A Brief Tour of "Query Space"

Nick Craswell

Microsoft Live Search / Microsoft Research Cambridge

Definitions

- **"Document space":** Retrieving documents, finding similarity between documents, gathering information from/about documents

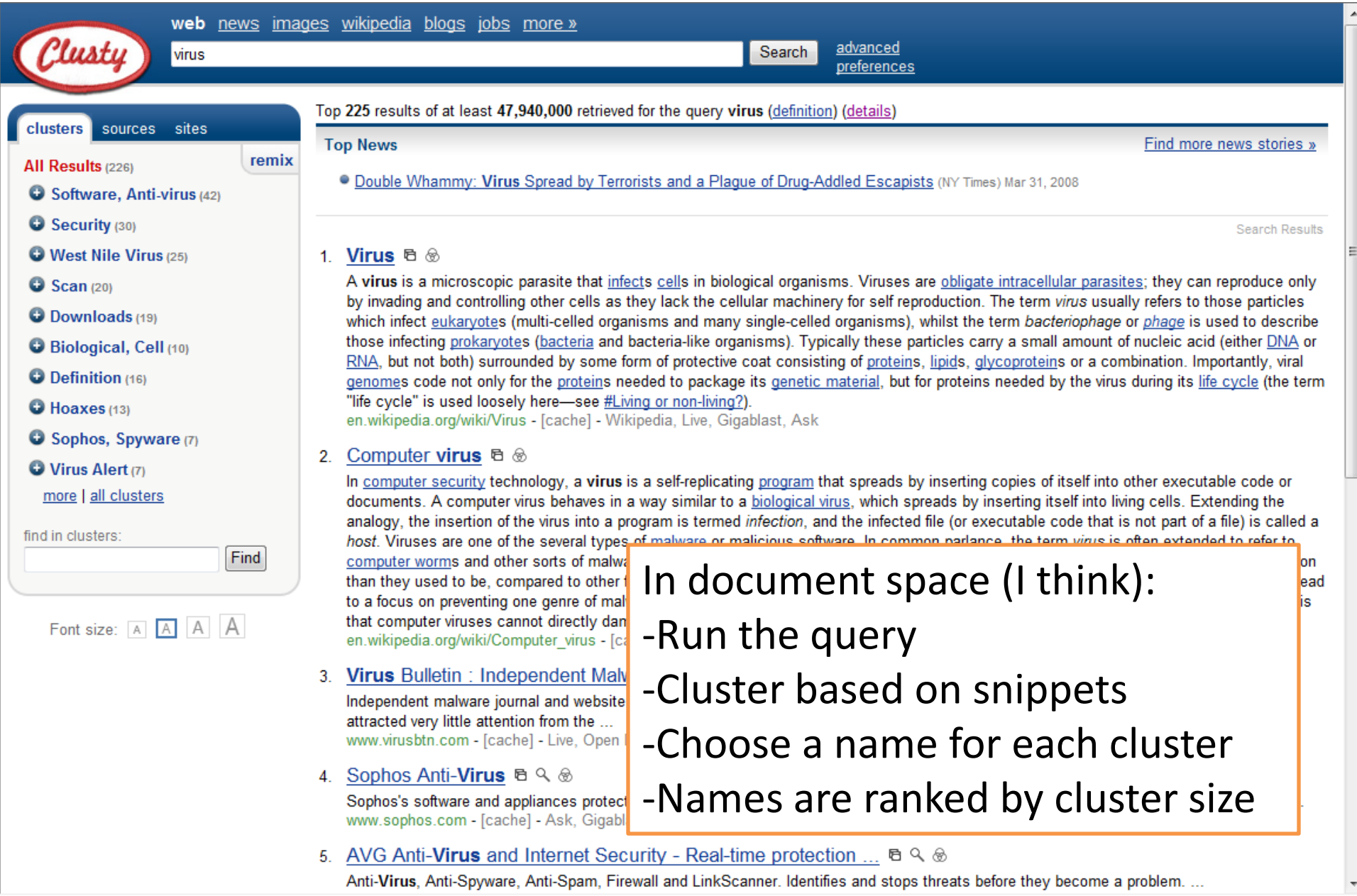
- **"Query space":** Retrieving queries, finding similarity between queries, gathering information from/about queries
  - Usage logs

Traditional IR has a focus on "Document Space"

- As seen in TREC Test Collection
  - 250,000 Documents
  - 50 Query Topics
  - Relevance judgments of query-document pairs
- Key techniques:
  - Search: Matching queries against documents
  - Clustering & dimensionality reduction in doc space
  - Query expansion: Using documents
  - "Document expansion": Using other documents

e.g. anchor text propagation

Ambiguous query ‘virus’



In document space (I think):

-Run the query

-Cluster based on snippets

-Choose a name for each cluster

-Names are ranked by cluster size

Comparison for ‘virus’

Clustering for ‘virus’ based on documents retrieved (+more?)

+ Software, Anti-virus (42)
+ Security (30)
+ West Nile Virus (25)
+ Scan (20)
+ Downloads (19)
+ Biological, Cell (10)
+ Definition (16)
+ Hoaxes (13)
+ Sophos, Spyware (7)
+ Virus Alert (7)

Popular queries containing ‘virus’

+ norton antivirus
+ antivirus
+ avg antivirus
+ free antivirus
+ west nile virus
+ norton anti virus
+ virus
+ free virus scan
+ anti virus
+ free virus protection

All I did so far is grep...

Comparison for ‘jaguar’

Clustering for ‘jaguar’ based on documents retrieved (+more?)

+ Jaguar Cars (42)
+ Photos (31)
+ Club (33)
+ Jaguar Dealership (29)
+ Cat, Panthera (16)
+ Jaguar and Land Rover (18)
+ Jacksonville Jaguars (9)
+ Apple, Mac (7)
+ Atari (8)
+ Classic Jaguar (8)

Popular queries containing ‘jaguar’

+ jacksonville jaguars
+ jaguars
+ jaguar cars
+ jaguar parts
+ jaguars.com
+ jaguar.com
+ jaguar animal
+ jaguares
+ jaguar usa
+ www.jaguar.com

grep is crude, but we can do more...

Information about “query space”

I. Query histogram: What queries are popular?

II. Session data: How do users reformulate?

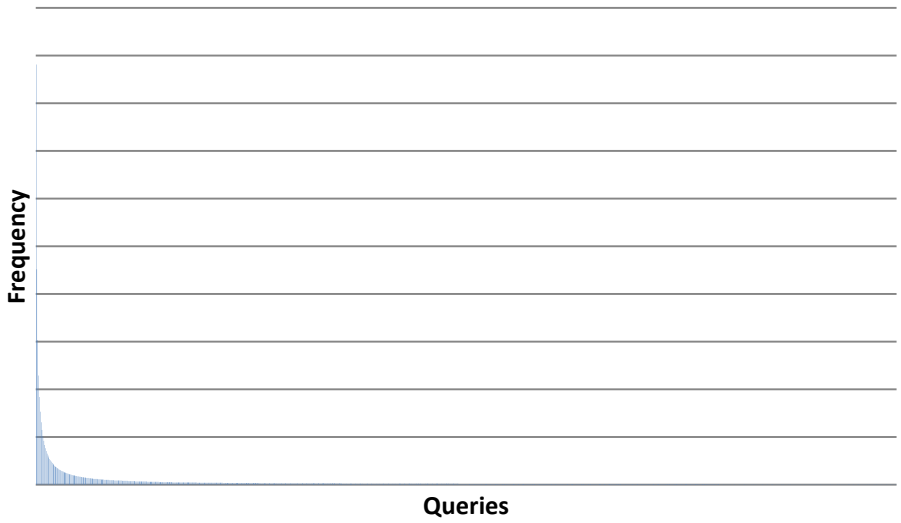
III. Co-click: What queries have same answers?

Conclusions:

a) Practitioners can use and share such data

b) Researchers, this is a fun area! (Data soon?)

I. Query Histogram



Head: ‘hotmail’, ‘jaguar’, ‘virus’, and all the queries I showed you

Tail: ‘xyz@hotmail.com’, ‘jaguar 1986model’, ‘viruses firefighting’

I think: Head queries are “better”?

Another Query Histogram Example

- Snoop Dog is a head query
  - Head: Snoop Dogg
  - Quite tail: Snoop Dogs (is rarer than Snop Dog)
- Boxer Dog is a head query
  - Head: Boxer Dogs
  - Very tail: Boxer Dogg

Weakness of query histogram: Don't really know that these queries are related
Strength of query histogram: Can find **lots** of these weak relationships...

glasgow

glasgow daily times | glasgow | glasgow rangers | glasgow herald | glasgow airport | glasgow coma scale | glasgow ky | glasgow scotland | university of glasgow | glascow coma scale | glasgow university | glasgow kentucky | glasgowdailytimes | www.glasgowdailytimes.com | glasgow city council | glasgow evening times | glasgowdailytimes.com | glasgow-ky.com | glasgow celtic | www.glasgow-ky.com | glasgow hotels | kings theatre glasgow | glasgow, ky | glasgow high school | glasgow montana | glasgow rangers fc | www.glasgow daily times | glasgow celtic fc | glasgow rangers football club | glasgow caledonian university | glasgow-ky | secc glasgow | glasgow, scotland | hotels in glasgow | the glasgow herald | new glasgow evening news | glasgow celtic football club | glasgow daily record | glascow | glasgow daily news | glasgow pram centre | glasgow montana national weather service | cineworld glasgow | glasgow,ky | map glasgow scotland | glasgow middle school | glasgow royal concert hall | celtic glasgow | glasgow international airport | glasgow fort | glasgow airport parking | glasgow, kentucky | david glasgow farragut | carling academy glasgow | glasgow daily times.com | glasgow mt | map of glasgow | glasgow school of art | glasgow map | glasgow angling centre | first bus glasgow | glasgow weather | richard glasgow | glasgow electric plant board | glasgow uni | theatre royal glasgow | glasgow coma score | arches glasgow | glasgow rangers home | hotels glasgow | the herald glasgow | glasgow, mt | glasgow caledonian | glasgow escorts | glasglow | gumtree glasgow | nights out in glasgow | glasgow independent schools | glasgow medical center | new glasgow | glasgow secc | glascow coma score | glasgow survival | glasgow courier | glasgow job service | glasgow, montana | pulitzer novelist glasgow | jobs in glasgow | pavilion theatre glasgow | flights to glasgow | new glasgow news | glasgow times | arnold clark glasgow | glasgow rangers songs | glasgow evening times uk | first community bank of glasgow | glasgow council | glasgow airport hotels | glasgow daily times .com | glasgow science centre | hilton glasgow | glasgow scale | holiday inn glasgow | www.glasgow daily record | glascow scotland | thistle hotel glasgow | new glasgow nova scotia | mctears glasgow | ugc cinema glasgow | highland cinema glasgow ky | glascow scale | glasgow montana real estate | baa glasgow | cheap hotels glasgow | abc glasgow | glasgow herald newspaper | port glasgow | glasgow mt weather | cheap west end hotels glasgow | daily record glasgow | glasgow newspaper | glasgow colleges | cheap flights to glasgow | ikea glasgow | "university of glasgow" | scotland glasgow rangers | glasgow gumtree | glasgow airport arrivals | travel lodge glasgow | university glasgow scotland | jurys inn glasgow | glasgow rangers pictures | pressure glasgow | first community bank glasgow | glasgow restaurants | glasgow airport departures | www.glasgow.gov.uk | glasgow dailey times | north glasgow college | "glasgow, dc 45 fr 45." | glasgow coma | glasgow news | glasgow missouri | glasgow daily | glasgow prestwick airport | www.glasgow.com | glasgow kentucky newspaper | glasgowguide | glasgowkentucky | glasgow chat | glasgow.com | evening times glasgow | glasgow theatres | glasgow uk | glasgowky | langs hotel glasgow | glasgow weather forecast | glasgow warriors | car hire glasgow | royal concert hall glasgow | glasgow,scotland | daily record glasgow scotland | menzies hotel glasgow | ellen glasgow | pavillion theatre glasgow | glasgow street map | glasgo | university glasgow | glasgow kentucky real estate | holiday inn glasgow airport | corus hotel glasgow | westfield way glasgow | the new glasgow news | glasgow yahoo.co.uk hotmail.co.uk aol.co.uk yahoo.com hotmail.com aol.com gmail.com | glasgow guide | glasgow estate agents | glasgow hilton | glasgow cinemas | glasgow ky daily times | glasgow jobs | glasgow, ky.com | glasgowairport | flights from glasgow to newyork | glasgow's river | glasgowrangers | glasgow barrowlands | weather glasgow scotland | city of glasgow | glasgow ky newspaper | flats to rent in glasgow | glasgow film theatre | glasgow concert hall | carlton george hotel glasgow | glasgow scotland weather | novelist glasgow | herald glasgow | millenium hotel glasgow | glasgow "city council" | glasgow furniture | odeon cinema glasgow | glasgow prestwick | weather glasgow | quality hotel glasgow | glasgow tourism | first glasgow | hilton hotel glasgow | estate agents glasgow | ugc glasgow | port glasgow scotland | glasgow royal infirmary | evening news new glasgow | dane glasgow | glasgow theatre | weather in glasgow | glasgow hotel | glasgow high school delaware | glasgow ky. | glasgow schools | street map glasgow scotland | glascow coma | glasgow ky website | glasgow chamber of commerce | glasgow scotland real estate | glasgow housing association | glasgow airport flight arrivals | airport parking glasgow | glasgow epb | strathclyde university glasgow | bed shed glasgow | city inn glasgow | glasgowmontana.com | glasgow celtic shop | crowne plaza glasgow | glasgow underground | glasgow.gov.uk | hotel glasgow | glasgow property | university of glascow | cheap hotels in glasgow | citizens theatre glasgow | argos uk glasgow | flights from southampton to glasgow | glasgow rangers screensavers | glasgowky.com | glasgow "high school" | obituaries new glasgow | cinema glasgow | secc glasgow ticket line | glasgow cathedral | colleges in glasgow | the arches glasgow | glasgow central station | glasgow gangs | glassgow |

Log data: Privacy concerns

- **Bad:** Show many queries from 1 user (and their IP address, and their email address)
  - People can be identified based on their queries [NYT]
- **Better:** Aggregate data, such as query histograms
  - Throw away user ids, IP addresses etc
  - No way of tracing 2 queries to the same user
  - But the user might have typed:
    - Their credit card number
    - Something incriminating
- **Even better:** Throw away the tail
  - Eliminate events that happened fewer than 5 times

In my work, and in this talk, we do these.

II. Session Data

- Sessions give us associations between queries:
  - User types query **A**
  - User types query **B** later in the same session
  - Record the association **A -> B**
  - Over many sessions that had A: **P(B | A)**
- Examples:
  - ,acys -> macys
  - amazon -> ebay
  - burtney spears -> ??

What happens **often** in session data?

Spelling correction

cam**brige** dictionary	cambridge dictionary

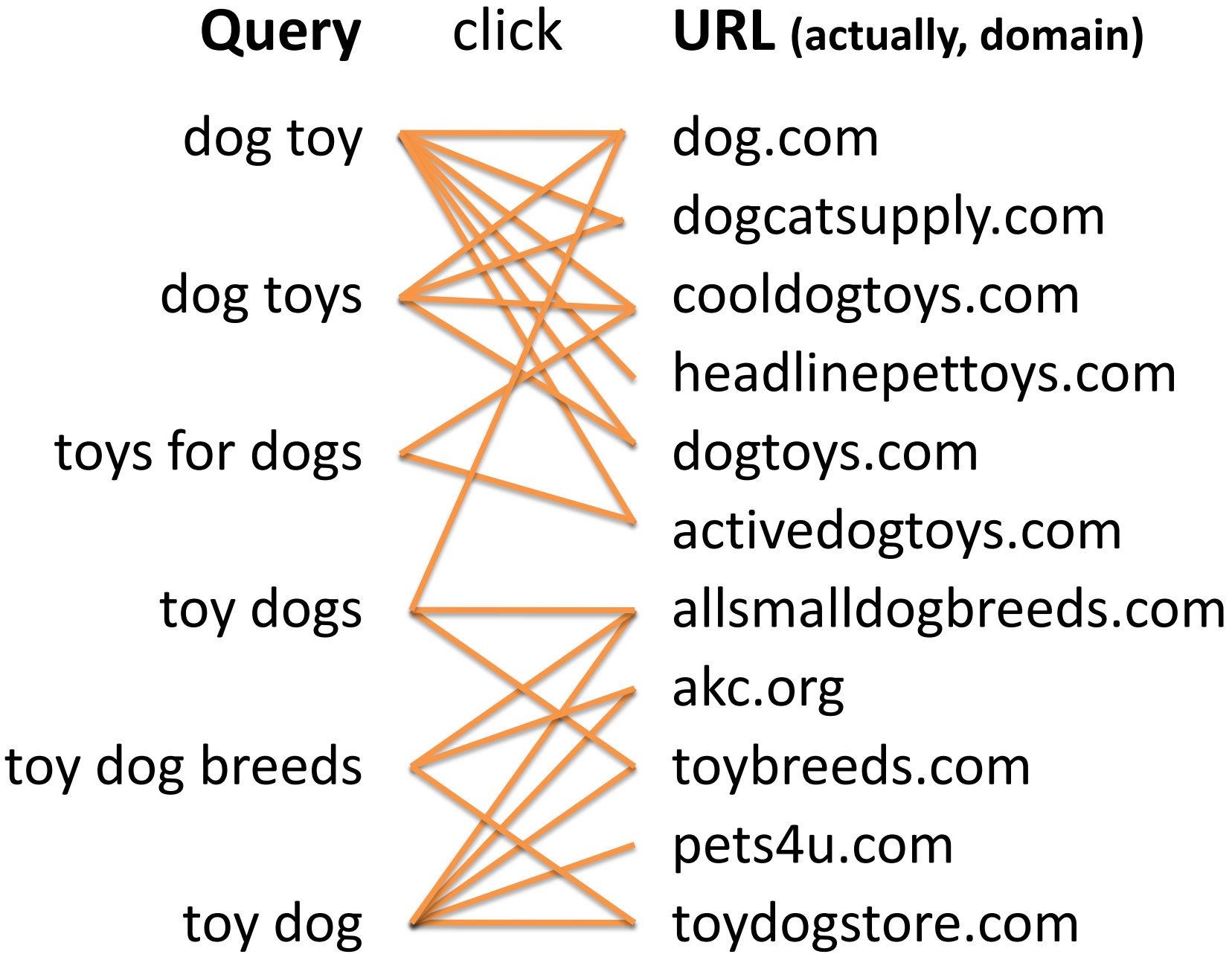
Related topics

cambridge dictionary	cambridge dictionary
cambridge dictionary	cambridge online dictionary
cambridge dictionary	dictionary
cambridge dictionary	oxford dictionary
cambridge dictionary	google
cambridge dictionary	wikipedia
cambridge dictionary	cambridge dictionary online
cambridge dictionary	webster dictionary

Useful! On the other hand it is sparse. A lot of spelling mistakes haven't been made and corrected yet. :-) Related topics are mostly within the head queries.

III. Co-click

- Another sort of association between queries
  - User types query **A** and clicks URL **U**
  - Another user types query **B** and clicks URL **U**
  - Queries A and B have similar answers
- Examples:
  - dog toy -- dog toys
  - dog toy -- toys for dogs
  - toy dog -- toy dogs
  - toy dog -- toy dog breeds



Craswell and Szummer. *Random walks on the click graph.* SIGIR 2007

What does ‘bill gates bio’ mean?

bill gates bio	bill gates biography	0.262538
bill gates bio	bill gates	0.193768
bill gates bio	bill gates bio	0.159985
bill gates bio	when was bill gates born	0.059319
bill gates bio	biography of bill gates	0.047038
bill gates bio	bill gate	0.038158
bill gates bio	bill gates net worth	0.036577
bill gates bio	richest man on earth	0.028249
bill gates bio	biography on bill gates	0.015176
bill gates bio	biography bill gates	0.013650
bill gates bio	bill gates foundation	0.011260

…

…

bill gates bio	gates	0.011112
bill gates bio	bill gates facts	0.010235
bill gates bio	bill gates home	0.009206
bill gates bio	bill and melinda gates	0.008341
bill gates bio	bill	0.006457
bill gates bio	pictures bill gates	0.006101
bill gates bio	bill gates pictures	0.005893
bill gates bio	bill gates house	0.005878
bill gates bio	bill and melinda gates foundation	0.005672
bill gates bio	info on bill gates	0.005583

What does ‘amazon’ mean?

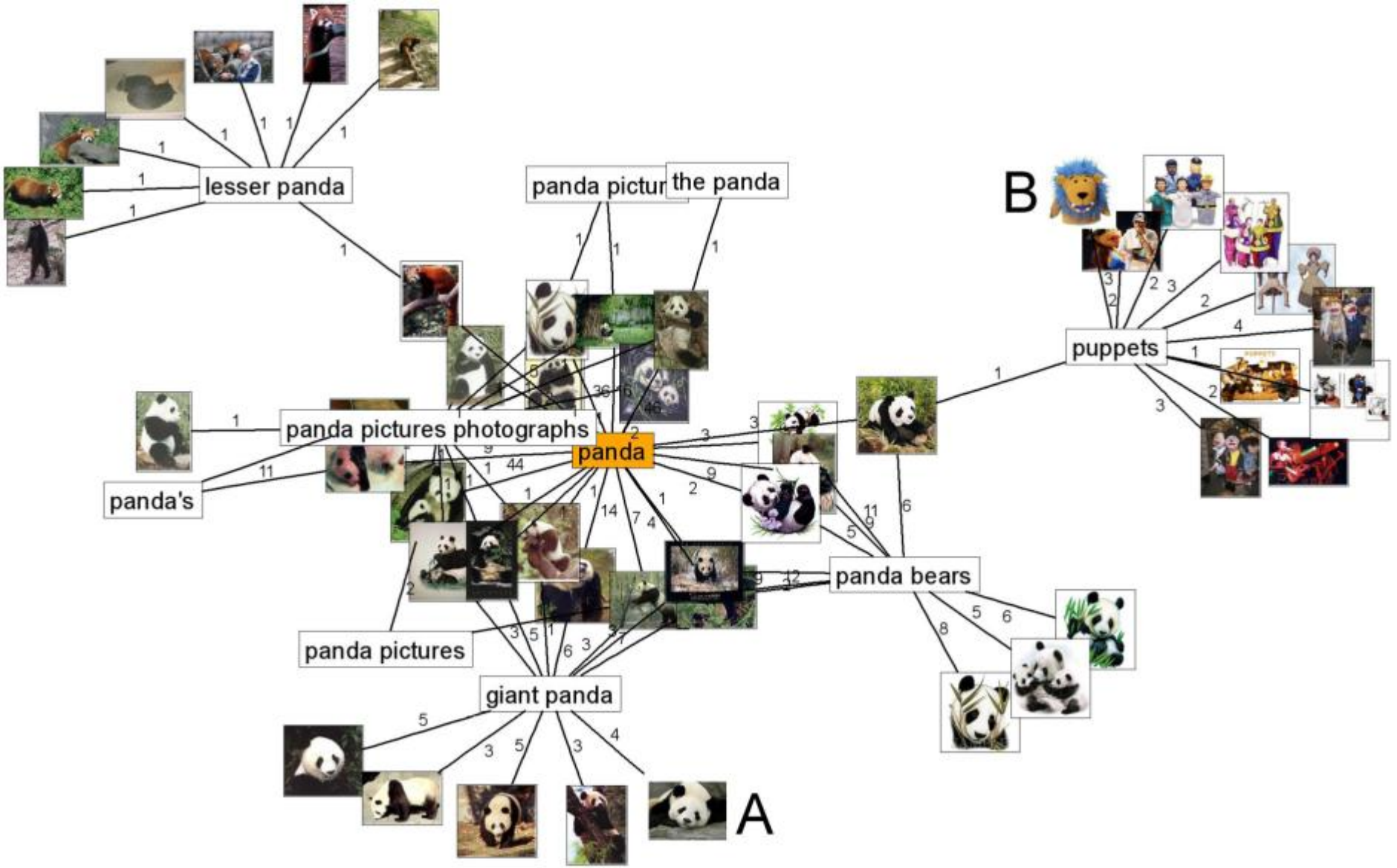
Co-click:

amazon	amazon	0.730631
amazon	amazon com	0.207988
amazon	www amazon com	0.030774
amazon	amazon books	0.007693
amazon	amazon uk	0.004771
amazon	amazon co uk	0.004196
amazon	www amazon co uk	0.002195
amazon	books	0.002163
amazon	amazon music	0.001506
amazon	www amazon	0.001044
amazon	amzon	0.000805
amazon	amizon com	0.000682
amazon	amizon	0.000618
amazon	buy book online	0.000491
amazon	http www amazon com	0.000361
amazon	amazon book uk	0.000271
amazon	amazon com uk	0.000190
amazon	amazon rainforest	0.000182

Session:

amazon	amazon
amazon	ebay
amazon	google
amazon	amazon books
amazon	yahoo
amazon	barnes and noble
amazon	walmart
amazon	hmv
amazon	target
amazon	best buy
amazon	amazon uk
amazon	borders
amazon	amazon music
amazon	overstock
amazon	argos
amazon	youtube
amazon	myspace
amazon	circuit city
amazon	play
amazon	toys r us
amazon	tesco

An Image Search Click Graph



Random walk: jaguar

Queries: jaguar jaguar cars jaguar.com jaguar cat jaguars jaguar parts www.jaguar.com jaguar car jaguar usa jaguar uk jaguar dealers jaguar canada jaguar automobile jaguar s type jaguar us jaguarcars jaguar cats jaguar.ca galpin jaguar jaguar autos jaguar xk8 jaguar animals jaguar cars uk http://www.jaguartechinfo.com jaguar convertible jaguar models jaguarusa jaguar wild cat 2007 jaguar jaguar dealer...

1 [1.02e-001] http://www.jaguar.com/uk jaguar
2 [8.75e-002] http://www.jaguar.com/ca/lang_select.htm jaguar
3 [8.43e-002] http://www.jaguar.com jaguar
4 [3.63e-002] http://en.wikipedia.org/wiki/Jaguar jaguar
5 [2.09e-002] http://www.jaguarusa.com/us/en/home.htm jaguar
6 [2.07e-002] http://autos.msn.com/browse/Jaguar.aspx jaguar
7 [1.61e-002] http://www.bigcatrescue.org/jaguar.htm jaguar
8 [1.30e-002] http://www.bluelion.org/jaguar.htm jaguar
9 [1.03e-002] http://www.kidsplanet.org/factsheets/jaguar.html jaguar
10 [7.43e-003] http://en.wikipedia.org/wiki/Jaguar_%28car%29 jaguar

Random walk: jaguar car

Queries: jaguar car jaguar jaguar.com new jaguar jaguar parts jaguar xj6 www.jaguar.com jaguar motor cars jaguar car parts jaguar automobile jaquar jaguar car covers 2008 jaguar jaquar cars new jaguar cars jaguar reviews jaguar car cover jaguar cars jaguar dealerships in virginia jaguar car dealers jaguar body parts jaguar x type parts jaguar x type performance part jaguar x-type accessories jaguar car mats jaguar dealers jaguar canada jaguar auto parts jaguars cars www.jaguar xj6 1970...

1 [6.88e-002] http://www.jaguar.com/ca/lang_select.htm jaguar

2 [6.87e-002] http://www.jaguar.com/global/default.htm jaguar.com

3 [5.19e-002] http://en.wikipedia.org/wiki/Jaguar_%28car%29 jaguar

4 [3.92e-002] http://www.automobilemag.com/reviews/jaguar new jaguar

5 [3.38e-002] http://www.cardomain.com/Make/Jaguar jaguar

6 [3.24e-002] http://www.jaguar.com/uk/en/vehicles/x-type/accessories/in_car_technology.htm jaguar car

7 [1.96e-002] http://www.motortrend.com/new_cars/01/jaguar/index.html jaguar car

8 [1.93e-002] http://www.carpartswholesale.com/cpw/jaguar-car-parts.html jaguar parts

9 [1.78e-002] http://www.jaguar.com/se/sv/vehicles/xj/reviews_awards/what_car.htm jaguar xj6

Random walk: jaguar cat

Queries: jaguar cat jaguar jaguar animal jaguar cats jaguar wild cat jaguar the animal jaguar facts jaguar animal facts jaguar animals panthera onca animal jaguar wild animal jaguars jaguar big cats jaguars the animal jaguar jaguar pics jaguar animal pics jaguars animals jaguar cars jaguar animal rain forest jaguar photo

1 [1.44e-001] http://www.bigcatrescue.org/jaguar.htm jaguar

2 [9.74e-002] http://www.indiantiger.org/wild-cats/jaguar.html jaguar animal

3 [6.74e-002] http://www.bluelion.org/jaguar.htm jaguar

4 [2.16e-002] http://www.bigcatrescue.org/jaguar_rescue.htm jaguar cat

5 [1.81e-002] http://dede.essortment.com/jaguaranimalca_rhvk.htm jaguar animal

6 [1.13e-002] http://www.ownajag.com/jaguar-cat.html jaguar cat

7 [1.07e-002] http://jaguarcat.co.za jaguar cat

8 [9.25e-003] http://www.zoo.org/pressroom/jag_htm/facts.htm jaguar facts

9 [8.93e-003] http://www.jaguarcat.co.za/gallery.htm jaguar cat

10 [7.14e-003] http://jaguarcat.us jaguar cat

Summary

1. Histogram: (query, count)
   - Least sparse of the three
   - Gives the weakest relationship between queries
2. Session: (query, query, count)
   - Sparse but strong relationship between queries
   - Related topics (and possibly spelling)
3. Co-click: (query, query, probability)
   - Synonyms
   - Random walk: (q/u, q/u, probability)

I don't know how Live speller works

Future work for the field:

We need to develop models for generating, combining and clustering queries.

schwarzenegger:

achwarzenegger (156) chwarzenegger (186) sarganager (177) sarinagerr (177) scgwarzenegger (187) schaneger (207) schanegger (194) schartsenegger (187) schartzanegger (164) schartzenegger (187) schartzneggar (187) scharwennger (187) scharwzenegger (187) scharzanegger (187) scharzenagger (187) scharzeneeger (187) scharzeneggar (187) scharzenegger (207) scharzennegger (187) scharznager (187) scharznegger (217) schawaneger (177) schawartzenegger (217) schawarzenegger (217) schawrtzenegger (187) schawrzenegger (187) schawzenegger (187) schazenegger (217) schearzenegger (187) schenager (177) scheneger (207) schenegger (217) schewarzenegger (217) schinager (177) schnagger (177) schnegger (207) schrarzenegger (187) schrazenegger (187) schrotzaniger (177) schuaseneger (187) schwaarzenegger (217) schwaenegger (217) schwaezenegger (187) schwaggner (177) schwagneer (177) schwagner (154) schwaizenegger (187) schwanager (177) schwanagger (164) schwaneggar (164) schwanzenegger (187) schwar (207) schwaraenegger (187) schwaranegger (187) schwarazenegger (217) schwarceneger (187) schwarcenegger (187) schwarchernnegger (187) schwarchnegger (187) schwarenager (187) schwarenagger (187) schwareneger (217) schwareneggar (187) schwarenegger (217) schwarenenegger (187) schwarengar (187) schwarengger (217) schwarenzager (187) schwarezegger (187) schwarezenegger (217) schwargenegger (187) schwargenner (187) schwargnegger (187) schwarinegger (187) schwarneger (194) schwarneggar (164) schwarnegger (217) schwarnenegger (187) schwarnezer (187) schwarr (207) schwarrenegger (187) schwarrzenegger (217) schwarseneggar (187) schwarsenegger (187) schwarsennegger (187) schwarsinegger (187) schwarsnegger (187) schwartanager (164) schwartchanager (177) schwartenegger (187) schwartenger (187) schwartennagger (187) schwartennegger (187) schwartnegar (164) schwartneger (164) schwartnegger (164) schwartsanegger (187) schwartsenegger (187) schwartsineger (187) schwartsiniger (177) schwartsniger (187) schwartz... (177) schwartzanagger (187) schwartzanegger (187) schwartzen (154) schwartzenagger (187) schwartzenarger (187) schwartzeneggar (187) schwartzenegger (269) schwartzeneggers (217) schwartzenegor (187) schwartzengar (187) schwartzenhager (164) schwartzenharger (187) schwartzennegger (217) schwartzenner (187) schwartzernegger (217) schwartzinegger (187) schwartznegger (187) schwartzner (177) schwarxenegger (187) schwarzaggter (187) schwarzanagger (187) schwarzaneger (187) schwarzaneggar (187) schwarzanegger (187) schwarzangger (187) schwarzaniger (164) schwarzannager (187) schwarzebegger (187) schwarzeeneger (187) schwarzeeneggr (164) schwarzegger (217) schwarzemegger (187) schwarzenaggar (187) schwarzenagger (187) schwarzenaggers (187) schwarzeneeger (187) schwarzenegar (187) schwarzenege (217) schwarzeneger (243) schwarzeneggar (187) schwarzeneggars (187) schwarzenegge (217) schwarzeneggeer (217) schwarzeneggerr (217) schwarzeneggers (120) schwarzeneggger (217) schwarzeneggr (217) schwarzenenegger (217) schwarzenengger (217) schwarzenerger (187) schwarzenergger (217) schwarzengeer (187) schwarzengegger (217) schwarzenger (217) schwarzengger (207) schwarzenhagger (187) schwarzenheger (187) schwarzennagger (187) schwarzenneger (229) schwarzenneggar (187) schwarzennegger (217) schwarzenneggers (217) schwarzenner (187) schwarzennerg (187) schwarzennger (187) schwarzerneger (187) schwarzernegger (217) schwarzinager (164) schwarzinagger (187) schwarzineger (187) schwarzinegger (187) schwarznagger (187) schwarzneggar (187) schwarznegger (217) schwarzneggers (187) schwarzrnegger (187) schwarzteneggar (187) schwarztenegger (217) schwarztnegger (187) schwarzzenager (164) schwarzzeneger (187) schwarzzeneggar (187) schwarzzenegger (217) schwarzzennegger (217) schwarzzneger (187) schwarzznegger (187) schwat (154) schwatrzenager (187) schwatrzenagger (187) schwatrzenegger (217) schwatzenagger (187) schwatzeneggar (187) schwatzenegger (187) schwatznager (187) schwatznagger (187) schwatznegger (187) schwawalui (177) schwazanegger (187) schwazenagger (187) schwazeneger (217) schwazeneggar (187) schwazenegger (207) schwazennegger (187) schwazernegger (187) schwaznegger (217) schwazzeneger (187) schwazzenegger (187) schwazzenger (187) schwearzenegger (217) schwenager (177) schwenegger (194) schwennegar (177) schwerneger (187) schwertzenegger (187) schwerzenegar (187) schwerzeneger (187) schwerzenegger (187) schwerznegger (187) schwienger (177) schwneger (207) schwnegger (217) schworzenegger (187) schwrzenegger (217) schwwarzenegger (217) schwzenegger (217) schzenegger (217) scwar (207) scwargenegger (187) scwarnagger (187) scwartinager (177) scwartnegger (187) scwartzenegger (187) scwartzneger (187) scwarzanegger (187) scwarzeenager (187) scwarzenagger (187) scwarzeneger (217) scwarzeneggar (187) scwarzenegger (217) scwarzennegger (187) scwarznegger (217) scwarzzenegger (187) scwatzenegger (187) scwharzenegger (187) sharsnagger (177) sharzenegger (217) sharzneeger (187) shawrtzenager (177) shcwartzenegger (187) shcwarzenegger (187) shcwazenegger (187) shuartiz (177) shvarzenegger (187) shwar (207) shwarcneger (187) shwarneggar (187) shwarsenegger (187) shwarsnager (177) shwarsnegger (164) shwartannager (177) shwartenager (164) shwartnegger (164)

Conclusions

Research has 50 queries, practice has query logs

a) Practitioners can use and share such data
   - Web search engines have a lot of data
   - But everyone with users has some user data
     - Even a small dataset is **specific to your users**
   - Researchers would partner with you (I think)

b) Researchers, this is a fun area!
   - New sources of evidence ==> Big gains.

I'm working with Microsoft (on two fronts) to release click data to the research community.

Thank You!