Privacy in Web Search Query Logs

Rosie Jones, Yahoo! Labs
ECML PKDD, Bled Slovenia
September 7th, 2009
Web Search is Informative

Facebook → Heading to Slovenia
Ljubljana
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

Privacy in Web Search Query Log Mining
Invited talk abstract:
Web search engines have changed our life to information about subjects that are born as well as existing. While the search engines search queries also log those queries, in

Correct spelling
related terms
Common interest
Location
Common interest

Web Search is Informative

Facebook → Heading to Slovenia
Ljubljana
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
A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher’s anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”
NYTimes Identification Method

• queries for
  – “landscapers in Lilburn, Ga,”
  – several people with the last name Arnold
  – “homes sold in Shadow Lake subdivision Gwinnett County Georgia.”

• “It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga.”
Why Care About Query Log Privacy?

• Security
  – Make sure noone can see the data

• Sharing
  – ML/KDD people want interesting data to work with
    – We want you to solve our problems!
    – AOL released data for this reason
    – MS limited release of data for this reason
Outline

• How identifiable are web searchers?
• Why do researchers want to store and study query logs anyway?
• Are there obfuscations to protect users’ identities in the event of a leak?
• What data can be safely shared?
Caveats

• I’m a scientist, not a policy person

• This talk based on published academic research

• No query logs were harmed for this talk
Quantifying Information in Query Logs
k-Anonymity [Samarati & Sweeney, 1998]

- Private data – medical records
- Names removed
- Postal code, gender, date of birth
- Join with public data – voter records
  - Uniquely identify 80% of people
- Identify medical records of then Governor of State of Massachusetts, USA

William Weld
### Anonymized Medical Records

<table>
<thead>
<tr>
<th>ID</th>
<th>DOB</th>
<th>Gender</th>
<th>Postal Code</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.03.69</td>
<td>Male</td>
<td>10011</td>
<td>Torn Ligament</td>
</tr>
<tr>
<td>2</td>
<td>18.08.76</td>
<td>Female</td>
<td>90210</td>
<td>HIV</td>
</tr>
<tr>
<td>3</td>
<td>02.22.48</td>
<td>Female</td>
<td>15213</td>
<td>Dementia</td>
</tr>
</tbody>
</table>
User could be any one of $k$
When $k$ is still too small, suppress sensitive information
Notions of Probabilistic k-Anonymity

• Beyond Suspicion
  – No more likely to be me than anyone else

• Probable Innocence
  – Less than 50% probability it is me

• Possible innocence
  – Non trivial probability it wasn’t me
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
Ljubljana
Hiking Slovenia
Via Alpina Slovenia
Trekking Slovenia

Women’s hiking boots $\rightarrow$ $P(\text{Gender} = \text{Female})$
ECML PKDD 2009 Rosie Jones
Dunja Mladenic

Clubbing in Bled
Golf hotel Bled
NIPS 2009

How to cover up grey hair $\rightarrow$ $P(\text{Age} = 29+/^-5)$
Latex tables
Yahoo stock price
YHOO

Weather Cambridge, MA $\rightarrow$ $P(\text{Postal code} = 02139)$
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

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

- $y_1 = \text{gender}$
- $y_2 = \text{age}$ [0..99..]
- $y_3 = \text{postal code}$

- Ground truth from registered users

- Learn $f(x) \rightarrow y$
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...
Not Everyone is a Stereotype

But the correctly identified individuals are at risk
Age classification

- Age $i$ similar to age $i+1$
- Regression with bag-of-unigrams predictors
  - $\text{Age} = \sum w_i f(w_i)$
  - $f(w_i)$ = 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

100,000s of users in bins of size 10
If We knew Gender, Age and Postal code

100,000 people in bins of size 100

100s of people in bins of size 10
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!

Here is search history data of 650,000 AOL users. It's interesting to view search history of particular person and analyze it for personality. Let's do it together!

More info:
- AOL Proudly Releases Massive Amounts of Private Data
- AOL apologizes for release of user search data

10 most interesting users

View search logs of AOL users and read what our visitors think of their personality.

These are 10 most interesting search logs:

1. 711391 (bad sex made me a lesbian)
2. 1879967 (disgusting)
3. 2708 (psycho_ex)
4. 59920 (JonBenet fan)
5. 98280 (Prayer Fighter)
6. 202765 (anol sex)

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
Vanity Queries

[Jones, Kumar, Pang, Tomkins, CIKM 2009]
Facebook

Ljubljana

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
Reported to be widespread

• “Almost half of all U.S. Internet users (47%) have searched for information about themselves online, up from 22% in 2002”
  – Pew Internet Study, 2007
Quantifying Vanity Queries in Search Logs

• Ground truth: Query traces coupled with people’s real names

• What we have: sequence of queries issued by a given user, paired with userIDs

• Quasi-ground-truth
  – Public user profile
  – Parsing userid according to popular firstname/lastname pairs
  – Automated, no manual inspection

  – rosiejones_au@yahoo.com
  – ghanirayid2006@hotmail.com
Extracting names

• In a sample of 700K users with query traces
• 23.4% Y+ profile (first or last) names found
• 88.57% a name-word is parsed out of the userid
• 16% with at least two names from both sources
• 1% the same two names from both sources
In search of ourselves…

• Out of the users with “reliable” names identified: over 10% issued a query containing both names

• But we also search for other names
  – Friends/coworkers/interviewee
  – Michael Jackson, Angela Merkel, Dunja Mladenic
Where do you Rank?

• Given a user, rank all the names issued by this user (tf/idf)
  – 90% query for their own name within top-10 names
• Given a name, rank all the users who issued the name
  – (modified tf/idf) 85% of the correct user rank at 1
Person Attack
Try to Find a Particular User’s Queries
Person Attack

- Given real-world person, try to find their trace
- Knowledge of (approx) age, Postal code, gender
- Knowledge of hobbies
- Seen queries on browser?
Known Unique Queries

• 50-64% of queries are unique (previous work)
• Knowing a single one identifies the user

• Scrub unique queries?
### Non-unique Query Guesses

<table>
<thead>
<tr>
<th>Category</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cars</td>
<td>volkswagen beetle (478)</td>
<td>triumph tr3 (23)</td>
</tr>
<tr>
<td></td>
<td>honda odyssey (1504)</td>
<td>e-type jaguar (5)</td>
</tr>
<tr>
<td></td>
<td>toyota prius (1070)</td>
<td></td>
</tr>
<tr>
<td>Sports</td>
<td>skiing (9618)</td>
<td>bassmaster (388)</td>
</tr>
<tr>
<td></td>
<td>football (123802)</td>
<td>Skulling (17)</td>
</tr>
<tr>
<td>Food</td>
<td>Pizza (104,888)</td>
<td>Assam (747)</td>
</tr>
<tr>
<td></td>
<td>Italian restaurant (4,998)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brie (39,325)</td>
<td></td>
</tr>
<tr>
<td>Books</td>
<td>Harry potter (27,838)</td>
<td>Holly Lisle (20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elizabeth Moon (27)</td>
</tr>
</tbody>
</table>
Conjunction of Query Guesses Reduces Bin Size Drastically

<table>
<thead>
<tr>
<th>Query Set</th>
<th>Bin Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry potter, pizza</td>
<td>4855</td>
</tr>
<tr>
<td>Football, harry potter, volkswagen beetle</td>
<td>3</td>
</tr>
<tr>
<td>Danielle steele, volkswagen beetle</td>
<td>1</td>
</tr>
<tr>
<td>Brie, holly lisle, pizza</td>
<td>1</td>
</tr>
</tbody>
</table>
Query Log Mining
Improve Search Engine

• Annotated result page
• Did You Mean?
• Related terms
• Document relevance based on clicks
Correct Spelling More Common than Misspelling in Query Logs

[Cucerzan and Brill, 2004]
Good and bad spellings point to same page

excite.com

- [Craswell et al 2001]
Reformulations from Bad to Good Spellings

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

[Jones & Fain, 2003]
## Semantic relationships between phrases

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>Example</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonym</td>
<td>low cost; cheap</td>
<td>4.2%</td>
</tr>
<tr>
<td>Hypernym</td>
<td>muscle car; mustang</td>
<td>2.0%</td>
</tr>
<tr>
<td>Hyponym</td>
<td>lotus; flowers</td>
<td>2.0%</td>
</tr>
<tr>
<td>Coordinate/Sibling</td>
<td>aquarius; gemini</td>
<td>13.9%</td>
</tr>
<tr>
<td>Generalization</td>
<td>lyrics; santana lyrics</td>
<td>4.8%</td>
</tr>
<tr>
<td>Specification</td>
<td>credit card; card</td>
<td>4.7%</td>
</tr>
<tr>
<td>Spelling change</td>
<td>peopl; people</td>
<td>14.9%</td>
</tr>
<tr>
<td>Stemmed form</td>
<td>ant; ants</td>
<td>3.4%</td>
</tr>
<tr>
<td>URL change</td>
<td>alliance; alliance.com</td>
<td>29.8%</td>
</tr>
<tr>
<td>Other relationship</td>
<td>flagpoles; flags</td>
<td>9.8%</td>
</tr>
<tr>
<td>No relationship</td>
<td>crypt; tree</td>
<td>10.4%</td>
</tr>
</tbody>
</table>

[Carterette et al, ACL 2006]
Relating Queries (Baeza-Yates, 2007)

- Common session
- Queries (q1, q2, q3, q4)
- Pages
- Clicks
- Common words
- Common clicks
- Common terms
- Links

Diagram:
- q1 ← q2
- q3 → q4
- w

Common terms:
- w

Common session:
- w
Topical Seasonality

christmas

christmas songs

christmas gifts

christmas trees

[Liu et al, CIKM 2006]
Personalization

• Location
  – Coffee shops [in Cambridge, MA]

• Gender
  – Winter jackets [for women]

• Age
  – Movies [in my demographic]
  – Science [for adult versus 10-year-old]
Identifying migraine causes from query logs

<table>
<thead>
<tr>
<th>Term</th>
<th>$\text{dep } \text{migraine}(q) \times 10^{-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>coffee</td>
<td>7.4</td>
</tr>
<tr>
<td>tea</td>
<td>8.2</td>
</tr>
<tr>
<td>coffee maker</td>
<td>10.1</td>
</tr>
<tr>
<td>caffeine</td>
<td>22.3</td>
</tr>
<tr>
<td>magnesium</td>
<td>24.7</td>
</tr>
<tr>
<td>dog</td>
<td>5.5</td>
</tr>
<tr>
<td>free</td>
<td>2.3</td>
</tr>
</tbody>
</table>

[Richardson, ACM TWEB 2008]
Sociology

[Richardson, ACM TWEB 2008]
Other Medical and Social Applications

- Identifying onset of H1N1 flu in a population
- Finding unknown links between behaviors and medical conditions
- Music interests and shopping habits

- Finding correlations depends on keeping within-user cooccurrences
Obfuscation
How Can We Protect Identity?
Obfuscation I: Remove Personally Identifying Information

- Remove names, placenames, numbers

- Trace attack
  - Gender: still works
  - Age: still works
  - Place ID: doesn’t work as well with place-names removed

- Unique query conjunction: still works
Obfuscation II

• Reset session identifiers periodically

– Can’t link my queries last year with my queries today
Days of Data: Postal code
Days of Data Needed? Gender

![Graph showing the accuracy of a gender classifier vs. the number of days in a query log.]

- Accuracy increases as the number of days in the query log increases.
- The number of days needed for high accuracy grows significantly with each additional day.
Days Of Data: Age

![Graph showing the relationship between the number of days in a query log and the average absolute error in age classification.](image-url)
Obfuscation III: Bundling to Provide Privacy

• removing key pieces of identifying information from its system every 18 to 24 months.

• IP addresses are altered, the information will be linked to clusters of 256 computers instead of just a single machine
  – IPs differing in last digits are often geographically close

• depersonalize computer "cookies" -- hidden files that enable Web sites to track the online preferences and travels of their visitors.
Risks with Bundles

• Bundle hunting
  – Can we tell which bundle a user is in?

• Bundle analysis
  – How much does a bundle tell us about the users in it?
Structure vulnerabilities inside bundles

• A bundle reveals significant information on its dominant user
  – About 3% of the bundles have a user that issued at least half of the queries
  – Privacy breach also exists for user who queries for a unique postal code with sensitive information

• Can individual users be reidentified?

• [Jones, Kumar, Pang and Tomkins, CIKM 2008]
Separating bundles into user-fibers

• Given a bundle of fibers (users), how to extract individual fibers from the bundle? (and efficiently?)
  – Link queries with the same geo locations identified (g-edges)
  – Link queries with word overlap
    • Also tried word topic classifiers
    • Word cooccurrence statistics
Evaluation

• Measures: f-measure computed over major users
• Baselines
  – Baseline 1: each query as a single cluster
  – Baseline 2: one cluster for the whole bundle
## Evaluation

<table>
<thead>
<tr>
<th>Mask Last n bits</th>
<th>Each Query One User</th>
<th>Whole Bundle as One User</th>
<th>geo-edges</th>
<th>geo, word-edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 bits</td>
<td>0.151</td>
<td>0.257</td>
<td>0.181</td>
<td>0.570</td>
</tr>
<tr>
<td>12 bits</td>
<td>0.160</td>
<td>0.105</td>
<td>0.187</td>
<td>0.562</td>
</tr>
<tr>
<td>16 bits</td>
<td>0.164</td>
<td>0.076</td>
<td>0.186</td>
<td>0.521</td>
</tr>
</tbody>
</table>
Summary – Query Log Bundles

- Studied the privacy implications of bundling when used as a tool to enhance the privacy of users in querylogs.

- User identity can be violated with query bundles:
  - Relations between users and bundles can be established via analysis of vanity search.
  - Structural vulnerabilities: privacy of dominant users in bundles violated.
  - Analytical vulnerabilities: bundles can be decomposed into individual sessions.

- There are significant challenges to using bundling alone to protect user anonymity.
External Estimation of Search Engine Query Logs

• Query suggestion services
  – Queries ordered by popularity, bad queries filtered
Power Law Distribution of Query Frequency

- Derive popularity from query rank
- Estimate query rank from shortest exposing prefix
  - Estimate how many other queries have the same prefix
  - Use sampling algorithm

[Gurevich et al, VLDB 2008]
Include only first $d$ queries per user
Include only queries seen $d_c$ times
Link queries only via co-click graph
Hundreds of Different Users Queried and Clicked

Inject noise in counts
Releasing Search Logs Privately

• Queries connected via co-click graph, not session
  – Cannot find sets of queries from a single user
  – Utility in cooccurrence information preserved via common clicks on documents

• Inject random noise into counts
  – Prevent any user from knowing exactly how many others issued the query

• Threshold on minimum numbers of users who issue query
  – Avoid queries issued by few users included in the sample

  – [Korolova et al WWW 2009]
Relating Queries (Baeza-Yates, 2007)

- **q1** ↔ **q2**
- **q3** ↔ **q4**
- **w**

- **common session**
- **common words**
- **common clicks**
- **common terms**

- **queries**
- **clicks**
- **pages**
- **links**
Korolova et al WWW 2009

- Release queries whose noisy counts exceed threshold
- For each query, top 10 results from a given search engine are public knowledge
  - Release noisy click counts for top 10 URLs

  - Algorithm provably private when
    - Threshold = d(1+ln(2/2delta()/epsilon)
    - Noise from Laplace distribution
    - Keeping the first d queries from each user
Open Research Problems

• How identifiable are web searchers?
• Why do researchers want to store and study query logs anyway?
  – Learning to Disambiguate Search Queries from Short Sessions Lilyana Mihalkova and Raymond Mooney
• Are there obfuscations to protect users’ identities in the event of a leak?
• What data can be safely shared?
Summary

• Query traces can reveal
  – Age
  – Gender
  – Location
  – Name

• Removing names, addresses insufficient

• One provable way of safely releasing co-click graphs

• Privacy of query sessions still open problem
  – Value in sessions for sociology, personalization, search engine improvement…. 
Questions?
Acknowledgements

• Ricardo Baeza-Yates, Bo Pang, Ravi Kumar, Andrew Tomkins
References

• [Samarati and Sweeney, 1998] Samarati and Sweeney, Generalizing Data to Provide Anonymity when Disclosing Information, SIGACT-SIGMOD-SIGART Symposium

• 1998,

• Measuring the Meaning in Time Series Clustering of Text [Liu et al, CIKM 2006]

• [Jones et al, SIGIR 2007]
• Richardson, Learning about the World through Long-Term Query Logs, in ACM Transactions on the Web, vol. 2, no. 4, Association for Computing Machinery, Inc., October 2008
• Vanity Fair: Privacy in Querylog Bundles [Jones et al, CIKM 2008]