February 18, 2025 Call with Google Engineer HJ Kim

Participants

- Google: Hyung-Jin Kim, Jack Mellyn (Google in house counsel)
- Expert Prof James Allan
- Williams and Connolly: John McGowan and Colette Connor
- Cornerstone: Kirti Gupta, Vivek Mani, Chanel O'Neill, and Elias Ilin

I. "Hand Crafting" of Signals

- Almost every signal, aside from RankBrain and DeepRank (which are LLM-based) are hand-crafted and thus able to be analyzed and adjusted by engineers.
    - To develop and use these signals, engineers look at data and then take a sigmoid or other function and figure out the threshold to use. So, the "hand crafting" means that Google takes all those sigmoids and figures out the thresholds.
        - In the extreme, hand-crafting means that Google looks at the relevant data and picks the mid-point manually.
        - For the majority of signals, Google takes the relevant data (e.g., webpage content and structure, user clicks, and label data from human raters) and then performs a regression.
- **Navboost.** This was HJ's second signal project at Google. HJ has many patents related to Navboost and he spent many years developing it.
- **ABC signals**. These are the three fundamental signals. All three were developed by engineers. They are raw, [redacted]
    - **Anchors (A)** – a source page pointing to a target page (links). [redacted]
    - **Body (B)** – terms in the document, [redacted]
    - **Clicks (C)** – historically, how long a user stayed at a particular linked page before bouncing back to the SERP. [redacted] Things like Navboost.
- ABC signals are the key components of **topicality** (or a base score), which is Google's determination of how the document is relevant to the query.
    - **T* (Topicality)** effectively combines (at least) these three signals in a relatively hand-crafted way. [redacted] Google uses to judge the relevance of the document based on the query terms.
    - It took a significant effort to move from topicality (which is at its core a standard "old style" information retrieval ("IR") metric) [redacted] signal. It was in a constant state of development from its origin until about 5 years ago. Now there is less change.

        - Ranking development (especially topicality) involves solving many complex mathematical problems.
        - For topicality, there might be a team of [redacted] engineers working continuously on these hard problems within a given project.
- The reason why the vast majority of signals are hand-crafted is that if anything breaks Google knows what to fix. Google wants their signals to be fully transparent so they can trouble-shoot them and improve upon them.
    - Microsoft builds very complex systems using ML techniques to optimize functions. So it's hard to fix things—e.g., to know where to go and how to fix the function. And deep learning has made that even worse.
- This is a big advantage of Google over Bing and others. Google faced many challenges and was able to respond.
    - Google can modify how a signal responds to edge cases, for example in response to various media/public attention challenges:
        - [redacted]
        - [redacted]
    - Finding the correct edges for these adjustments is difficult, but would be easy to reverse engineer and copy from looking at the data.

II. Ranking Signals "Curves"

- Google engineers plot ranking signal curves
- The curve-fitting is happening at every single level of signals. For example:
    - [redacted]
- If Google is forced to give information on clicks, URLs, and the query, it would be easy for competitors to figure out the high-level buckets that compose the final IR score. High-level buckets are:
    - ABC – topicality
        - Topicality is connected to a given query
    - Navboost

- Quality
  - Generally static across multiple queries and not connected to a specific query.
  - However, in some cases Quality signal incorporates information from the query in addition to the static signal. For example, a site may have high quality but general information so a query interpreted as seeking very narrow/technical information may be used to direct to a quality site that is more technical.

- Q* (page quality (i.e., the notion of trustworthiness)) is incredibly important. If competitors see the logs, then they have a notion of "authority" for a given site.
- Quality score is hugely important even today. Page quality is something people complain about the most
  - HJ started the page quality team ~17 years ago
  - That was around the time when the issue with content farms appeared.
    - Content farms paid students 50 cents per article and they wrote 1000s of articles on each topic. Google had a huge problem with that. That's why Google started the team to figure out the authoritative source
    - Nowadays, people still complain about the quality and AI makes it worse.
- Q* is about [redacted]
  - This was and continues to be a lot of work but could be easily reverse engineered because Q is largely static and largely related to the site rather than the query.

III. Other Signals

- **eDeepRank.** eDeepRank is an LLM system that uses BERT, transformers. Essentially, eDeepRank tries to take LLM-based signals and decompose them into components to make them more transparent. HJ doesn't have much knowledge on the details of eDeepRank.
- **PageRank.** This is a single signal relating to distance from a known good source, and it is used as an input to the Quality score.
- [redacted](popularity) signal that uses Chrome data.

IV. Search Index

- HJ's definition is that search index is composed of the actual content that is crawled - titles and bodies and nothing else, i.e., the inverted index.
- There are also other separate specialized inverted indexes for other things, such as feeds from Twitter, Macy's etc. They are stored separately from the index for the organic results. When HJ says index, he means only for the 10 blue links, but as noted below, some signals are stored for convenience within the search index.
- Query-based signals are not stored, but computed at the time of query.

  - [redacted]
  - Q* - largely static but in certain instances affected by the query and has to be computed online (see above)
- Query-based signals are often stored in separate tables off to the side of the index and looked up separately, but for convenience Google stores some signals in the search index.
  - This way of storing the signals allowed Google to [redacted]

V. User-Side Data

- By User Side Data, Google's search engineers mean user interaction data, not the content/data that was created by users. E.g., links between pages that are created by people are not User Side data.

VI. Search Features

- There are different search features - 10 blue links as well as other verticals (knowledge panels, etc). They all have their own ranking.
- **Tangram (fka Tetris).** HJ started the project to create Tangram to apply the basic principle of search to all of the features. [redacted]
- Tangram/Tetris is another algorithm that was difficult to figure out how to do well but would be easy to reverse engineer if Google were required to disclose its click/query data. By observing the log data, it is easy to reverse engineer and to determine when to show the features and when to not.
- **Knowledge Graph.** Separate team (not HJ's) was involved in its development.
- Knowledge Graph is used beyond being shown on the SERP panel.
  - Example – "porky pig" feature. If people query about the relation of a famous person, Knowledge Graph tells traditional search the name of the relation and the famous person, to improve search results - Barack Obama's wife's height query example.
- Self-help suicide box example. Incredibly important to figure it out right, and tons of work went into it, figuring out the curves, threshold, etc. With the log data, this could be easily figured out and reverse engineered, without having to do any of the work that Google did.

VII. Reverse Engineering of Signals

- There was a leak of Google documents which named certain components of Google's ranking system, but the documents don't go into specifics of the curves and thresholds.
- For example, [redacted]

[redacted]

[redacted]The documents alone do not give you enough details to figure it out, but the data likely does.