Measuring Scale Economies in Search

June 2, 2015
Preston McAfee
Microsoft

With Justin Rao, Aadharsh Kannan
Di He, Tao Qin, Tie-Yan Liu
Email From the Texas Higher Education Coordinating Board:

Your last name is your password.

If you have any questions or have forgotten your password, please contact the Coordinating Board.
Self Assessment

Type in your name.

Preston McAfee

What is your relationship to the person noted above?

- Self
- Direct Report
- Peer
- Manager

Self need not answer below question
Value of Data in Algorithmic Search

• Search engines do not answer queries (mostly)
• Search engines are enormous matching programs, matching billions of URLs to hundreds of billions of queries
• Process uses sophisticated algorithms, probably the most complicated algorithms ever built
• Data is used to initialize and update algorithms
• How important is data at modern web scale?
• With hundreds of billions of observations, does a doubling or even a 20X increase actually matter?
How Large are the Scale Economies in Search?

- Statistically speaking, a trillion observations, a billion right hand side variables is still a trillion degrees of freedom
- Most queries are rare
  - 50% of Bing queries unique in 2014, 8% of searches
- But rare queries have related queries
  - Pasadena Ethiopian Restaurant related to Pasadena Restaurant, Ethiopian Restaurant
- $1/\sqrt{n}$ errors, where $n$ is the amount of data – but what is $n$?
- Hasn’t anyone measured this before?
  - Need billions of observations
Analytic Problems to Overcome

• Data used indirectly, makes whole search engine better
  • Not directly visible in results data
• Common queries are often easier
  • Common queries have less ambiguity
  • Common queries more likely to be navigational rather than informational
Overview and Summary

• **Rare query trend analysis.** As data on rare queries comes in, the quality of search page results improves, for Bing and Google.
  • Because of Google's greater scale, it improves faster.

• **Direct and indirect view count analysis.** Data on related (indirect) queries can help improve the quality of response to new queries. Indirect data has similar effect to direct; but many queries have little indirect data.
  • Google's greater scale means that it acquires data on new queries more quickly and that it has more data on indirect queries on which it can draw.

• **Click position analysis.** Quality of ranking improves with more data.
  • Not just improvements in website quality
Rare Query Trend analysis

• Even if scale does not have a significant effect on the quality of responses to common queries, additional data on rare queries may improve the quality of algo search results.

• If so, then a search engine with greater scale will learn faster.

• Conservative: should understate data scale effects
  • Missing indirect data and external effects
Experimental Design #1: Rare Query Trend

• For a search engine, consider its IE logs in 2014
• Use 1/2014 to 3/2014 as benchmark data
• Use 4/2014 to 12/2014 as predicting data
  • #Clicks to define historical clicks
  • Use sliding windows in the next period to compute future CTR
• Definition of rare queries
  • #Clicks<200 in benchmark period
  • 1000<=#Clicks<2000 in prediction period
• With this setting, for the buckets [100, 900], we have exactly the same set of queries
  • Start getting different mixes as number goes up.
Examples of the Rare Query Trend

• Madam Secretary – a CBS show
• Letter Garden – an online game
CTR v.s Historical Data Volume: US

- Overall, we observe CTR growth for rare queries for both Bing and Google
- Non-monotonicities not statistically significant
- Holds query mix constant
- Levels not comparable between Google and Bing
  - Only observe portion of Google data but all of Bing
  - Different queries for Bing and Google (meeting the 200/1000 requirement)
CTR v.s Historical Data Volume: EU

- Aggregated results
- Overall, we observe CTR growth for rare queries for both Bing and Google
How Much Data Do We Have?

• >50% of queries are unique in the year
  • That does not mean there is no data!
  • Historical data on “Pasadena restaurant” useful for “Pasadena Ethiopian restaurant”
  • Learn authoritative sites from past queries and then do text matching

• How many queries have “little” relevant data?
• This is a second, independent approach to assessing the importance of data
Direct and indirect view count analysis

• If data on related (indirect) queries are useful in responding to new queries, then a search engine with greater scale will acquire data on new queries faster and it will also have more indirect data that it can use to improve the quality of responses to new queries.

• We can assess this empirically with Bing data

• We can estimate the effects of direct and indirect data on Bing quality
Approach

• Use data from Bing/Yahoo only
  • Query trend used IE logs; this is Bing logs

• Build semantic graph
  • Using click similarity
  • Exclude observations where graph not completed
  • Conservative because true singletons are excluded

• Assess how much indirect data is available
Semantic Graph Illustration

- Antonio Conte
- Gianluigi Buffon
- Italy national football
- Antonio Conte Salary
- Antonio Conte botches the world cup
- Gianluigi Buffon new contract

- 100%
- 25%
- 9%
- 2%
- 15%
- 5%
- 30%
- 20%
• Queries leading the same URL frequencies are similar
• URLs with the same queries are similar
• Recognized methodology: Baeza-Yates and Tiberi, 2007, KDD.
Stats

100B searches = 4.5B queries

<table>
<thead>
<tr>
<th>Type of query</th>
<th>Number</th>
<th>% of queries</th>
<th>% of traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queries that could be clustered</td>
<td>2.6B</td>
<td>53.3%</td>
<td>92.3%</td>
</tr>
<tr>
<td>Unclustered queries</td>
<td>1.9B</td>
<td>42.7%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Total</td>
<td>4.5B</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

2.6B queries mapped to 128M clusters
Many Queries Have Limited Indirect Data: Indirect Data is Not a Panacea

Circle Radius \( \equiv \) #Query Instances
Total (Direct + Indirect) Data Frequency
17% of Queries, 10% of Searches Have < 1000
Data Accumulation

• Define new queries
  • Zero observed in 2012
  • Received 10+ per month in 2013
  • Results in 17K queries; 7800 after deduplication

• Build Graphs successively over time
  • New data allows cluster-joining
**Examples**

<table>
<thead>
<tr>
<th>Query</th>
<th>Reason it appears (based on human judgment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>minecraft miniplex</td>
<td>The actual search is for Minecraft Mineplex, a Minecraft online server service launched in Oct 2013.</td>
</tr>
<tr>
<td>despicable me training wheels</td>
<td>This short movie was released in Oct 2013.</td>
</tr>
<tr>
<td>wwe 2k14 xbox 360 controls</td>
<td>The game’s release date was 29th Oct 2013.</td>
</tr>
<tr>
<td>ipad air pictures</td>
<td>iPad air was launched in Nov 2013 while pictures were released/leaked around Oct 2013.</td>
</tr>
<tr>
<td>gta 5 online funny moments</td>
<td>GTA 5 game was launched on 17th Sep 2013 and it took some time for folks to get online and make funny moments. This query started appearing around Oct 2013 timeframe.</td>
</tr>
</tbody>
</table>
Illustrative graphics for construction of progressive clusters

We built the cluster progressively with data increasing each month for 12 months
Effect of direct & indirect view count on success (long) click through rate

We regressed the line given below for every query in the sample and averaged the coefficients and constant

\[ y_{\text{SuccessCTR}} = \alpha + \beta_1 \times \text{IndirectViewCount} + \beta_2 \times \text{DirectViewCount} \]

\[ \beta_1 = 2.251 \times 10^{-5} \quad [2.79 \times 10^{-7} \text{ to } 4.48 \times 10^{-5}] * \]

\[ \beta_2 = 1.109 \times 10^{-5} \quad [5.28 \times 10^{-6} \text{ to } 1.69 \times 10^{-5}] * \]

\[ \alpha = 0.742 \quad [0.740 \text{ to } 0.745] * \]

Inference 1:
Both the view counts contribute positively to the increase in Success CTR.

* We took queries with Pearson’s coefficient of 0.7 or lesser deriving reliable coefficients for multiple regressions. 85% CI for all intervals
Effect of direct view count on Indirect view count (Collinearity)

\[ x_{IndirectViewCount} = 2.221 x^{0.7025}_{DirectViewCount} \]

Inference 2:
For direct view count increase, a positive growth of the indirect view count occurs. The growth in indirect view count is faster during the initial increase of the direct view count.
Effect of direct view count on Indirect view count

Inference 2:
For direct view count increase, a positive growth of the indirect view count occurs. The growth in indirect view count is faster during the initial increase of the direct view count.
Effect of direct view count on Indirect view count

Inference 3:
Even after adjusting for indirect data, many low data query clusters remain.
Bing Gets Better as More Data Becomes Available

We regressed the line given below for every query in the sample and averaged the coefficients and constant

Average click position rises as more data becomes available

\[ y_{\text{AverageClickPosition}} = \alpha_{\text{ClickPos}} + \beta_{\text{ClickPos}} \times \text{DirectViewCount} \]

\[ \beta_{\text{ClickPos}} = (-)1.034131 \times 10^{-3} \quad [(-)1.07 \times 10^{-3} \text{ to } (-)1.00 \times 10^{-3}] \]
\[ \alpha_{\text{ClickPos}} = 2.597 \quad [2.524 - 2.670] \]

Inference 4: Means Click Position decreases as View Count increases, a ranking improvement.
Does Search Improve Only Because Content Improved?

• Increased data is increasing clicks, suggesting better performance. But in principle, search engines could be static, and just the available results are improving.

• Mostly existing URLs, not newly created ones

• Clicks migrate upward, showing better search results, not just better content
IE Logs: Few New URLs in Rare Queries

• Perhaps improvements due to discovery of new URLs?
• No, most URLs already existed (note Google has different query mix)
Same in the EU

- The same analysis for EU market, with similar observations
  - The percentages for EU market are around 97~98.5%;
  - The real percentages should be even higher due to low coverage of IE logs.

Historical clicks

Percentages of Seen Clicked URLs in Bing

Percentages of Seen Clicked URLs in Google
Bing: Effect of direct view count on URL Position

We regressed the lines given below for every query in the sample and averaged the coefficients and constant

\[ y_{\text{DistributionOfClickPosition}} = \alpha_{\text{ClickPos}} + \beta_{\text{ClickPos}} x_{\text{DirectViewCount}} \]

\[
\beta_{\text{ClickPos1}} = 3.836 \times 10^{-4} \\
\alpha_{\text{ClickPos1}} = 0.5681
\]

\[
\beta_{\text{ClickPos2}} = (-)1.5911 \times 10^{-5} \\
\alpha_{\text{ClickPos2}} = 0.2952
\]

\[
\beta_{\text{ClickPos3}} = (-)5.996 \times 10^{-6} \\
\alpha_{\text{ClickPos3}} = 0.2698
\]

\[
\beta_{\text{ClickPos3}} = (-)4.321 \times 10^{-5} \\
\alpha_{\text{ClickPos3}} = 0.1800
\]

**Inference 5:**

Means **Click Position decreases** as **View Count increases**.

This means **better URLs are pushed to the top**.
Conclusion

• We measured effects of more data on new queries
  • Both Google and Bing
  • More data makes both search engines get better

• We examined related queries and websites
  • Proxied by similarity
  • Dropped singletons as a conservative measure
  • Many queries have little indirect data
  • More data makes the results better (regression)

• We provided evidence that URL position rises with more data
  • Search engine results improve with data
Conclusion, Continued

• Web scale involves billions of searches
  • Nearly unlimited degrees of freedom
• But, webscale involves matching billions of queries to billions of websites, searching a space with $10^{20}$ possibilities
• Even at web scale, more data makes search better
Thank You!

Trust
Partnership
Innovation
Performance

OUR VALUES
Strongly Correlated (99.8%) From tylervigen.com

US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation
87% Correlated

Age of Miss America correlates with Murders by steam, hot vapours and hot objects

Murders by steam: 8, 6, 4, 2
Age of Miss America: 25, 23.75, 22.5, 21.25, 20, 18.75
Backup Slides
## Rare Query Trend Examples: Data

<table>
<thead>
<tr>
<th>query</th>
<th>Bing</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q</td>
<td>search</td>
</tr>
<tr>
<td>letter garden</td>
<td>1</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>339</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>508</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1305</td>
</tr>
<tr>
<td>madam secretary</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>487</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1283</td>
</tr>
</tbody>
</table>