I Search, Therefore I Am
Defining Identity in Anonymous Search Logs

Michael Pennisi
05-834: Applied Machine Learning
Instructor: Carolyn Rose
Carnegie Mellon University

December 4, 2009

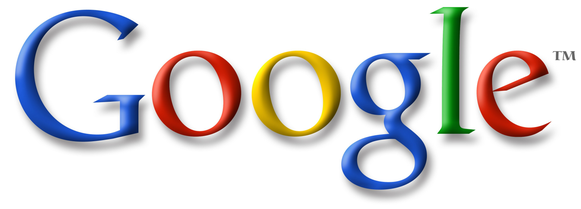
MOTIVATION

Source: http://tsvetabah.wordpress.com/2008/09/14/feeling-alone-world-wide-web-will-connect-you-with-everyone-everywhere-and-in-every-way-you-want/img_27761_aol-logo/

In 2006, the US government demanded search query logs from many search engines. AOL not only provided this data but made it publicly available on the web. After numerous complaints, AOL ceased hosting the logs. At this point, numerous mirror sites had begun hosting the data, and it is freely available to this day.

Machine learning is well-suited to finding trends in search habits. By investigating these trends, we may begin to understand how well our searching behavior identifies us.

Today, over 10 million people have accounts with the search engine Google[1]. Similar to AOL, Google logs each user's request along with identification information. Using similar machine learning methods, search engines like these can build profiles of registered users. This research may also show how search engines might "recognize" users, even when they are not logged in.



Source: http://www.psi.toronto.edu/~inmar/wiml/2009/mailinglist.php

THE DATA SET

20 Million Queries.

650 Thousand Users.

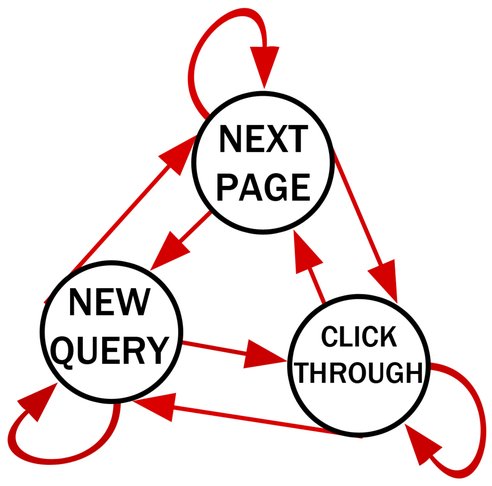
The dataset contains 20 million queries from 650 thousand anonymous users. Each query contains the following information: Anonymous ID, Query, Query Time, Item Rank, and Clicked URL. "Item Rank" and "Clicked URL" are only present in queries where the user clicked a search result.

Anonymous ID
Query
Query Time
Item Rank
Clicked URL

530	2708	bravin home services	3/2/2006 18:27	1 http://www.servicemagic.com
531	2708	how to humiliate someone	3/3/2006 7:47	1 http://www.rateitall.com
532	2708	cd's to order and buy	3/3/2006 16:47	4 http://www.musichristian.com
533	2708	12 cd's for the price of one	3/3/2006 16:49	
534	2708	bill me pay later for cd's	3/3/2006 16:50	9 http://www.shopping.com
535	2708	bill me pay later for cd's	3/3/2006 16:52	20 http://forums.dealofday.com
536	2708	scams to play on people	3/3/2006 16:53	3 http://itim.tamu.edu
537	2708	playing tricks on someone	3/3/2006 16:55	
538	2708	how humiliate someone	3/3/2006 16:56	4 http://www.iron-rose.com
539	2708	how humiliate someone	3/3/2006 17:03	
540	2708	how humiliate someone	3/3/2006 17:04	10 http://www.sysopt.com
541	2708	how to make someone misreable	3/3/2006 17:05	2 http://www.sketchyorigins.com
542	2708	how to make someone misreable	3/3/2006 17:05	5 http://www.ojar.com
543	2708	how to make someone misreable	3/3/2006 17:05	10 http://www.depnet.ae
544	2708	how to drive someone crazy	3/3/2006 17:09	1 http://www.23nlpeople.com
545	2708	how to drive someone crazy	3/3/2006 17:09	2 http://www.answers.com
546	2708	how to drive someone crazy	3/3/2006 17:14	
547	2708	how to get revenge on an old lover	3/3/2006 17:15	1 http://www.urdumped.co.uk
548	2708	how to get revenge on an old lover	3/3/2006 17:15	9 http://www.moviecall.org
549	2708	how to get revenge on an old lover	3/3/2006 17:19	20 http://www.sfgate.com
550	2708	i hate my ex boyfriend	3/3/2006 17:20	2 http://www.angry.net
551	2708	how to really make someone hurt for the pain they ca	3/3/2006 17:22	6 http://www.halcyon.com
552	2708	sean p parsons	3/3/2006 18:14	9 http://www.seanparsons.net
553	2708	things to send to emails that are free	3/3/2006 18:25	5 http://www.bored.com

FEATURE SPACE DESIGN

To get a better view of search patterns, each user's queries were collapsed into sessions. Previous work has defined sessions based on query similarity [Janson et al., 1998], but other approaches consider only the time between queries [Silverstein et al., 1998]. Because the latter method is able to capture sessions that "wander", a time-based approach was chosen. Sessions were defined to include queries submitted within 15 minutes of each other.



Sessions could be modeled as finite state machines with the following states: submit new query, request more results, and select a single result ("click-through"). The user's transitions between these states could be modeled with a vector of probabilities.

This vector was realized as a set of nine numeric attributes.

An informal study of a held-off subset of data inspired a number of other features. These include:

- Query Count
- Average Query Length
- Session Duration
- Uses Question Words
- Uses URLs
- Searched for Another Engine
- Tried Another Engine
- Gave Up (session ended without a click-through)

OPTIMIZATION

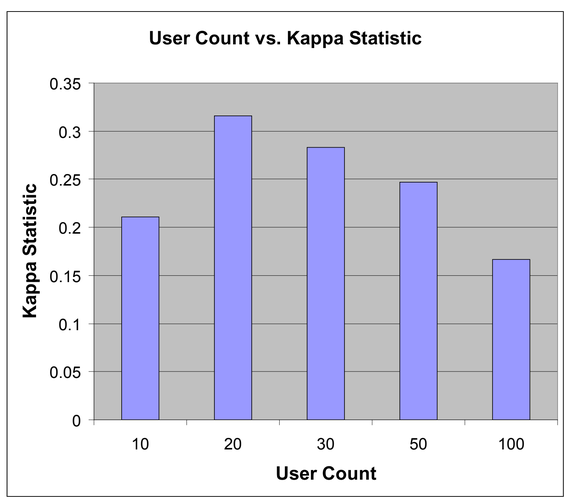
In a rough initial survey of the data, multiple machine learning algorithms were applied to a reserved subset. This survey indicated that C4.5 decision trees would be well-suited to the classification task.

The parameters of C4.5 were optimized using a subset of the data held-off for tuning purposes.

Certain parameters were specifically ignored in the optimization process. Because the dataset does not contain any nominal attributes, enabling binary splits would have no effect. Smoothing was not considered because the process of combining queries into sessions is itself a form of smoothing.

In total, three parameters of C4.5 were tuned: confidence factor, use of subtree raising, and use of pruning. In disabling pruning, the other parameters are disregarded, so seven models were trained in total.

There was no significant difference between the seven models trained, so the default configuration of the algorithm was selected:
Confidence Factor: 0.25
Subtree Raising: On
Pruning: On



SCALABILITY

Because real-life applications of such techniques would span thousands of users, models were trained and tested on datasets containing 10 users, 20 users, 30 users, 50 users, and 100 users.

As expected, performance degrades as the number of users increases. Somewhat surprisingly, better performance was achieved over 20 users than over 10 users. A possible explanation is that a dataset comprising 10 users is too easily influenced by idiosyncrasies in the data.

ERROR ANALYSIS

The probability for the "Query-to-query" transition was significantly higher among incorrectly-classified instances. This seems to indicate that users who frequently modify their query are difficult to distinguish. This may include users who struggle with creating appropriate search terms or who are prone to typos.

The probability for the "Query-to-click" transition was significantly lower among incorrectly-classified instances. This seems to indicate that users who are more skilled at using a search engine are easier to distinguish.

REFERENCES

[Silverstein et al., 1999] Silverstein, A., Marais, H., Henzinger, M., and Moricz, M. 1999 Analysis of a Very Large Web Search Engine Query Log, SIGIR FORUM, 33 (1):6-12

[Jansen et al., 1998] Jansen, B.J., Spink, A., Bateman, J., and Saracevic, T. 1998 Real Life Information Retrieval: A Study of User Queries on the Web, SIGIR FORUM, 32 (1):5-17.

[1] http://www.google.com/intl/en/press/pressrel/20090114_googleapps_reseller.html

FUTURE IMPROVEMENTS

A number of improvements might be made to improve the performance and scalability of these techniques.

To recognize how users' queries change over time, sessions could include information about the average change in their queries. This value could be computed with the simple Manhattan distance, or the more robust (and expensive) Levenshtein distance.

To recognize that users' search terms and search habits change over time (i.e. with mood, familiarity, tastes, etc.), instance-based learning could be employed. Such a technique may pick up on subtle contextual trends.

To recognize the distinctions between search meta-data and query content, co-training may be employed. This could be applied on feature space pairs such as meta-data and query content, or meta-data and click-through content.