Searching, fast and slow
A tech perspective

October 12th, 2020
Prof. dr. ir. Arjen P. de Vries
arjen@acm.org
Several factors render Google's power in online search generally immune to competition or threat of entry.

General online search strongly favors scale due to (1) the **high fixed costs of servers needed for crawling and indexing the entire web**, and (2) the **self-reinforcing advantages of click- and query data**, which let a search engine constantly improve the relevance of search results.

Even search engines that choose to syndicate their search results rather than create their own index and algorithm face major obstacles. This is primarily because Google — through both integration and contractual agreements — has established itself as the **default search provider on 87% of desktop browsers and the vast majority of mobile devices**.
ADVANTAGES OF CLICK- AND QUERY DATA

Without the log data, web search isn't as good

This hinders retrieval experiments in our lab, and academia in general!

Note:
Wednesday October 14th, 11:00, presentation by Djoerd Hiemstra
Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers.

Google VP of data centers Joe Kava's presentation at Google Cloud Next 2017 in San Francisco:
GCP Infrastructure
6 regions, 18 zones, over 100 points of presence, and a well-provisioned global network comprised of hundreds of thousands of miles of fiber optic cable.
Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers.

Google VP of data centers Joe Kava's presentation at Google Cloud Next 2017 in San Francisco:

The Dalles in Oregon: investment $1.8 billion
Pryor Creek, Oklahoma: investment $2 billion

Times 15...

The new data center under construction in 2016 in Eemshaven, Netherlands, is expected to cost $773 million.

Overall, Google's capital expenditures for 2016 were just under $10.2 billion. Most of that can be accounted for by its data centers and land acquisitions.

https://www.datacenterknowledge.com/google-data-center-faq/
Inexpensive... yet, privacy invasive!
Very fast!
Not that relevant!

Human-in-the-loop!
instance c5.8xlarge
install pihole
install pivpn
install raspbian
install curl
instagram
InShared
Insidious
instant gaming
instance c5.8xlarge
install pihole
install pivpn
install raspbian
install curl
instagram
instant gaming
instagram login
instagram account verwijderen
instagram fonts
instance c5.8xlarge
install pihole
install pivpn
install raspbian
install curl
install instagram
install account verwijderen
insta
install account verwijderen
install inloggen
instant

instant gaming
instant gist
instant pot
Instant Apeldoorn B.V.
Point of Interest • Laan van de Dierenriem 32, Apeldoorn
Instant noodle
instant camera
Instant Family
2016 film
instant ink
instant payments
Instant coffee
Beverage

Google Search I'm Feeling Lucky
Google

instant s

- instant streetview
- instant stresser
- instant stooge
- instant sports switch
- Instant Sports Summer Games
  Video game
- instant skateboards
- Instant soup
- instant snow
- instant smile
- instant steiger
  Instant Apeldoorn B.V. - Laan van de Dierenkem 32, Apeldoorn

Google Search  I'm Feeling Lucky
Finally!
Very fast!

Human-in-the-loop!

Not that relevant!
Search anno 2020:
- Snippets
- Verticals
- Knowledge Graph
- Instant Answers
- Mobile
- ...

Google Instant search

Google Instant is a feature that predicts what you’re searching for and shows results as you type. It uses Google’s autocomplete technology to show predicted search terms in a drop-down box, and begins to display search results below the drop-down. Sep 8, 2010

searchengineland.com › google-instant-complete-users...


People also ask

- How do I turn on Google Instant Search?
- What happened Google Instant?
- What are the suggestions in Google search?
- How does Google autocomplete?

searchengineland.com › google-dropped-google-instant...

Google has dropped Google Instant Search

Jul 25, 2017 — Several years after Google launched Google Instant, they are killing the default search feature to bring search more inline with mobile devices.
Google has dropped Google Instant Search

Several years after Google launched Google Instant, they are killing the default search feature to bring search more inline with mobile devices.

Barry Schwartz on July 26, 2017 at 10:31 am
Above $10 million in expenses for research on GPT-3 and training the final model

Tens of thousands of dollars in monthly cloud computing or server and electricity costs for running the model

Possibly more than a million dollars in yearly retraining costs due to model decay

Additional costs of customer support, marketing, IT, security, legal and other requirements of running a product. This could be in the tens of thousands of dollars based on the number and size of customers OpenAI acquires.

https://bdtechtalks.com/2020/09/24/microsoft-openai-gpt-3-license/
Result:
OpenAI is giving Microsoft exclusive access to its GPT-3 language model (in exchange for 1B$)

https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/
BARRIERS TO ENTRY

WHAT TO DO?!

Make a different product; not “Web Search” as we know it!

We can never beat them at what they do very well if we try to do the same

Our / EU pockets are not deep enough

Even if we get as good as Bing, we are not so likely to get more than 6% market share
(Google has the “first mover advantage” and – so far – ample resources to stay the entry point to the Web)

Right strategy by OSF:
Create a European Crawl Index first!

Generic European Web search engine second?
WHAT TO DO?
PUSH SEARCH TO THE EDGE!

Decentralize Web Search

yacy.net/
WHAT TO DO?

DECENTRALIZE WEB SEARCH?

“Is that what he meant with slow search?!”
Viewpoint

Slow Search

Seeking to enrich the search experience by allowing for extra time and alternate resources.

We live in a world where the pace of everything from communication to transportation is getting faster. In re-
Safely gain access to rich personal data:
   Email
   Browsing history
   Documents read
   Contents of the user's home directory (i.e., documents written as well!)

Is this new? Well, we used to call this “Personal Information Management”. I like human-centric search!

Do we still need that log data if we can have all that?!
Can high quality evidence about an individual's recurring long-term interests replace the shallow information of many?
“Even more broadly than trying to get people the right content based on their context, we as a community need to be thinking about how to support people through the entire search experience.”

Jaime Teevan on “Slow Search”

Search as a short- and long-term dialogue (with or without “conversational search”)
A Personal Search Engine, in the Edge
WEB INDEX AT HOME

REALISTIC?

Clueweb 2012: 80TB
Recent CommonCrawl (August 2017): 3.28B pages, 280TB

Average web page takes up 320 KB
  Large sample collected with Googlebot, May 26th, 2010
  Reported 4.2B pages (would require ~1.3 Petabyte)

De Kunder & Van de Bosch estimate an upper bound of ~50B pages
  http://www.worldwidewebsize.com/

Also considering continuing growth (claimed in unpublished work)
  Andrew Trotman, Jinglan

Who actually needs all of the Web if their search engine is truly personal?
   E.g., I cannot read more than 4 or 5 languages (and even those...)
   E.g., I do not need the club league for soccer

And...
I could always fall back to using an “out-dated, non-personalized” Web Search engine...
   ... should I suddenly feel an urgent need to search for the soccer club league!
How to get the web data on the personal search engine?

How to replace the lack of usage data from many?

Wednesday October 14th, 11:00, presentation by Djoerd Hiemstra
Idea:
Organize the web crawl in **topically related bundles**
Apply bittorrent-like decentralization to share & update bundles
  - webtorrent.io
  - IPFS.io
  - academictorrents.com

Use techniques inspired by query obfuscation to hide the real user’s interests when downloading
**Idea:**
Web Archives already store the data that the personal search engine would need
Just not (yet) organized in topical and temporary bundles

Win-win situation:
A business model for archiving?
A way to enrich the (rarely used) web archives with usage data?
A way to crowd-source seed-lists for crawling?

“Rescue the Web Archives”

Radboud University
“… communication and media limitations, due to the distance between Earth and Mars, resulting in time delays: they will have to request the movies or news broadcasts they want to see in advance.

[...]

Easy Internet access will be limited to their preferred sites that are constantly updated on the local Mars web server. Other websites will take between 6 and 45 minutes to appear on their screen - first 3-22 minutes for your click to reach Earth, and then another 3-22 minutes for the website data to reach Mars.”

AN ANALOGY

PUSH SEARCH TO THE EDGE / GETTING THE DATA

Web Archive

Personal Search Engine
PUSH SEARCH TO THE EDGE / GETTING THE DATA

PRE-FETCHING & CACHING

Hide latencies of getting the data from the live web:

- Pre-fetch pages linked from initial query results page
- Pre-fetch additional related pages
- Pre-fetches expanded with those from query suggestions

Cache web data to avoid accessing the live web

Related work:

http://dx.doi.org/10.1109/MIC.2016.2


Caching Web data at home

Build a Personal Web Archive (PWA) while browsing

WASP, with WebIS:

github.com/webis-de/wasp/

Prizm by Jimmy Lin (personal Web archiving on a Raspberry Pi)

Extend the PWA, considering this as a seed

P-o-C in student project extending WASP (by Gijs Hendriksen)
BLUEPRINT
THE PERSONAL SEARCH ENGINE

Push Search to the Edge
Human-centric Search
   Exploit the rich source data that can be processed safely locally
Webarchives to the rescue
   Super-peers in a P2P network of personal search engines