Google Search Litigation

January 31, 2025 Call with Google Engineer Pandu Nayak

Participants

- Google: Pandu Nayak
- Expert Prof. James Allan
- Williams and Connolly: John McGowan and Colette Connor
- Cornerstone: Vivek Mani, Chanel ONeill, and Blake Boswell

I. Overview/Intro

Preview of terminology and Google search structure referenced during the call

- **Document:** Google's name for a webpage. "Document" can refer to the webpage itself or the constructed webpage version Google keeps in its database.
- **Signal**: Google uses many signals to come up with the SERP. Top-level signals are aggregates of other signals, called raw signals. Google uses over 100 raw signals. Some signals are developed using machine learning models vs. others that are called, or thought of as, traditional signals. Examples of signals discussed on the call include:
    - Q* (pronounced "Q star"), Google's measure of quality of a document;
    - Navboost, a measure of how frequently users (subset by location and device type) click on a particular document for a particular query is a traditional signal. Uses most recent 13 months of data;
    - RankEmbed, one of Google's primary LLM-trained signals;
    - Twiddlers – re-rank a set of already selected results
    - PageRank, one of Google's original signals, continues to be one of the factors that goes into the quality signals for pages.
- [redacted] combines signals into a single score, which then determines document rank in the list of blue links on the SERP. [redacted] was an early Google use of a machine learning model in its search algorithm. [redacted] was taught using a process of [redacted]

Review of Debugging Interface

Pandu typed "james allan umass" into a Google search window and then called up the internal debugger window showing:

- [redacted] shows query expansion and decomposition process
    - E.g. umass to be re-written as University of Massachusetts, James is a first name, Allan might also be spelled Allen, etc.
- [redacted] and [redacted]
- [redacted] containing a table with list of 10 blue links and corresponding score for each top-level signal as well as corresponding total score "Final IR" across all

Ex. No.
PXR0357
1:20-cv-03010-APM
1:20-cv-03715-APM

[redacted]

REDACTED PUBLIC VERSION

signals. Top-level signals are a linear combination of log of individual raw signals. Signals are formulated such that their impact on ranking is monotonic relative to the signal.

- E.g., to compare just two of the signals, [redacted]

- [redacted] provides comparison of any two selected blue links to see the difference in how they scored across individual signals [redacted]

II. Google and LLM

History of Google's Addition of Learning Models to Search

- Google's traditional approach to ranking was in the style of Okapi BM25, a ranking function used to estimate the relevance of documents to a given search query.
- The first move away from the traditional way of doing things was [redacted] (defined in terminology above).
- Then the transition to greater reliance on machine learning / deep learning - RankEmbed (discussed below), RankBrain, DeepRank - to generate signals.
- Google found that the BERT-based DeepRank ML model signals could be decomposed into signals that resembled the traditional signals and that combining the traditional and "predicted" signals produced by ML resulted in better outcomes
- Google avoids simply "predicting clicks" because clicks are easily manipulated and are a poor proxy for enhancing user experience

RankEmbed and How Disclosure of Google's Information Could Be Used By Competitors To Reverse Engineer Google Search Innovations

Obtaining Google's ranking signals is not a necessary step to closely approximating Google's ranking given the availability of deep learning. Deep learning-based distillation is proposed as a specific example as is Google's own RankEmbed signal.

- RankEmbed is a dual encoder model that embeds both query and document into embedding space. Embedding space considers semantic properties of query and document in addition to other signals. Retrieval and ranking are then a dot product (distance measure in the embedding space)
- Extremely fast; high quality on common queries but can perform poorly for tail queries
- Google trained RankEmbed on a sample from a single month of search data. This has implications for understanding what a competitor could do with synthetic query remedy.

- Quality of RankEmbed is demonstrated by Google FastSearch product, [redacted] Google uses FastSearch as a RAG (retrieval-augmented generation) mechanism on Vertex AI, Google's cloud offering, and the Gemini app, where it can ground responses.

Data Size as it relates to Mimicking Google, Training

- Even just hundreds of query/result combinations would allow for an approximation of certain Google signals, which could be used to train a model, such that competitors could begin to recreate Google search
- For ML models Google has increasingly been using less and less data (90 days, 60 days, etc.). However, in general, Google's rule is it wants to deliver the best product for users. Therefore, if there is a benefit to using more data, that data should be used until the point where using more data negatively impacts Google's ability to deliver the best product for users.

LLMs for Ranking and Retrieval

- LLMs can improve portions of Google's search stack (e.g. query interpretation and summarizing presentation of results).
- Google is currently re-thinking their search stack from the ground-up with LLM taking a more prominent role. They are thinking about how fundamental components of search (ranking, retrieval, displaying SERP) can be reimagined given the availability of LLMs. One consideration is the computation time of LLMs, depending on the use case.

III. Use of User Query Data in Google Search

- How many signals are impacted by user-side data? Difficult to quantify. For signals that are impacted by user-side data, some are impacted more than others.

Navboost

- Described as a QD table, a query-to-document lookup table, used in both directions containing counts/frequencies of user query activity by document

IV. Examples of Google Innovation That Proposed Remedies Could Reveal

Google's development process is responsive to Google's mission to meet the needs of users. Google identifies issues and then debugs/reviews to understand what's missing in their signals and what can be done to incorporate new information or otherwise ameliorate their ranking process.

- **Anecdote**: An early signal, [redacted] measured how many times a link was shown vs how many times it was clicked. It was found that this measure was biased by link position. For example, the query "currency conversion" may return a perfectly reasonable link in position 1, but it might not actually be *better* than the link displayed in position 2. The

use of [redacted] would create a system that reinforced this ranking. [redacted] is a modification developed by a Google engineer that calculated [redacted] while considering (and avoiding bias created by) link position.

- **Anecdote**: Innovation currently in development: [redacted]
- **Anecdote:** An issue that was faced by search engines in the past is content farms. Google signal Anchor identified a document's quality in part by how many other documents linked to it. Advertisers created large websites where useless content would be created specifically to be included in search results, receive traffic, and be able to show ads. This problem required Google engineers to develop new signals to promote higher quality content.
- **Anecdote:** It was reported that Google search would return a Holocaust denier as the top link for the query "did the Holocaust occur". Google innovated so that reliable, quality results appear first. Search engines must consider topics such as diversity of search results in a nuanced way – e.g. diversity of results when the query is "did the Holocaust occur?" vs diversity of results when the query is "Michael Jordan".
- Google discontinues some Signals. For example, when a new better signal is developed, if the signal performs poorly, or if the efficacy of the signal degrades overtime due to the evolving nature of the internet.

Contribution to Web Ecosystem

- **Anecdote**: Mobile vs Desktop – As mobile devices became more popular Google signaled to website developers that they would rank websites that offer mobile optimized deployments highly for searches performed from mobile devices
- Google also provides information to website developers to help them with SEO (such as ways to markup their sites)
- Google also provided a tool to help website developers optimize load times