

Experiment with the AOL search-query dataset

Marek Kumpošt

<http://www.BUSLab.org>

Data mining workshop – Cikháj

Faculty of informatics
Masaryk university
Brno



Introduction and the story

- AOL released a list of 21 million web search queries on 1. August 06
- Online version <http://www.aolsearchdatabase.com>
- Focused on 658 000 subscribers
- Search queries during a three-month period
- UserIDs were anonymized
- Released on AOL Research site – for academic purposes
- Examples of queries:
 - ▶ find family by social security number
 - ▶ how to secretly poison your ex
 - ▶ learning to be single

Introduction and the story

- AOL released a list of 21 million web search queries on 1. August 06
- Online version <http://www.aolsearchdatabase.com>
- Focused on 658 000 subscribers
- Search queries during a three-month period
- UserIDs were anonymized
- Released on AOL Research site – for academic purposes
- Examples of queries:
 - ▶ find family by social security number
 - ▶ how to secretly poison your ex
 - ▶ learning to be single
- Allows for user profiling – e.g. AOL user 311045 possibly owns a Scion XB automobile in need of new brake pads. User is possibly a Florida resident. . .
- User 710794 is possibly an overweight golfer, owner of a 1986 Porsche 944 and 1998 Cadillac SLS, and a fan of University of Tennessee Basketball team.

Identification of a real person

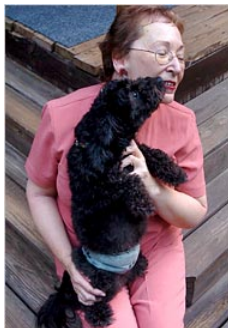
Full identification of a real individual

User No. 4417749 (Thelma Arnold) was identified

Examples of her queries:

- 60 single men
- dog that urinates on everything
- landscapers in Lilburn, Ga
- dogs-related queries

She agreed to discuss her searches with a reporter and was shocked to hear that AOL had saved and published her searches.



Identification of a real person

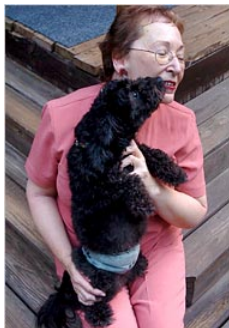
Full identification of a real individual

User No. 4417749 (Thelma Arnold) was identified

Examples of her queries:

- 60 single men
- dog that urinates on everything
- landscapers in Lilburn, Ga
- dogs-related queries

She agreed to discuss her searches with a reporter and was shocked to hear that AOL had saved and published her searches.



How many times did you search your name with Google? :-)

AOL has apologized

- “An innocent attempt to reach out to the academic community”
- “Analysis with new research tools”
- Privacy advocates said that identification will be possible (even if the data is pseudonymized)
- Under normal conditions data is stored for one-month period
- Three-month period was for the purposes of this disclosure only
- AOL has apologized for this disclosure
- This type of thing will never happen again
- Search-queries is a very sensitive information
- How about the other search engines?

How to protect users from search-query profiling

- Do not use any authenticated search engines
- Use several search engines
- Hide your searches – dummy search queries
 - ▶ TrackMeNot
 - ▶ Firefox browser plugin
 - ▶ Support for AOL, Yahoo!, Google and MSN
 - ▶ Periodically issues randomized search-queries
 - ▶ Hides users' actual search trails in a cloud
 - ▶ Static word list has been replaced with a dynamic query mechanism
 - ▶ 'Evolves' each client (uniquely) over time
 - ▶ <http://mr1.nyu.edu/~dhowe/trackmenot/>

How to protect users from search-query profiling

- Do not use any authenticated search engines
- Use several search engines
- Hide your searches – dummy search queries
 - ▶ TrackMeNot
 - ▶ Firefox browser plugin
 - ▶ Support for AOL, Yahoo!, Google and MSN
 - ▶ Periodically issues randomized search-queries
 - ▶ Hides users' actual search trails in a cloud
 - ▶ Static word list has been replaced with a dynamic query mechanism
 - ▶ 'Evolves' each client (uniquely) over time
 - ▶ <http://mr1.nyu.edu/~dhowe/trackmenot/>
- Drawbacks of TrackMeNot
 - ▶ If government wants to know who's been searching for "al Qaeda"
 - ▶ 1673 search terms in the dictionary
 - ▶ One search every 12 seconds – randomly chosen pair
 - ▶ Additional traffic – about 60 MB/day

Our experiment with the dataset

- Goal of the experiment
 - ▶ Identification based only on query dataset
 - ▶ Concentrate on reidentification
 - ★ Training and testing datasets
 - ▶ Try to find some reidentification rate
 - ★ User profiles based on N queries
 - ▶ How often users have to change their pseudonym
 - ▶ ... to keep themselves nonreidentifiable
 - ★ After N queries a user should change to a new user ID
 - ▶ Manipulate with the N threshold

Our experiment with the dataset

- Goal of the experiment
 - ▶ Identification based only on query dataset
 - ▶ Concentrate on reidentification
 - ★ Training and testing datasets
 - ▶ Try to find some reidentification rate
 - ★ User profiles based on N queries
 - ▶ How often users have to change their pseudonym
 - ▶ ... to keep themselves nonreidentifiable
 - ★ After N queries a user should change to a new user ID
 - ▶ Manipulate with the N threshold
- Input data
 - ▶ AOL search query dataset
 - ★ user pseudonym, query terms, date, returned URL, page rank
 - ★ grouped by user IDs, ordered by time
 - ▶ List of English “stop words”

Work done so far

- Semantic database extracted from a usenet of 26 topics
- Probability for each word to occur in the 26 groups
- Build the Vectors of Interest (VOI) of each query in the database
- Build the VOIs for all users and their queries
 - ▶ Resulting vector is the user's profile
- Used with all ten parts of the AOL data – evaluation of the similarities
 - ▶ Cosine similarity used (compare the similarity of two vectors)
 - ▶ 1 – same; 0 – different; -1 – oposite
 - ▶ Threshold set to 0.9
- Results
 - ▶ 440 users have no similar data (distinguishable among the others)
 - ▶ 2315 users have 55% similarity to the others
 - ▶ So far, very few users are clearly identifiable
 - ▶ Most users are pretty similar to other user profiles
 - ▶ All users have pretty similar interests: computers, girls...

Steps of the experiment

- 1 Create word list of all words in the dataset
 - ▶ Separate the dataset (training, testing) – same users, different time
 - ▶ List all words in the training set, delete all the “stop words”
- 2 Build up query-key-words matrix
 - ▶ For each query – probabilities of query words occurrences
 - ▶ Size of each vector = number of words in the word list
- 3 Build up user profiles
 - ▶ Based on 2.
- 4 Compare a testing profile with all other profiles
 - ▶ Find the most similar one
 - ▶ Cosine similarity
 - ▶ Cluster analysis
- 5 Find the optimal value of N
 - ▶ Check what happens for different values of N
 - ▶ With small N reidentification becomes harder

Any questions or suggestions?

Thanks for your attention!