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 7:
Display Advertising Targeting

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 007 plan

- Targeting overview
  - Main tasks
- Traditional targeting
  - Demo, geo, behavioral, retargeting
  - A few words on privacy
- Current trends in targeting
  - Personas and look-a-likes
  - A few words on social targeting
- User profile generation – case studies (time permitting)
  - IR-style language models for user profiling (CIKM11)
  - Behavioral targeting using click data
  - Information-theoretic approach to profiling (not published yet)

Display ad targeting

- Place ads on pages viewed by the users
- The current event in display targeting is a page view
  - Versus the search query in Sponsored Search
- Browsing a lot less intentional than querying
  - "**High entropy**" of user intent
- CTR/CVR Rates in display are several orders of magnitude lower than in Sponsored Search
- Reduce the entropy by:
  - Using past user activity, especially activity with higher intentionality (search, transactions)
  - Comparing to "similar" users
- Asks for different techniques than in Sponsored Search

The tasks in display targeting

- Targeting can be decomposed into three (related) tasks:
- **User profile generation:** describe the user
- **Audience selection**: find the best audience for a given ad
- **Performance prediction**: find the best ad for a given impression (we have discussed this before)

User profile generation

- Understanding the user based on all available data:
  - Registration, online activity, offline activity (supermarkets, panels, etc.)
- Three basic data item types:
  - **Meta data**: name, zip, income bracket, profile interests
  - **Activities** (events) with a timestamp: purchases, searches, page views, clicks, …
  - **Connections:** friends and others
- Effective and efficient profile to support the targeting tasks
  - Effective – right granularity for best modeling performance
  - Efficient – from TBs of data accumulated daily, get a succint representation
- Profile generation**: user data → intents and interests**

Audience selection

- Given a campaign find the right audience
- Audience selection and user profile generation are related tasks
  - However distinct objectives
  - Can be done by different companies
- Challenges:
  - Define the objective
  - Training data
  - Modeling methods
- Audience selection in relation to ad selection

Ad Selection ← duality → Audience selection

ad = search (context, user) **user = search (context, ad)**

Traditional targeting techniques

Overview – Traditional targeting

- Demographic targeting
- Geo targeting
- Behavioral targeting
- Retargeting

Demographic Targeting

Using demographics in advertising

- Demographic targeting is widely used in traditional media
- A must for any display or textual ad network
- Important indicator of people's interest and potential of a conversion
  - Imagine you want to sell a $50K sports car. Who do you target?
- Used widely in traditional advertising:
  - TV, magazines, etc. maintain very detailed statistics of their audience
- Common classic dimensions:
  - Age
  - Gender
  - Income bracket
  - Location
  - Interests ("Golf enthusiast")
  - ….
- Each dimension has multiple values

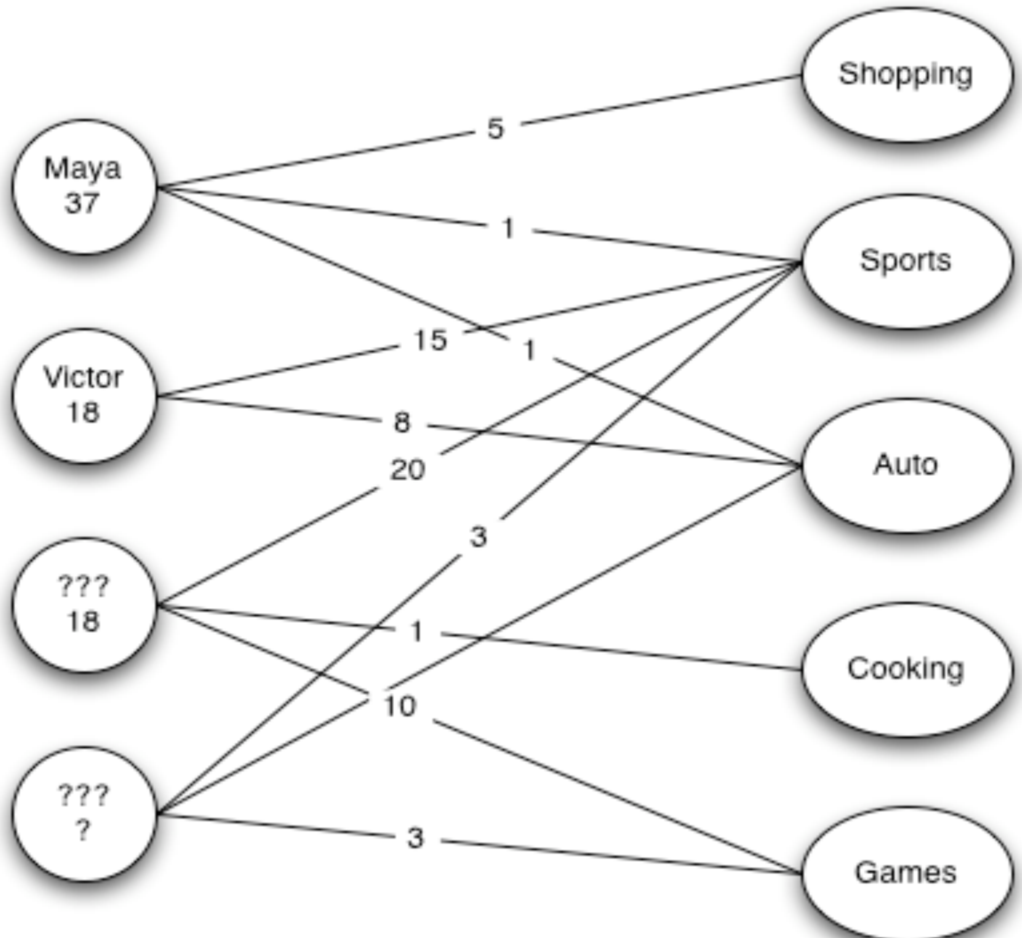
Key Challenge: Obtaining Demographic Information

- **User supplied demographic information**
  - Most reliable – if filled correctly
    - In some cases 15-20% of users born on 1st of January
  - Most users see very little incentive to fill the form
  - Privacy concerns
  - But credit card data, shipping address, etc, are almost 100% reliable.
- **Inferred demographic information**
  - Guess the demographics based on user browsing/querying behavior
    - 74% women/ 58% of men seek health or medical info online
    - 34% women/ 25% men seek religious info online
  - Wider reach – virtually every user

Inferring demographics

- How to infer the demographics from past behavior?
- **Classification**: e.g. regression model on the top of features extracted from history
- **Bipartite graph approach**:
  - Analyze the bipartite graph of users and their web pages/searches
  - Seed the graph with some demographic information
  - Infer demographics of users without the info
- **Combined approach**
  - "Demographic Prediction Based on User's Browsing Behavior" Hu et al, WWW 2007

Bipartite graph: Users and Web page visits



Bootstrap: get page attributes

- Each web page assigned tendency – probability distribution over the space of possible demographics attributes (gender & age)
- Training set construction:
  - Textual features extracted from the pages
  - Training labels based on the available user information:

$$\Pr(c \mid w_j) = \sum_{i=1}^{I} r_{ij} u_i(c) / \sum_{i=1}^{I} r_{ij}$$

  - Feature selection based on
    - Distribution grade on pages (Only discriminative pages kept)
    - Information gain for words on pages.
  - One binary classification SVM for each demographic dimension value
  - Normalize the outputs for each demographic dimension to translate to probabilities
- Smooth the page probability based on similar page by content
  - Reduced dimensional space (LSI – based on SVD)
  - The original feature space

Geo Targeting

Geo targeting

- Goal: determine user location
  - Home, work, …
  - Today: Often wrong ☹
- Inputs
  - Registration data
  - IP
  - GPS
  - Browser default language, search language, …
- Fast developing in the era of mobile
- Lots of papers/results, but no time to discuss …

Example of IP location – my stay in Glasgow

IP:	86.188.169.10
Decimal:	1455204618
Hostname:	86.188.169.10
ISP:	British Telecommunications
Organization:	Tidewell Solution Ltd
Services:	None detected
Type:	Broadband
Assignment:	Static IP
Blacklist:	Blacklist Check

Geolocation Information

Country:	United Kingdom
State/Region:	Windsor and Maidenhead
City:	Heston
Latitude:	51.4833

Behavioral Targeting (BT)

What is BT?

- A technique used by publishers and advertisers to increase campaign effectiveness based on a given user's historical behavior:
  - **Previous searches/search sessions**
  - **Previous browsing activity**
  - **Previous ad-clicks**
  - Previous conversions
  - Declared demographics data
  - Etc.

Main techniques

- Utility – everyone wins! (at least in theory ☺)
  - Advertisers: get a more appropriate/receptive audience, increased conversion rate, better ROI
  - Publishers: can ask for a premium
  - Users: see more interesting ads

Example categories: Exelate.com

[redacted]				
Finance - Small Business	30%	Read small business management content on relevant websites Search / interact with small business support documentation Interact with small business reference directories	Leading nationally branded small business finance publications Small business reference and listing directories	Interest
[redacted]				
Shoppers - Babies and Kids - Books - Fashion - Fashion - Jewelry & Watches - Health & Beauty - Home & Garden - Appliances - Home & Garden - Furniture - Music - Musical Instruments - Personal Tech - Personal Tech - Cameras - Personal Tech - Computers - Personal Tech - Mobile - Personal Tech - Software - Personal Tech - Video Games - Sports and Recreation	30%	In market shoppers looking for pricing, ratings and purchase information on specific products as well as enthusiasts commenting on and shopping for niche-interest items.	Comscore Top 10 shopping sites and vertical enthusiast sites that track product interest and drive specific purchasing behavior.	Intent
[redacted]				
Sports Enthusiasts	20%	Avid sports team followers visiting websites oriented around pro sports leagues and teams. Members and creaters of sports-based micro communities	Sports content websites Sports social networks	Interest + Registration
[redacted]				
Hispanic	20%	U.S. visitors to Hispanic / Spanish speaking websites; dual language sites; Hispanic news sites Registered visitors to Hispanic social networks	Branded U.S. hispanic websites Leading Hispanic social networks	Interest + Registration
[redacted]				
Singles / Daters	20%	Registered visitors to dating / social websites that have declared themselves as "unattached" in the last 30 days	Branded dating and social network sites	Interest + Registration
[redacted]				
Offer Seekers	20%	Users that have responded to online offers such as coupons, free trials, giveaways or newsletters	Registration Paths, discount sites, and lead gen sites	Interest
[redacted]				
Home Improvement	20%	Consumers seeking tips and advice on how best to upgrade their homes, inside and out	Leading home décor publications	Interest
[redacted]				

Current example: holiday season has started!

Interest

Event

- Gift-giving holidays
- Trips & getaways
- Travel enthusiasts
- Parenting
- Tech Enthusiasts
  - Mobile
    - *Droid, Samsung, HTC, LG, Nokia, Motorola*

Winter Seasonal

The holidays are the busiest season of the year, so make sure you have all your bases covered! Get ready for an audience that's looking for anything and everything - from ailment relief to New Year's Resolutions!

- Cold & Flu
- Winter Activity Enthusiasts
- Holiday Shoppers
- Holiday Dessert Baking
- Holiday Entertainers
- New Year's Resolutions

Intent

Shopping

Reach our audience of over 90M unique users!

- Holiday Presents
- Toys
- Babies and Kids
- Books
- CPG*
- Mobile
- Movies
- Music
- Personal Technology*
- Sports and Recreation
- Fashion
- Accessories
- Apparel
- Jewelry and Watches
- Men
- Shoes – *(*By Type)*
- Sneakers – *(*By Brand)*
- Women

Travel

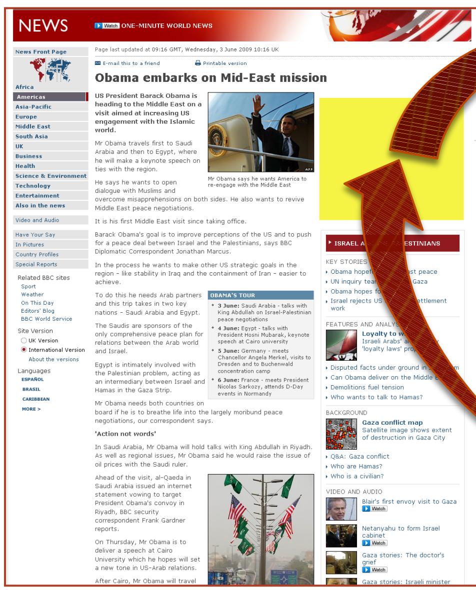
Reach our audience of 50M unique users!

- Accommodations
- Vacation Packages
- Cruises
- Car Rentals
- Flights

Repeat: Basic search retargeting scheme

- User searches for shoes on the XYZ engine

Site ABC sends ad request + XYZ cookie to XYZ



XYZ creates shoes ad based on XYZ cookie that remembers “shoes”

How does retargeting compare to other targeting

- Much more detailed information
  - The product you searched
  - The query you issued
  - Items put in the cart
- Retargeting is a key technique today, and more companies are coming around with ideas to expand on it

Is this delightful or creepy?

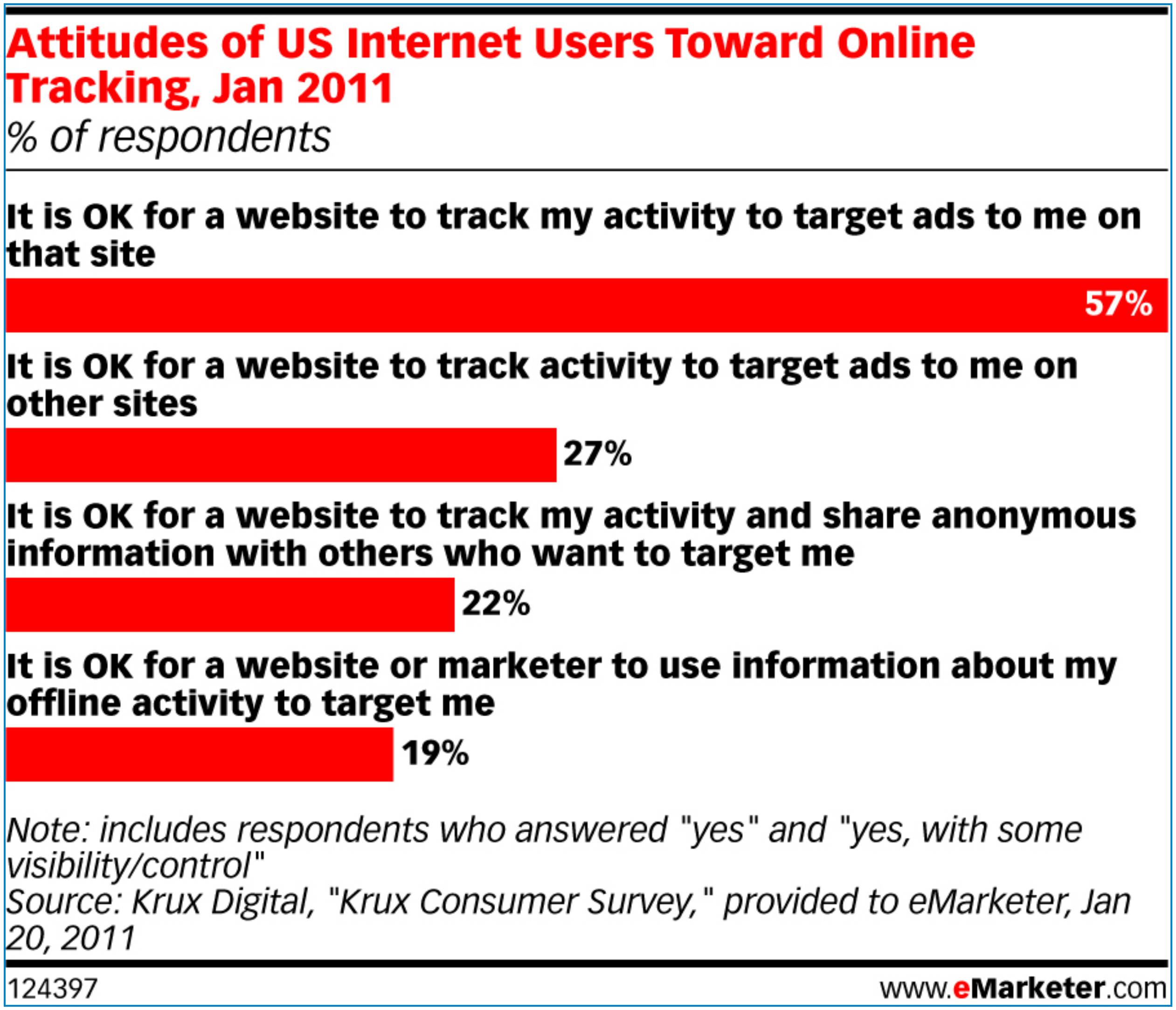
- Waiter:
  - Mr. Broder, welcome back! You always like to sit at the corner table -- we kept it for you!

- Waiter:
  - Mr. Broder, welcome back! You ate 11 times at one of our restaurants, never order cocktails, never order lobster. You usually spend about $75 on 2 appetizers, one split main, & one split dessert + $30 on wine, and tip twice the tax. Here is a special menu where we removed all items you never order or outside your price range …

Detour: Does privacy still matter?

- Privacy is not dead
- As of now, the impact of data leaks online is relatively small
  - Probably smaller than in the real world
  - This could change through technology changes, hacking etc.
  - This will change the public's view on privacy and induce legislation
- Algorithmic privacy is almost unattainable
  - Very little usable in practice
- Users must be motivated to provide their data (companies already are motivated to collect data)
  - Clear value for disclosure
  - Self promotion (facebook)
  - New experiences (apps, ads, etc.)
- If things go wrong here, the net effect would be setting back the development of deep personalization for decade(s)

Users perceptions



How to address privacy concerns

- Users do not understand the cookie mechanisms
- Difficult to turn off – many sites stop being functional without cookies
- If you accept cookies from XYZ, XYZ can become aware of your visits to any site where XYZ has a visible or invisible presence on the page.
- Many proposals / regulations / “trust-me” solutions
  - E.g. Phorm (http://www.phorm.com/) promises to collect only category data and keep cookies anonymous (not linked to IP, name, etc)
  - Most companies have data retention policies (90 days for Yahoo!)
  - Most companies allow user control over stored data
  - Opt-out BT
  - Etc

Recent:
Network advertizing initiative

- Signed on by all major players + trade groups incl Google, Yahoo!, etc
- Icon on all ads:
- Users click on it and get general "opt out" page

Opt out status

Opt-Out Status

Select all | Clear | Submit

Member Company	Status	Opt-Out
aCerno More Information	**Active Cookie** You have not opted out and you have an active cookie from this network.	Opt-Out ☐
AdBrite More Information	**Active Cookie** You have not opted out and you have an active cookie from this network.	Opt-Out ☐
AdChemy More Information	**No Cookie** You have not opted out and you have no cookie from this network.	Opt-Out ☐
Adconion More Information	**Active Cookie** You have not opted out and you have an active cookie from this network.	Opt-Out ☐
Adara Media More Information	**No Cookie** You have not opted out and you have no cookie from this network.	Opt-Out ☐
Adify Media More Information	**Active Cookie** You have not opted out and you have	Opt-Out ☐

Social targeting: the power of the graph

Based on slides from Kun et al CIKM11

Social targeting

- (The) Social network(s) becoming the prime destination for display advertising
- A mixture of display advertising and textual advertising
- Proprietary formats
- Two types of user information
  - Profiles attributes
  - Graph
- Where does the key value come from?
- Seems like most of the advertising is targeted by explicit selection of 'likes' and other user attributes
  - Very precise demographics: e.g. get Chevy owners
- Sophisticated method vs. easy, understandable and data rich rule language

The value of the graph

- The social graph can be used to smooth the data available about the user
  - If you are friend with many people that mountain bike, you are likely to do so as well
- How much added value can we get from the graph?
- Challenges:
  - Dilute the profile information
  - Differentiation of different connections
  - How far you need to go to get all the benefit

Study: the mail graph

- Friends are similar along a variety of dimensions is a long-observed empirical regularity (*homophily*), but…
  - Whether and how can we leverage friends' activities for behavioral targeting?
  - Whether forecasts derived from the social graph is more accurate than standard behavioral targeting models?
- Investigated a wide array of supervised and unsupervised machine-learning approaches to utilize social data for BT models.
- Evaluated the predictive power of social data across 60 consumer domains on a network of over 180 million users in a period of 2.5 months.
  - the most comprehensive study of social data for BT

Homophily – BT qualifications

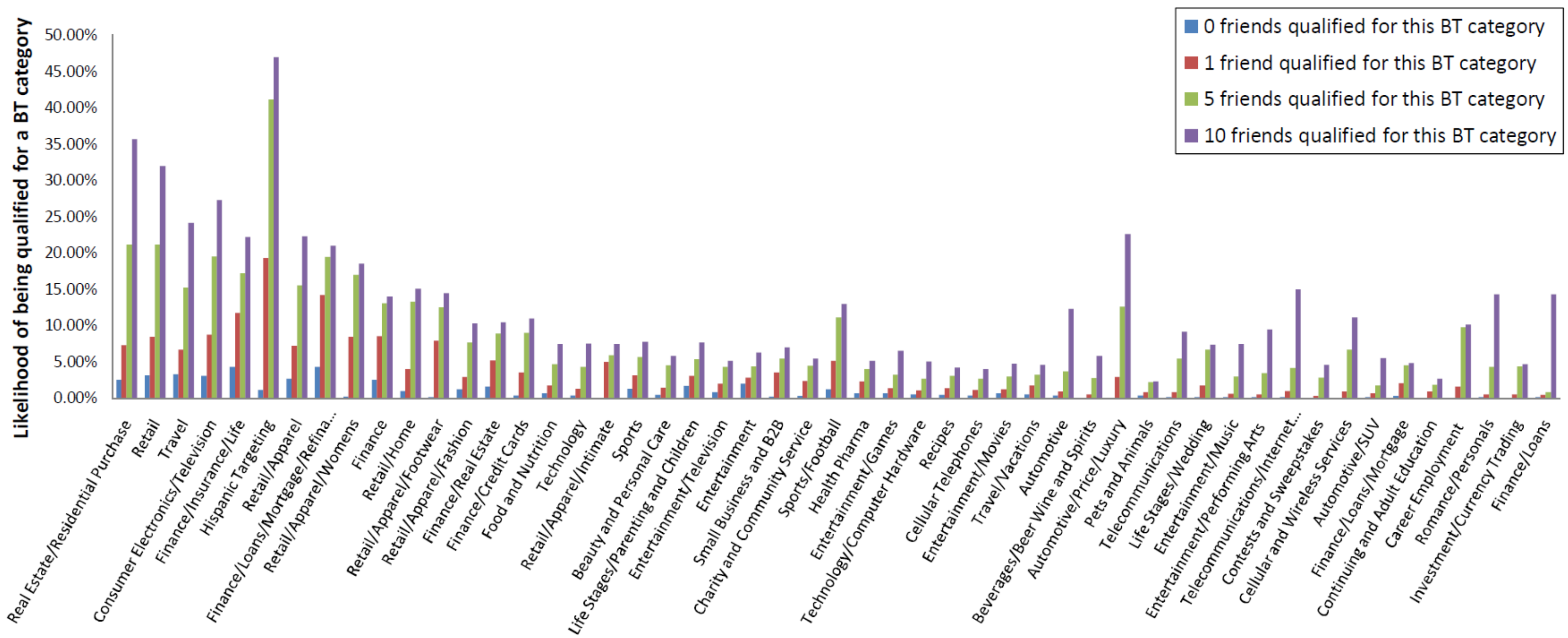


Fig. 1 Likelihood of being qualified for a BT category as a function of having social contacts who are also qualified for the same category.

Homophily – ad clicks

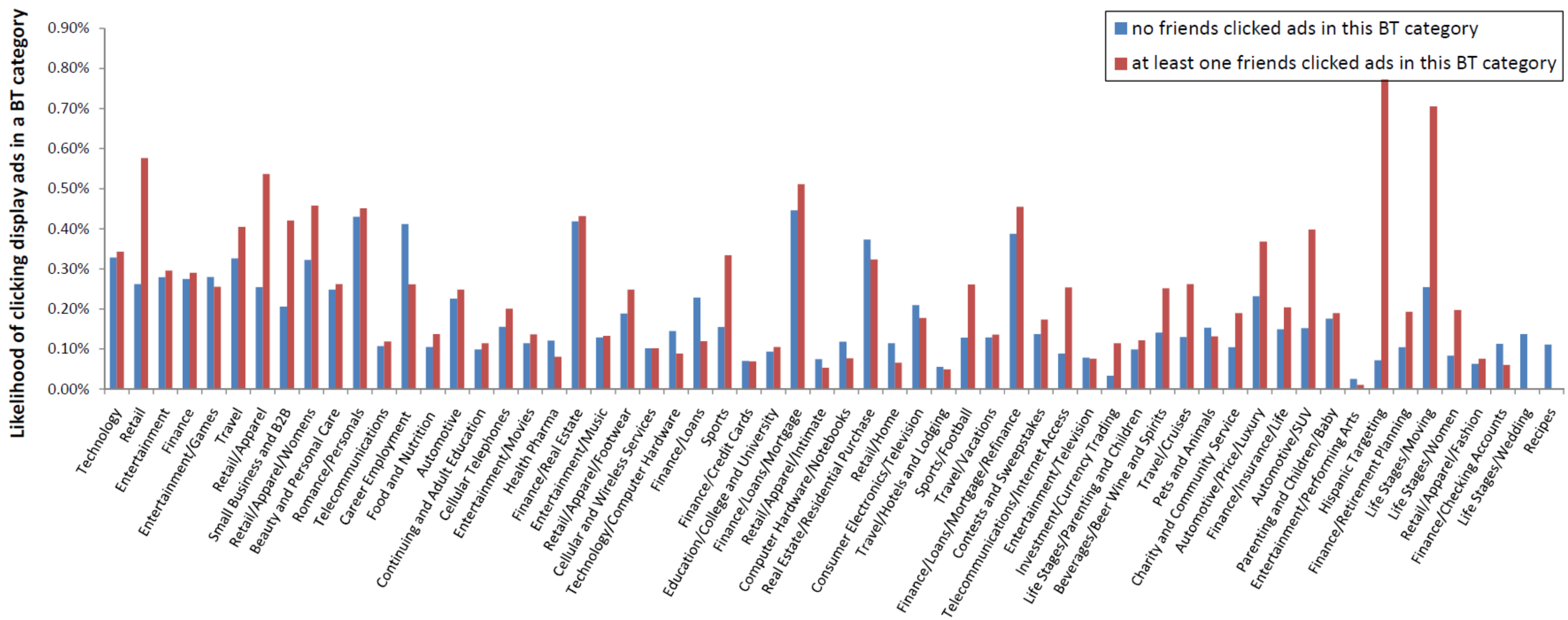


Fig. 2 Likelihood of clicking on display advertisements in a BT category as a function of having social contacts who have done the same.

Leveraging social data for BT

- Classification: append social features to standard BT features, then train models (supervised approach)

- Ensemble: combine BT model and social model (supervised approach)

- Network propagation: use propagation to infer BT scores from one's neighbors (mostly unsupervised approach)

Experiments

Features	Performance Improvement
BT baseline	0.00%
Random targeting	**-24.25%**
Neighborhood feature I	-16.58%
Neighborhood feature II	-13.34%
Community feature	-15.84%
BT + Neighborhood feature I	0.97%
BT + Neighborhood feature II	0.86%
BT + Community feature	0.08%
Ensemble	0.001%

Table 1. Performance improvement over BT baseline, measured by the lift of view-weighted average AUC across all 60 BT categories on all users.

Category-level improvement

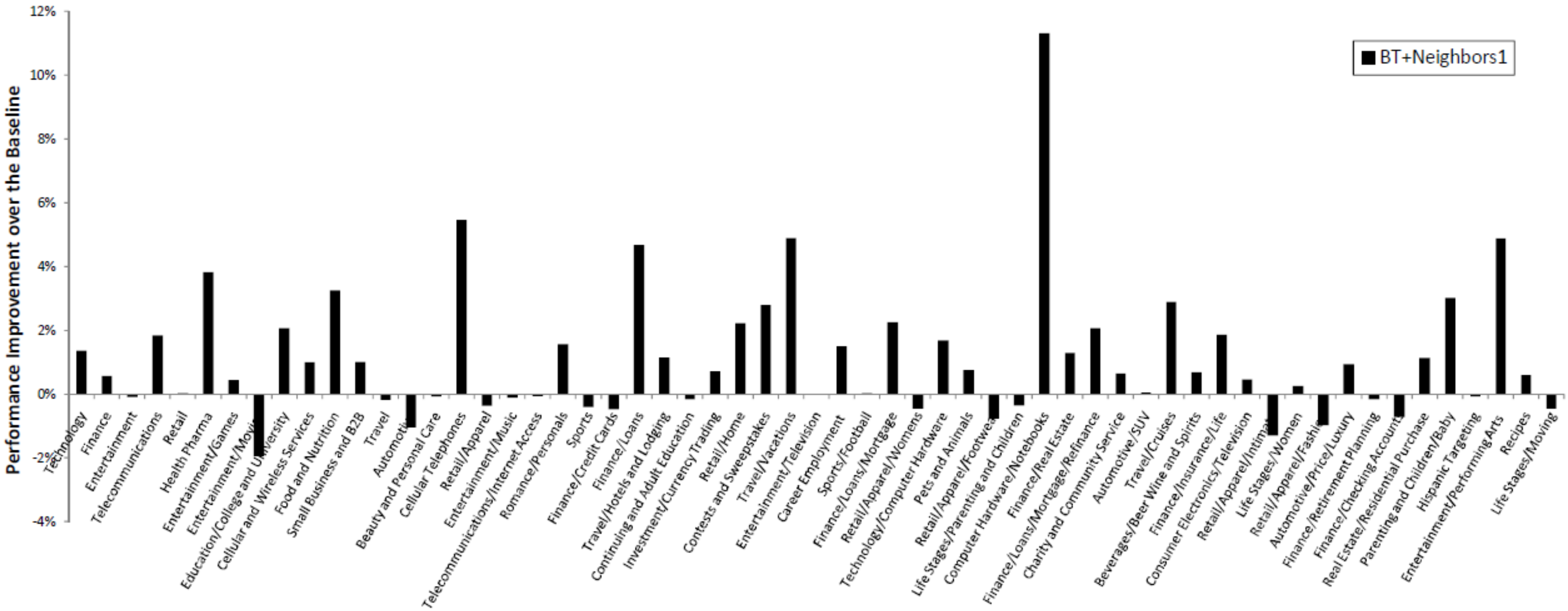


Fig. 5 Category-level performance improvement, measured by the lift of AUC, i.e. (AUC_c / $AUC_{baseline}$ -1) x 100%.

What we have learned?

- Social features do carry certain informative signals.

- BT baseline model substantially outperforms all other models trained from social features only; indicating that individual's own behavioral is much more useful than their friends.
- Note this study uses the email graph
  - Implicit feedback from the user
- In social networks the graph induces different behavior
  - Explicit and implicit feedback
- Still, the value of the graph per se is not clear
  - Find out in your projects!

Current trends in targeting

The world from the standpoint of the advertiser

Market needs: Advertisers want to target ‘personas’

Targeting Personas: Advertisers want to target a specific persona – that may not be available through a standard Demo or a publisher defined category

For cosmetics company, XYZ, the segments “**Women between 35-54**” or “**Interest Beauty-Cosmetics**” are not narrow enough as they **don’t capture the unique beauty needs of women with different persona** in the same age group

“Frazzled Mom”

≠

“High Flying career woman”

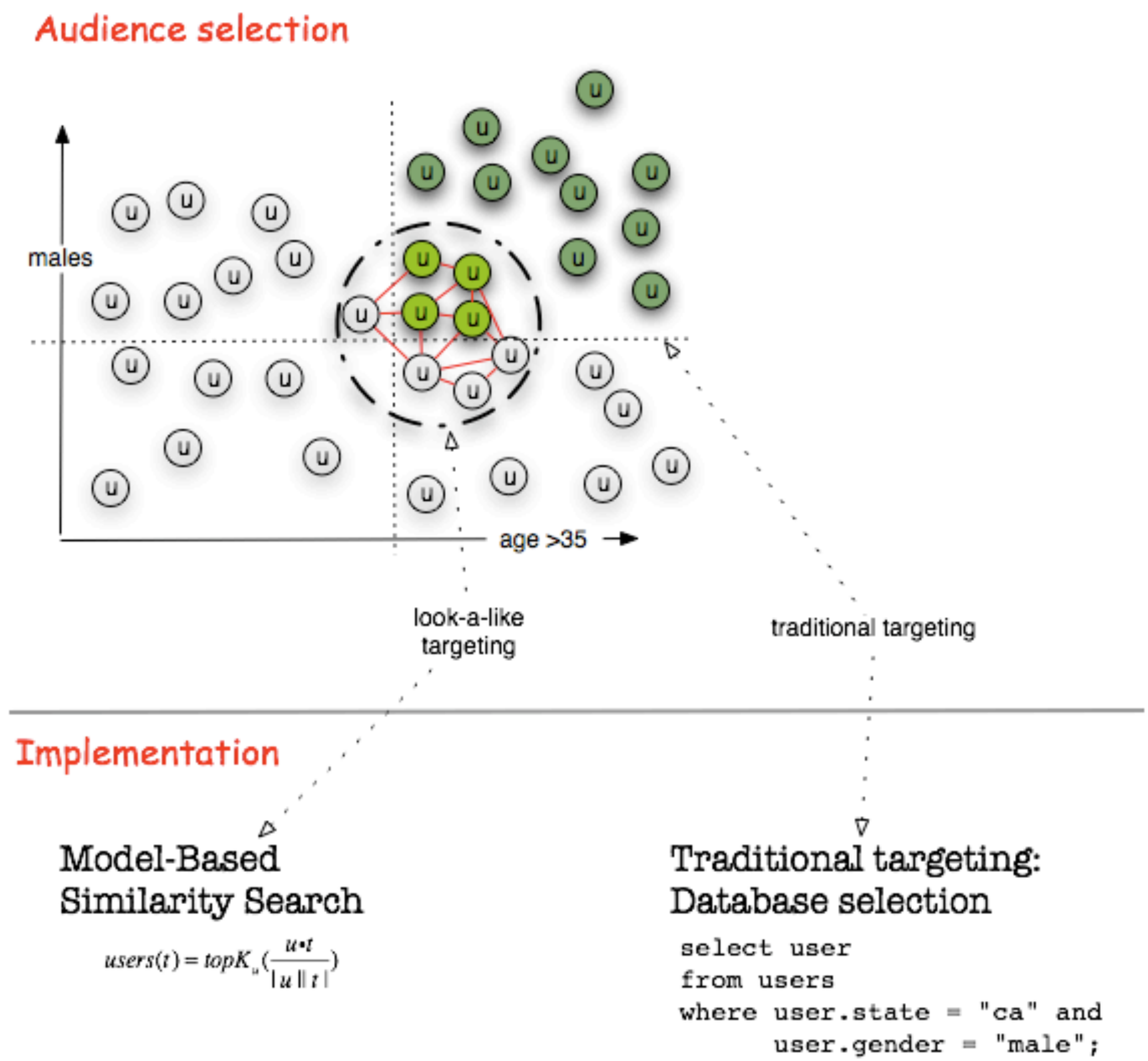
≠

“School Teacher”

More on personas

- Online users perform a sequence of (overlapping) tasks
- Personas share interests and behavior
- Usually pre-defined segments based on real world behavior
- Same user can have multiple personas: **personas are facets of personality**
  - Vanja is a computer geek
  - Vanja is also
    - A father
    - A rollerblader
    - An electronic music listener
    - A skier
    - Etc
- NB: "is" here means "behaves like"
- NB: Personas & interest may vary over time

Advanced audience selection: look-a-like modeling



User profile generation: interests and intents

Working with high entropy of intent

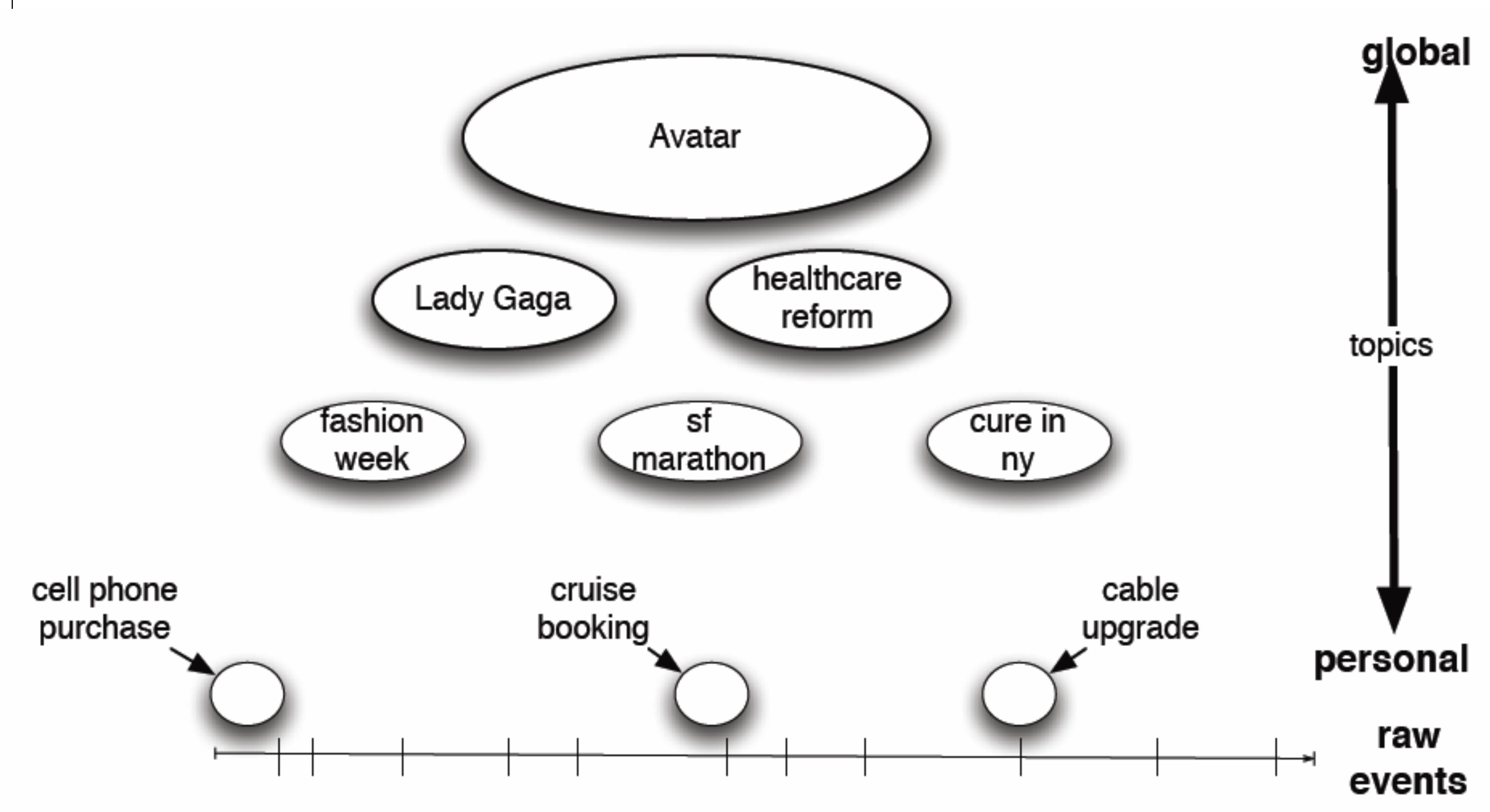
Display ad targeting vs. Sponsored Search ad selection

- The current event in display targeting is a page view
  - Versus the search query in Sponsored Search
- Browsing a lot less intentional
  - "**High entropy**" of user intent
  - More of "**interest**" than intent
- CTR/CVR Rates in display are several orders of magnitude lower than in Sponsored Search
- Reduce the entropy by:
  - Using past user activity, especially activity with higher intentionality (search, transactions)
  - Comparing to "similar" users
- Asks for different techniques than in Sponsored Search
- **Profiling:** Induce the interest and intent of the user

Dichotomy of user interests and intents

- Personal vs. global
- Short term (current) vs. long term interest
- General vs. specific (sports vs. Inter Milan)
- Commercial vs. non-commercial
- Distinguishing between these would allow for more effective advertising
  - Commercial interest probably better aligned with interaction with advertising

Global vs. personal interests



Short term vs. long term interest

- Online tasks have limited time span
- What is of interest now, might not be in a few minutes/ hours/days:
  - Buy a cruise, don't need another one for a while
- Other interest persist
  - Skiing, Inter Milan
- Which are better suited for advertising?

General vs. specific interests

- Can we decide on the generality vs. specificity of an interest of a user?
- Specific interest can be targeted much better
  - Require adequate ad supply
- General interest better coverage in the ad supply
- What is the relationship between general-specific and short-long term interests
  - Short term interest more specific (Lady Gaga) [Kim et al IUI2003]
  - Long term interest more general (pop music)
- Help differentiate short term interests by first understanding the long term interests and use those as prior in the interpretation of recent events

Generative language models for user profiles

Tyler et al CIKM2011

Audience selection task: intuition

- **Given**: seed set of users that have performed a desired action (e.g. bought a product) for a campaign
- **Find**: more users that are likely to do so.
- Finding such users can be formulated as a **retrieval task**
  - Top-k similarity search over the user space
  - Query composed from the seed set

Problem definition

- Set of users U = { u_i }
- Subset of users are **converters** $U = U_c \bigcup U_{nc}; U_c \bigcap U_{nc} = \varnothing$
- Each user u_i has associated a sequence of events <e_j, timestamp_j>
  - Search terms, page views, ad clicks, etc.
- Learn to distinguish converters from non-converters
  - Using events before the conversion
- Non-converters are not really negative examples!
  - Many reasons why users would not click on an ad
  - Suitable task for unsupervised approach (phase 1 of retrieval)
- Transform the user into set of feature-weight pairs
  - u_i = { <feature_j, weight_j>}
  - Features extracted from events

Generative Models

- Generate an item (i.e. document, search query, user activity) based on some rules from a predefined distribution.
- Example: multinomial distribution:
  - (tick:0.5; tock 0.25; tack 0.25)
  - tick, tock, tock, tick, tick, tack, tick, tick, tack
- How to find the distribution corpus C– **the language model** θ^{C}
  - $p(w) = n_w/n$ – maximum likelihood estimate
  - n = sum of lengths of all documents in the corpus
  - n_w = number of occurrences of the word **w**
  - Can produce a model for the whole corpus as well as for a single document
- Issues
  - Documents, activity are not generated by drawing from a distribution – a much more complex process
  - Independence of the words – many ways to improve here

Retrieval models for audience selection

- In IR: model of the queries and documents, and their similarity
- Successfully applied in IR
  - Vector space
  - Language models
- Can we use similar tools to formalize the audience selection task in display advertising targeting?
  - Users instead of documents
  - Query derived from the seed set

Language Model for Audience Selection

- Generate the user from some underlying distribution
- User is a sequence of events $u = \{e_1, e_2 \ldots e_n\}$
  - event is a triplet of <type, time interval, content>,

$$e_i =< type_i, int_i, c_i >$$

- Under a multinomial model over the space of events and the independence assumption

$$\begin{aligned} p(u) &= p(e_1, e_2 \ldots e_n) \\ &= p(e_1)p(e_2|e_1)p(e_3|e_2e_1)\ldots p(e_n|e_1e_2\ldots e_{n-1}) \\ &\sim \prod_{i\in 1\ldots n} p(e_i) \end{aligned}$$

Generating an event

$$p(u) \sim \prod_{i \in 1 \ldots n} p(e_i) \sim \prod_{i \in 1 \ldots n} \{p(int_i) \cdot p(type_i|int_i) \cdot p(c_i|int_i, type_i)\}$$

$p(int_i)$	prob of observing an event in a given time interval; assume it to be proportional to the length of the interval
$p(type_i \mid int_i)$	prob of observing in that interval an event of the given type; assume that the event mix is independent of the interval
$p(c_i \mid type_i, int_i)$	prob of observing a specific event content given the interval and the event type

$$p(c_i|int_i, type_i) = \prod_{w \in c_i} p(w|int_i, type_i)$$

Seed set representation

- What would be a good representation of the seed set?
  - A language model of the seed set - probability distribution
  - Close to the seed set
  - Far from the background
- To formalize we can use KL divergence D:

$$D(p\|q) = \sum_t p(t) \log \frac{p(t)}{q(t)}$$

- Measure of divergence between two distributions:
- Not symmetrical
- Delta between the cross entropy and entropy of P:
  - D = H(P|Q) – H(P)
- The more similar the distributions are, the less is the difference in the encoding lengths for each item

How to compose the query

- Formalize

$$\theta^q = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} D(\theta \mid \theta^i) - \mu D(\theta \mid \theta^C)$$

- Solution exists:

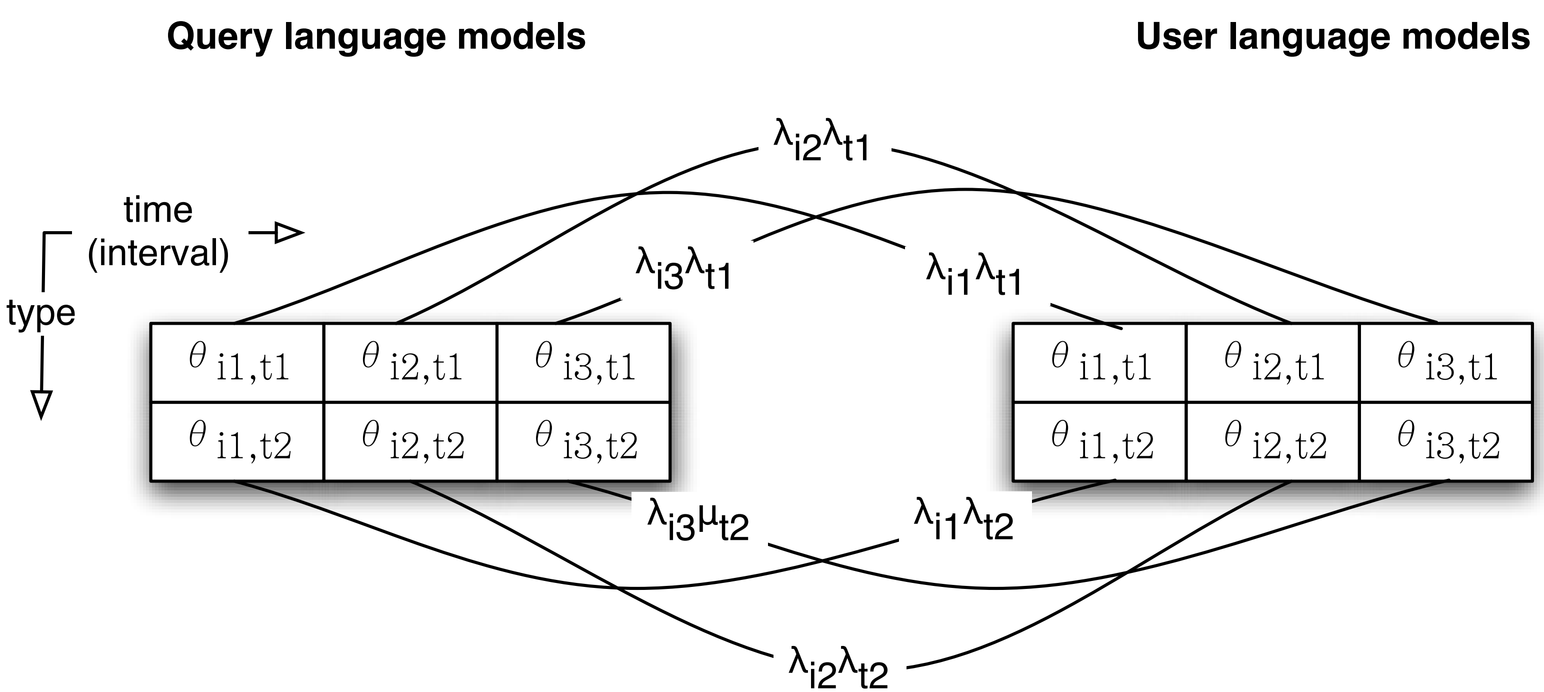
$$p(w|\theta^q) = \exp\left(\frac{1}{1-\mu} \frac{1}{n} \sum_{i=1}^{n} log\ p(w|\theta^i) - \frac{1}{1-\mu} log\ p(w|C)\right)$$

- Compare the query and the user: use KL Divergence again

Temporal aspects of the user history

- The time when the user has done a particular action correlated to the action
  - Search for cruises 1-2 weeks before departure
- So far, there is not time aspect in what we did – how to introduce time?
  - Ignore time
  - Sequence of events – each event in its own time interval (as in Markov processes)
  - Something in between - intervals

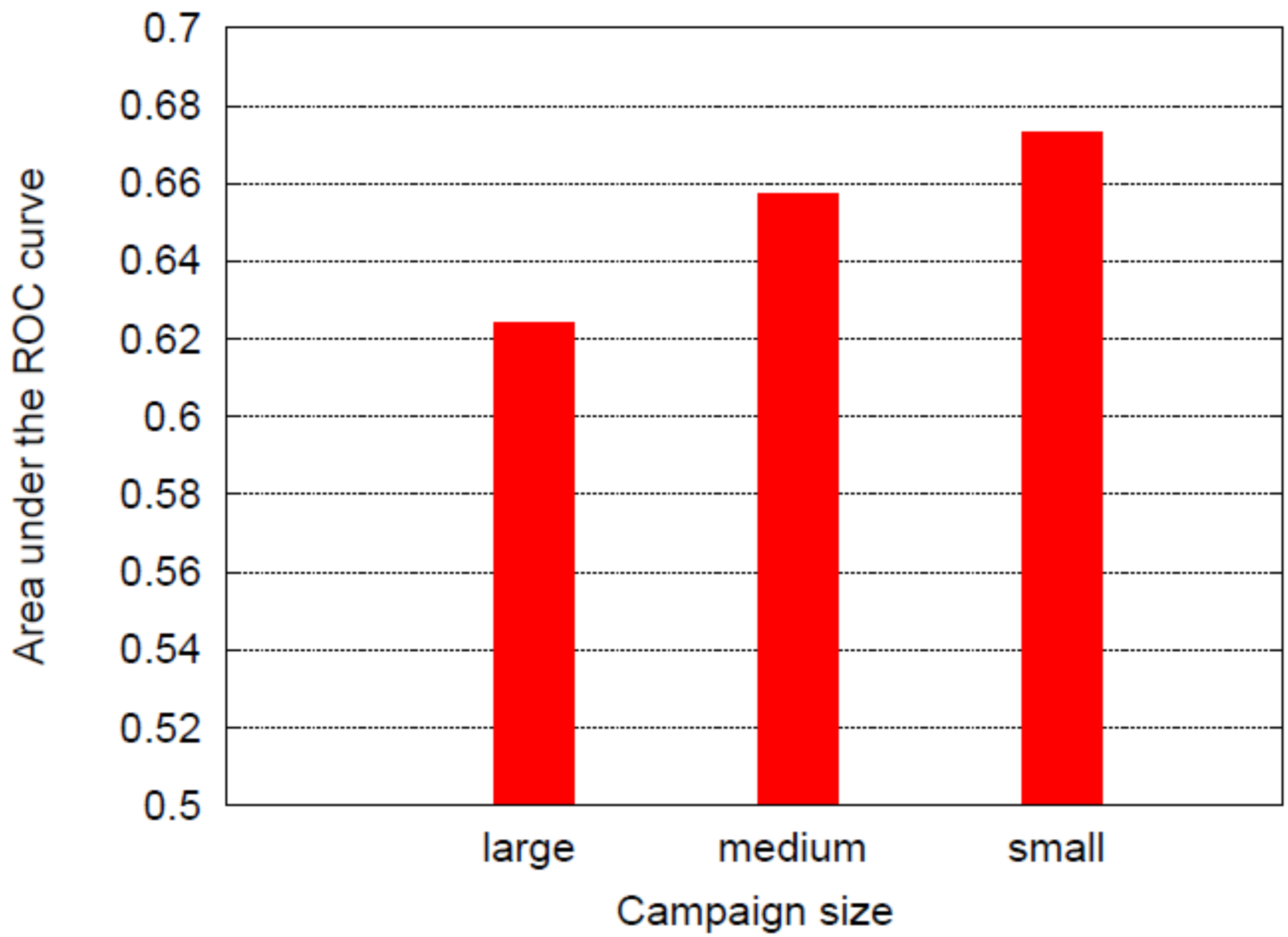
Comparing different intervals



Experiments

- Dataset: 34 different ad campaigns from Yahoo advertising network
- Seed set: users who converted on these campaigns between 02/04/2010 and 02/18/2010
- Test set: ad impressions between 02/19/2010 and 02/24/2010
- Each user profile has 4 weeks of her online activity prior to the ad impression

Performance on Campaigns of Different Sizes



Retrieval models for audience selection

- Provide solid formalization for audience selection
- Can use multiple models
  - Language modeling
  - Vector space model
- Time intervals to capture the temporal aspect at good granularity
- Compare the seed set activity within the same window of the target users
- Users with similar activity at the similar interval with the seed set are ranked higher
- What if the user is similar, but only in certain aspect?
  - Look at resolving this in one of the subsequent studies
- Time shift in key events before conversions
  - Smoothing between the intervals

Learning BT categories from click data

Y. Chen, D. Pavlov and J. Canny: Large-Scale Behavioral Targeting, KDD 2009

Problem definition

- Track activity for each user
- Put activity into one or more of the classes
- For each user **i**, class **c**:
    - y_i=count(i,c): the count of events in that class
- Represent each user with a set of features
    - Bag-of-words from page titles and query unigrams
- Predict the click, view counts for each user for each class, using these features
- One class at the time

Model

- Poisson distribution natural fit for modeling counts of rare events
- Fixing the class, for each user i:
- Lambda is the mean of the distribution
- Estimate lambda from the features in each instance:

$$p(y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

$$\lambda = W^T x_i$$

  - W – vector of weights for the features
  - Linear estimation. Alternative exponential relationship:

$$\lambda = e^{W^T x_i}$$

  - Weights smaller, sparser;
  - Does 3 times more views of a page with the word "baseball" indicate 3 times vs. 10 times higher propensity for sports ads?

Parameter optimization

- Maximize the log likelihood of the data

$$l = \sum_i l_i = \sum_i y_i log(\lambda_i) - \lambda_i - log(y_i!)$$

- Derivative per feature: $$\frac{\partial l}{\partial w_j} = \sum_i (\frac{y_i}{\lambda_i} x_{ij} - x_{ij})$$
- Assume non negative weights:
- Tune weights so that the count estimates (λ_i) equal actual counts (y_i)
- Iterate until convergence $$w'_j \leftarrow w_j \frac{\sum_i \frac{y_i}{\lambda_i} x_{ij}}{\sum_i x_{ij}}$$

Estimation of actual CTR

- Build a model to predict
  - Clicks
  - Views
  - Smooth the parameters α and β for new users without enough data. α / β = average CTR

$$CTR_{ci} = \frac{w_{ci}^{clicks} x_{ij} + \alpha}{w_{ci}^{views} x_{ij} + \beta}$$

- Online recalculation:
  - Produce a user profile in batch daily (see paper for description of scalable methods for calculating the profiles)
  - Incrementally update the user profiles:
  - δ - decay
  - Δ t – the time interval
  - Δ x_j - features from the new events

$$\lambda'_{ci} = \lambda_{ci}\delta^{\Delta t} + w_j \Delta x_j$$

Information-Theoretic User Profile Generation

Xiaoxiao Shi, Kevin Chang, Vijay K. Narayanan, Vanja Josifovski, Alex Smola

Compression of user profiles

- Which user activity characterizes the user?
- What activity of this user is different than the activities of other users?
- Compare the user activity with the general population, other groups of users
- Reconstruct the user from the general population with least amount of data
  - Similar to Wavelet denoising

Concepts

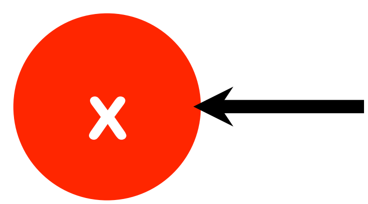
- Bag of tokens (unordered set of activities)
  - "page views about banking"
  - "queries about cars"
- Represent user as distribution over tokens
  - Proportional to number of activities associated with token.
  - Background distribution is uniform over users
- Kullback Leibler divergence weighting (Konopnicki et al., 2010)
  - KL divergence between user distribution and background (#bits to encode u)
  - Term weight proportional to KL contribution

$$D(p||q) = \sum_t p(t) \log \frac{p(t)}{q(t)}$$

term weight

Step one

- Basic Idea
  - Use a hierarchical generative model for users
  - Encode objects by most meaningful subset of tokens relative to background (upper level)
- One Stage Encoding: Konopnicki et al. 2010
  - No model for generating user distribution X.



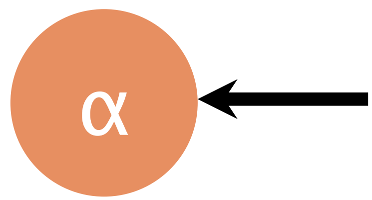
  - Represent X by tokens where X diverges significantly from background, B.
    - Based on D(X | | B).
  - If X is very similar to background B, then keep few terms.
    - Lossy compression.

Two Stage Compression

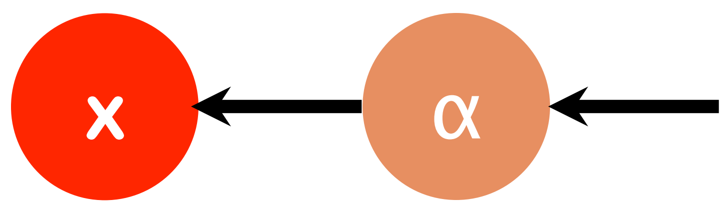
- Our approach: Generative Clustering
  - User distribution X drawn according to cluster with parameter alpha.

$$p(X) = \int p(X \mid \alpha) d\alpha$$

  - Cluster represented as a distribution over terms, alpha



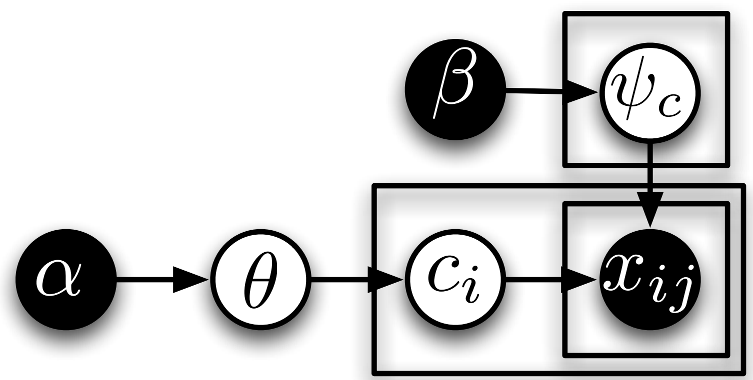
  - Encode cluster in terms of most divergent tokens, relative to background.



  - Encode user as most divergent tokens, relative to cluster.

Dirichlet Process Clustering

- Flexibility and Simplicity
  - No need to specify #clusters --- data speaks by itself
  - Easy to implement with Chinese Restaurant Process
- Nice Statistical properties
  - Nonparametric
  - Infinite mixture of models
  - Parameters: #iterations, background smooth parameters, initial #clusters, maximum #clus

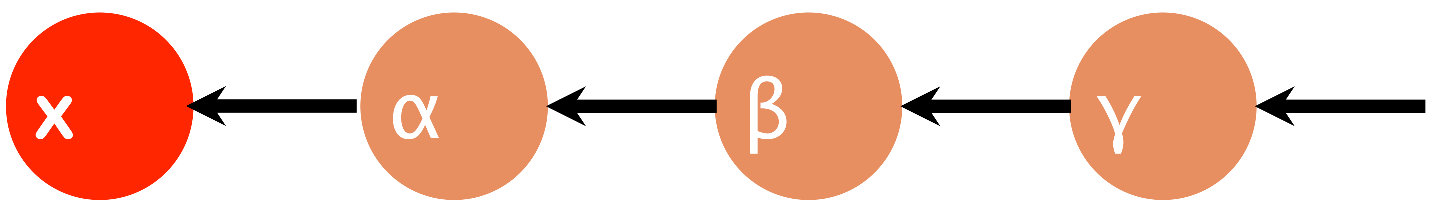


General Framework

- Hierarchical Generative Clustering
  - User distribution X drawn according to a hierarchy of clusters.

$$p(x) = \int p(x|\alpha)p(\alpha|\beta)p(\beta|\gamma)\ldots d\alpha d\beta d\gamma$$

- Stage-wise compression for key terms



  - Encode user distribution X in terms of cluster parameter α.
  - Encode α in terms of cluster parameter β .
  - …

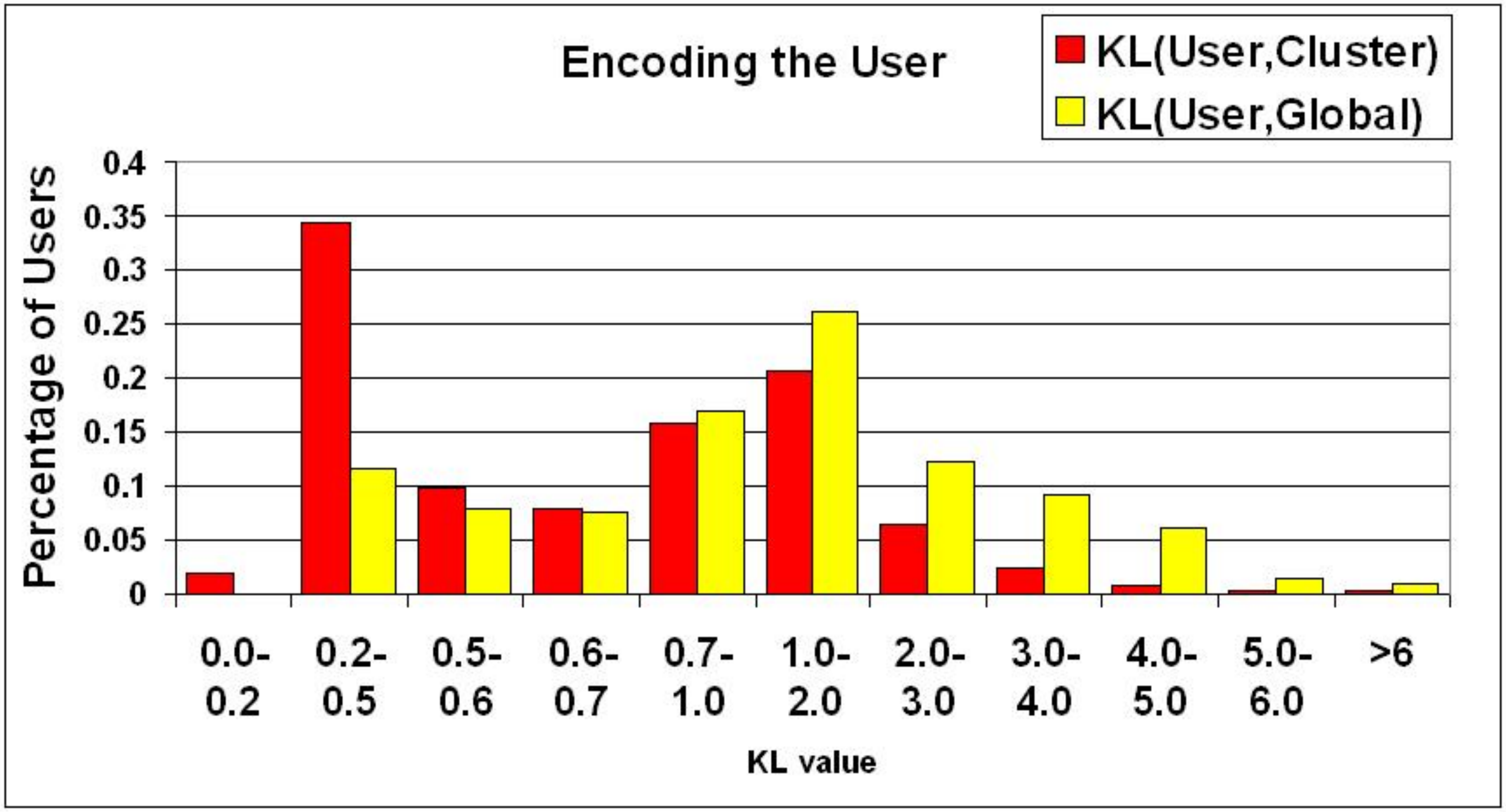
This approach

- Meaningful features
  - Cluster ID is often meaningless
  - Activity / tokens often user interpretable
  - Retain user interpretable representation
  - ‘tokens are the best features’
- Flexibility
  - Extend this to general hierarchical models
  - Extend this in a streaming environment with evolutional clustering
- Data reduction
  - We keep only the most divergent features.
  - Number retained is tunable. Two different strategies:
    - Top k most divergent features
    - Features with divergence term > fixed threshold.

Experiments

- Constructed user profiles
  - 30 million anonymous users at Yahoo!
  - 55 day period of activity
  - Captured page views, display ad clicks, views, searches.
    - Tagged with relevant categories. Use categories only for privacy
  - Example tokens
    - cpv_Finance (page view on finance page)
    - adc_Telecomm (display ad click on telecomm ad)
    - sch_Home Improvement (search for home improvement related query).
  - Constructed 586 clusters, using Dirichlet Process Clustering.
- Applied user profiles to predictive modeling for advertising.

The effect of clustering



Meaningful user profiles

Cluster specific (Bold red)

Personalized

Raw Data	cpv_Miscellaneous/News:0.16827 cpv_Miscellaneous:0.158012 **cpv_Sports:0.147968** **cpv_Miscellaneous/News/Sports:0.141748** cpv_Miscellaneous/News/Science:0.058412 **cpv_Sports/Golf:0.058054** ……
KL(Cluster; Global)	cpv_Sports:0.331078 cpv_Sports/Basketball:0.196653 cpv_Miscellaneous/News/Sports:0.166676 cpv_Sports/Football:0.057338 ……
KL(User; Cluster)	cpv_Miscellaneous/News/Science:0.180663 cpv_Miscellaneous/Science:0.17687 cpv_Miscellaneous/News/Business and Finance:0.126494 ……

- Features ordered by term-level divergence from background.
- Represent cluster well
  - sparse encoding

Meaningful user profiles

Raw Data	cpv_Automotive:0.079874 cpv_Automotive/Used:0.076244 cpv_Automotive/US/Chrysler/Jeep:0.07500 cpv_Automotive/US/Chrysler:0.067948 cpv_Miscellaneous:0.067087 cpv_Miscellaneous/News/World:0.013403 cpv_Miscellaneous/News/Science:0.0123 ……
KL(Cluster; Global)	cpv_Automotive:0.136555 cpv_Automotive/Used:0.124304 cpv_Automotive/Used/Certified Pre Owned:0.097341 cpv_Automotive/Non US:0.075489 ……
KL(User; Cluster)	cpv_Miscellaneous:0.083574 cpv_Miscellaneous/News:0.055034 cpv_Miscellaneous/News/World:0.050208 ……

Cluster Specific

Personalized

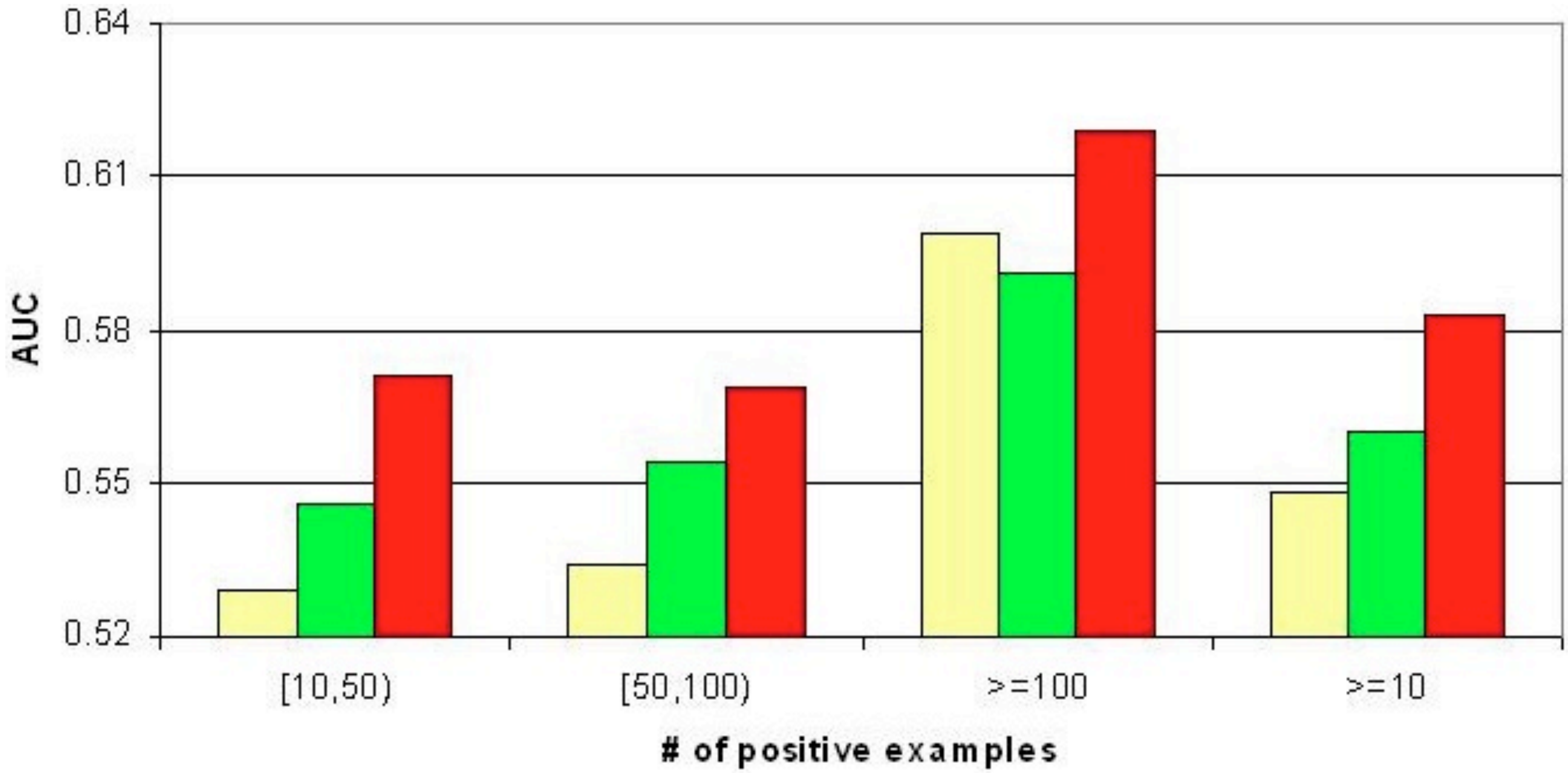
Conversion and click prediction

- Modeling task
  - Campaign conversion prediction – approximately 200 NGD campaigns
  - Anonymous users, features also anonymized
  - Train SVM to predict whether user will convert on specific campaign.
    - Use different user profiles as input features.
- Metrics: how well we separate interested users from others
  - Compute area under ROC curve for each campaign
  - Evaluate modeling methodology by average AUC of all ads.

Experiments

- Comparison of three different feature sets
  - User Profile: Features from compression framework representation.
    - Top most divergent activity tokens in D(user || cluster).
    - Top most divergent activity tokens in D(cluster || global).
    - Feature weights given by Divergence term for token.
  - Baseline I: Raw user activities.
  - Baseline || : Use only Konopnicki framework.
    - Top most divergent tokens in D(user || global).

Effect of training set size on conversion prediction



- **Improvements across the spectrum**

Summary: user profile compression

- Each user can have many different events
- Not all activity is meaningful for prediction of future activity
- Find the distinguishing user features
- Compare the user with its neighborhood
- Define neighborhood hierarchically:
  - Largest: the whole user population
  - Closest: few closest users
  - Everything in-between
- Improvements for users with noisy raw data
  - Smoothing from the similar users
- Improvements across the spectrum of campaigns sizes

Summary

Key points

- Targeting is a key step in differentiation of impressions and extracting value!
- Traditional targeting: demo, geo, BT
  - How to get the data from the user?
  - Infer the data from historical activity
- One of the key step in targeting is user profile generation
  - Generative models to assign probability of a sequence of events
  - Weighting based on time, event type and content
  - Predict the counts of events in certain categories
  - Clustering and other unsupervised techniques useful – more to come in the next lecture

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