Query Log Anonymization

Recuperação de Informação

Doutoramento em Engenharia Informática e Computadores

José Portêlo – 52096

Instituto Superior Técnico
Universidade Técnica de Lisboa

Outline

Introduction (1)

Search engines collect detailed information on query and result click logs.

Introduction (1)

Search engines collect detailed information on query and result click logs.

This large amount of information offers immense opportunities for the development and improvement of Information Retrieval techniques.

Introduction (1)

Search engines collect detailed information on query and result click logs.

This large amount of information offers immense opportunities for the development and improvement of Information Retrieval techniques.

Therefore, it should be made publicly available to researchers around the world.

Introduction (2)

Problem: query logs contain sensitive information about search engine users.

Introduction (2)

Problem: query logs contain sensitive information about search engine users.

Query Log Anonymization

The art of maximizing the utility of the information presented in query logs while preserving individual privacy.

Motivation

AOL released in 2006 a large query log extracted over 3 months, containing 20 million queries from ∼650k users.

Using only the information in the query logs, the New York Times was able to match some user IDs to real persons.

Although it is difficult to quantify how many people could be identified this way, the mere possibility raises some important questions.

Data availability

A useful query log contains a user ID (or user session ID), query terms, time stamp, URLs of clicked results and the results positions.

Availability of these logs grants academic researchers access to relevant and realistic data for performing information retrieval experiments.

Applications

- Implicit feedback for web search ranking
  Requires session level information.
- Query spelling correction
  Requires only aggregated query frequency statistics.
- Query suggestion and refinement
  Requires session level information.
- Automatic monitoring and evaluation
  Requires session level information.
- Web search personalization
  Requires persistent user ID across sessions.

Sensitive information on query terms

- Information ownership
  - Personal information
  - Business information
- Subject entity
  - Query user
  - Third-party individual
- Type of sensitive information
  - Identifying information
  - Financial information
  - Health information
  - Political viewpoints

Privacy breach of query logs

An adversary who has access to query logs will try to mine it for personal information, either regarding the users who performed the queries or third-party entities.

Query terms in a single query or in a single session are likely to be related to each other, so any identifying information in one query or one session is likely to be related to the same entity.

However, identifying information from two different sessions are less likely to be related to the same entity than if they were found in the same session.

Query log anonymization

Two possible ways to anonymize query logs:

- Query log grouping
  - Different levels of granularity regarding query entry information from both application and privacy perspectives.
- Query de-identification
  - Different levels of granularity regarding anonymity strictness from both data privacy and data utility perspectives.

Query log grouping (1)

- User

  Keep both user ID and session ID.
  Provides maximal utility of the data.
  High chance of linking the data to the query user.

- Session

  Remove user ID, group queries with a unique session ID.
  Some chance of linking the data to an identifiable entity.
  Less chance of linking the data to the query user itself.

Query log grouping (2)

- Query session

  Combine set of clicks and follow-up queries (click sequence).
  Increased privacy, as very little data links to a specific entity.

- Query

  Only individual queries are preserved.
  Provides similar level of privacy to query session approach.
  Sacrifices data utility because of lack of query sequences.

- Aggregate

  Keep only aggregate and statistics information.
  Provides maximal privacy.
  Applications that do not require user or session information.

Query de-identification

- Full de-identification

  All identifiers are removed.
  The remaining information cannot be used for identification.
  User ID can be preserved.

- Partial de-identification

  Remove only direct identifiers.
  May be subjected to data linkage attacks.

- Statistical de-identification

  Interesting trade-off between data privacy and usefulness.
  Different approaches: k-anonymity, l-diversity, t-closeness.

- No de-identification

  Provides maximal query log coverage by sacrificing privacy.

Research Directions

- Detecting identifying information

  Pattern detection techniques fail because of lack of context.
- Entity mapping

  Map relevant query terms to the same entity efficiently.
- Metrics for privacy and utility

  Determine if a certain approach provides sufficient privacy.
  Data usability must be measured with the applications.