

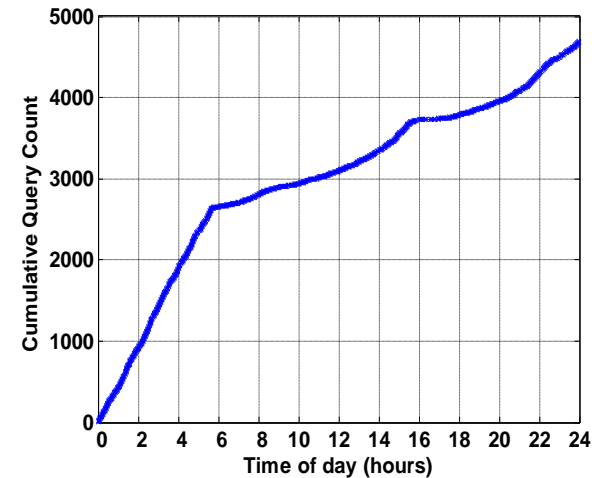
A Large-scale Study of Automated Web Search Traffic

Greg Buehrer¹, Jack Stokes² and Kumar Chellapilla¹

¹Microsoft Live Labs, ²Microsoft Research

Problem Statement

- Goal
 - Distinguish search queries as either automated or by a human
- Motivation
 - Improve QoS for humans
 - Increase/improve data for relevance
- Caveats
 - Currently next-day analysis
 - Requires sessionization
 - Currently on a per-day basis, could analyze over longer time periods



Why automate query traffic?

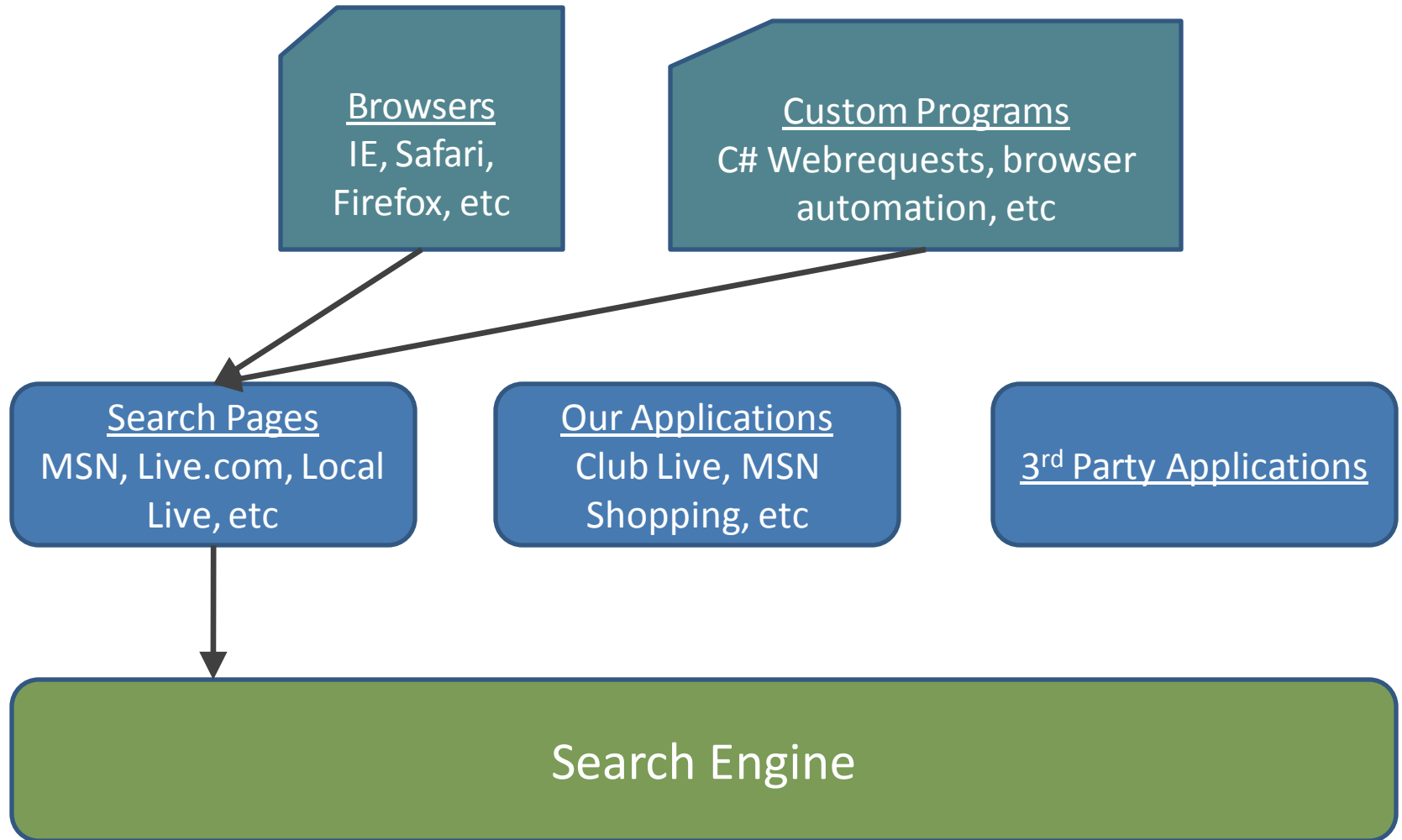
- To collect information
 - About the search engine
 - SEOs will query to check URL presence, rankings, find low result queries
 - For personal gain
 - Easy stock quotes, business news, etc
 - Scrape for email addresses, phone numbers, good spam queries
- To commit click fraud
 - Click on ads of competitors

Exploring the Query Logs

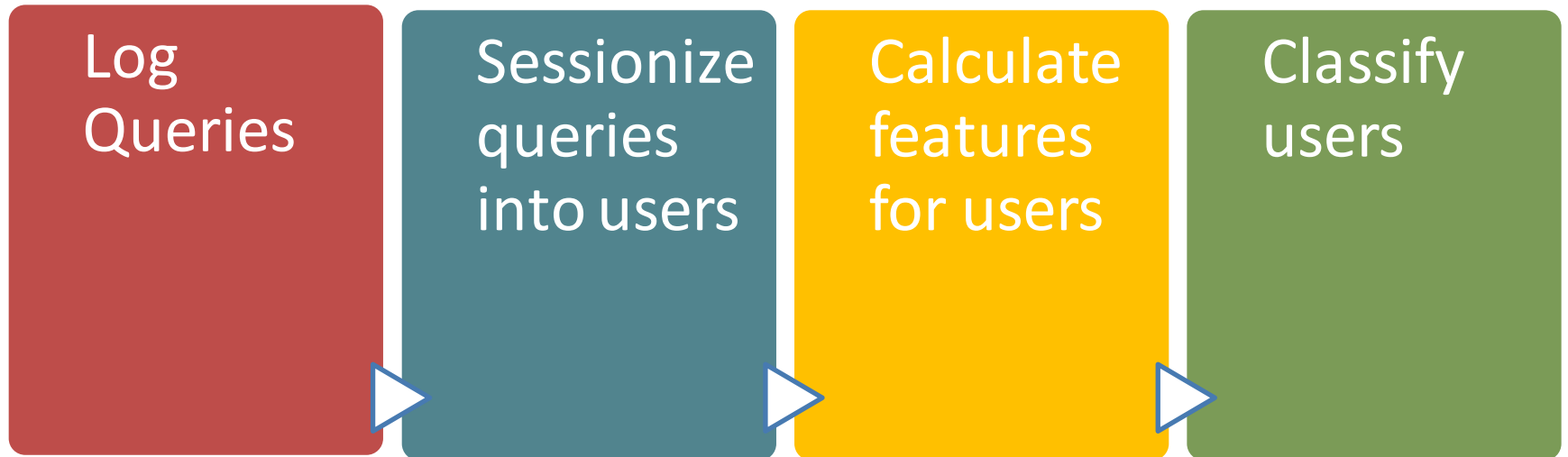
Top Queries of the Day

1. ""
2. "google"
3. "yahoo"
4. "fire+department+-location%3ajp"
5. "youtube"
11. "microsoft"

Search Traffic Flow



Query Stream Classification Process



Focus of this paper

Feature Set

- Physical limits - time and space bound
 - Volume
 - Number of queries, clicks, etc (sustained)
 - Rate
 - Maximum interactions in a small time frame
 - Space
 - Distinct locations in a given time frame
- Behavioral Signals
 - Entropy/chaos bound
 - Entropy of keywords, lengths, temporal ordering, periodicity, query category
 - Signatures
 - Spam score of keywords, adult score of keywords, etc
 - CTR, dwell time, etc
 - Blacklisted IPs, User agents, locales, etc

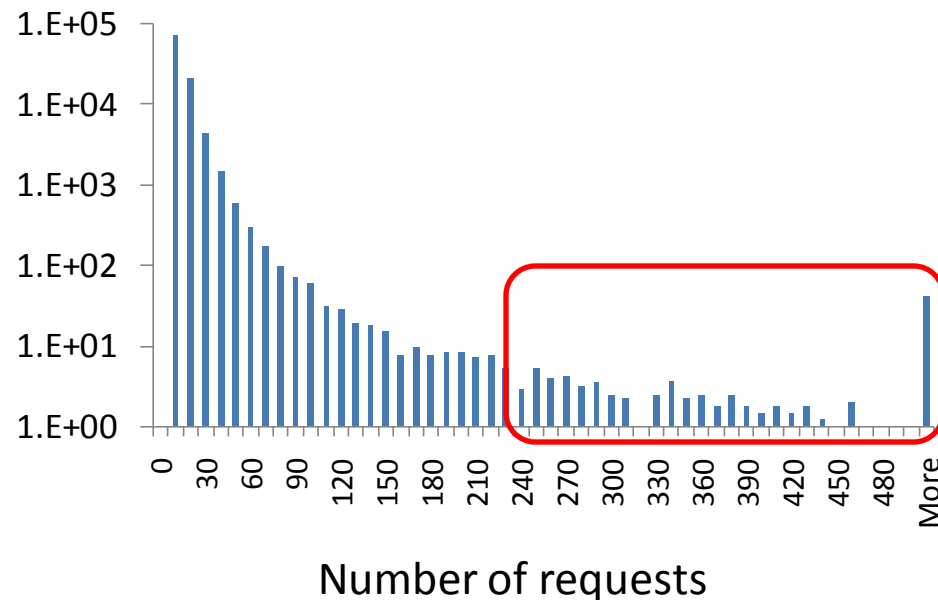
Features are simple calculations and require little time for full data

Data Set

- First we sampled 100M requests (all requests for a chosen user are included using cookies)
- Then we pruned it to those that had at least 5 interactions, totally 46M requests

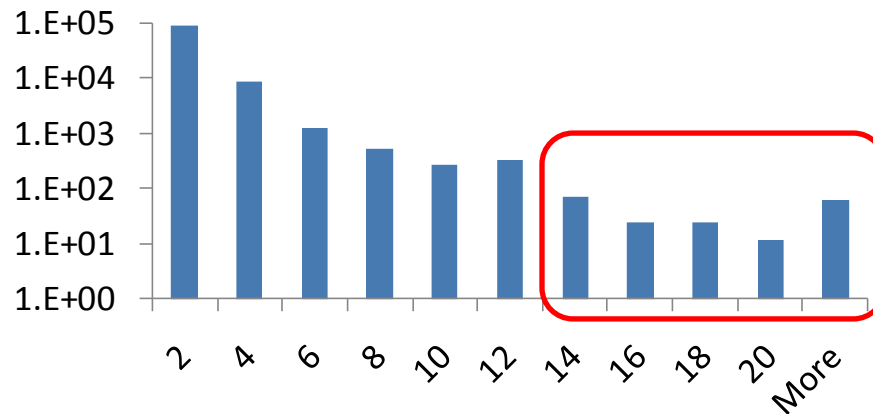
PL: Volume

- Total Requests, Queries, Clicks, Keywords, etc
 - Most discriminating feature class
 - One user queried for “mynet” 12,061 times



PL: Rate

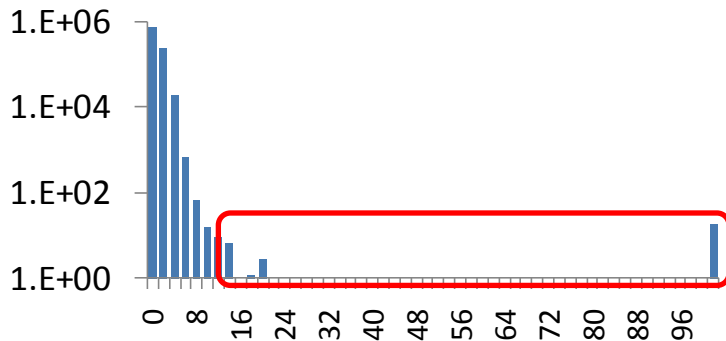
- Number of events per (small) time period
 - Requests, clicks



Max number of requests per 10 second period

PL: Geography

- Distinct IP address, considering only the first two octets
 - One user had 38 different cities in 4 hrs (428 queries)



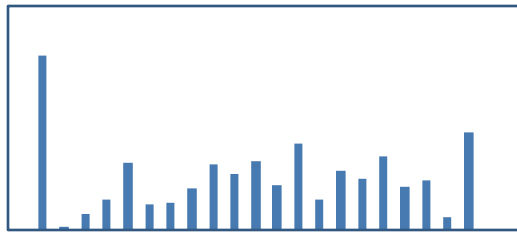
Number of IP addresses (first two octets)

Example

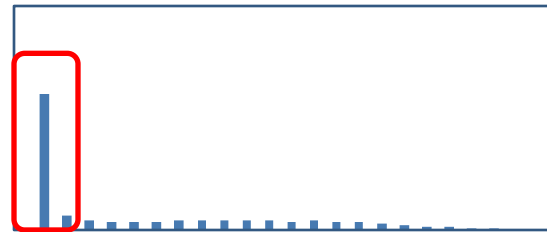
| | | |
|------------|-----|---------------------------|
| 4:18:34 AM | IP1 | Charlottesville, Virginia |
| 4:18:47 AM | IP2 | Tampa, Florida |
| 4:18:52 AM | IP3 | Los Angeles, California |
| 4:19:13 AM | IP4 | Johnson City, Tennessee |
| 4:22:15 AM | IP5 | Delhi, Delhi |
| 4:22:58 AM | IP6 | Pittsburgh, Pennsylvania |
| 4:23:03 AM | IP7 | Canton, Georgia |
| 4:23:17 AM | IP8 | Saint peter, Minnesota |

B: Click-through Rate

- Histogram for light and heavy users



Users with 5+ requests

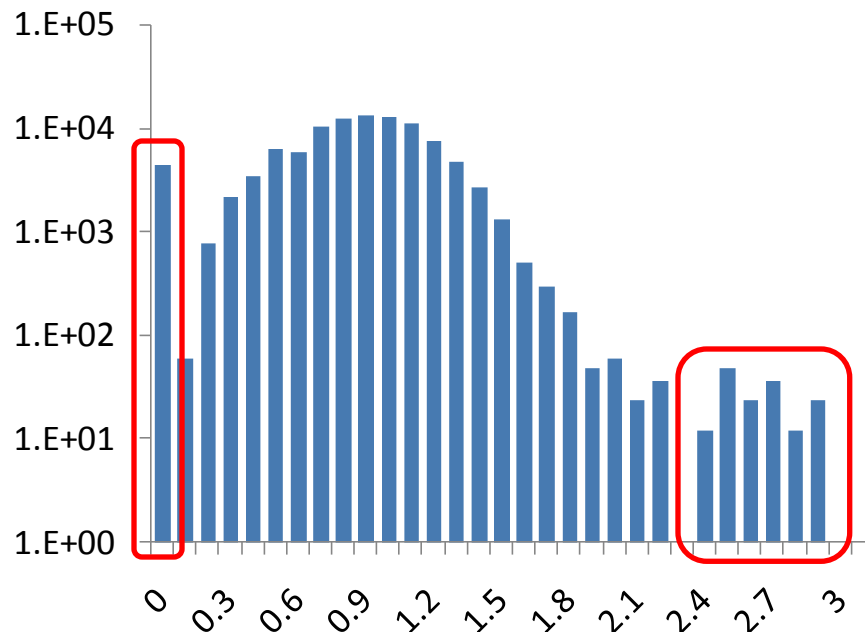


Users with 50+ requests

- Histograms show many more zero-click users when the volume is high
 - Rank checking does not require a click
 - Scraping top URLs for a query does not require a click

B: Keyword, Query Entropy

- Calculated as informational entropy where the token is either a keyword or the whole query

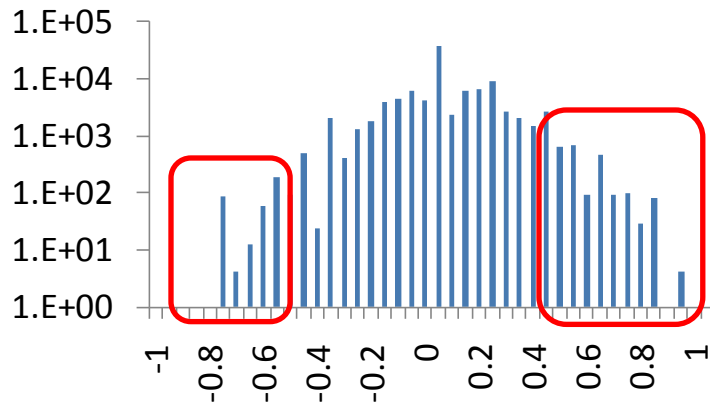


Example

06:20:59 2007 :financial+trade+cycle,
06:24:14 2007 :blue+letter+bible,
06:25:30 2007 :should+know+before,
06:27:40 2007 :individuals+cannot+adequately,
06:30:23 2007 :representing+several+bareboat+companies,
06:31:52 2007 :following+provisions+that,
06:33:22 2007 :post+jobs+with+careerbuilder,
06:34:38 2007 :edit+keyboard+shortcuts,
06:35:15 2007 :ways+consumer+knowledge+test,
06:36:28 2007 :like+writing+good+code,
06:39:19 2007 :save+money+with+road+runner,
06:41:00 2007 :featured+inquiry+logo+when+does,
06:43:03 2007 :asylum+lake+controversy,
06:44:40 2007 :introduced,
06:45:11 2007 :abdominal+wall+pathway,
06:46:51 2007 :calendars,
06:47:44 2007 :free+press+release+distribution,
06:49:25 2007 :early+double+knits+were,
07:03:27 2007 :serves+audiobook+professionals,

B: Alphabetical Ordering

- Some users issue their queries in alphabetical order



Example 1

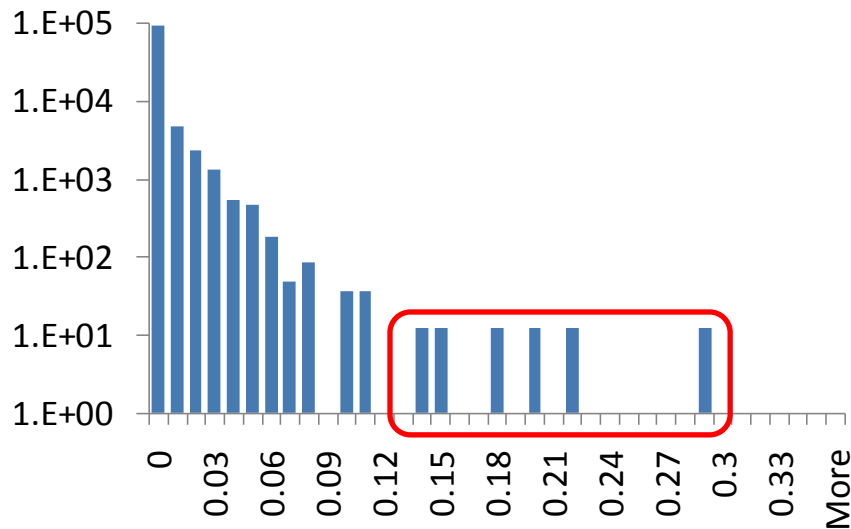
2102manpuku,
2103manpuku,
2104manpuku, ...

Example 2

http astro stanford
http adulthealth lo
http www bigdrugsto
http www cheap diet pills online ...
http www generic vi
http contrib cgi cl
http www e insaat b
http buy tramadol o
http cialis raulserrano info ciallis ...
http englishgrad cas ilstu edu files ...

B: Spam & Adult Scores

- A small dictionary of spam (or adult) keywords and weights is used (normalized sum)

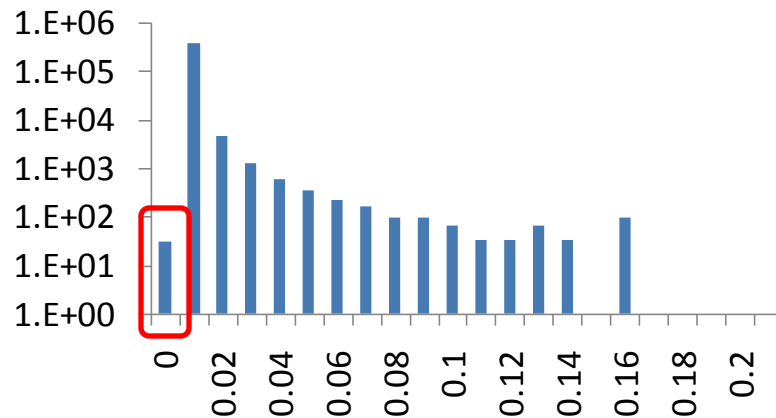


Example

Managing your internal communities
based group captive convert video from
book your mountain resort
agreement forms online
find your true love
products from thousands
mtge market share slips
mailing list archives
studnet loan bill
your dream major
computer degrees from home
free shipping coupon offers

B: Length Entropy

- Length of each keyword, length of each query

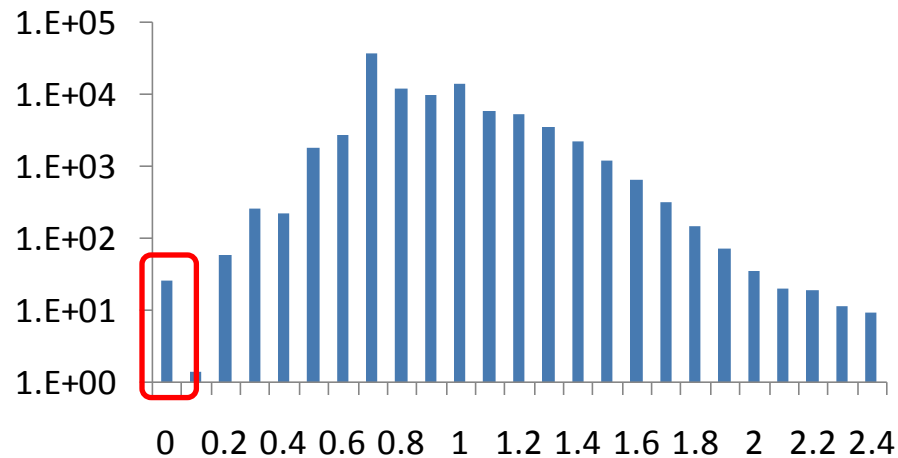


Example

| | |
|------|------|
| pae | nex |
| cln | intc |
| eu3 | tei |
| eem | wfr |
| olv | ssg |
| oj | sqi |
| lqde | nq |
| igf | trf |
| ief | cl |
| nzd | dax |
| rib | ewl |
| xil | bbdb |
| nex | csc |

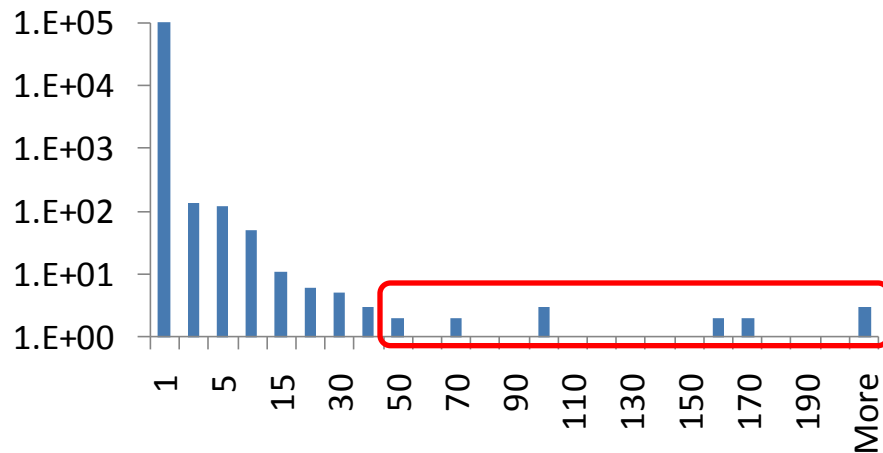
B: Query Periodicity

- Entropy of elapsed time between successive requests (or clicks, for dwell time)
 - Could also use FFT



B: Advanced Query Terms

- Scan index for “title:”, “link:”, “url:”, etc and keep a count of the total number of occurrences



Others

- Reputations
 - Use bags of values that represent black lists (or white lists) for particular fields
 - IP address
 - User agent
 - User ID (was previously tagged as automated)
 - Country code / locale
- CLR boost
 - % clr gain afforded by user Id, day, etc
- Ranks of the queries

Preliminary Classification Results

- Weka – 320 Labeled data points
 - Not chosen randomly (Active Learner)
 - Search page entry points
 - Didn't include reputations

| Classifier | TP | TN | FP | FN | % |
|-------------|-----|-----|----|----|----|
| Bayes Net | 183 | 120 | 11 | 6 | 95 |
| Naïve Bayes | 185 | 106 | 25 | 4 | 91 |
| AdaBoost | 179 | 119 | 12 | 10 | 93 |
| Bagging | 185 | 115 | 16 | 4 | 94 |
| ADTree | 182 | 121 | 10 | 7 | 95 |
| PART | 184 | 120 | 11 | 5 | 95 |

| Rank | Field |
|------|---------------|
| 1 | Query Count |
| 2 | Query Entropy |
| 3 | Max interval |
| 4 | CTR |
| 5 | Spam Score |

Mixed Signals

- It is not uncommon to have automated traffic and human traffic on the same user Id
 - 6,534 queries, first five (4 clicks) were
 - Pottery barn
 - Pottery barn kids
 - Pottery barn kids outlet
 - Pottery barn kids outlet store
 - Pier 1 ...
 - Then 6529 queries without a click (mostly blank)

Conclusion

- Feature set to distinguish between human search query traffic and automated query traffic
 - Divided into two groups, physical limits and behavioral signals
 - Initial results suggest the features can be used to classify traffic effectively

Exploring the Query Logs

Future Work

How many IP addresses have no cookies at all?

19.3M

How many of these 19.3M have < 100 queries?

19.1M

Can we sessionize these into users?

Questions?