
Towards Privacy Preserving Query Log Publishing

Li Xiong
Eugene Agichtein

Department of Math and Computer Science
Emory University



Introduction

- Query log analysis vs. privacy concerns
- Notion of privacy
 - HIPAA, NSF guidelines, Library records guidelines
- Our position: Query logs should be published but should be anonymized
 - Query log analysis applications
 - Sensitive information
 - Some dimensions for query log anonymization

Query Log Applications

AnonID	Query	QueryTime	ItemRank	ClickURL
217	lottery	2006-03-01 11:58:51	1	http://www.calottery.com
217	lottery	2006-03-27 14:10:38	1	http://www.calottery.com
1268	gall stones	2006-05-11 02:12:51		
1268	gallstones	2006-05-11 02:13:02	1	http://www.niddk.nih.gov
1268	ozark horse blankets	2006-03-01 17:39:28	8	http://www.blanketsnmore.com

(Source: AOL Query Log)

- Implicit feedback for web search ranking
- Query spelling correction
- Query suggestion and refinement
- Automatic monitoring and evaluation
- Web search personalization

What degree of granularity of query log do we need?

Query Log Privacy

- Type of sensitive information
 - Personal
 - Business
- Owner of sensitive information (subject entity)
 - Query user
 - Third-party
- Sensitive information in query logs
 - Identifying information
 - Financial, medical, political information

AnonID	Query
142	westchester.gov
142	space.comhttp
142	207 ad2d 530
217	lottery
217	ameriprise.com
217	buddylis
217	lottery
217	ask.com
217	weather.com
993	chasebadkids.net
1268	John Smith
1268	gall stones
1268	gallstones
1268	ozark horse blankets

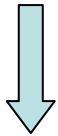
What are the potential privacy breach?

Query Log Anonymization

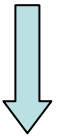
- Central challenge
 - Maximize utility for query log applications
 - Minimize potential privacy breach
- Proposed approach
 - Query log grouping (minimize identity linking)
 - Query de-identification (minimize identifying information)

Query Log Grouping

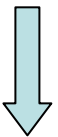
User



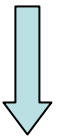
Session



Query Session



Query



Aggregate

UserID	Query
1268	John Smith
1268	gall stones
1268	gallstones
1268	ozark horse blankets

SessID	Query
234	John Smith
234	gall stones
234	gallstones
235	ozark horse blankets

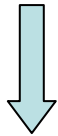
QSessID	Query
567	John Smith
568	gall stones
568	gallstones
569	ozark horse blankets

QueryID	Query
12345	John Smith
12346	gall stones
12347	gallstones
12348	ozark horse blankets

statistics

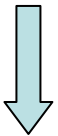
Query De-identification

No De-identification



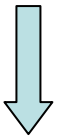
UserID	Query
1268	John Smith
1268	Atlanta
1268	gall stones
1268	gallstones

Partial De-identification



UserID	Query
1268	John Smith
1268	Atlanta
1268	gall stones
1268	gallstones

Statistical De-identification



UserID	Query
1268	John Smith
1xx8	Georgia
1xx8	gall stones
1xx8	gallstones

Full De-identification

UserID	Query
1268	John Smith
1268	Atlanta
1268	gall stones
1268	gallstones

Research Directions

- Detecting identifying information from query terms
 - Short queries
 - Rare and foreign names
 - Misspellings
- Entity mapping/linking
 - Linking query terms from multiple queries to a single entity
- Metrics for privacy and utility
 - Privacy: discernibility, information entropy
 - Utility: application dependent

Work of Interest in WWW 2007

- Privacy-enhancing personalized web search, Xu et al.
- Exposing private information by timing web applications, Bortz & Boneh
- On Anonymizing query logs via token-based hashing, Kumar et al.
- Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography, Backstrom et al.

Thank you