Subject: search lingo
From: "John Giannandrea" <Redacted@apple.com>
Received(Date): Thu, 08 Nov 2018 08:03:22 +0000
To: "Adrian Perica" <Redacted@apple.com>, "Steve Smith" <Redacted@apple.com>
Bcc: "Kelsey Peterson" <Redacted@apple.com>
Date: Thu, 08 Nov 2018 08:03:22 +0000

Privileged and Confidential.

Adrian,

Since Kevin and I were talking in code at last week's meeting here is a primer on the key parts or a modern search product. It may help explain why there are so few serious contenders in this space because of the depth of R&D needed.

-jg

Crawl/Index

A search index is essentially a copy of the web

But the web is really big and its impossible to make a copy of all of it

The issues are size and freshness, its a tradeoff

Also many docs are duplicates so you need some way to deduplicate content at scale (hard)

If you have 1T documents then probably some are only updated order months

There is a shortage of bandwidth to be able to copy the web and keep it all fresh at scale

But if you index Twitter or NYT you want order seconds

The tech to make the tradeoff is called crawl scheduling and its super hard (because the web doesnt tell you what is important)

Once you have your copy of the web you need to turn it into a searchable database, the "index"

You can index documents by query term (term indexed) or by document (doc indexed).

One gets you recall, the other gets you precision. Again a tradeoff.

Given that you are crawling and refreshing millions of docs per minute you need a super fancy db to be the index

Query reformulation/Search ranking

Given a query, you want to find candidate docs quickly (the search part)

But first you need to understand if the query typed is the query that was intended

Spelling is the biggest problem here

Followed by synonyms, user types A but they meant A' (gaga -> Stefani Germanotta)

Usually search engines turn the query into a structured set of queries with scores for possible interpretations

There is a tradeoff here between retreival (not missing a doc) and accuracy (ranking the right doc at position @1)

Then you need to score this interpretation "query" and then rank the retreived docs.

This is what MSFT called the Algo bit and Google calls search ranking

There are broadly three kinds of result kinds, navigational [facebook.com], authoritative [should I get a flu shot], and categorical [best finance books].

The most powerful signals here are click though on previously presented docs.

If you show the right answer at position @3 and people click on it more than @1 then you know that you should be ranking it higher and you can learn from this.

Its machine learning a ranking signal by raw counting clicks!

But in the long tail you dont have enough clicks so you have to find ways of 'smearing' click signal to more obscure queries/docs.

But lots of static signals like PageRank and Titles and Anchors (the words on the links that point to a doc) are important too

There is a huge question here about how much ML to use. Bing uses a lot, Google doesnt like to.

The way that you figure out whether a search algo change was good is that you use human raters to rate thousands of random queries with and without the change and you consider the overall impact of the change.

The raters are not rating the algorithm, just the effect it had on top 5 results for a statistical sample of queries.

We can assume that Bing spends $Redacted yearly on rating alone.

Google's rater guidlines are public, published here.

An example Google weekly launch commitee meeting is on Youtube. Its old but illustrative of how these changes are deliberated on. (Aside, most of the people in this video are still doing this same thing a decade later). (BTW, Steve Baker from Luna speaks up with a key insight in this video at 2:50)

Serving

You have to do all this ranking work at a high number of queries per second (presumably 100K+ of QPS)

So you need elaborate serving systems and cost is a factor

Almost a decade ago Google decided to launch 'universal search' where any given query was run past multiple search engines

So you could mix in web search, video search, book search and more all in one search page. This was presumably 5X more expensive per query

Then they rolled out something called "instant search' which was the idea that for every character you typed in the search box they would run the search and render the result. This was again 5-10X more expensive per query. It was a huge deal, like they were not sure they could buy that many servers!

Its since been unlaunched in favour of more cycles per query (see below).

Search features

Most of the above is about search "quality", the quality of the 10 blue links and the continuous effort to get the best link in the top 3 results.

As search became more mobile, there was a pivot to giving you the answer first, not ten blue links. [where did obama go to school]

The array of things in this investment is generally called search "features" or "domains" or "verticals"

Bing tried very hard in the beginning to make this their differentiation. Even today they have better results for college tuition than Google for example.

Features are even more incrementally expensive. You need UI per domain, data feeds per domain and then you need to multiply by the number of languages and locales.

So Cricket is a no brainer in India and UK. Australian rules football is a critical vertical down under. Its all "by the yard" i.e. linear in R&D expense.

There is a long tail of domains and verticals. My favourite at Google wasPokeman characters and Chickens breeds, both experiments in how low you could get the curation task per domain.

I18n Scale and Coverage

And of course you get to do it all of this in as many languages and platforms as you want! CJK is harder than Latin languages, for example in Korean there is no esy way to segment queries into words. Doing all this for low memory feature phones or old iPhones might mean varients of all these tradeoffs and features. All of which has to be QAed and rated on a daily basis.

Hopefully this explains why a world class search engine is at least a $[Redacted]B/year R&D investment and that is before you build a search ads business to pay for it.

-jg