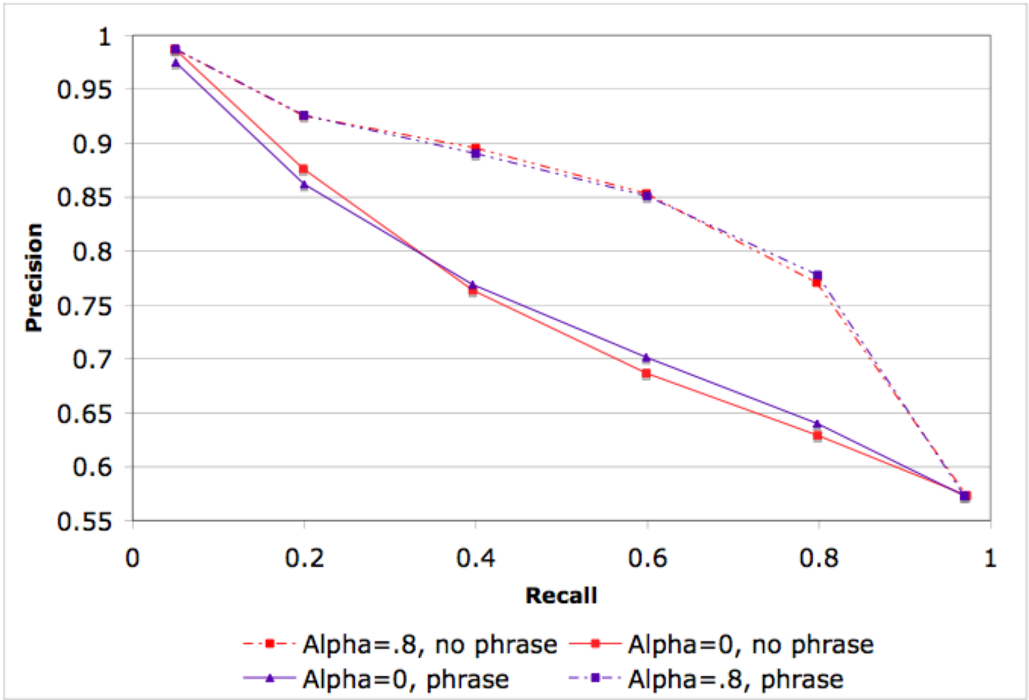
Dataset

- Ad-page pairs manually evaluated 3 times by human editors as: (1) Relevant; (2) Somewhat relevant; and (3) Irrelevant
- Average judgments and round to the closest integer
- 3 x 3K judgments for a set of 105 pages
- The pages sampled from a set of over 20M pages that are enabled for contextual advertising
- Ads selected from a set of over 10M ads

Pooling: using data from previous evaluation

- Faster turnaround, lower cost
- Essentially reordering of the prior results
- Could be off if the new method would select substantially different ads
- For each page consider only the judged ads
  - Did not have the exact ad set used in the original experiments
- Rank the ads by each method
- Precision/recall and K-tau to compare different orderings
- Precision at 1,3,5
- Evaluate relative performance of the methods

Some Results - using past editorial judgments



alpha	K-tau
0.00	0.086
0.25	0.155
0.5	0.158
0.75	0.166
1.00	0.136

Conclusions

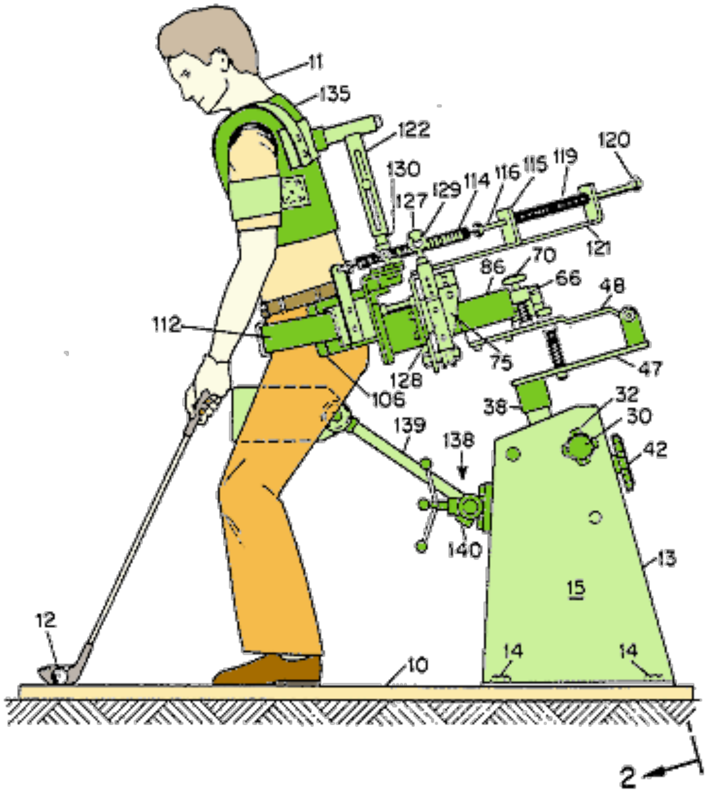
- Contextual advertising is the economic engine for a large number of non-transactional sites
- Novel way to match ads to pages
- Topical (semantic) similarity is a major component of the relevance score (~80%)
- Evaluation showing results for different alpha values

When to advertise

To Swing or not to Swing: Predicting when (not) to Advertise.
Broder et al, CIKM 2008

The “Swing” Problem

- Repeatedly showing non-relevant ads can have detrimental long-term effects
- Want to be able to predict when (not) to show individual ads or a set of ads (“swing”)
- Modeling actual short and long term costs of showing non-relevant ads is very difficult



Two Approaches

- Thresholding Approach:
  - Rank the ads by score; cut-off at certain rank or score
  - Decision made on individual ads
  - Only based on ad scores
- Ad Set Machine Learning Approach
  - Decision made on **sets** of ads
  - Based on a variety of features
- Applies to both Sponsored Search and Contextual Advertising

Thresholding Approach

- Set a global score threshold
- Only retrieve ads with scores above threshold
- If none of the ad scores are above the threshold, then no ads are retrieved ("no swing")

Ad Set Approach

- Learn a binary prediction model ("swing" or "no swing") for an entire **set** of ads
- If we swing, then all ads are retrieved
- If we do not swing, then no ads are retrieved
- Must extract features defined over sets of ads, rather than individual ads
- Use support vector machines (SVMs)

Features

- Relevance features
  - Word overlap
  - Cosine similarity
- Vocabulary mismatch features
  - Translation models
  - Point-wise mutual information
  - Chi-squared
- Ad-based features
  - Bid price
  - Coefficient of variation of ad scores
- Result set cohesiveness features ✓
  - Result set clarity
  - Entropy

Ad set language model

- Language model: relative frequency of words conditioned on a given query:

$$\theta_w = \sum_{A \in Ads} P(w \mid A) P(A \mid Q)$$

$$P(w \mid A) = \frac{tf_{w,A}}{\mid A \mid}$$

$$P(A \mid Q) = \frac{score(q,A)}{\sum_{A' \in ads(q)} score(q,A')}$$

Clarity and entropy of the language model as features

$$H(\theta) = \sum_{w \in V} \theta_w log(\theta_w)$$

$$D_{KL}(P,Q) = H(P,Q) - H(P) = \sum_j p(j)log(q(j)) - \sum_j p(j)log(p(j))$$

$$CLARITY(\theta) = D_{KL}(\theta,\hat{\theta})$$

$$\hat{\theta}_w = \frac{tf_w}{|Corpus|}$$

- Intuition: how much is the distribution of words different from noise (aggregate over all ads)
- Entropy: in every domain there is a set of core words that describe the domain

Conclusion

- Two approaches to determine when to show ads
- Thresholding approach
  - Only shows ads above some global score threshold
  - Most effective for sponsored search
- Machine learning approach
  - Predicts over entire set of ads
  - Semantic class features important for prediction
  - Effective for both sponsored search and content match
- In practice we can combine both approaches

Search-based ad selection for sponsored search

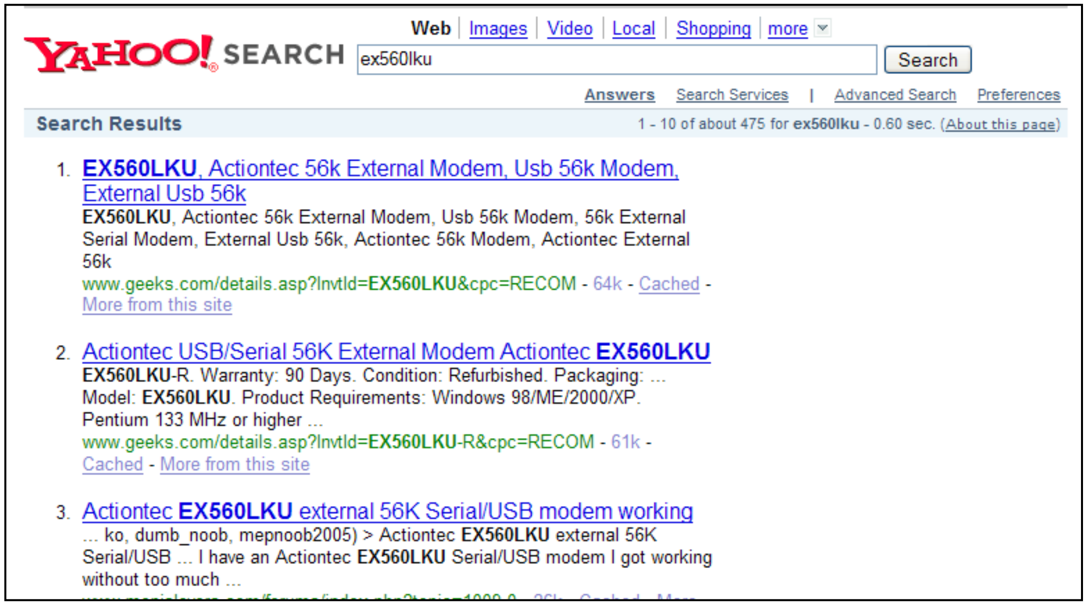
Search Advertising Using Web Relevance Feedback: AB, P. Ciccolo, M. Fontoura, E. Gabrilovich, VJ, L. Riedel. ACM CIKM 2008

An alternative view of Search Advertising

- A lesson from Content Match
  - View Search Advertising as CA on the web search result page
  - More general: use the web search results as a basis for ad selection
- What are the benefits?
  - Uniform look of the result page – improved user experience
  - Re-use of the web search technology
  - Circumstantial evidence for Search Advertising
- The approach
  - Web search results as (pseudo) feedback for the web search query
  - Expanded web search query used as a long ad query
  - Evaluate the ad query to select the ads

Where to look for features?

Snippets or full pages?



Web | Images | Video | Local | Shopping | more
YAHOO! SEARCH
Search
Answers
Search Services | Advanced Search
Preferences

ex560lku

Search Results 1 - 10 of about 475 for **ex560lku** - 0.60 sec. (About this page)

1. **EX560LKU**, Actiontec 56k External Modem, Usb 56k Modem, External Usb 56k
   EX560LKU, Actiontec 56k External Modem, Usb 56k Modem, 56k External Serial Modem, External Usb 56k, Actiontec 56k Modem, Actiontec External 56k
   www.geeks.com/details.asp?InvtId=**EX560LKU**&cpc=RECOM - 64k - Cached - More from this site

2. Actiontec USB/Serial 56K External Modem Actiontec **EX560LKU**
   EX560LKU-R. Warranty: 90 Days. Condition: Refurbished. Packaging: ... Model: **EX560LKU**. Product Requirements: Windows 98/ME/2000/XP. Pentium 133 MHz or higher ...
   www.geeks.com/details.asp?InvtId=**EX560LKU**-R&cpc=RECOM - 61k - Cached - More from this site

3. Actiontec **EX560LKU** external 56K Serial/USB modem working
   ... ko, dumb_noob, mepnoob2005) > Actiontec **EX560LKU** external 56K Serial/USB ... I have an Actiontec **EX560LKU** Serial/USB modem I got working without too much ...

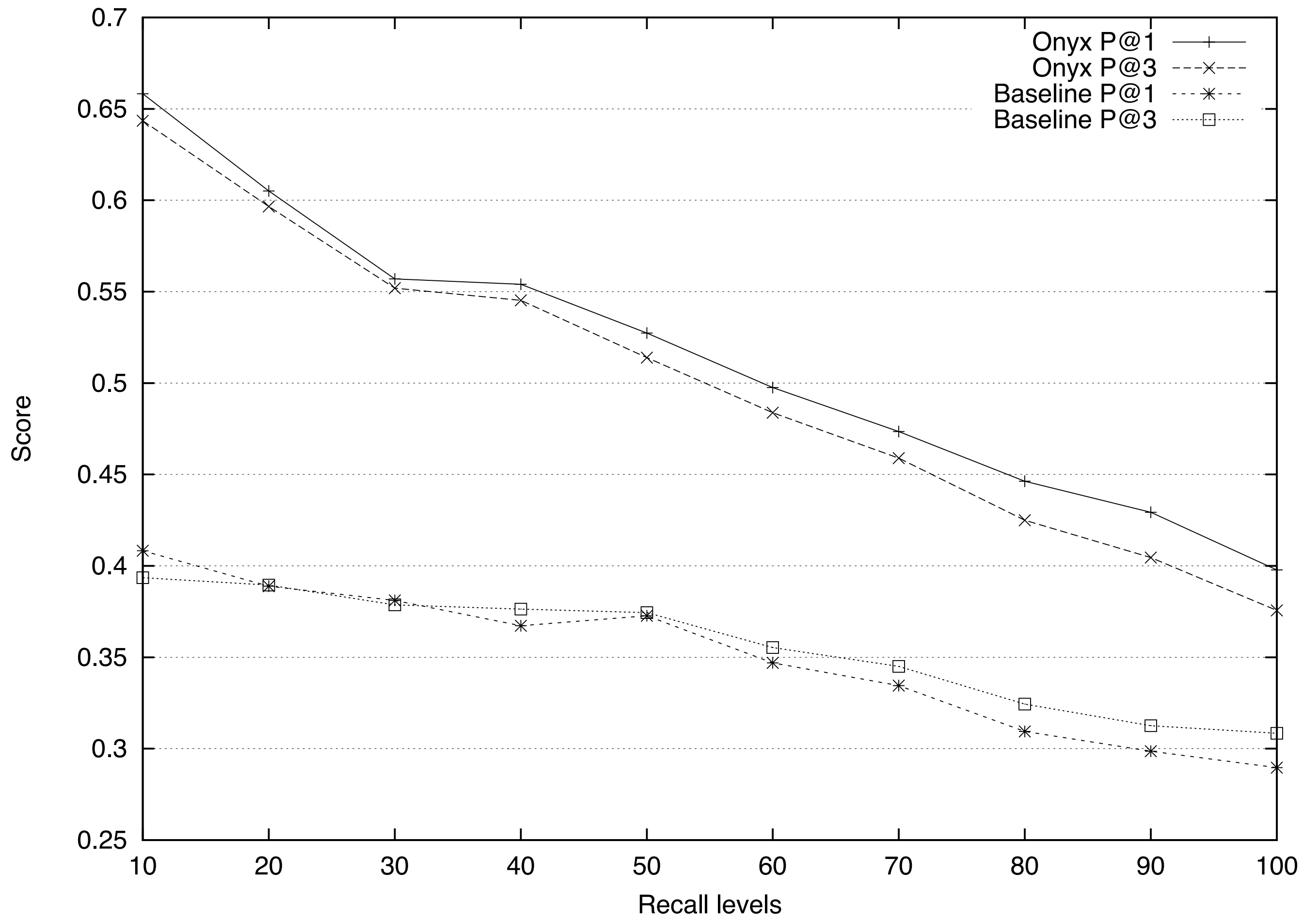
Number of search results to obtain

Number of features per search result

Aggregation:

bundling or voting?

Precision-Recall



Click prediction in Sponsored Search

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

Predicting Clicks: beyond the bid phrase

Predicting Clicks: Estimating the Clickthrough Rate for New Ads: M. Richardson, E. Dominowska, R. Ragno, WWW2008

How to predict the CTR of a sponsored search ad

- Lets start with the simplest scenario: exact match
  - the query is equal to the ad bid phrase
- Try 1: average ctr per query
- Some queries this will not work (which?)
- Try 2: cluster queries (bid phrases), smooth the estimates toward the cluster based on the query volume

$$eCTR = \frac{\alpha CTR_{query} + n CTR_{cluster}}{\alpha + n}$$

- Is there information beyond the query?

Using Ad Features to Predict CTR

- Still assuming **exact match**: the CTR depends solely on the information in the ad
- Predict CTR by extracting features from the ad and building a model based on a training set
- Logistic regression as a model:

$$z = \sum_i w_i \cdot f_i(ad) \qquad ctr = \frac{1}{1 + e^{-z}}$$

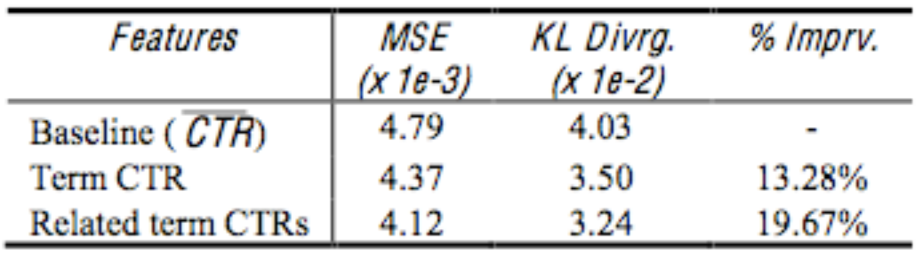
- Cross entropy loss function
- For each feature
  - Normalize to zero mean, unit standard dev, crop outliers to 5S
  - Add derived features log(f+1) and f^2

Features: bid phrase based

- Ads with the same bid phrase $f_0(ad) = \frac{\alpha \overline{CTR} + N(ad_{term}) CTR(ad_{term})}{\alpha + N(ad_{term})}$
- Ads with similar bid phrases – logit of CTR and counts:

$$R_{mn}(t) = \left\{ ad : \begin{array}{l} |ad_{term} \cap t| > 0 \text{ and} \\ |t - ad_{term}| = m \text{ and} \\ |ad_{term} - t| = n \end{array} \right\} \qquad CTR_{mn}(term) = \frac{1}{|R_{mn}(term)|} \sum_{x \in R_{mn}(term)} CTR_x$$

Table 1: *Term* and *Related Term* Results



Features	*MSE (x 1e-3)*	*KL Divrg. (x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
Term CTR	4.37	3.50	13.28%
Related term CTRs	4.12	3.24	19.67%

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

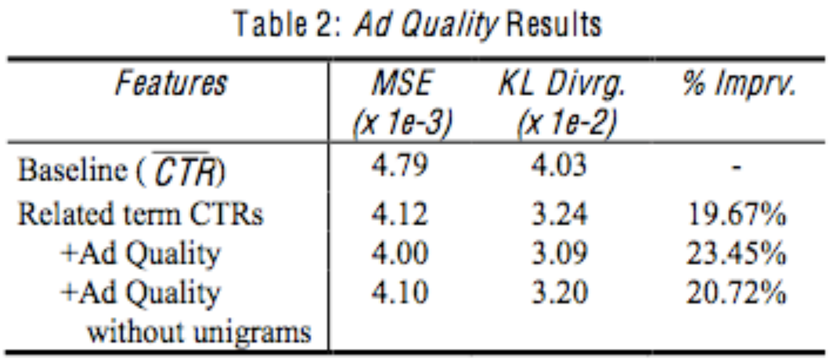


Table 2: *Ad Quality* Results

Features	*MSE* *(x 1e-3)*	*KL Divrg.* *(x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
Related term CTRs	4.12	3.24	19.67%
+Ad Quality	4.00	3.09	23.45%
+Ad Quality without unigrams	4.10	3.20	20.72%

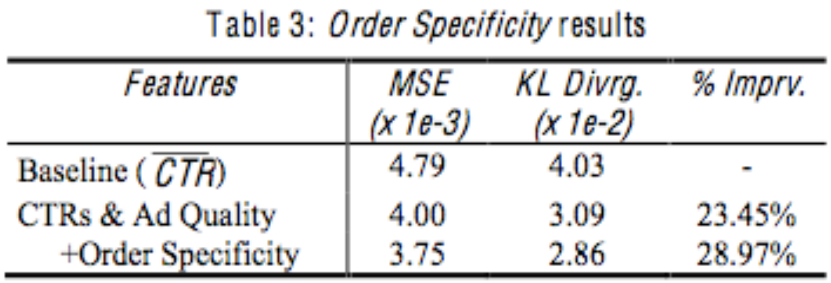


Table 3: *Order Specificity* results

Features	*MSE* *(x 1e-3)*	*KL Divrg.* *(x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
CTRs & Ad Quality	4.00	3.09	23.45%
+Order Specificity	3.75	2.86	28.97%

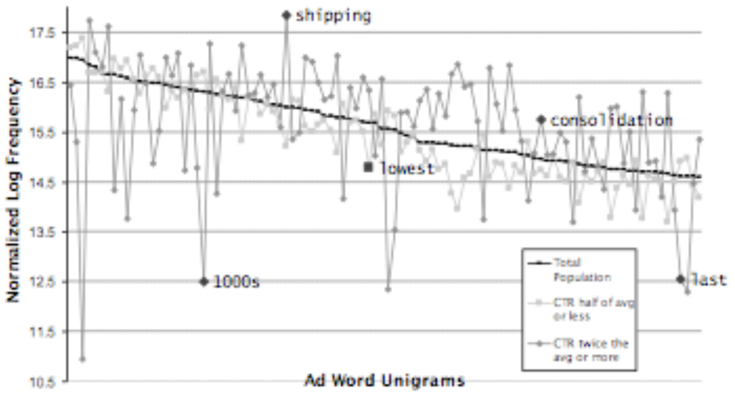


Figure 4. Frequency of advertisement word unigrams, sorted by overall frequency. The light and dark gray lines give the relative frequency of unigrams in low and high CTR ads.

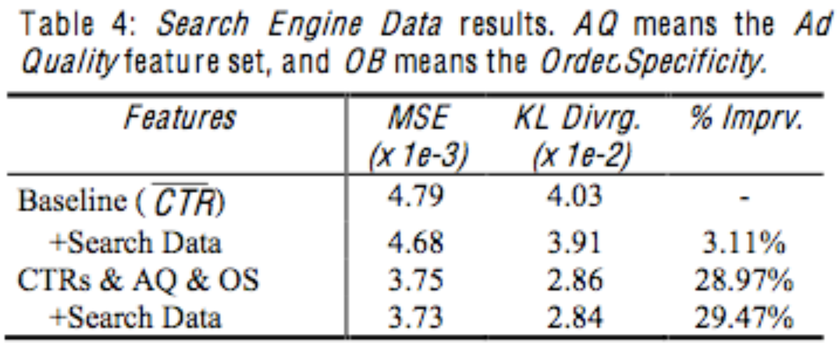


Table 4: *Search Engine Data* results. *AQ* means the *Ad Quality* feature set, and *OB* means the *Order Specificity*.

Features	*MSE* *(x 1e-3)*	*KL Divrg.* *(x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
+Search Data	4.68	3.91	3.11%
CTRs & AQ & OS	3.75	2.86	28.97%
+Search Data	3.73	2.84	29.47%

Richardson et al. WWW2008

Empirical vs. estimated CTR with regard to ad views

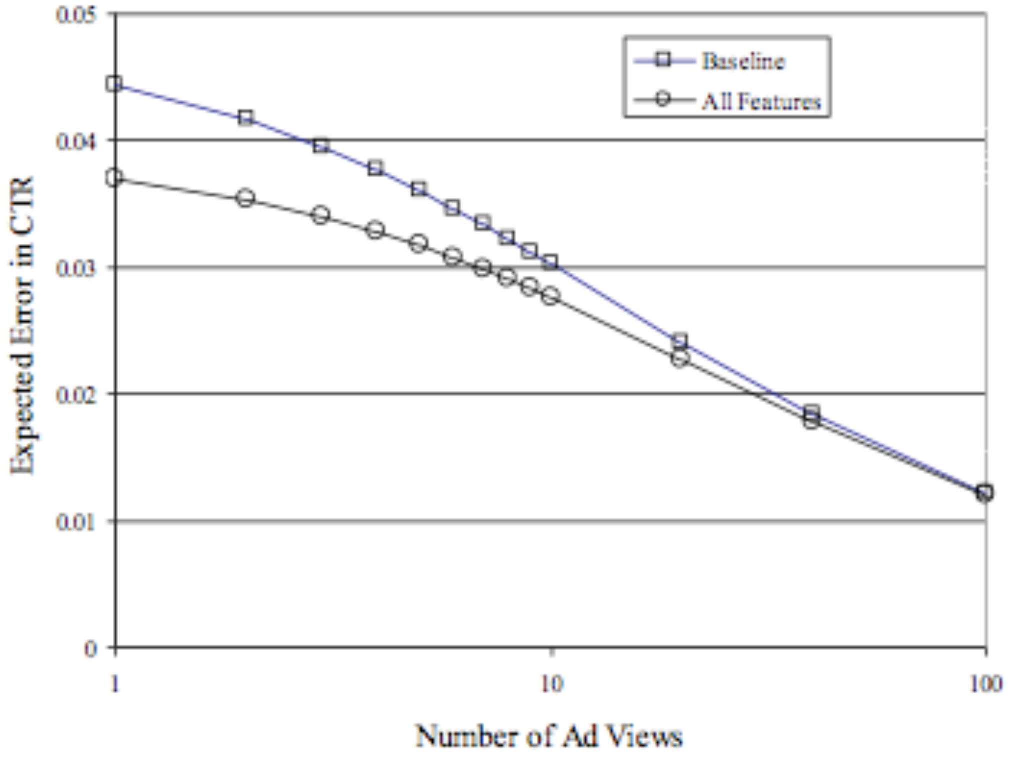


Figure 6: Expected mean absolute error in CTR as a function of the number of times an ad is viewed.

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

Summary

Contextual Advertising - summary

- One of the two textual advertising channels on the web
- Supports a large swath of the web eco system
- Challenging from both business and tech side
  - No clear intent
  - Lower ctr/conversion
  - Higher volume
  - Share revenue with publisher
- Two types of ad placement mechanisms:
  - Phrase extraction form the publisher pages
  - IR-style matching of the page content to the ads
  - Use of clic
- Industrial systems likely using a combination of technologies
- Space for improvement in today's state-of-the-art

Questions?

We welcome suggestions about all aspects of the course: msande239-aut0910-staff

Thank you!

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

http://research.yahoo.com