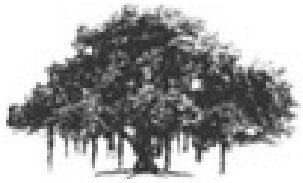


IIIT-H

Web Mining
Lecture 25: User Understanding by Log Mining

Manish Gupta

13th Nov 2013

Slides borrowed (and modified) from

Daxin Jiang, Jian Pei, Hang Li. Web Search/Browse Log Mining: Challenges, Methods, and Applications. Tutorial at WWW 2010

Rosie Jones. Privacy in Web Search Query Logs . Invited Talk at ECML PKDD 2009

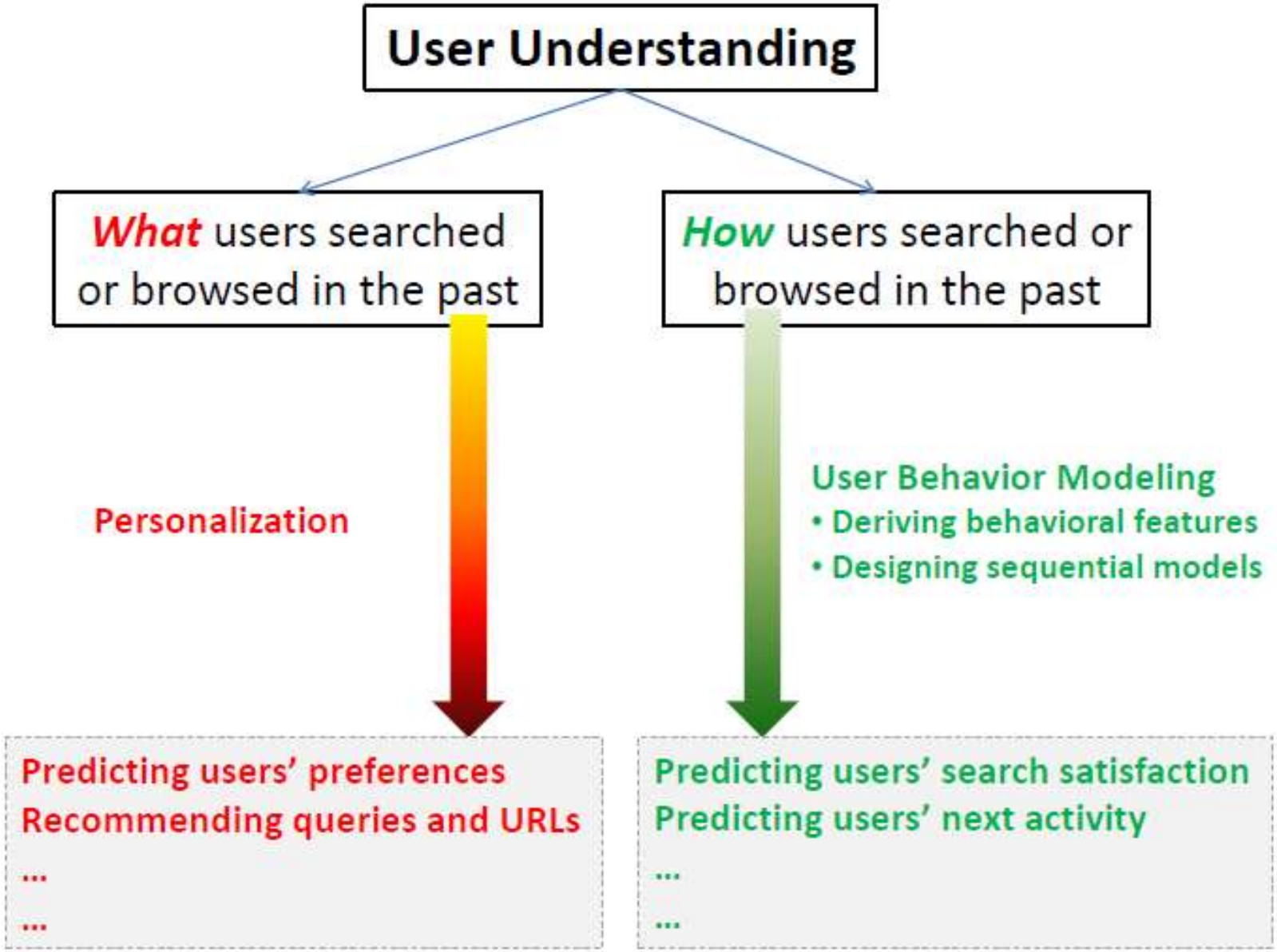
Recap of Lecture 24: Query-Document Matching by Log Mining

- Learning user preferences from logs
- Modeling and predicting clicks

Announcements

Today's Agenda

- Personalized search
- User behavior modeling
- Privacy in Web Search Query Logs


User Understanding
What users searched or browsed in the past
How users searched or browsed in the past
Personalization
User Behavior Modeling
• Deriving behavioral features
• Designing sequential models
Predicting users' preferences
Recommending queries and URLs
...
...
Predicting users' search satisfaction
Predicting users' next activity
...
...

Today's Agenda

- **Personalized search**
- User behavior modeling
- Privacy in Web Search Query Logs

Outline for Personalized Search

- **Introduction to personalized search**
- Three questions for personalized search
  - Which personalization methods work better?
  - How much is personalization feasible in Web search?
  - When is personalization effective in Web search?

Personalized Search

- Different users may have different intents behind the same query
  - Example “GMC”
  - The mix-and-for-all method may not be an optimal solution

General Medical Council

Regulating doctors
Ensuring good medical practice

Contextualization and Individualization

- Personalized search and context-aware search
- Following Piktow, et al., personalization includes *individualization* and *contextualization*
  - Individualization: the totality of characteristics that distinguishes an individual
- Often creates a profile for each user from long history
  - Contextualization: the interrelated conditions that occur within an activity
- Often leverages the user's previous search/browse information within the same session
- Pitkow, J. et al. Personalized search. Commun. ACM, 45(9):50-55, 2002.

How Personalization Helps Search

- Consider the query “GMC”
  - Individualization: if the user profile shows that the user often raises medical-related queries or browses medical-related Web pages, it is more likely the user is searching for General Medical Council
  - Contextualization: if the user inputs query “Honda” and “Nissan” before “GMC” in the same session, it is more likely the user is searching for GMC cars

Approaches to Personalized Search

- Individualization
  - Create a profile for each user from a long history
    - Topic-based profile, e.g., [Pretschner99][Liu02][Pitkow02][Speretta05][Qiu06]
    - Term-based profile, e.g., [Teevan05][Tan06]
    - Click-based profile, e.g., [Teevan07]
- Contextualization
  - Use previous search/browse info in the same session
    - Use previous queries/clicks in the same session, e.g., [Shen05][Cao09]
    - Use previous browsed pages in the same session, e.g., [White09]

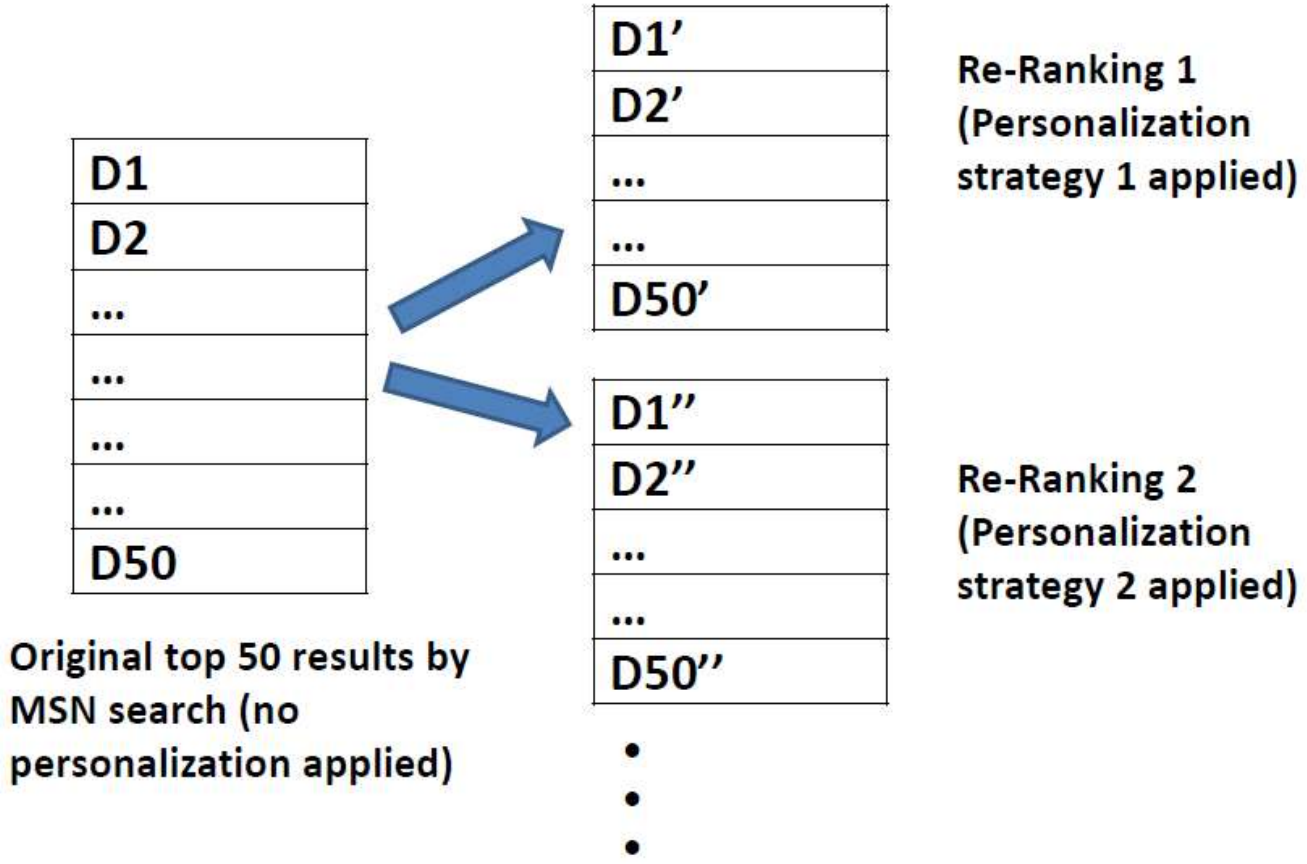
Outline for Personalized Search

- Introduction to personalized search
- Three questions for personalized search
  - **Which personalization methods work better?**
  - How much is personalization feasible in Web search?
  - When is personalization effective in Web search?

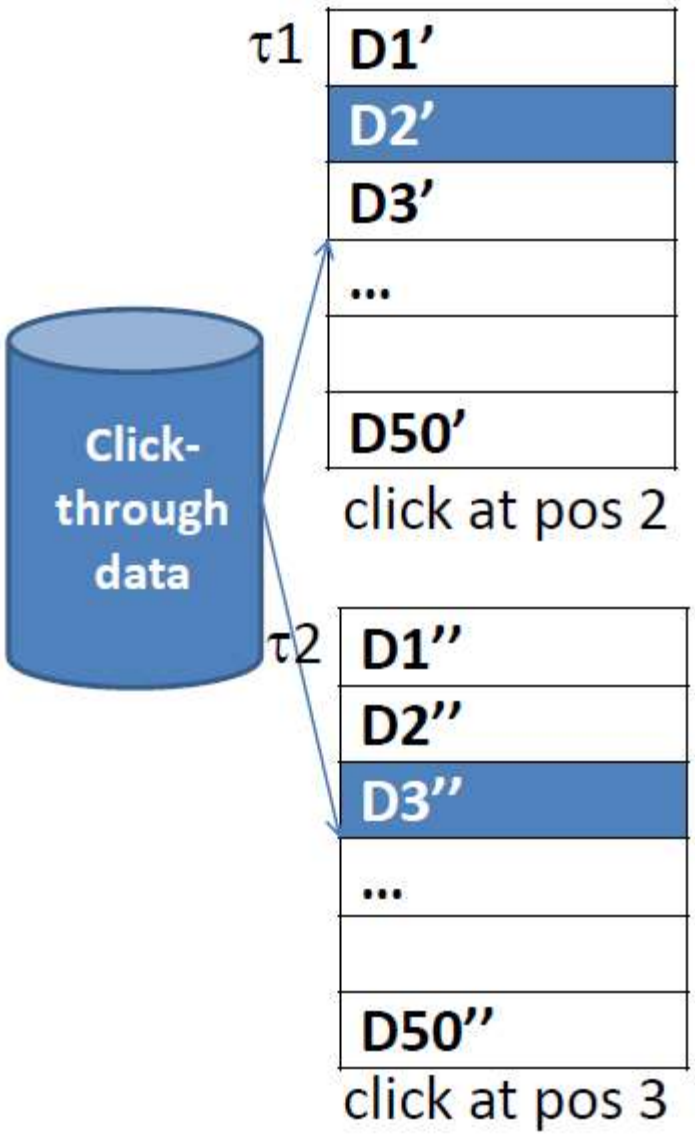
- A Large-scale Evaluation and Analysis of Personalized Search Strategies. Zhicheng Dou, Ruihua Song and Ji-Rong Wen, WWW' 2007
- 12 days MSN search logs
  - 11 days for training and 1 day for testing
- 10,000 randomly sampled users
- 4,639 test queries

Methodology (Step 1): Re-Ranking

- Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW' 2007.



Methodology (Step 2): Evaluation



- Evaluation Measurements
  - Given two re-ranked lists τ_1 and τ_2, if the clicked documents are ranked higher in τ_1 than in τ_2, then τ_1 is better than τ_2
  - Metrics
- Rank scoring (the larger the better)
- Average Rank (the smaller the better)
- Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW' 2007.

Five Strategies

- Topic-based strategies
  - Strategy 1: Topic-based individualization
  - Strategy 2: Topic-based contextualization
  - Strategy 3: A combination of topic-based individualization and contextualization
- Click-based strategies
  - Strategy 4: Click-based individualization
  - Strategy 5: A smoothing method of click-based individualization
- Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW' 2007.

Topic-Based Strategies

- Individualization
  - Create a topic profile cl(u) for each user u

0.1	0	0.5	0	0	0	0	…	0	0.2

    - The profile is aggregated from all visited pages of u
  - For each result p of query q, create a topic vector c(p)

0.05	0	0.2	0	0.3	0	0	…	0	0.1

  - Personalized score: $S^L(q,p,u) = \frac{c_l(u) \cdot c(p)}{\|c_l(u)\|\|c(p)\|}$

Probability of user interested in a topic

Probability of page belonging to a topic

- Contextualization
  - Replace the topic profile with topic context
- The context is aggregated from the visited pages of u only within the current session
- Combination $S^{LS}(q,p,u) = \theta S^L(q,p,u) + (1-\theta) S^S(q,p,u)$
- Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW' 2007

Click-Based Strategies

- Users tend to click on the results they clicked before
  - Personalized score
    - $S^{C}(q,p,u) = \frac{|Clicks(q,p,u)|}{\beta+|Clicks(q,\cdot,u)|}$
  - A user's history could be sparse
- Users tend to click on the results which were clicked by similar users before
  - Personalized score
    - $S^{GC}(q,p,u) = \frac{\sum_{u'} Sim(u,u')|Clicks(q,p,u')|}{\beta+\sum_{u'} Sim(u,u')|Clicks(q,\cdot,u')|}$
    - Sim(u,u') is the similarity between the topic profiles of u and u'
  - Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW' 2007.

Experimental Results

	Method	Ranking Score	Average Rank
Baseline	MSN search	69.4669	3.9240
Click-based	Click-based	70.4350	3.7338
	Group click-based	70.4168	3.7361
Topic-based	Long history	66.7378	4.5466
	Short history	66.7822	4.4244
	Combined profile	68.5958	4.1322

- Click-based strategies have better performance
- A combined approach is better than purely individualization or contextualization
- Topic-based strategies do not perform well
  - Possible reasons: simple implementation, simple user profiles, insufficient search histories, noises in user search histories
- Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW' 2007.

Remaining Questions

- With better implementation, how will the topic-based strategies perform?
- Why a combination of long and short history works better?
- How is the coverage for different strategies?
- How is the correlation between the implicit measures from click-through data and the explicit user labeling?
- **Need more works comparing different personalization strategies.**

Outline for Personalized Search

- Introduction to personalized search
- Three questions for personalized search
  - Which personalization methods work better?
  - **How much is personalization feasible in Web search?**
  - When is personalization effective in Web search?

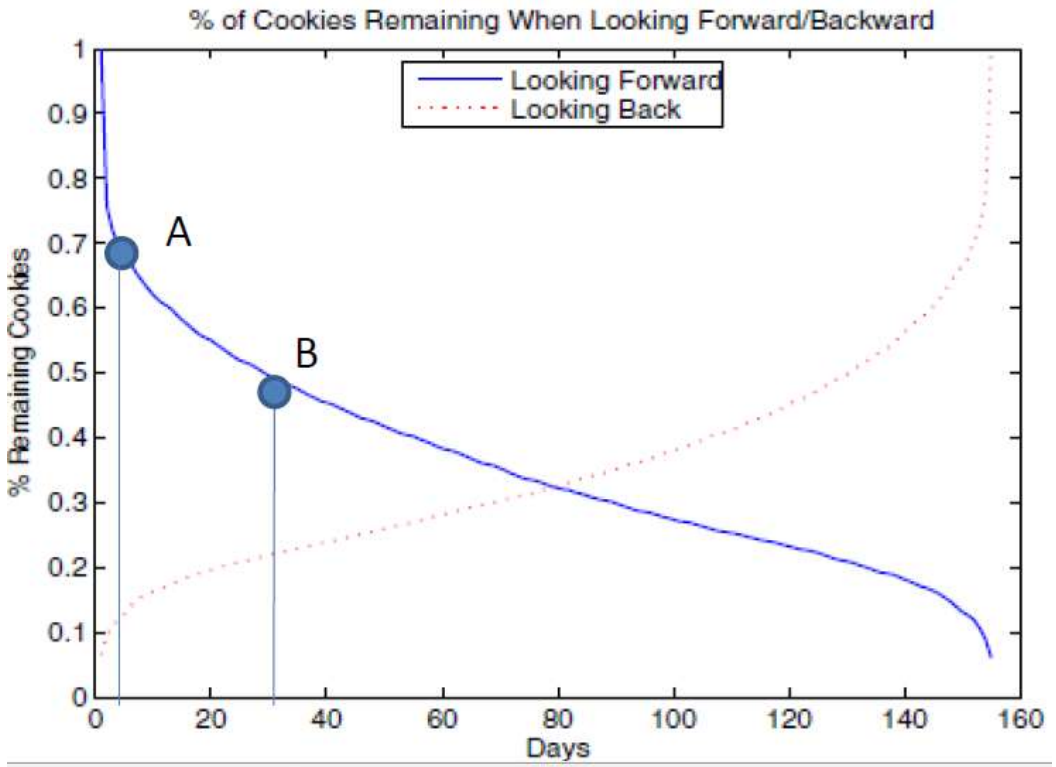
Assumptions in Personalized Search

- Individualization
  - Each user has sufficiently long history to create user profile
  - Each user has consistent interests over time, while different users' interests vary
- Contextualization
  - A short history of a user is available as context information
  - A user's information need does not change within a session

User Persistence

- Each user has sufficiently long history to create user profile?

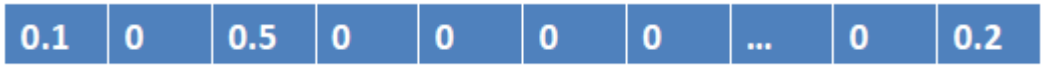
Although 30% cookies expire after the first day (point A) over 40% cookies persist for at least a month (point B)



- Wedig S. and Madani, O. A large-scale analysis of query logs for assessing personalization opportunities. KDD' 06

User Topic Interests

- Each user has consistent interests over time, while different users' interests vary?
  - 22 general topics ("Travel", "Computing"…)
  - Create a topic vector for each query q

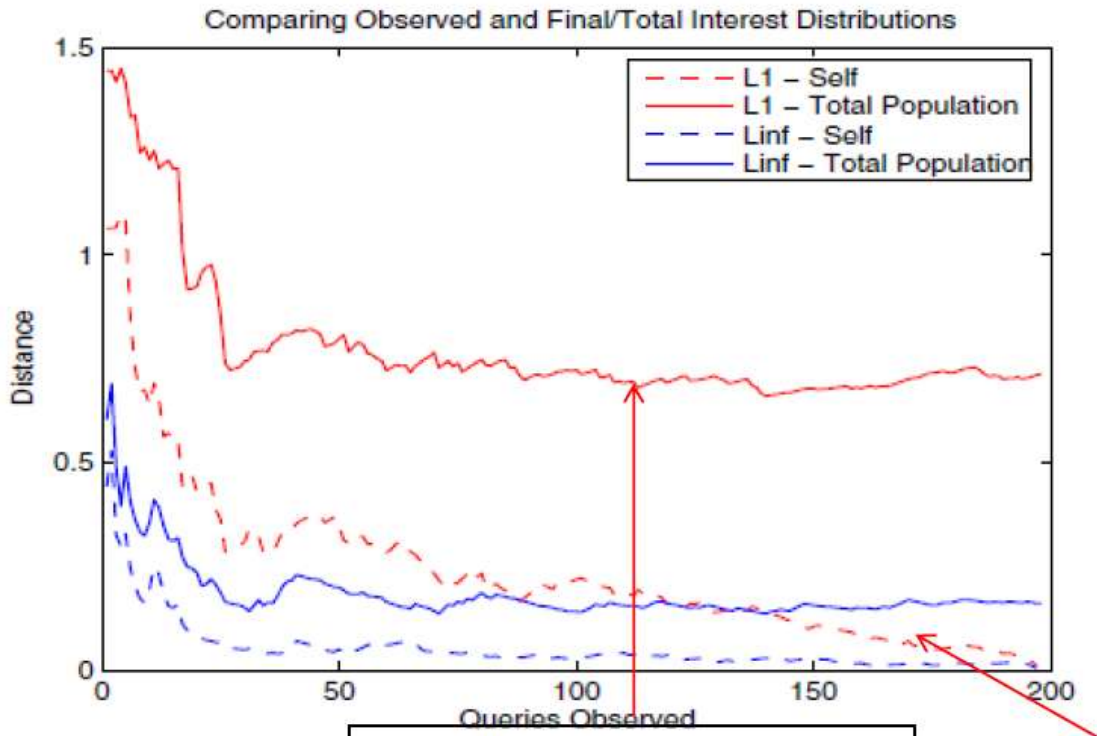


Probability of q belonging to a topic

  - User distribution: F
    - Aggregate from all queries of a user
  - Cumulative distribution: C(K)
    - Aggregate from the first K queries of a user
  - Global distribution: G
    - Aggregate from all queries of all users

Wedig S. and Madani, O. A large-scale analysis of query logs for assessing personalization opportunities. KDD' 06.

Consistence and Distinctiveness



- F: from all queries of a user
- C(K): from the first K queries of a user
- G: from all queries of all users

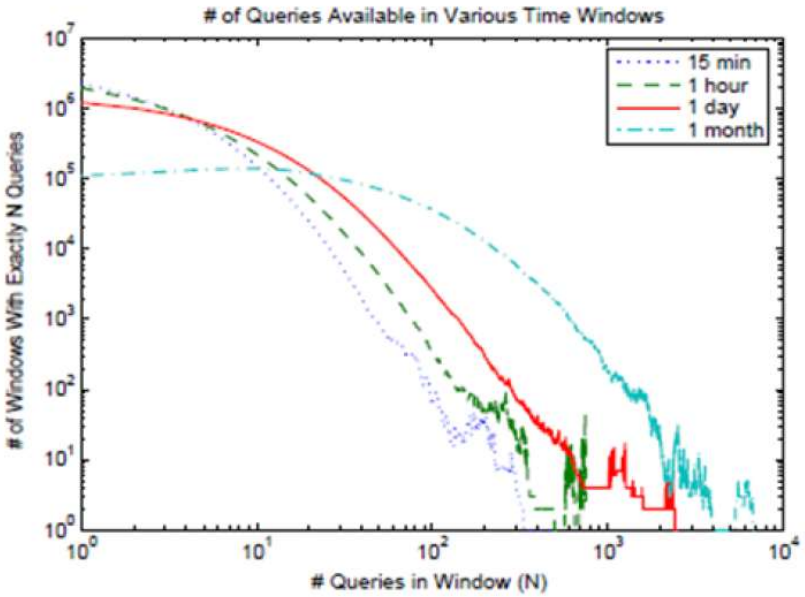
Queries issued by different users have different topics

Queries issued by the same user have consistent topics

The red dotted curve: how the L1 distance between F and C(K) changes with K
The red solid curve: how the L1 distance between G and C(K) changes with K
Wedig S. and Madani, O. A large-scale analysis of query logs for assessing personalization opportunities. KDD' 06.

Context Availability

- A short history of a user is available as context information?



60% of 15-mintues time window contains at least two queries [Wedig06]
About 50% of sessions (30 minutes time out) have at least 2 queries [Cao09]
Average session length is 1.6-3.0

- 30%~60% of queries have previous searches in the same session as context information

Topic Consistency in Sessions

- A user's information need does not change within a session?
- Depending on how sessions are derived
  - Simple timeout threshold: ~70% accuracy
  - More complex features: >90% accuracy

- Jones, R. and Klinkner K.L. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. CIKM'08.

Assumptions in Personalized Search

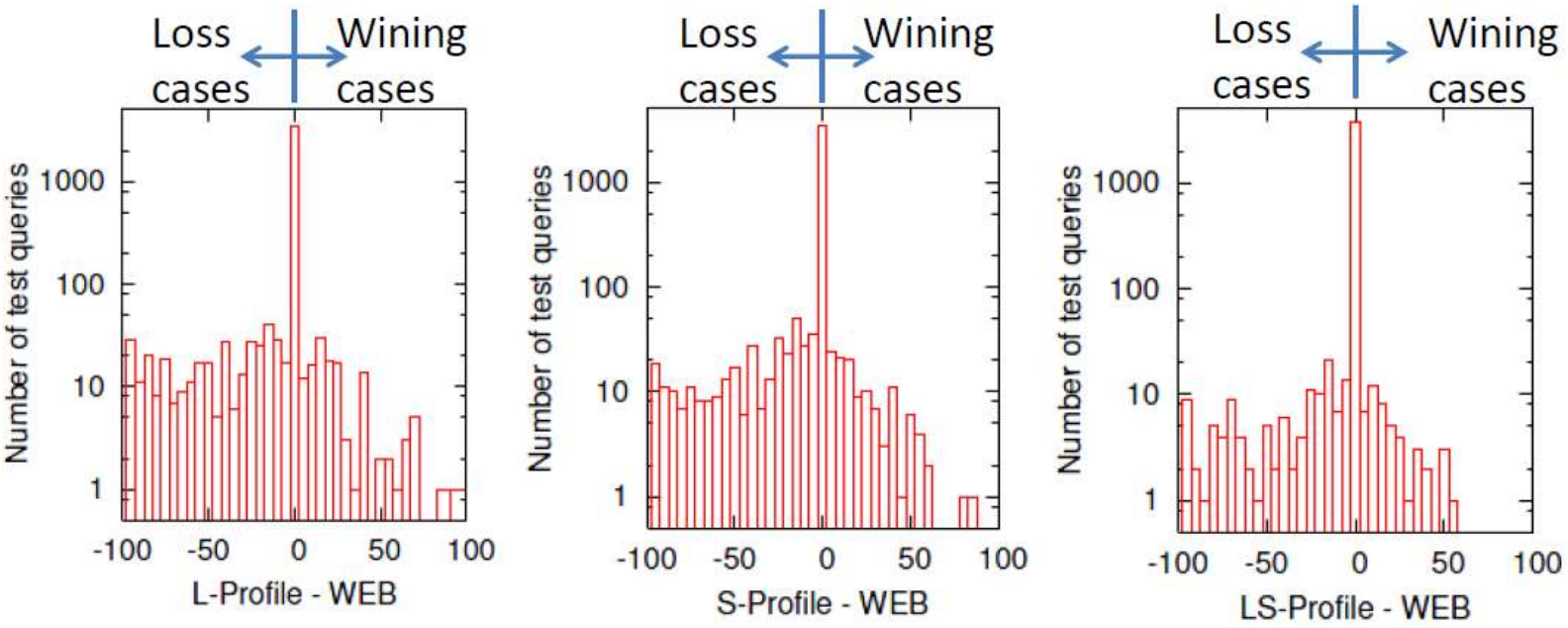
- Individualization
  - Each user has sufficiently long history to create user profile (>40% [Wedig06])
  - Each user has consistent interests over time, while different users' interests vary (Yes [Wedig06])
- Contextualization
  - A short history of a user is available as context information (30%-60% [Wedig06][Cao09])
  - A user's information need does not change within a session (depending on session segmentation: 70%~90% [Jones08])

Outline for Personalized Search

- Introduction to personalized search
- Three questions for personalized search
  - Which personalization methods work better?
  - How much is personalization feasible in Web search?
  - When is personalization effective in Web search?

Case Studies on Personalization Strategies

- Each personalization strategy benefits some queries (wining cases), but harms others (loss cases)
- Can we automatically recognize the wining cases and apply personalization only to those cases?



Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW'07.

Click Entropy

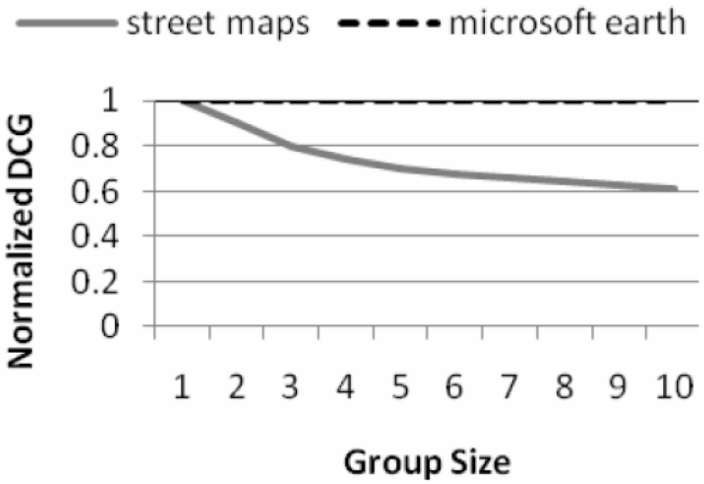
- Major observation from the case studies
  - Personalization only works well on ambiguous queries such as "GMC"
- How to recognize ambiguous queries?
  - Idea: Ambiguous query ⇒ different intents ⇒ click on different search results
  - Click entropy: indicate click diversity of a query

$$H(q) = \sum_{p \in Clicks(q)} -P(p \mid q) \log_2 P(p \mid q) \text{ , where } P(p \mid q) = \frac{|Clicks(q, p, \bullet)|}{|Click(q, \bullet, \bullet|}$$

Dou, Z. et al. A large-scale evaluation and analysis of personalized search strategies. WWW'07.

Building a Predictive Model

- Major features to identify ambiguous queries
  - Click entropy
  - Click-based potential for personalization
- Given the click-through information of several users
  - Specify a random user's clicks as the "ground truth"
  - Calculate the average NDCG of other users' clicks
- The curve for an unambiguous query like "microsoft earth" is flat, but dips with group size for a more ambiguous query like "street maps".
- Teevan. , J. et al. To personalize or not to personalize: modeling queries with variation in user intent. SIGIR'08.

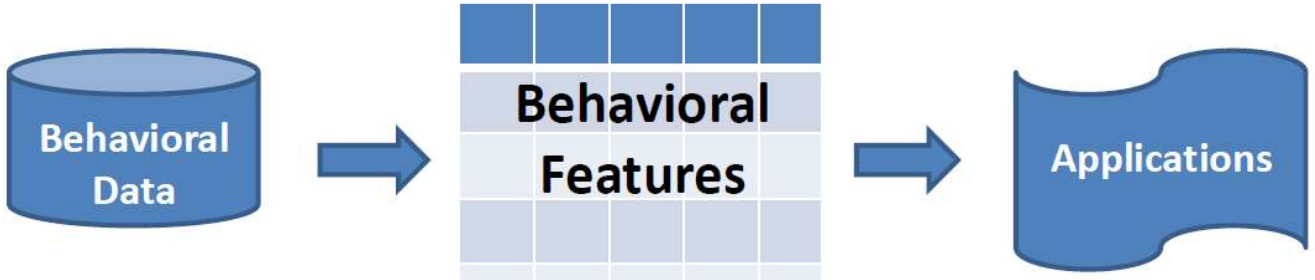


Today's Agenda

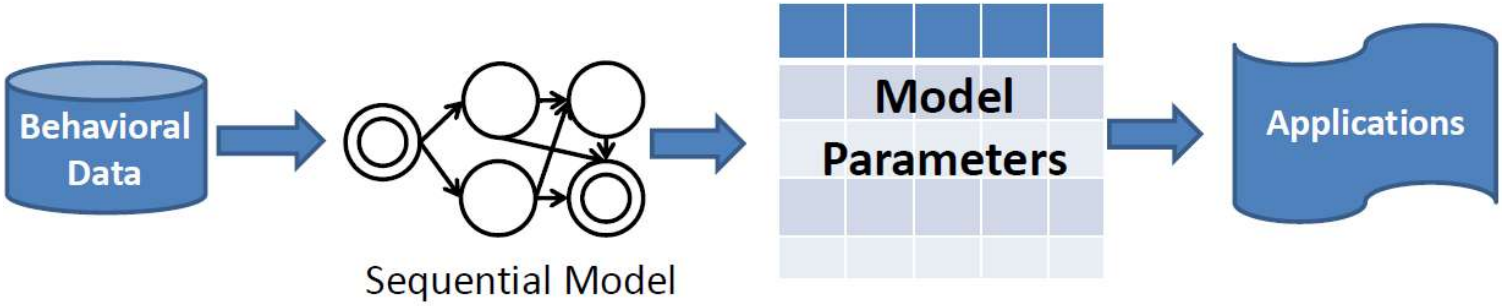
- Personalized search
- **User behavior modeling**
- Privacy in Web Search Query Logs

Two Approaches to User Behavior Modeling

- Deriving behavioral features



- Designing sequential models



Deriving Behavioral Features

- Basic pageview-level user behavior
  - Viewing search results
  - Clicking on search results
- More pageview-level user behavior
  - Dwell-time on pages, mouse movement, screen scrolling, printing, adding to favorite
- Session-level user behavior
  - Post-search browsing behavior
  - Behavior in query chains and search trails

Features Correlated to User Satisfaction

- [Fox05] Apply a user study to correlate behavioral features with user satisfaction
  - Besides click-through, **dwell time** and **exit type** of a search result page strong predictors of user satisfaction
- Exit type: kill browser window; new query; navigate using history, favorites, or URL entry; or time out.
  - Printing and Adding to Favorites highly predictive of user satisfaction
  - Combining various behavior predicts user satisfaction better than click-through alone
- Fox et al. Evaluating implicit measures to improve web search. TOIS, 2005

Features Correlated to Page Relevance

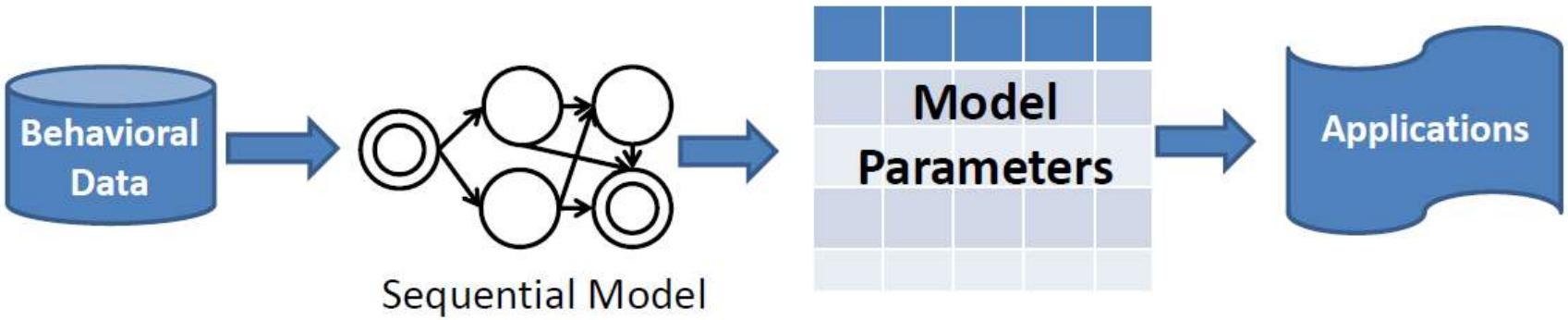
- [Agichtein06] and [Agichtein06a]
- DwellTime
  - Result page dwell time
- DwellTimeDeviation
  - Deviation from expected dwell time for query
- Agichtein, E, et al. Learning user interaction models for predicting web search result preferences. SIGIR'06
- Agichtein, E. et al. Improving Web Search Ranking by Incorporating User Behavior Information. SIGIR'06.

More Features in Previous Studies

- Post-search browsing behavior for URL recommendation ([White07] and [Bilenko08])
- Behavior in sessions to categorize users as navigators and explorers ([White07a])
- Six categories of features for search results pre-fetching ([Downey07])

Two Approaches to User Behavior Modeling

- Deriving behavioral features
- Designing sequential models
  - Sequential patterns
  - Markov chains
  - Layered Bayesian model



Sequential Patterns for User Satisfaction

- Describe sessions with a vocabulary of five letters [Fox05]
- Explore correlations between sequential patterns and user satisfaction

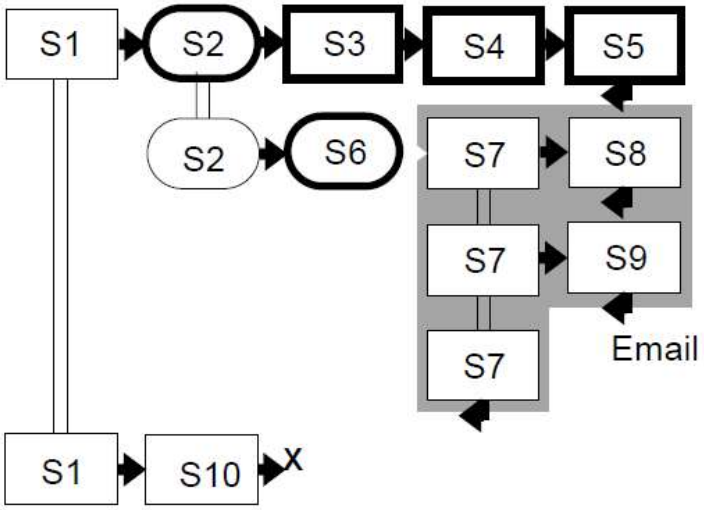
Pattern	Freq.	%SAT	%PSAT	%DSAT
SqLrZ	509	81	10	7
SqLrLZ	117	75	15	9
SqLrLrZ	82	73	13	13
SqLrqLr*	70	64	25	10
SqLrLrLrZ	61	57	22	19
SqLrLr*	362	23	39	36
SqLrLrLr*	129	20	37	42
SqLrLrLrLr*	114	13	35	51

- Session starts (S)
- Submit a query (q)
- Result list returned (L)
- Click a result (r)
- Exit on result (Z)

- Fox et al. Evaluating implicit measures to improve web search. TOIS, 2005.

Sequential Patterns for User Types

- Described user behavior with a vocabulary of three letters [White07a]
- Explored correlations between behavioral sequences and types of users (navigators vs. explorers)



S = search
B = browse
b = back

S B B B b S S

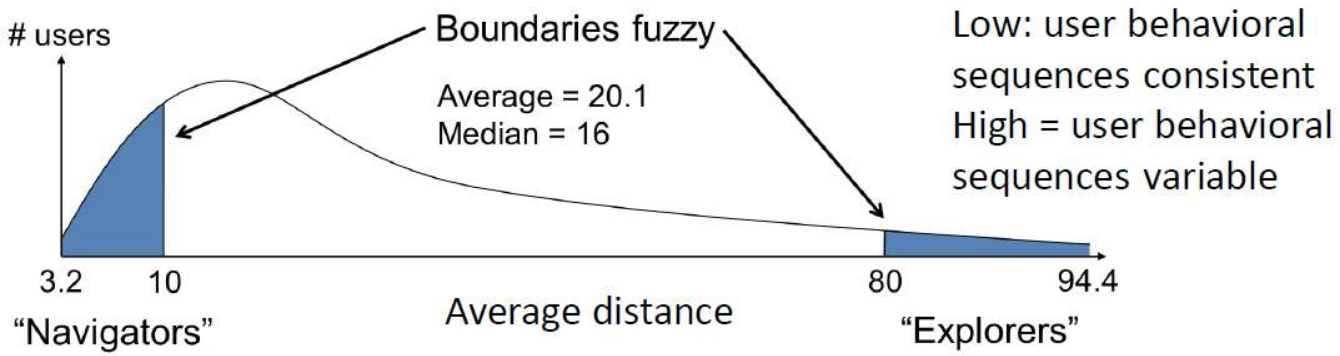
An example from a user session to a sequence

- White, R.W. and Drucker S.M. Investigating behavioral variability in web search. WWW'07.

Distinguishing Navigators and Explorers

- Calculate the average distance of the behavioral sequences for each user
- Suppose a user has three sessions

	Behavioral Sequences	Pair-wise distance	Average Distance
S1	S S B b S B S	ED(1,2) = 4	(4+4+5)/3 = 4.33
S2	S B B b B S b S	ED(1,3) = 4 ED(2,3) = 5	
S3	S B B B B		



Limitations of Sequential Patterns

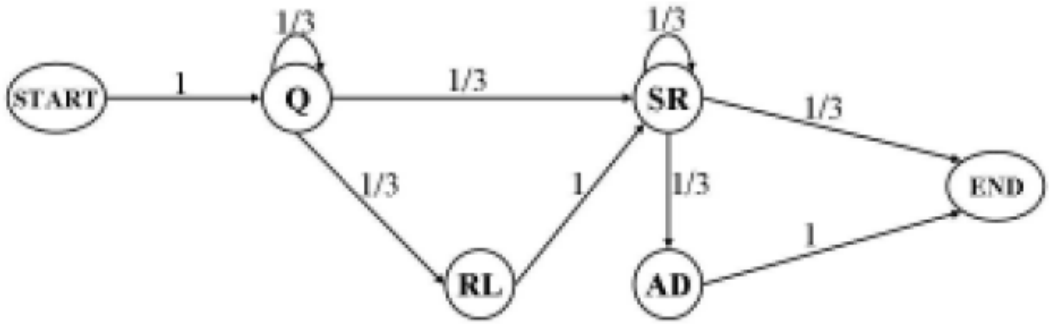
- The previous sequential pattern models have several limitations
  - Only two or three types of behavior modeled
- May not capture user behavior well
  - No aggregate on patterns
- May harm the generalization power on new sessions
  - Hard to incorporate other features for user activities
- E.g., dwell time, whether first click in the session, etc

A Markov Chain Model

- Model rich types of user clicks with a Markov chain [Hassan10]
  - START: the user starts a new goal
  - A query (Q)
  - A click of any of the following types:
    - Algorithmic Search Click (SR)
    - Sponsored Search Click (AD)
    - Related Search Click (RL)
    - Spelling Suggestion Click (SP)
    - Shortcut Click (SC)
    - Any Other Click (OTH), such as a click on one of the tabs
  - END: the user ends the search goal
- Hassan, A. et al. Beyond DCG: user behavior as a predictor of a successful search, WSDM'10.

User Activity Markov Model

- The Markov model is defined as G = (V, E, w)
  - V = {Q,SR,AD,RL, SP,SC,OTH} is the set of possible user actions during the session
  - $E \subseteq V \times V$ is the set of possible transitions between any two actions
  - w:E→[0..1] is the transition probability from state s_i to state s_j
    $w(s_i, s_j) = \frac{N_{s_i,s_j}}{N_{s_i}}$
- Hassan, A. et al. Beyond DCG: user behavior as a predictor of a successful search, WSDM'10

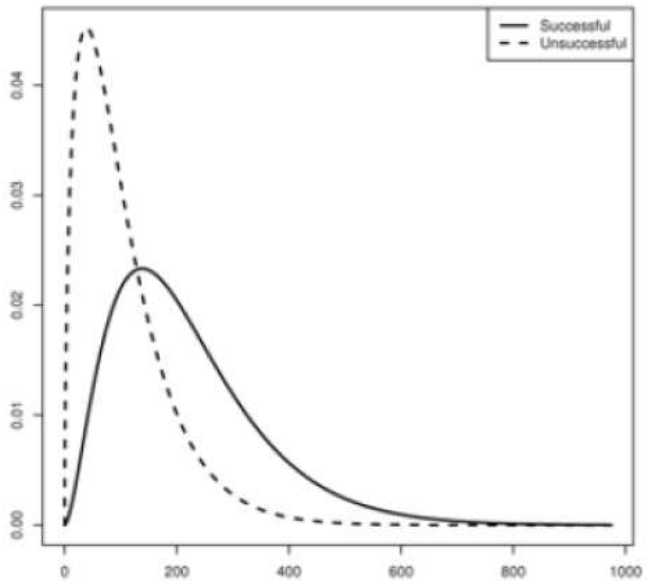


Predict Search Success

- Offline: train two Markov models
  - M_s model: trained by success sessions
  - M_f model: trained by failure sessions
- Online: given a session S=(s_1,…s_n)
  - Calculate the likelihood with M_s and M_f, respectively
- $LL_M(S) = \sum_{i=2}^{n} W(S_{i-1}, S_i)$
- $Pred(S) = 1 \left(if \frac{LL_{M_s}(S)}{LL_{M_f}(S)} > \tau \right)$
  $= 0 \qquad otherwise$
- Hassan, A. et al. Beyond DCG: user behavior as a predictor of a successful search, WSDM'10.

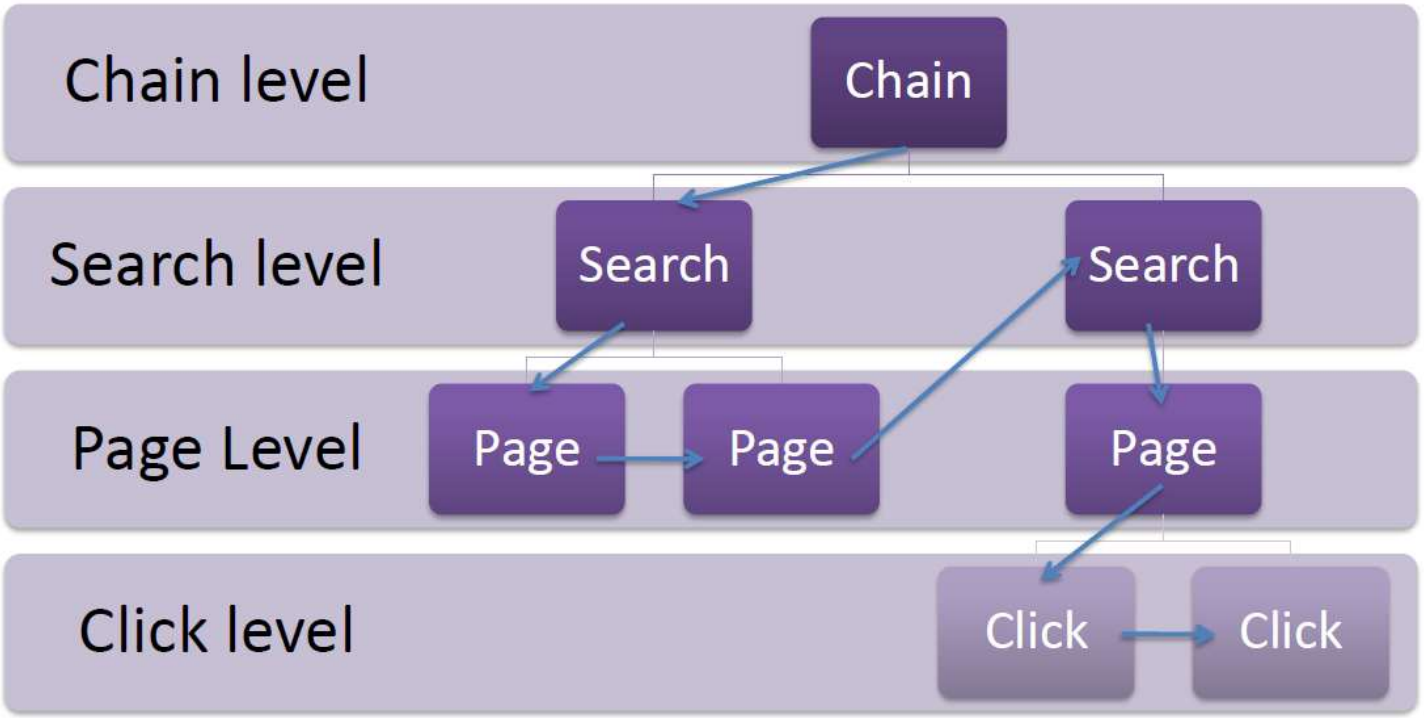
Adding Time to Model

- Time distributions of state transitions are different in successful and unsuccessful sessions
- Assume the transition time follows gamma distribution
- Time distributions incorporated into the transition probabilities of the Markov models
- Hassan, A. et al. Beyond DCG: user behavior as a predictor of a successful search, WSDM'10.



Time distributions of SR → Q transitions for successful and unsuccessful sessions

From Flat Model to Hierarchical Model

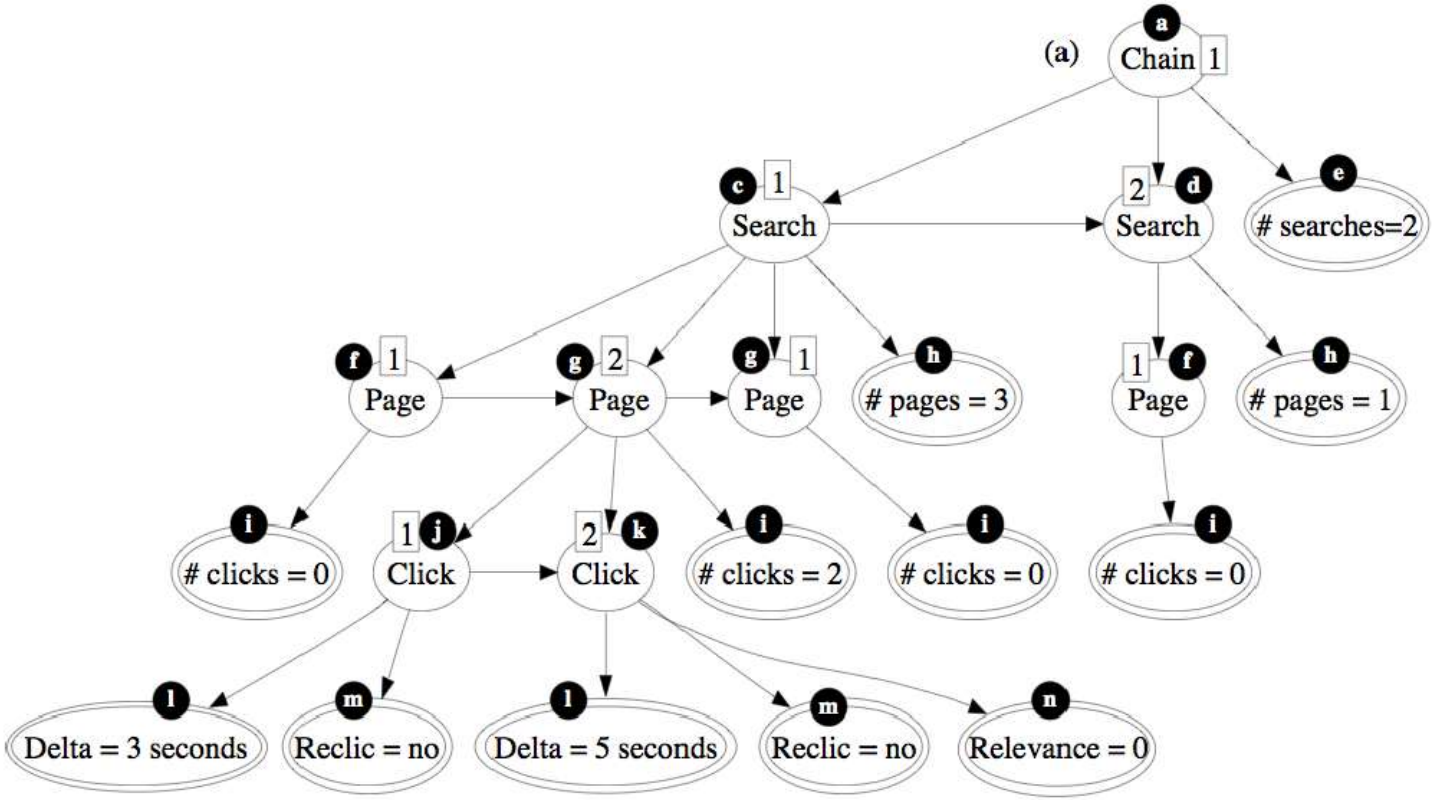


- Piwowarski, B., et al. Mining user web search activity with layered Bayesian networks or how to capture a click in its context. WSDM'09.

A Bayesian Network Model

- Four hidden variables
  - Chain, search, page, click
  - Each hidden variable has a predefined number of states
- Each hidden variable is associated with some observed features
  - Chain: # searches
  - Search: # pages requested
  - Page: # clicks
  - Click: a) dwell time; b) whether “re-click”; c) relevance judgment if available
- Piwowarski, B., et al. Mining user web search activity with layered Bayesian networks or how to capture a click in its context. WSDM'09.

Example of BN



- Piwowarski, B., et al. Mining user web search activity with layered Bayesian networks or how to capture a click in its context. WSDM'09.

Using BN to Predict Page Relevance

- In the offline stage
  - Learning model parameters
- In the online stage
  - Given the observed values of a user session, infer the distributions of states of the hidden variables
  - Extract BN features
- Distribution of the states for each hidden variable
- Maximal likelihood of the BN
- Combining BN features with other features
- E.g., Position of search result, etc.
- Piwowarski, B., et al. Mining user web search activity with layered Bayesian networks or how to capture a click in its context. WSDM'09.

Summary of User Understanding

- What users searched or browsed in the past
  - Personalization: better understand users' intent
- How users searched or browsed in the past
  - User behavior modeling
    - Deriving behavioral features
    - Designing sequential models

Today's Agenda

- Personalized search
- User behavior modeling
- **Privacy in Web Search Query Logs**

Intuitive Understanding of k-Anonymity

- How much anonymity do we need?
- How much gives us plausible deniability?
- Muddier waters with query logs since other information available may be hard to quantify

K-Anonymity in Query Logs

Facebook

Ljubjlana
Hiking Slovenia
Via Alpina Slovenia
Trekking Slovenia

Women's hiking boots

ECML PKDD 2009 Rosie Jones
Dunja Mladenic

Clubbing in Bled

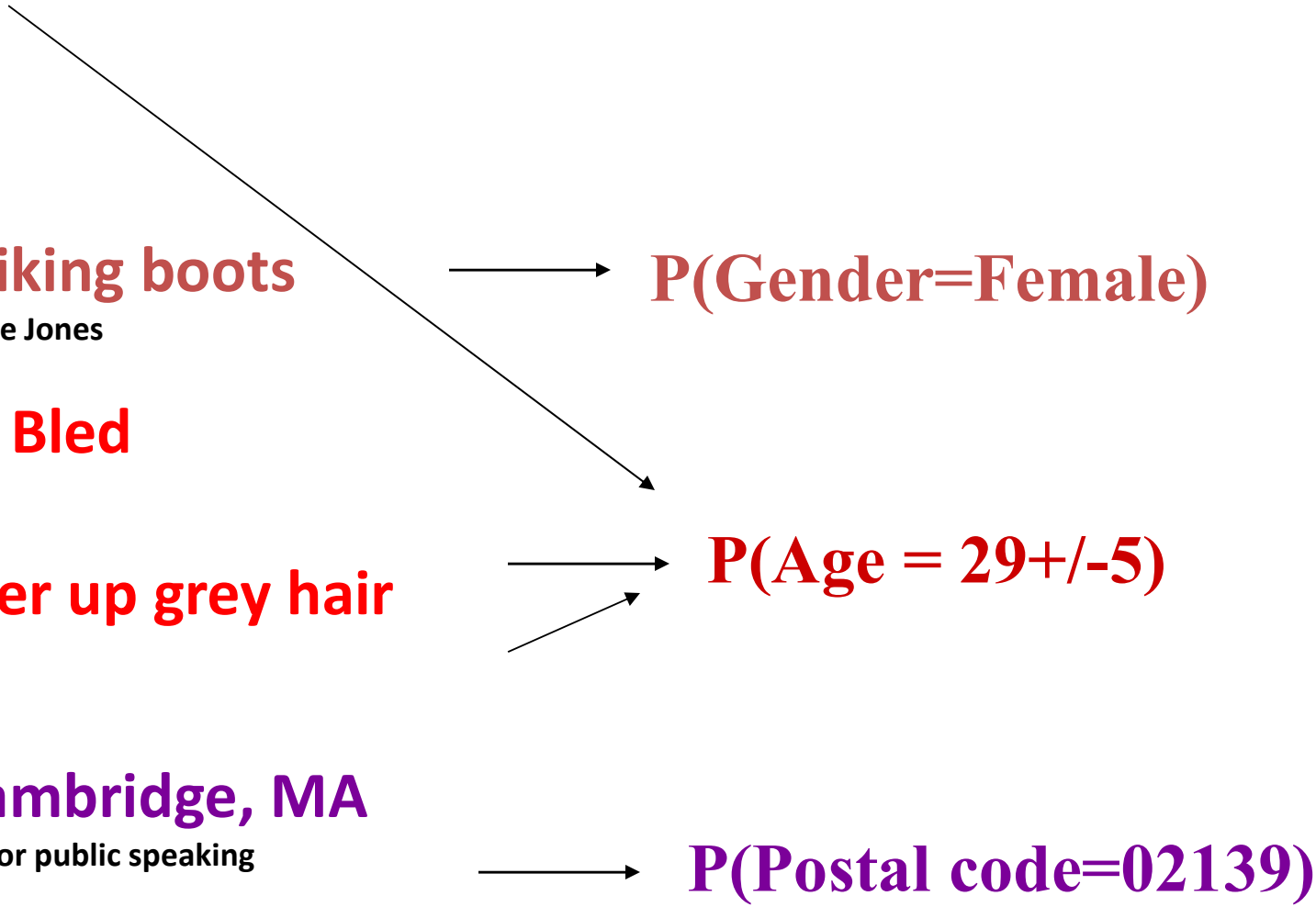
Golf hotel Bled
NIPS 2009

How to cover up grey hair

Latex tables
Yahoo stock price
YHOO

Weather Cambridge, MA

Overcoming shyness for public speaking



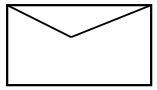
K-Anonymity in Query Logs

- What proportion of users can be uniquely identified from (statistical properties of) their queries?

- [Jones et al, CIKM 2007]

Frame as Supervised Machine Learning Problems

- x = {query1,query2 query3, ...queryn}
  - Queries from a single user: query trace
  - Minimum of 100 queries / included user
  - |X| = 750k

- y1 = gender 
- y2 = age [0..99..]
- y3 = postal code 

- Ground truth from registered users
- Learn f(x) $\longrightarrow$ y

Classifiers Illustrative, Not Optimized

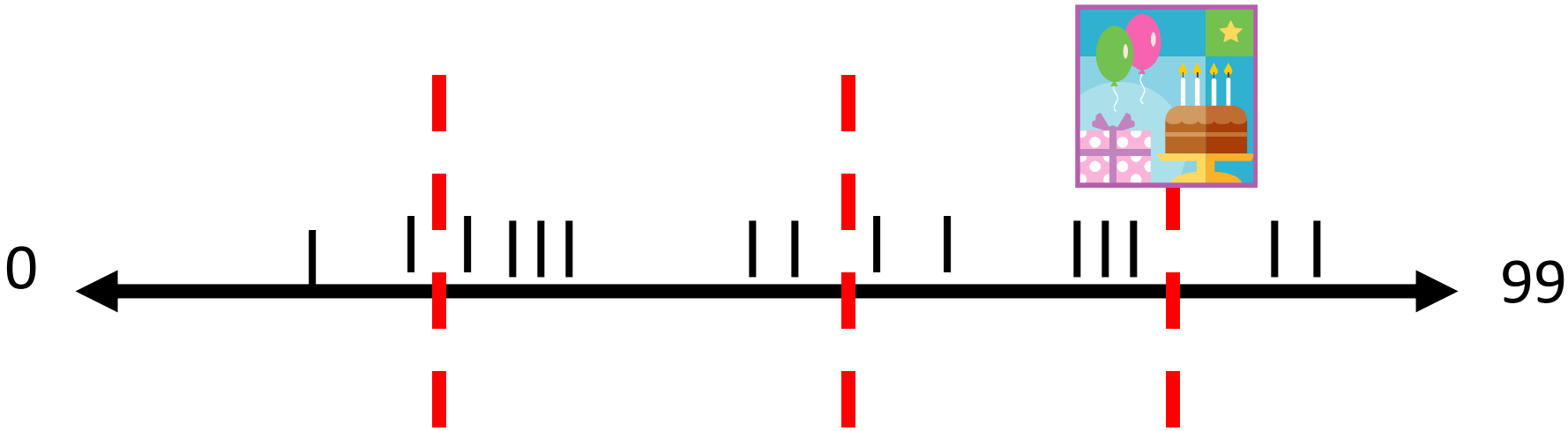
- How much can we learn given pretty good classifiers?
  - Lower bound on attacker's power

Gender Classification – Binary Text Classification

- bag-of-words classifier on query unigrams
- SVM light
- 83% accuracy
- Top terms
  - Female: fanfiction, bridal, makeup, women's, knitting, hair, ecards, glitter, yoga, diet, divorce
  - Male: nfl, poker, espn, ufc, railroad, prostate, football, golf, male, wrestling, compusa, saddam, a variety of adult terms

- Possible improvements: bigrams, fetching webpages...

Age classification



- Age i similar to age i+1
- Regression with bag-of-unigrams predictors
  - Age = SUM w_i f(w_i)
  - Where f(w_i) = frequency of word i in query trace
- SVM light

Age Classification

- 65% of users within 7 years of true age
- Indicators of relative youth: myspace, pregnancy, wikipedia, lyrics, quotes, apartments, torrent, baby, wedding, mall, soundtrack;
- Indicators for older age: aarp, telephone, lottery, amazon.com, retirement, funeral, senior, mapquest, medicare, newspapers, repair
- Improvements: bigrams, fetch pages, query length

US Postal code Codes

- US 5-digit Postal codes: > 42,000 of them
- Cambridge, MA: **02138, 02139, 02140, 02141, 02142, 02163, 02238, 02239**
  - All querying for "Cambridge weather"
- Nearby places have nearby Postal codes
- Postal code3/Zip3 = 021XX ~= Cambridge, MA
- Boston: 02101..02455
- Postal code 2/Zip2 = 02XXX … near Boston, MA

Location Identification

- In-house system to extract placenames
- Sum probs over all placenames found
- 35% correct postal code-3 (1000 class problem!)
- 52% correct postal code-3 in top-3 guesses
- Improvements: topic filtering (high school, restaurants), page fetching, data cleaning (match IP and profile Postal code)
- Outperforms bag-of-words (data sparsity)

Attack of the Mechanical Turk!

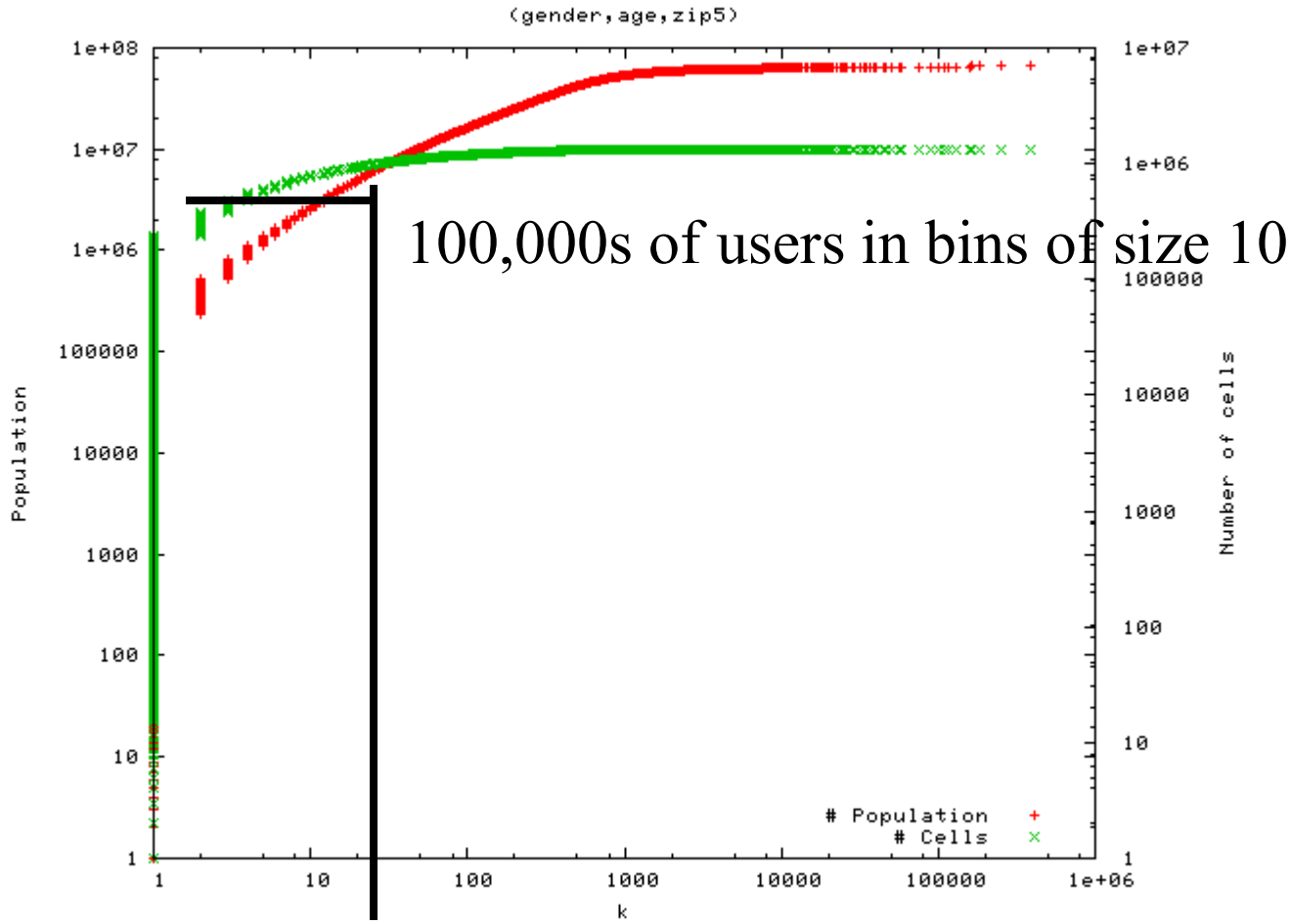
Cheap, fast and good [Snow et al, 2008]

http://www.mturk.com/

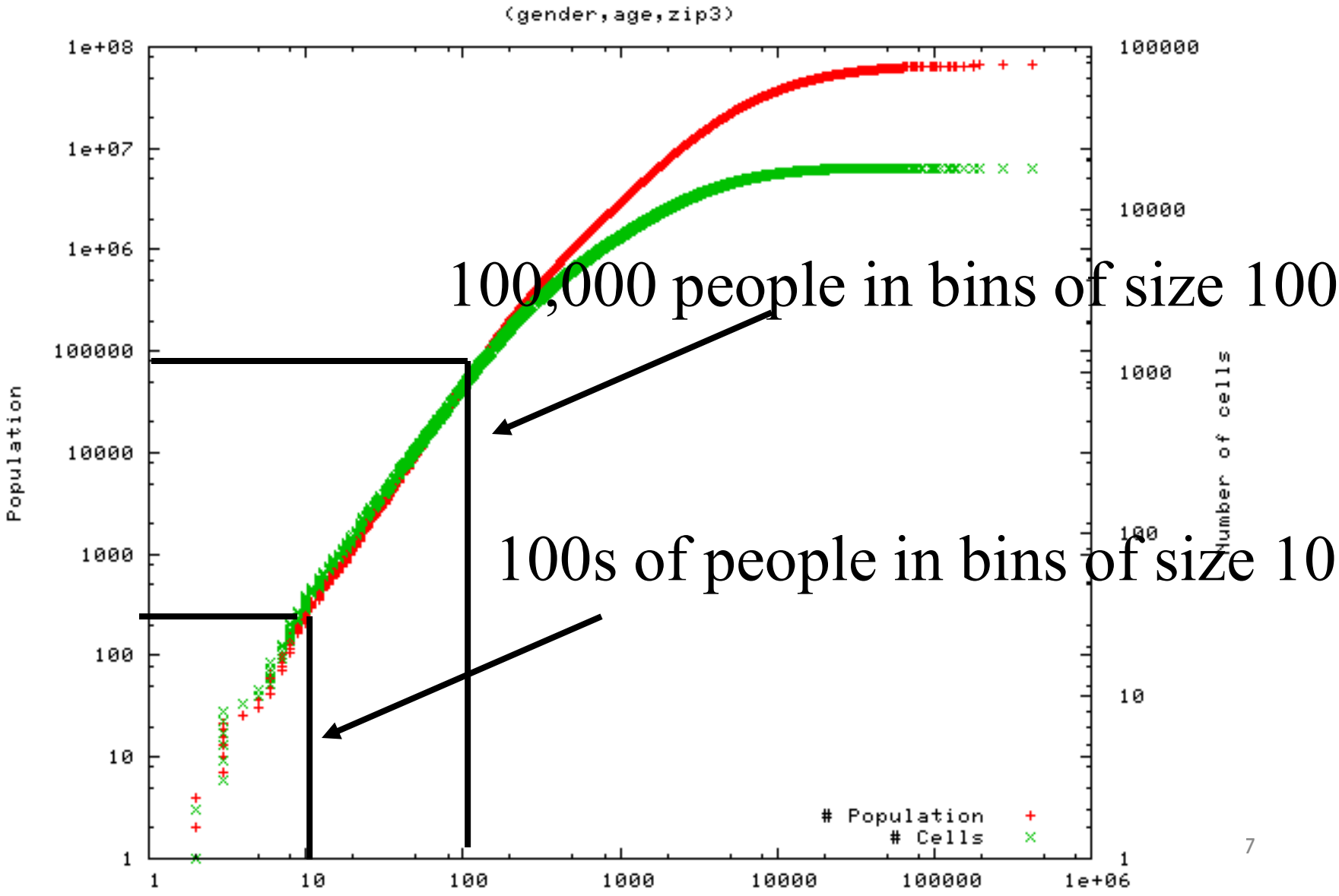
Attack Scenario

- Logs from 750,000 users leaked
- Attacker tries to identify true user among sample of 66.5M registered user profiles
- Uses volunteers and Mechanical Turk to get labeled training data
- (analogous to identifying leaked user as member of US population)

Oracle Classifier



If We knew Gender, Age and Postal code3



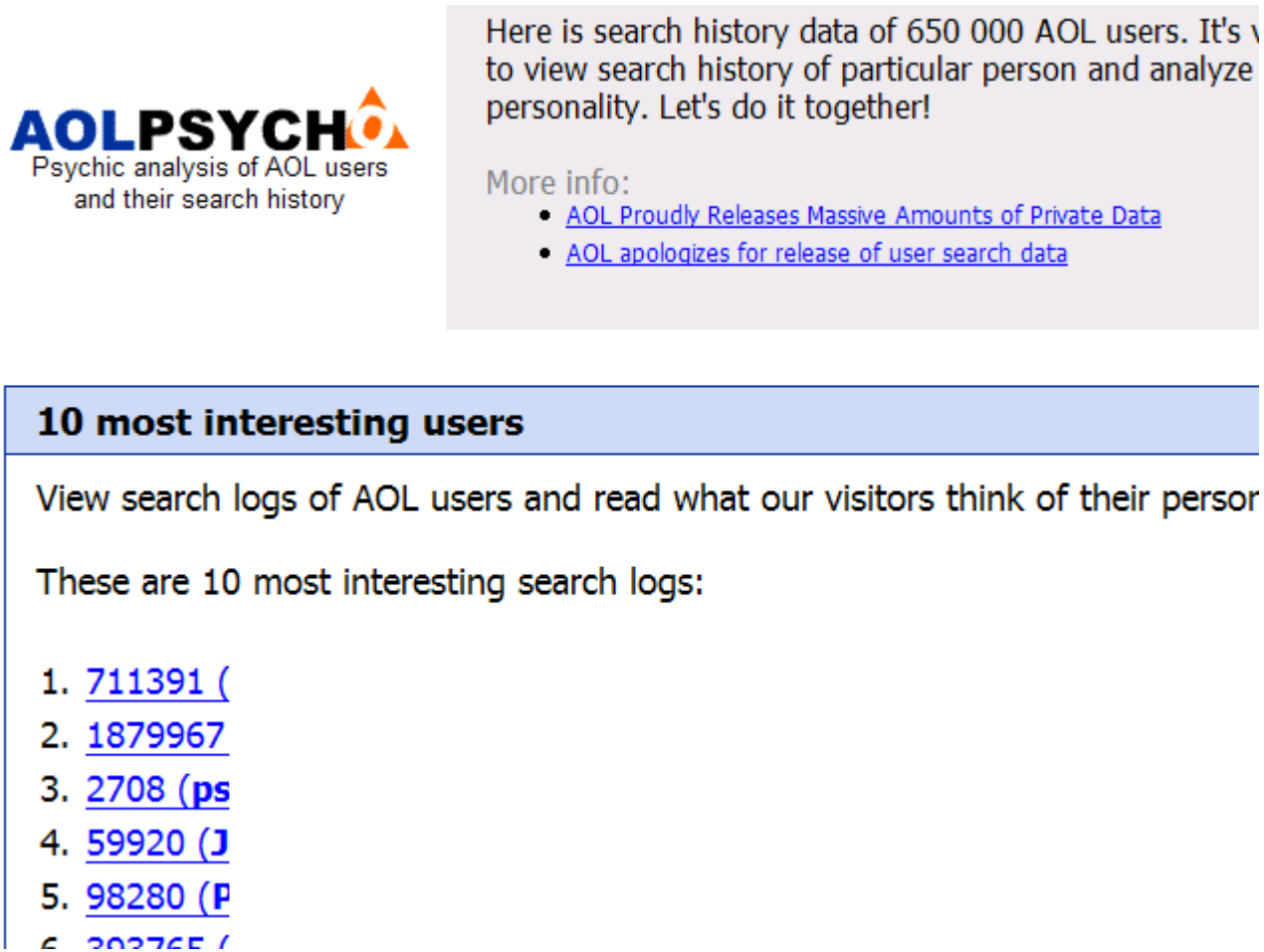
Small Bins Can Be Manually Browsed

- Names, hobbies, etc
- Visit each person…

Trace Attack Model

1. Attacker is willing to sort through all users in a bucket of size k
2. k can vary depending on how specific we are with age, Postal code
3. Take a trace, classify it into bucket
4. If user classified into the correct bucket, by (1) , attacker finds them
5. Number of users found in this way depends on bucket distribution and classifier accuracy

Many Hands Make Light Work!



http://www.aolpsycho.com/

Using Classifiers

- 300 times more likely to find a user than by chance
- This was just predicting age, gender, location
- Lots of other information available in the query trace

Take-away Messages

- Personalized search
- User behavior modeling
- Privacy in Web Search Query Logs
  - Query traces can reveal
    - Age
    - Gender
    - Location
    - Name
  - Removing names, addresses insufficient
  - Privacy of query sessions still open problem
    - Value in sessions for sociology, personalization, search engine improvement….

Further Reading

- Daxin Jiang, Jian Pei, Hang Li. Web Search/Browse Log Mining: Challenges, Methods, and Applications. Tutorial at WWW 2010
- Daxin Jiang, Jian Pei, Hang Li. Mining Search and Browse Logs for Web Search: A Survey. ACM Transactions on Computational Logic, Vol. V, No. N, February 2013, Pages 1–42.
- Maristella Agosti, Franco Crivellari, Giorgio Maria Di Nunzio. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Min Knowl Disc (2012) 24:663–696
- Fabrizio Silvestri. Mining Query Logs: Turning Search Usage Data into Knowledge. Foundations and Trends in Information Retrieval. Vol. 4, Nos. 1–2 (2010) 1–174
- Marius Pasca. Tutorial. Web Search Queries as a Corpus. ACL 2011
- Ricardo Baeza-Yates, Fabrizio Silvestri. Query Log Mining.
- Rosie Jones. Privacy in Web Search Query Logs . Invited Talk at ECML PKDD 2009

Preview of Lecture 26: Crowdsourcing

- Introduction to Crowdsourcing
- Applications of Crowdsourcing
- Challenges in Crowdsourcing
- Managing Quality for Simple Tasks

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!