Using search results to microaggregate query logs semantically

Arnau Erola[1], Jordi Castellà-Roca[1]

[1] Departament d'Enginyeria Informàtica i Matemàtiques, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili

September 2013

Outline

1. Introduction
2. Background
3. Proposal
4. Results
5. Conclusions

Web Search Engines

- Information society: access to services and information anywhere anytime.
- Web Search Engines (WSE) are one of the most successful services on Internet.
  - Easy way to access the web.
  - During 2011 Google received 5000 million transactions per day.
  - All these transactions are stored in search logs.

Search logs

A standard search log from a WSE is composed of lines of the form:

(id, q, ts, r, url)

```
24963762 myspace codes  2006-05-31 23:00:52  2  http://www.myspace-codes.com
24964082 bank of america  2006-05-31 19:41:07  1  http://www.bankofamerica.com
24967641 donut pillow  2006-05-31 14:08:53
24967641 dicontinued dishes  2006-05-31 14:29:38
24969374 orioles tickets  2006-05-31 12:31:57  2  http://www.greatseats.com
24969374 baltimore marinas  2006-05-31 12:43:40
```

Search logs utility

- Personalization.
  - Results relevant to the users.
    - 68% clicks in the first page.
    - 92% clicks within the first three pages.

Search logs utility

- Personalization.
  - Results relevant to the users.
    - 68% clicks in the first page.
    - 92% clicks within the first three pages.
  - Disambiguation.

Example

mercury

Search logs utility

- Personalization.
  - Results relevant to the users.
    - 68% clicks in the first page.
    - 92% clicks within the first three pages.
  - Disambiguation.

Example

mercury

Search logs utility

- Personalization.
  - Results relevant to the users.
    - 68% clicks in the first page.
    - 92% clicks within the first three pages.
  - Disambiguation.

Example

mercury

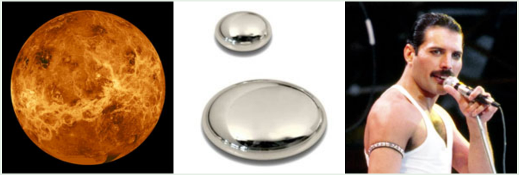
- Interests of the user and query context.

Search logs utility

- Personalization.
    - Results relevant to the users.
        - 68% clicks in the first page.
        - 92% clicks within the first three pages.
    - Disambiguation.

Example

mercury



-
    -
        - Interests of the user and query context.
    - Advertisements.
        - Google had a revenue of $43,686$ million dollars from advertisements in 2012.

Search logs utility

- Improving search.
    - Improve ranking algorithms.
    - Suggest reformulated queries.

Search logs utility

- Improving search.
  - Improve ranking algorithms.
  - Suggest reformulated queries.
- Sharing data.
  - Researchers.
    - IR algorithms, users needs, use of language in queries...
  - Marketing companies.
    - Characterize profiles, behavior and search habits, improve keyword advertising campaigns, extract market tendencies and trending topics...

Search logs privacy

- Query logs clearly contain valuable information.
- Logs can also contain personal information.

Example

- A user searched for a certain place.
- A user searched for a disease.
- A user can make a vanity query.
- Various information: *Drug Clinic in Portland*

- Queries can disclose private information about the user.

Search logs privacy

- Privacy disclosure risks
  - Identity disclosure.

    User is re-identified.
  - Attribute disclosure.

    Information about the user is retrieved.
- Main threat: link user's queries with user's identity.

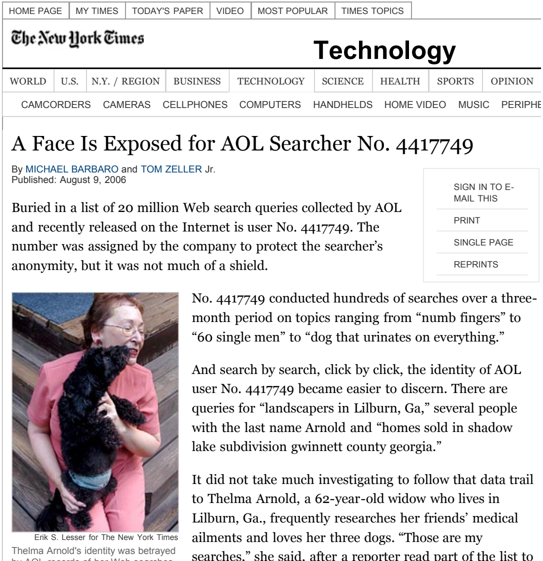
Search logs anonymization

- In order to limit disclosure risks, search logs should be anonymized: Data modifications which limit the privacy disclosure risks and reduce the data utility.
  - Utility is conditional to the ability of performing a latter analysis with the data.
  - Privacy is conditional to the ability of disclosing information about the users.
- Once anonymized, search logs can be stored or outsourced.

Search logs anonymization

- The unbounded nature of queries make it difficult to detect the sensitive information.
  - Not constitute well-defined sets of attributes (several subsets of queries could play the role of quasi-identifiers)
  - Variable length and high dimensionality.
  - Free text.
- Important trade-off between the privacy and the utility.
- Although the search logs are anonymized, there is no absolute guarantee of anonymity.

AOL



HOME PAGE | MY TIMES | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS
The New York Times
Technology
WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION
CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHE
SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS

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."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "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., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches

Background

- Deletion of specific queries or logs.
  - Remove infrequent queries.
  - Select the queries to preserve an acceptable degree of privacy
  - Choose the publishable queries.
- Microaggregation to anonymize search logs.
  - Ensures a high degree of privacy (k-anonymity).
  - Preserves some of the data utility.
  - Navarro-Arribas et al. (2009)
  - Erola et al. (2010)

Microaggregation *broadly* explained

- Microaggregation is divided in three steps:
  - Partition: create clusters of individuals.
  - Aggregation: calculate a representative individual for each cluster.
  - Replace original data by the representative.

Microaggregation *broadly* explained

- Microaggregation is divided in three steps:
    - Partition: create clusters of individuals.
    - Aggregation: calculate a representative individual for each cluster.
    - Replace original data by the representative.
- Microaggregation can be defined as an optimization problem.
    - Minimize information loss: find optimal partition.

Erola et al.

- Semantic Microaggregation: take into account semantics of the queries.
    - WSE need to know users' interests, which are represented by queries' semantics.

Erola et al.

- Semantic Microaggregation: take into account semantics of the queries.
  - WSE need to know users' interests, which are represented by queries' semantics.

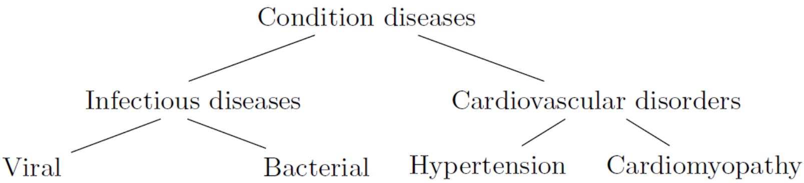
Example

Freddie Mercury VS. Queen Singer

  - Utility of search logs is related to the preservation of user's interests.

Erola et al.

- Interpret query terms in ODP (Open Directory Project) in order to extract their semantics.
- ODP is distributed data base of Web content classified by humans.
- Hierarchically structured in categories.



Condition diseases

Infectious diseases

Cardiovascular disorders

Viral

Bacterial

Hypertension

Cardiomyopathy

Erola et al.

- Similarity coefficient ODP_{sim} between two given users u_i and u_j:

$$OPD_{sim}(u_i, u_j) = \sum_{l=1}^{L} \{|c_l| : c_l \in \{C_l(u_i) \cap C_l(u_j)\}\}$$

- The representative is composed of random queries of all query logs in the cluster.

Erola et al.

- Drawbacks:
  - Fail to retain their meaning of the complex queries with several words or nouns. *For instance: water sports.*

  - Fail to retain the meaning of various nouns. *For instance: windsurfing in the Mediterranean.*
  - The size of the hierarchy is limited: although we find a classification in ODP it can be a non precisse one.
  - We need to disambiguate queries again.

Erola et al.

- Drawbacks:
  - Fail to retain their meaning of the complex queries with several words or nouns. *For instance: water sports.*

  - Fail to retain the meaning of various nouns. *For instance: windsurfing in the Mediterranean.*
  - The size of the hierarchy is limited: although we find a classification in ODP it can be a non precisse one.
  - We need to disambiguate queries again.
- All them are cause of the interpretation of the queries on the knowledge base.

Our proposal

- We consider that selected results can better represent the users' interests.
- We propose a microaggregtion method that uses selected results insted of queries in order to anonymize the data.
    - We use ODP as metric space.
    - In this way, we can compare our proposal with Erola et al. proposal.

Our proposal

- Our proposal is divided in four steps:
  - Search results for all the queries in a WSE.
    - We select the first result.
  - Classify selected results in ODP tree.
  - Partition: we use ODP_{sim}.
  - Aggregation: calculate a representative for each cluster.

- We select 840 users from the AOL files, which correspond to $400,000$ queries, to test our proposal.
- We compare our proposal with Erola et al. proposal and a random microaggregation.

Evaluation

Coefficient	**Formula**
Jaccard	$\frac{\lvert S_1 \cap S_2 \rvert}{\lvert S_1 \cup S_2 \rvert}$
Sokal and Sneath	$\frac{\lvert S_1 \cap S_2 \rvert}{2 \times (\lvert S_1 \rvert + \lvert S_2 \rvert) - 3 \times \lvert S_1 \cap S_2 \rvert}$
Dice	$\frac{2 \times \lvert S_1 \cap S_2 \rvert}{\lvert S_1 \rvert + \lvert S_2 \rvert}$

Table : Similarity coefficients between two sets S_1 and S_2

Evaluation

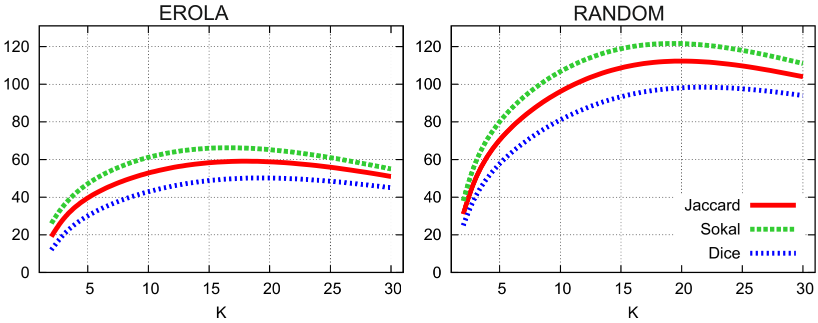
- **Utility**

$$Utility_{Coef_y}(C_{prop}, C_x) = \frac{\sum_{i=1}^{n} Coef_y(C_{orig}(u_i), C_{prop}(u_i))}{\sum_{i=1}^{n} Coef_y(C_{orig}(u_i), C_x(u_i))} - 1$$

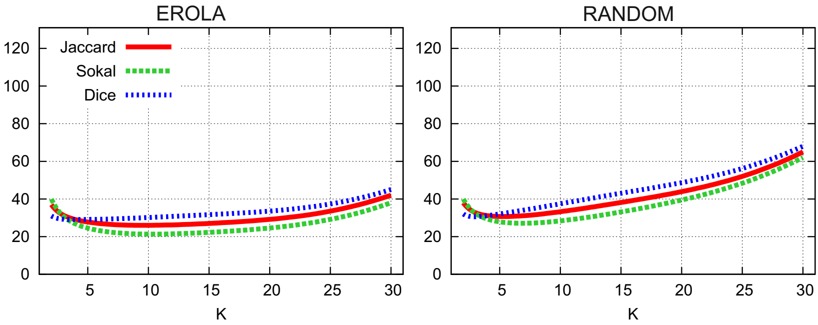
- **Disclosure risk**

$$Linkability_{Coef_y}(Q_{prop}, Q_x) = \frac{\sum_{i=1}^{n} Coef_y(Q_{orig}(u_i), Q_x(u_i))}{\sum_{i=1}^{n} Coef_y(Q_{orig}(u_i), Q_{prop}(u_i))} - 1$$

Results: Utility improvement



Results: Linkability reduction

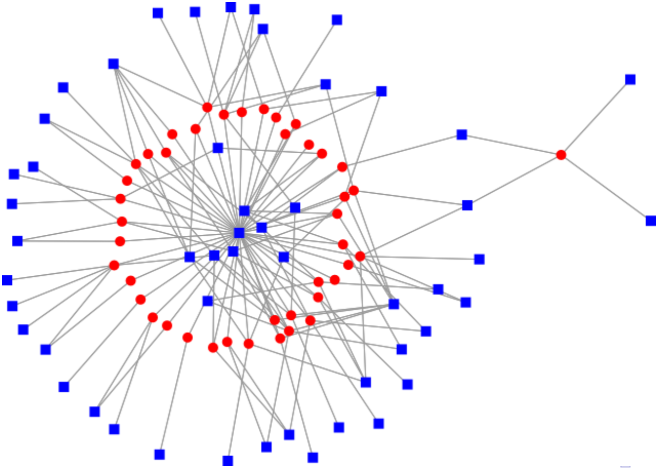


Conclusions

- The present work studies the maximization of the utility in search logs anonymization.
- We propose a microaggregation method that uses the selected results to interpret the users' interests.
- We compared our proposal with the Erola et al. proposal and a random microaggregation. Results shows that using the selected results:
  - Information loss is reduced.
  - The record linkage is reduced.
- Search results can better represent the users' interests.

Conclusions

- Alternative representation: a bipartit graph.



Thanks for your attention.

Using search results to microaggregate query logs semantically

Arnau Erola[1], Jordi Castellà-Roca[1]

[1] Departament d'Enginyeria Informàtica i Matemàtiques,
UNESCO Chair in Data Privacy, Universitat Rovira i Virgili

September 2013