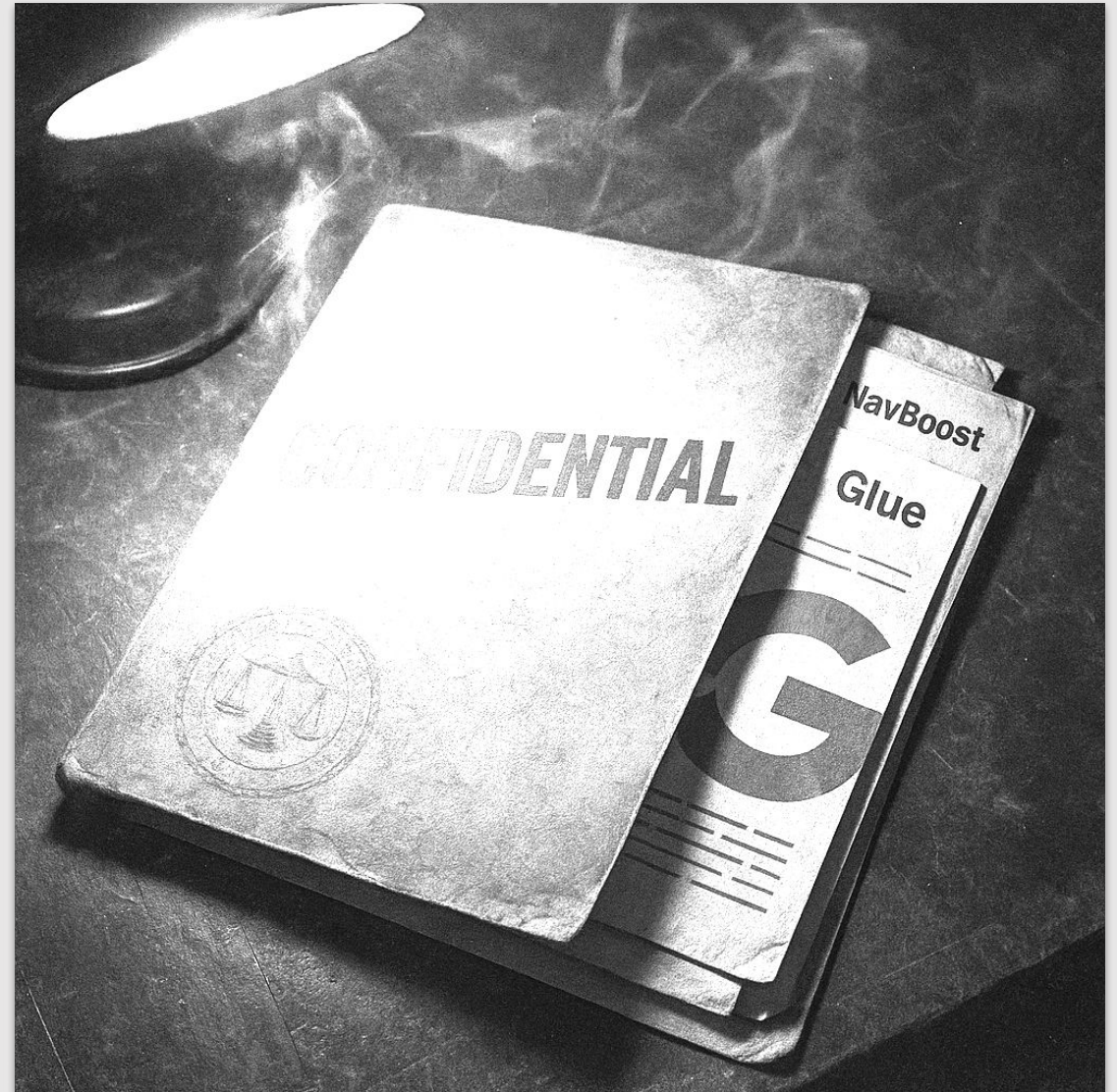


# Dossier Google

Filtraciones y  
revelaciones sobre  
el sistema de  
ranking de Google

Juan González Villa

@seostratega



# ¿Cuáles son mis fuentes?

Cantidad

- Filtración de documentación de APIs internas de Google (+2500 documentos). Fuente primaria, pero fuera de contexto.
- Documentos y testimonios del juicio antimonopolio en EEUU (2023-2025). Fuente primaria, pero parte de la información está tachada.
- Endpoints de APIs internas de Google encontradas por Mark Williams-Cook. Fuente primaria y reconocida por Google (pagaron una recompensa de 13.337 dólares)

Contexto

# La “metafuente”: documento del juicio sobre la filtración

## 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]

Page 4 of 5

**USDOJ-GOOG-00195745**

**REDACTED PUBLIC VERSION**

**CONTAINS HIGHLY CONFIDENTIAL INFORMATION – SUBJECT TO PROTECTIVE ORDER**

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

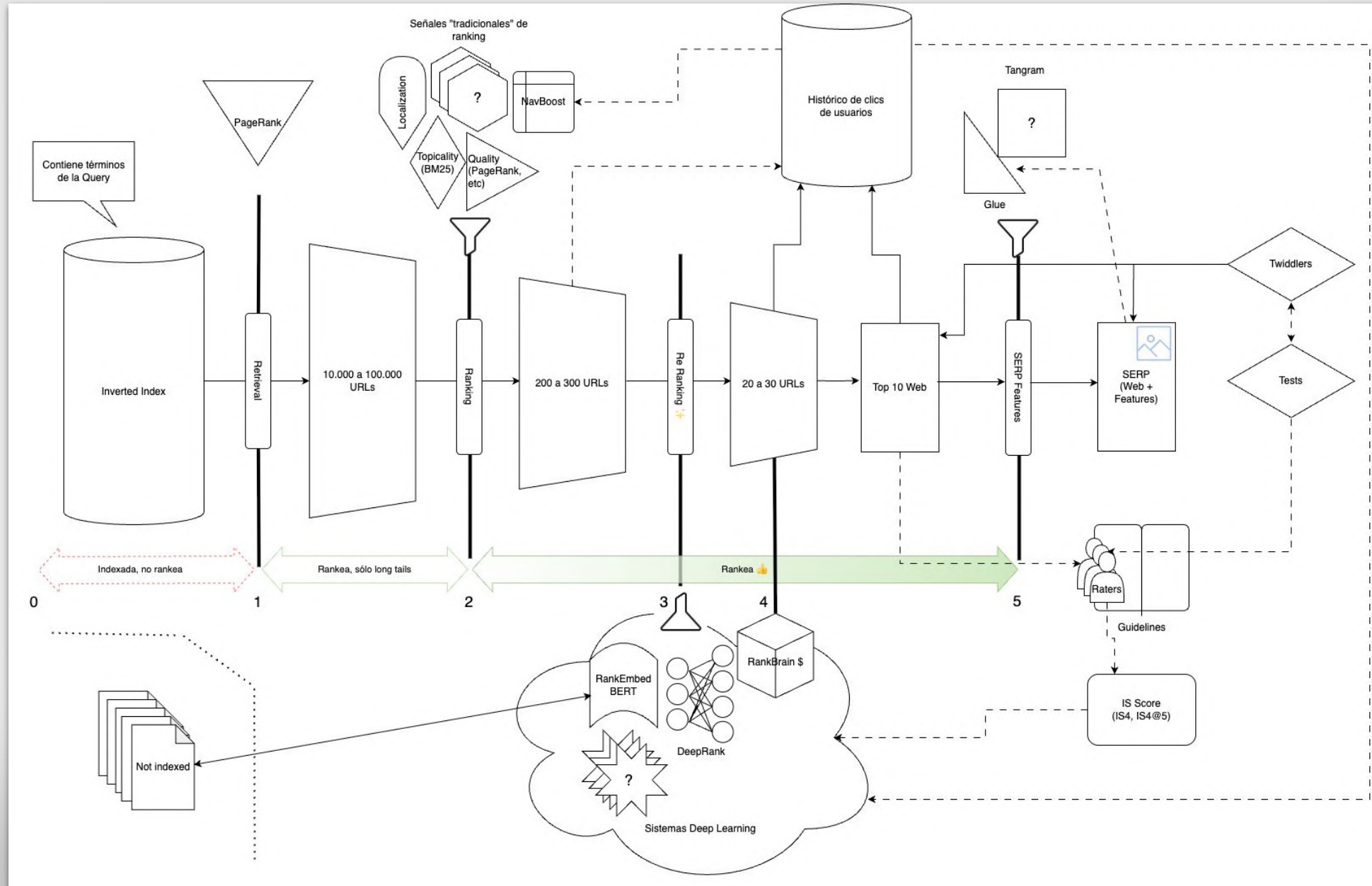


**Pandu Nayak**  
VP, Search

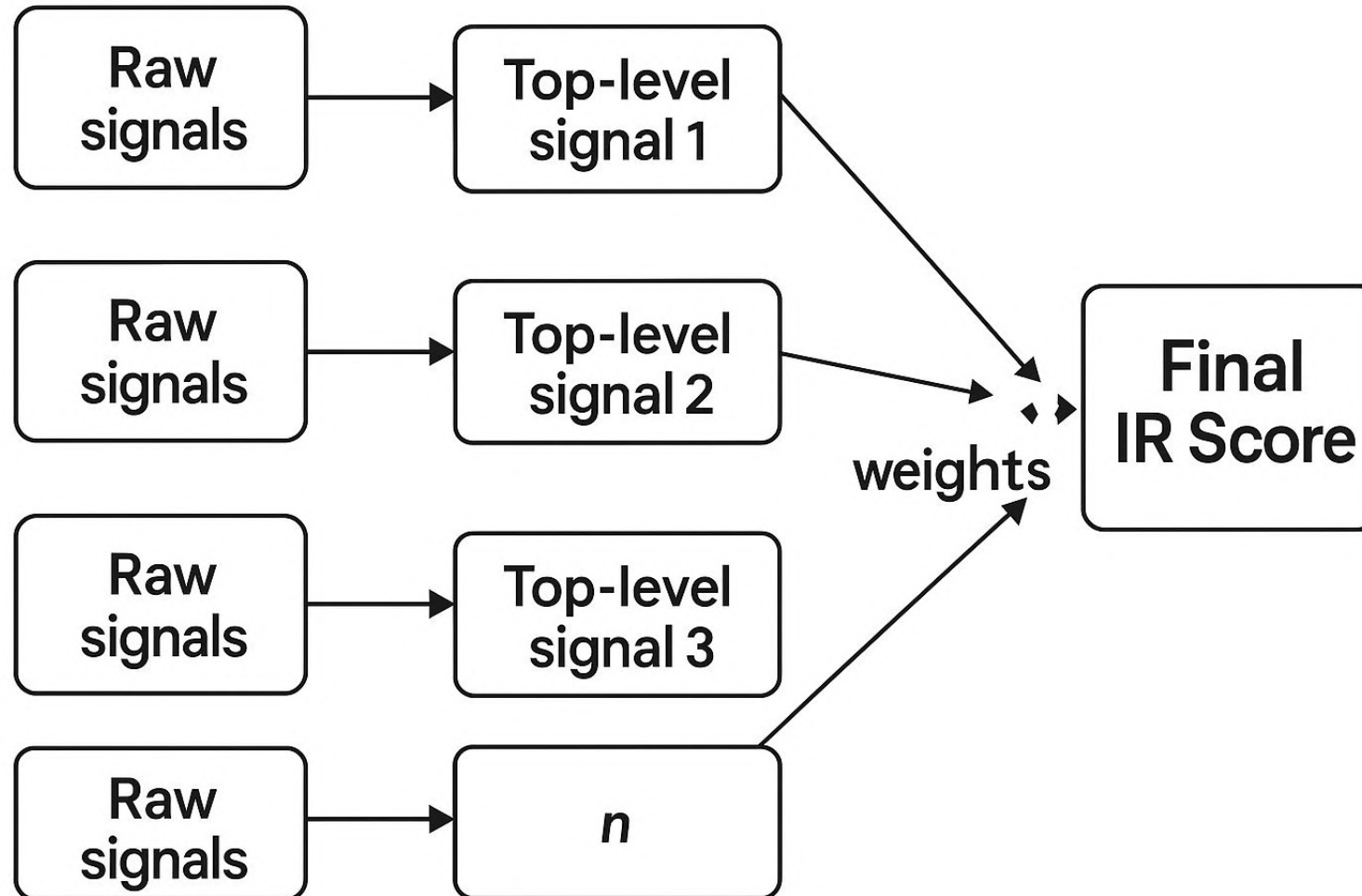


**Paul Haahr** (He/Him) · 3er  
Software engineer at Google

Y muchos más...



# Raw signals, Top-level signals, IR (Information Retrieval) Score



**T\***

**Topicality**  
*(dependiente  
de la query)*

**Q\***

**Quality**  
*(independiente  
de la query\*)*

*Probabilidad de que una página sea relevante para lo que está buscando el usuario.*

**A****Anchors**

Enlaces hacia esa página

**B****Body**

Términos incluidos en el texto de la página

**C****Clicks**

Y otras señales de interacción del usuario

## Tipos de datos de interacción (en la SERP)

- Si el usuario hace clic o no
- Si vuelve a la SERP, y cuánto tarda
- En qué resultados se detiene y sus patrones de scroll

88. There are different types of user data. Click data, for example, includes the search results on which a user clicks; whether the user returns to the SERP and how quickly; how long a user hovers over SERP results; and the user's scrolling patterns on the SERP. *See* UPX4 at .004. From such data, a GSE learns not only about the user's interests but also the relevance of the search results and quality of the webpages that the user visits. Tr. at 1767:215–1771:22 (Lehman) (discussing UPX4 at .004).

- Es una tabla gigantesca que guarda el histórico de datos de clics de los últimos 13 meses
- Activo desde 2005
- Evalúa pares query-URL. No puede contener datos sobre una URL nueva, Google necesita probarla
- Tiene en cuenta el sesgo de posición
- Datos separados para desktop / móvil, y por ubicaciones
- Puede ayudar a detectar intenciones de búsqueda
- Da mejores resultados para "freshness" que sistemas de Deep Learning más modernos



**Pandu Nayak**  
VP, Search



**Q.** Now, the navboost system memorizes past clicks that have been issued for past queries, right?

**A.** Yes.

**Q.** It's trained on user data?

**A.** Yes, it is.

**Q.** And navboost memorizes this information for all queries received in the past 13 months?

**A.** That is correct.

**Q.** And before 2017, navboost memorized this information for all queries received for 18 months?

**A.** Yes.

**Q.** And navboost, navboost slices -- I think you used the word "slice" in your deposition. Navboost slices the locale information; is that right?

**A.** The navboost does slice locale information, yes.

**Q.** And it slices the data information that it has in it by locale?

**A.** Yes.

**Q.** Navboost also slices the data information by mobile versus desktop?

**A.** Yes, it does.

**Q.** And slicing here means that navboost is creating different datasets for each of these categories?

**A.** That is correct.

**Q.** And there are differences in user intent on mobile and desktop. I think you talked about this in your direct?

**A.** Yes.

**Q.** You used the Bank of America example?

**A.** Yes, I did.

**Q.** To some extent, that's navboost, right, picking the -- oh, we're looking for local branches, we're not looking for online banking?

# useo Y si hubiéramos sabido qué buscar, en 2017



**Paul Haahr** (He/Him) · 3er  
Software engineer at Google

## Experiencia



### Distinguished Engineer

Google

mar. 2002 - actualidad · 22 años  
Mountain View, California, United

I work on Search.

11/29/2017

Paul Haahr: Google Resume

## Paul Haahr

Senior Staff Software Engineer and Manager, Search Quality

[\[redacted\]@google.com](#)

I joined Google in March, 2002. (I was on paternity leave from November, 2005 through March, 2006.)

## Accomplishments & Responsibilities

### Clicks Team (April, 2006 - present)

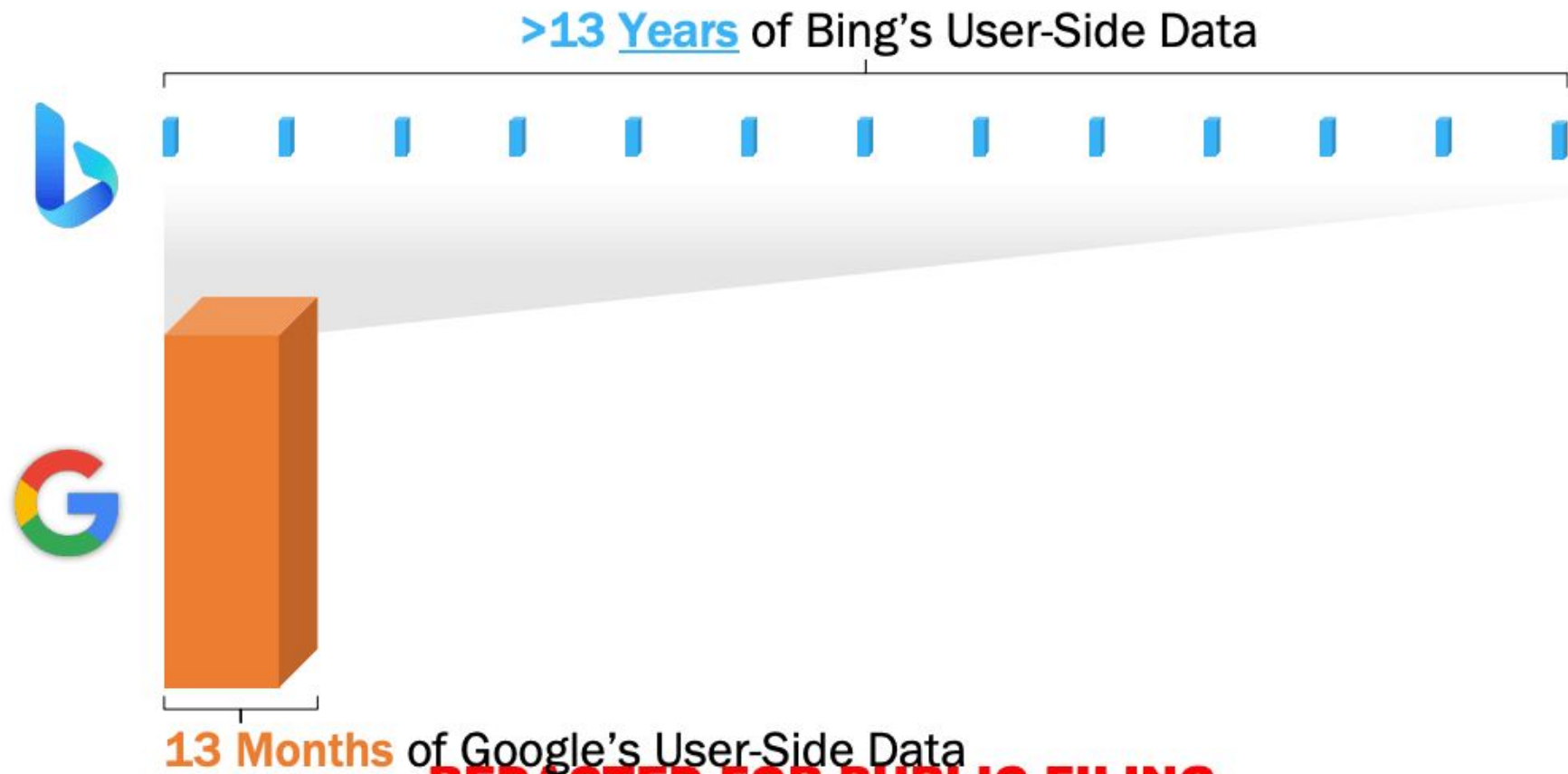
I'm the manager for logs-based ranking projects. The team's efforts are currently split among four areas:

- Navboost. This is already one of Google's strongest ranking signals. Current work is on automation in building new navboost data; recency and anti-grandfathering; and building an improved model of user behavior, taking into account impressions and display bias.
- Generalizing navboost: extending the system to help with (query, document) pairs for which we have no direct evidence.
- Mergeserver and navseek: These are really two sides of the same coin. The former is an experimental framework for comparing user behavior in the presence of changes to ranking or UI; the latter, a system for exploring new results for existing queries.
- Logging improvements for quality purposes.

useo

13 meses de datos Google = 13 años Bing

## Google Clearly Gets “Returns” from User-Side Data

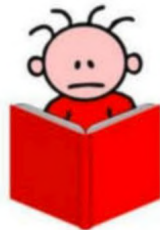


**REDACTED FOR PUBLIC FILING**

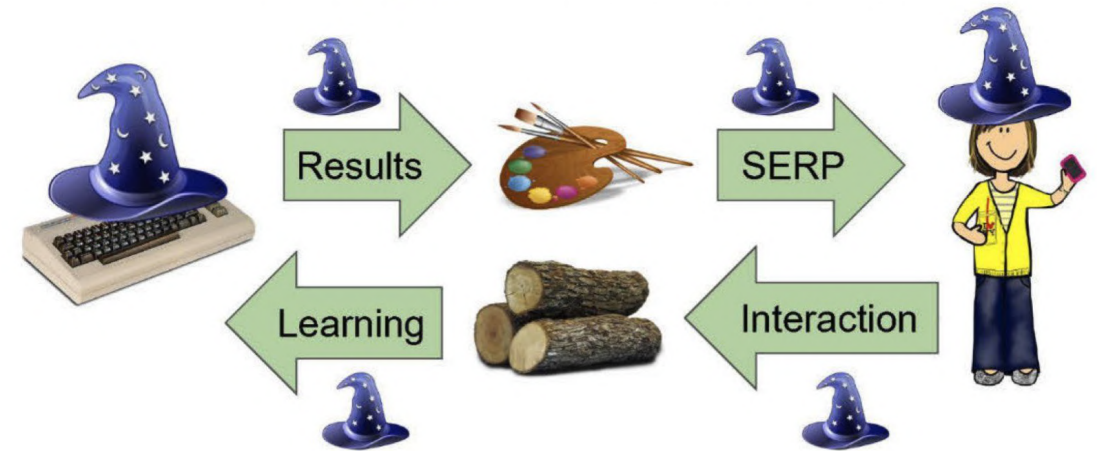
# ¿Era esta la magia de Google? (2017)

## We do not understand documents. We fake it.

- Today, our ability to understand documents directly is minimal.
- So we watch how people react to documents and memorize their responses.



This **dialogue** is the source of magic.



Let's start with some background..

A billion times a day, people ask us to find documents relevant to a query.

What's crazy is that we don't actually understand documents. Beyond some basic stuff, we hardly look at documents. We look at people.

If a document gets a positive reaction, we figure it is good. If the reaction is negative, it is probably bad.

Grossly simplified, this is the source of Google's magic.

- El cálculo de relevancia era "tradicional": similar a Okapi **BM25** (un TF-IDF ligeramente más sofisticado)
- En 2015 se inicia una transición hacia señales generadas por sistemas de Machine Learning y Deep Learning (primero **RankBrain**, luego otros)

## 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 **RankBrain** (defined in terminology above).
- Then the transition to greater reliance on machine learning / deep learning - RankEmbed (discussed below), RankBrain, DeepRank - to generate signals.

- Su objetivo es "generalizar", no depender de datos de usuarios para cada par query/url, aunque están entrenados con datos de usuarios
- Caros de ejecutar, por eso sólo actúan sobre 200 resultados mejor posicionados (RankBrain sólo 20-30)



**Juan González Villa** ✓

@seostratega



...

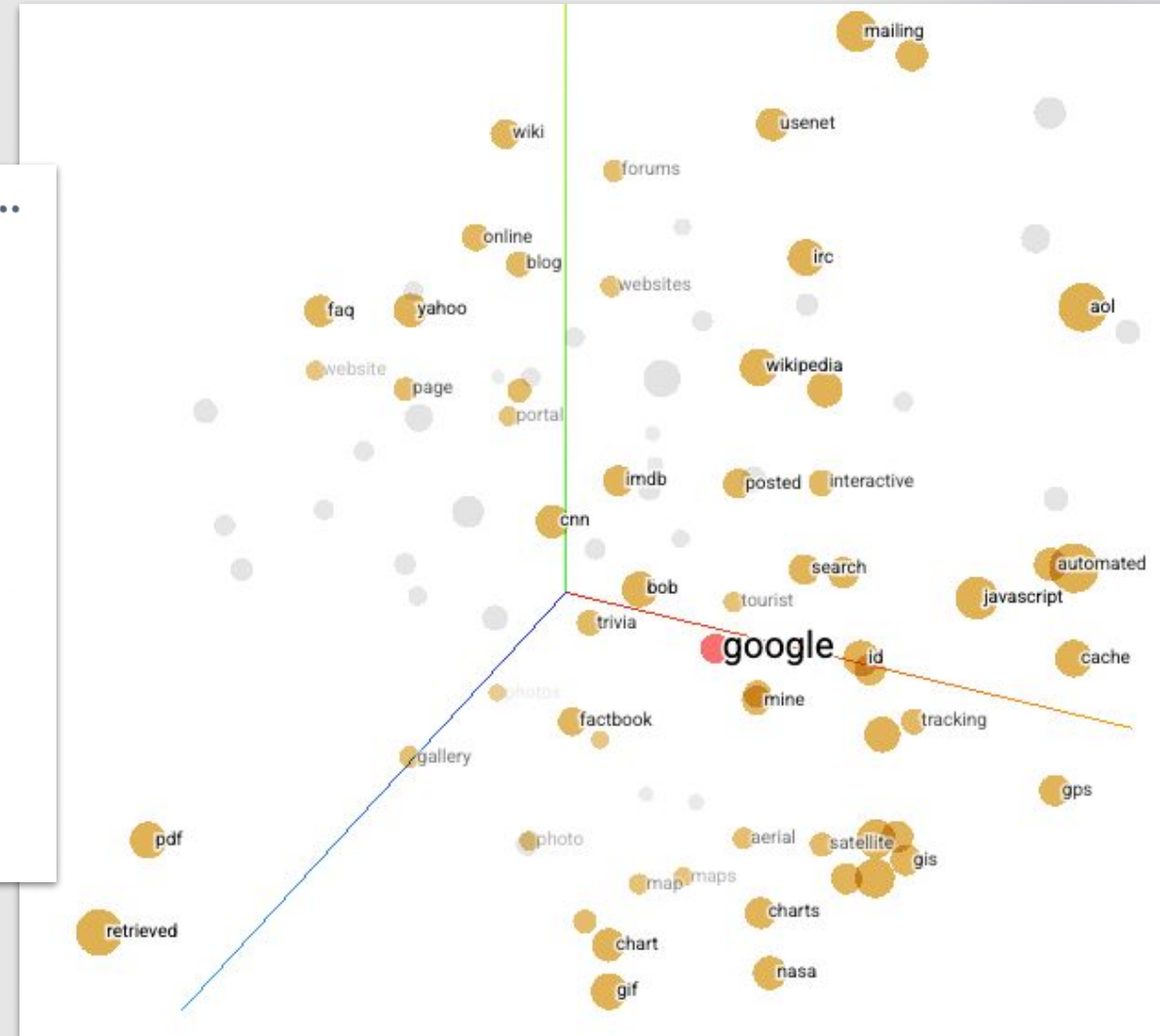
## 1. Embeddings

Representación numérica de un texto, frase, palabra o token.  
Normalmente es un vector con muchas dimensiones.

Para dos palabras con significado parecido o muy relacionadas entre sí (aparecen a menudo juntas en un corpus) sus embeddings tendrán valores cercanos.

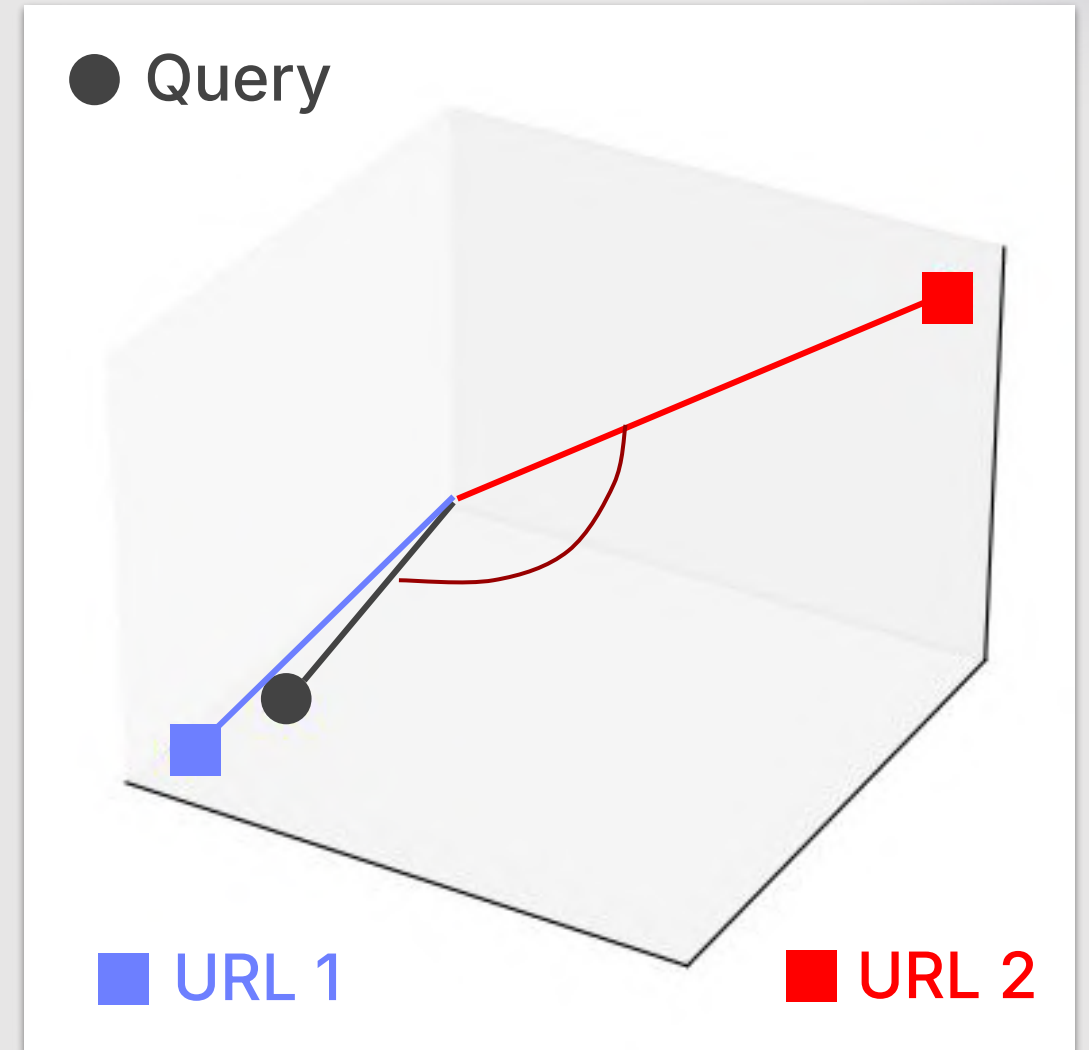
[Translate post](#)

2:39 PM · May 30, 2024 · **1,373** Views



## RankEmbed: Embeddings de query y página

- Genera embeddings para la query y para cada página, y calcula la distancia (cuanto más cerca, más similares/relevantes)
- Entrenado con un mes de datos de búsqueda
- Muy rápido, y muy efectivo para búsquedas frecuentes
- Rinde peor con las búsquedas menos frecuentes (*long tails*)
- Puede repescar páginas no consideradas por señales tradicionales



## Combinación señales tradicionales + DL

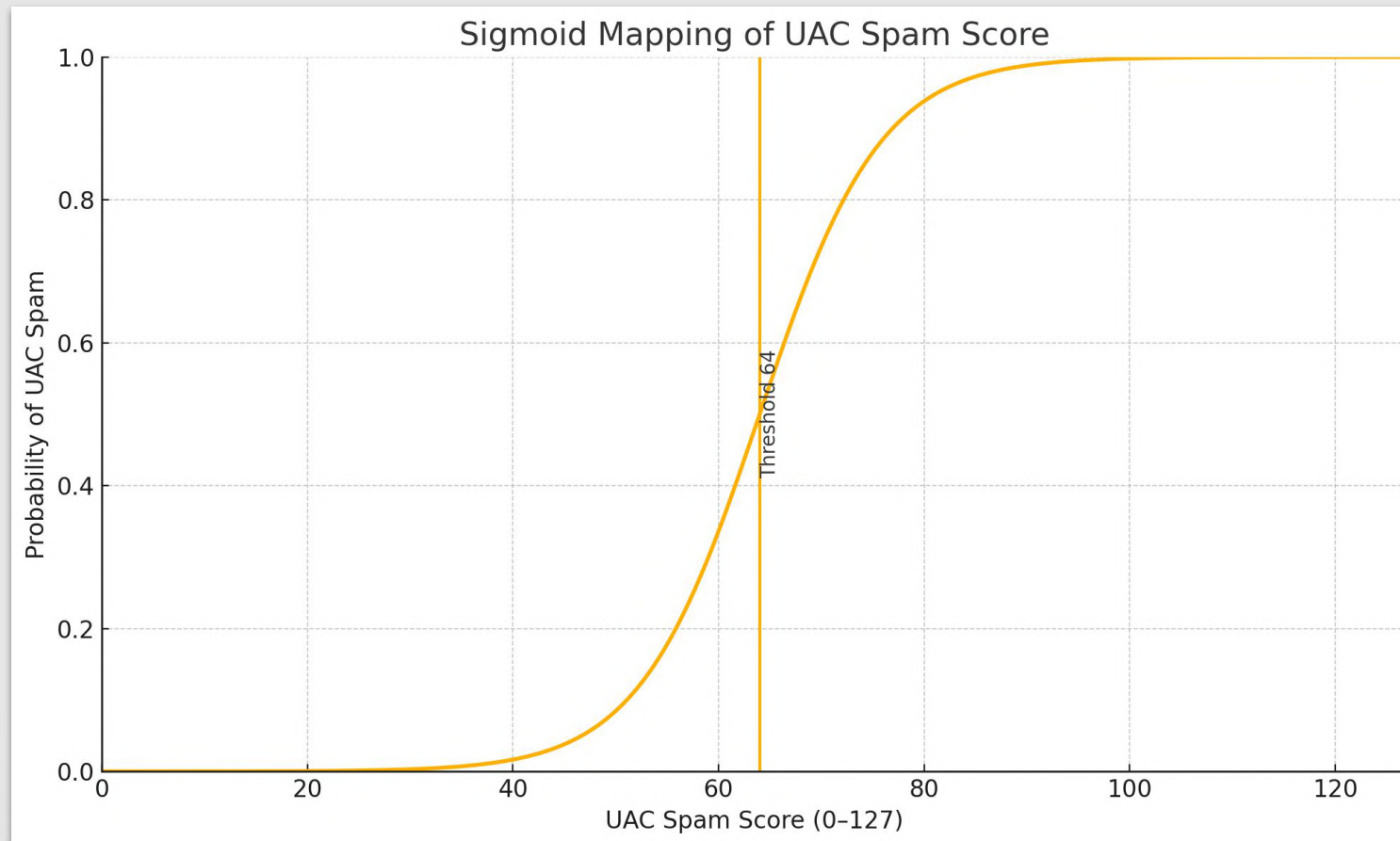
- Then the transition to greater reliance on machine learning / deep learning - RankEmbed (discussed below), RankBrain, DeepRank - to generate signals.
- Google found that the BERT-based DeepRank ML model signals could be decomposed into signals that resembled the traditional signals and that combining the traditional and “predicted” signals produced by ML resulted in better outcomes

# Los pesos o umbrales de cada señal, elegidos “a mano”

## 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.
- `uacSpamScore` (type: `integer()`, default: `nil`) - The uac spam score is represented in 7 bits, going from 0 to 127. Threshold is 64. Score  $\geq 64$  is considered as uac spam.

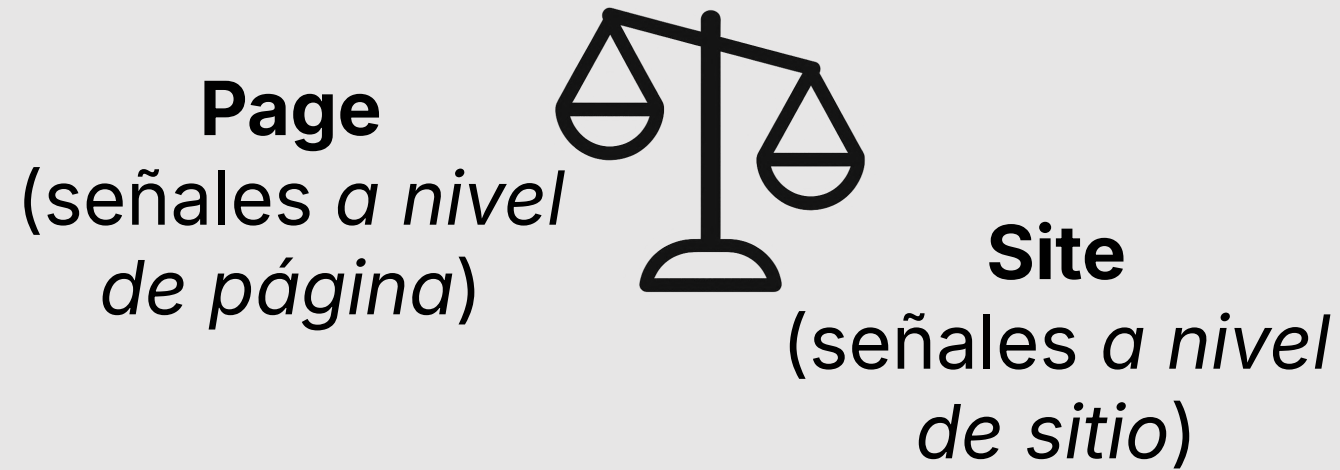
- `uacSpamScore` (type: `integer()`, default: `nil`) - The uac spam score is represented in 7 bits, going from 0 to 127. Threshold is 64. Score  $\geq 64$  is considered as uac spam.



# Page Quality = Trustworthiness / Authority

- 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.

## Señales de calidad: a nivel página o sitio



# Model.QualityNsrNsrData

## Attributes

site|

89/109

^ v x

- `ugcScore` (type: `number()`, default: `nil`) -
- `spambrainLvcScores` (type: `list(GoogleApi.ContentWarehouse.V1.Model.QualityNsrVersionedFloatSignal.t)`, default: `nil`) -
- `titlematchScore` (type: `number()`, default: `nil`) - Titlematch score of the `site`, a signal that tells how well titles are matching user queries.
- `site2vecEmbeddingEncoded` (type: `list(GoogleApi.ContentWarehouse.V1.Model.QualityNsrNsrDataEncodedEmbedding.t)`, default: `nil`) - Encoded `site`2vec embedding (to be used in superroot) since the full embeddings take too much space.
- `smallPersonalSite` (type: `number()`, default: `nil`) - Score of small personal `site` promotion go/promoting-personal-blogs-v1
- `pnavClicks` (type: `number()`, default: `nil`) - denominator for the pnav computation
- `siteChunkSource` (type: `String.t`, default: `nil`) - These are only annotated in the Goldmine NSR annotator.
- `clusterUplift` (type: `GoogleApi.ContentWarehouse.V1.Model.QualityNsrNsrDataClusterUplift.t`, default: `nil`) -
- `siteAutopilotScore` (type: `number()`, default: `nil`) - Aggregated value of url autopilot scores for this `site`chunk.
- `isVideoFocusedSite` (type: `boolean()`, default: `nil`) - Bit to determine whether the `site` has mostly video content, but is not hosted on any known video-hosting domains. `Site` is considered to be video-focused, if it has > 50% of the URLs with watch pages (with smoothing prior). ariane/4045246
- `metadata` (type: `GoogleApi.ContentWarehouse.V1.Model.QualityNsrNsrDataMetadata.t`, default: `nil`) -
- `chromeInTotal` (type: `number()`, default: `nil`) - `Site`-level Chrome views.
- `chardVariance` (type: `number()`, default: `nil`) -
- `chardScoreVariance` (type: `list(GoogleApi.ContentWarehouse.V1.Model.QualityNsrVersionedFloatSignal.t)`, default: `nil`) - `Site`-level Chard Variance for all pages of a `site`.
- `chardScoreEncoded` (type: `list(GoogleApi.ContentWarehouse.V1.Model.QualityNsrVersionedIntSignal.t)`, default: `nil`) - `Site`-level Chard (encoded as an int).

&gt;

# Algunas señales de calidad a nivel de sitio

- `siteAuthority` (*type:* `integer()`, *default:* `nil`) – `site__authority`: converted from `quality__nsr.SiteAuthority`, applied in Qstar.
- `titlematchScore` (*type:* `number()`, *default:* `nil`) – Titlematch score of the site, a signal that tells how well titles are matching user queries.
- `chromeInTotal` (*type:* `number()`, *default:* `nil`) – Site-level Chrome views.
- `siteQualityStddev` (*type:* `number()`, *default:* `nil`) – Estimate of site's PQ rating stddev – spread of the page-level PQ ratings of a site. Note this is different from `nsr__variance` which predicts error of NSR itself from the aggregated site-level rating.

useo

Mark Williams-Cook extrajo datos reales  
de Site Quality (800.000 sitios)

```
},  
"click_age_probability": 0.14759482,  
"long_answer_extension": {  
  "answer_fringe": {  
    "site_quality": 0.80060625  
  },  
},
```

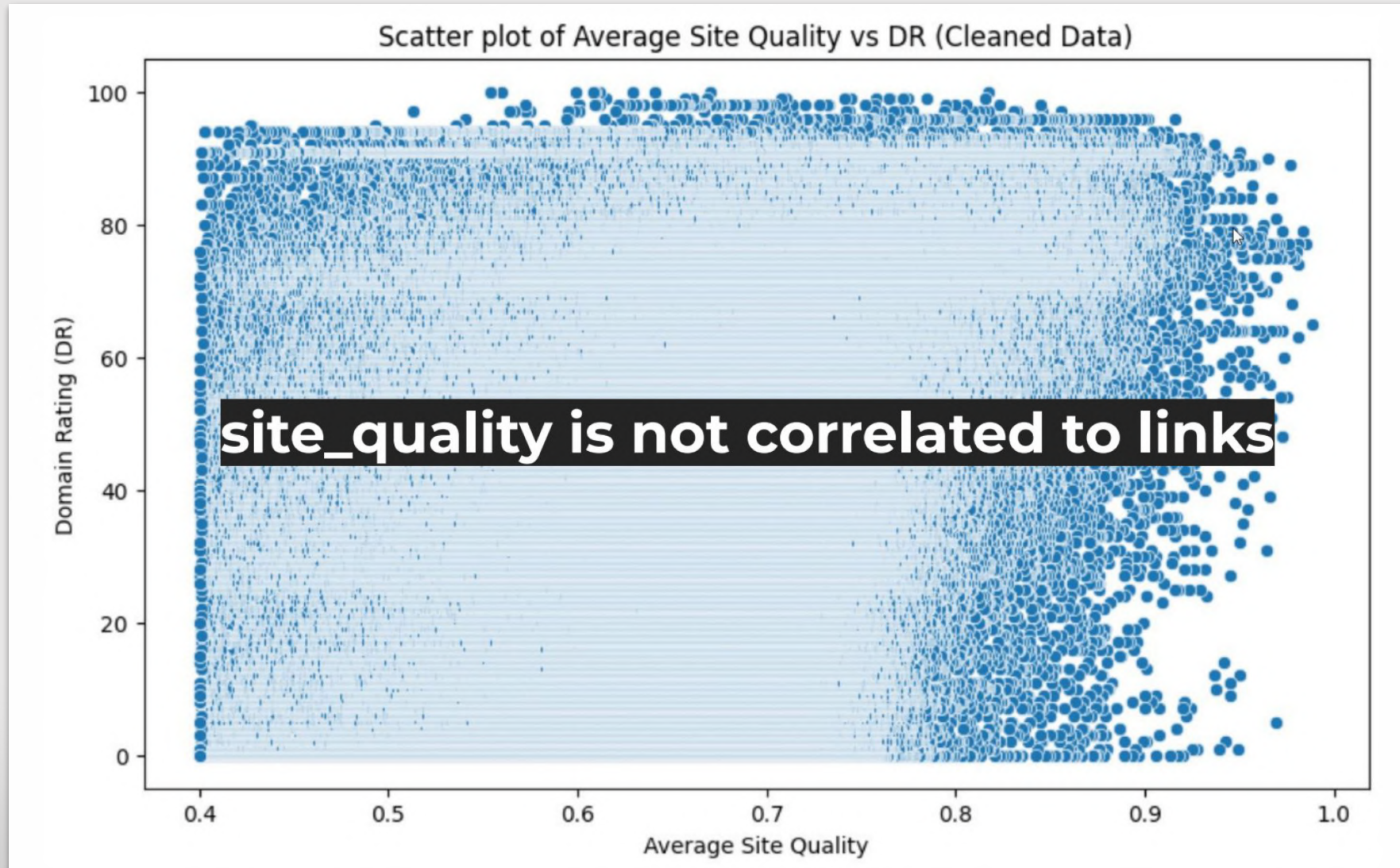
One of these was 'site\_quality'





👁️ Por debajo de 0.4 es imposible aparecer en Fragmentos Destacados / People Also Ask

No hay correlación entre site\_quality  
y los links de un sitio (DR)



## Tampoco hay correlación con tráfico o popularidad

| domain                                                     | site_quality |
|------------------------------------------------------------|--------------|
| brandservices.amazon.com                                   | 0.860030217  |
| watch.amazon.com                                           | 0.85251013   |
| logistics.amazon.com                                       | 0.843636478  |
| flex.amazon.com                                            | 0.836095889  |
| <b>Specific sites score much higher than popular sites</b> |              |
| www.amazon.com                                             | 0.7011381    |

# ¿Cómo podría funcionar Site Quality?

Un **ratio** entre:

- Número de búsquedas navegacionales o de marca (buscan específicamente ese sitio), y
- Todas las búsquedas genéricas que han supuesto al menos un clic hacia ese sitio

## Site quality score

### Abstract

Methods, systems, and apparatus, including computer programs encoded on computer storage media, for determining a first count of unique queries, received by a search engine, that are categorized as referring to a particular site; determining a second count of unique queries, received by the search engine, that are associated with the particular site, wherein a query is associated with the particular site when the query is followed by a user selection of a search result that (a) was presented, by the search engine, in response to the query and (b) identifies a resource in the particular site; and determining, based on the first and second counts, a site quality score for the particular site.

Images (4)

US9760641B1

United States

 Download PDF

 Find Prior Art  Similar

**Inventor:** [April R. Lehman](#), [Navneet](#)

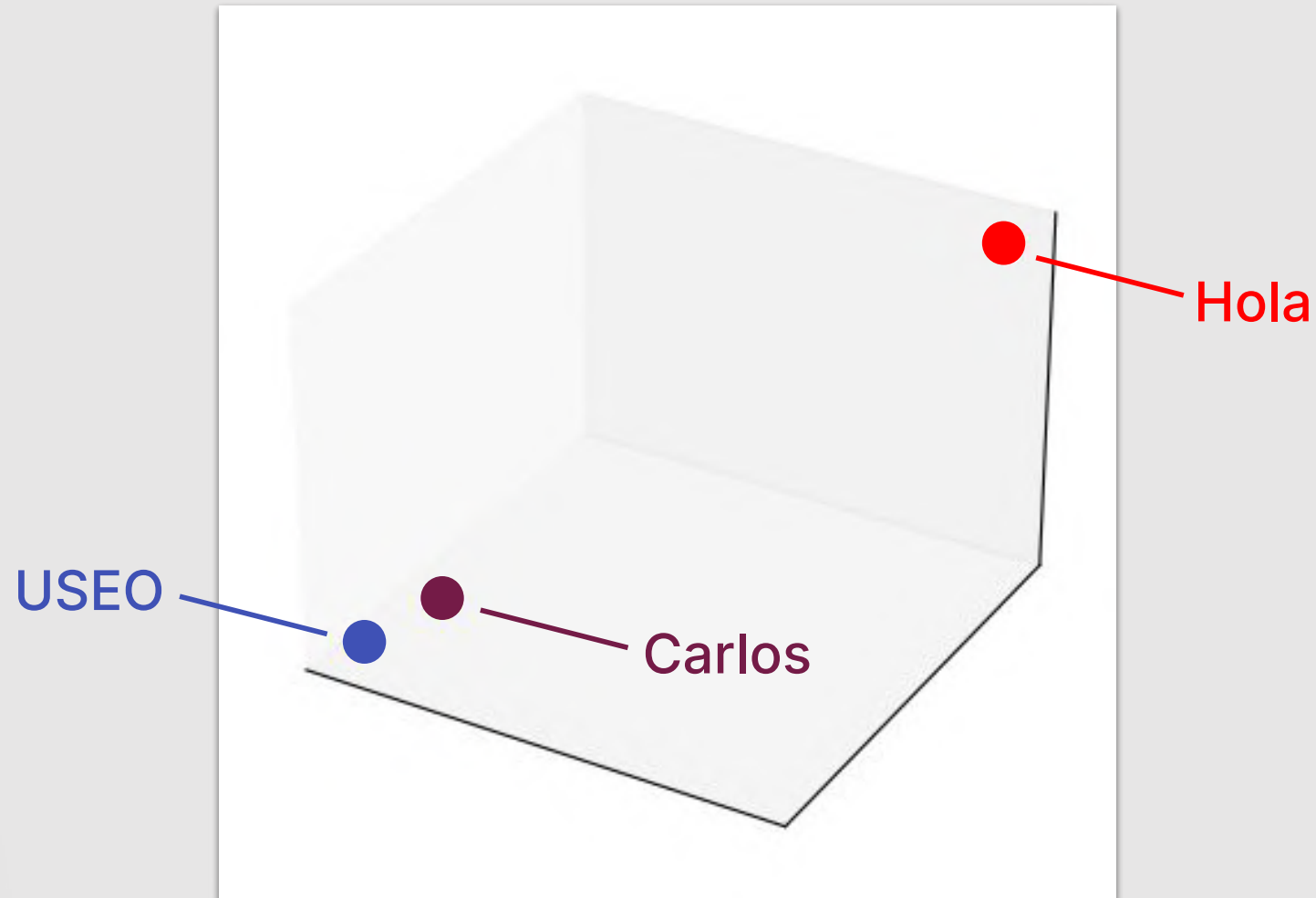
[Panda](#)

**Current Assignee:** [Google LLC](#)

Worldwide applications

2012 • [US](#) 2015 • [US](#)

- `site2vecEmbedding` (`type: list(GoogleApi.ContentWarehouse.V1.Model.QualityNsrNsrDataEmbedding.t)`, `default: nil`) – Site2vec embeddings.



Q\* es estática: no depende de la query, salvo en casos especiales

- 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 independiente + Q dependiente**

# ¿Es la autoridad temática una de estas excepciones?

## GoogleApi.ContentWarehouse.V1.Model.QualityAuthorityTopicEmbeddingsVersionedItem

Proto populated into shards and copied to superroot. Message storing a versioned TopicEmbeddings scores. This is copied from TopicEmbeddings in docjoins.

### Attributes

- `pageEmbedding` (type: `String.t`, default: `nil`) -
- `siteEmbedding` (type: `String.t`, default: `nil`) - Compressed site/page embeddings.
- `siteFocusScore` (type: `number()`, default: `nil`) - Number denoting how much a site is focused on one topic.
- `siteRadius` (type: `number()`, default: `nil`) - The measure of how far page\_embeddings deviate from the site\_embedding.
- `versionId` (type: `integer()`, default: `nil`) -

# Topic Authority: Factor reconocido para noticias

## Información sobre la autoridad de los temas de noticias

Martes, 23 de mayo del 2023

### Cómo funciona la autoridad de los temas

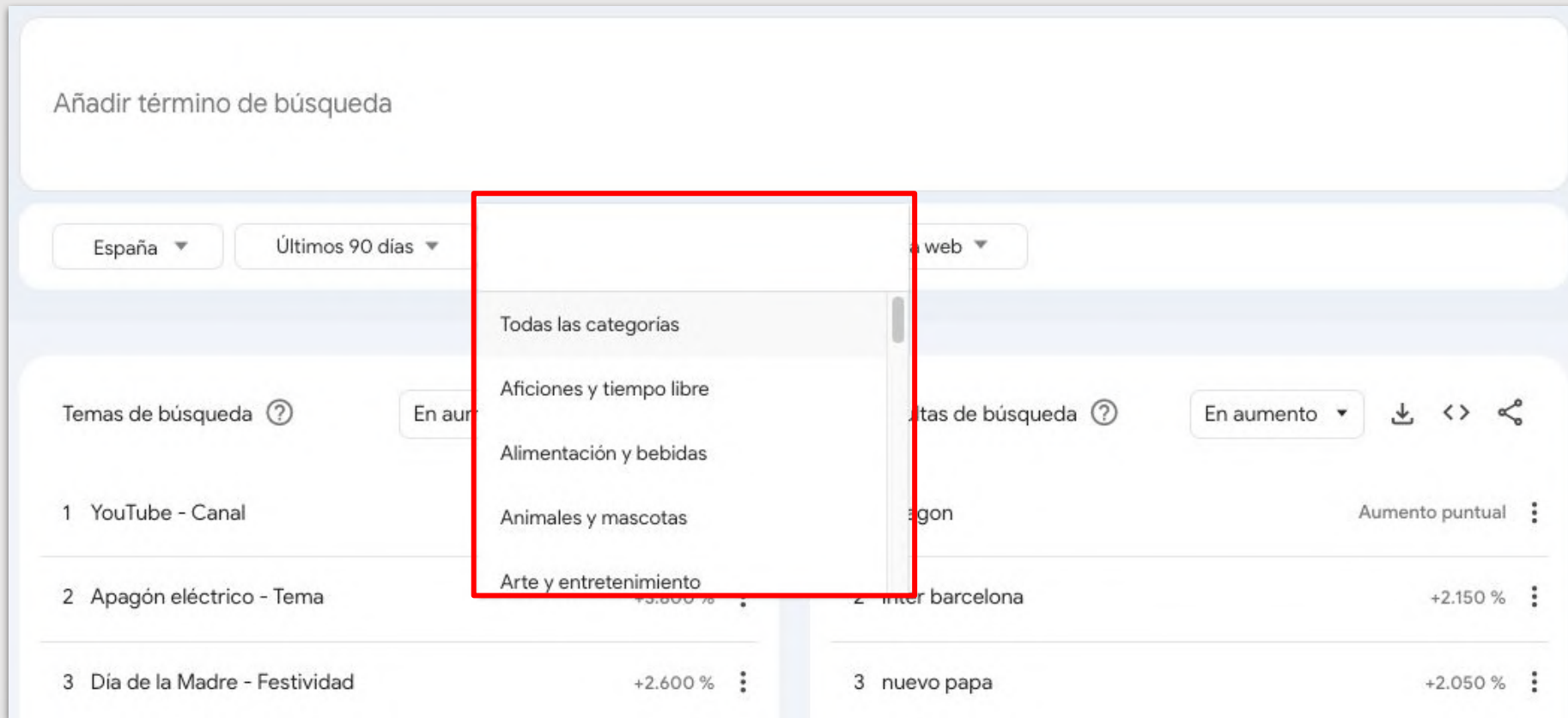
El sistema de autoridad de los temas analiza varias señales para determinar el nivel de especialización que tiene una publicación en determinadas áreas. Estas son algunas de las señales más importantes:

- **El nivel de importancia que tiene una fuente para un tema o ubicación:** nuestros sistemas entienden las publicaciones que parecen ser especialmente relevantes para ciertos temas o ubicaciones. Por ejemplo, pueden saber que los usuarios que buscan noticias sobre el fútbol americano del instituto de Nashville suelen consultar publicaciones como [The Tennessean](#) para obtener cobertura local.

## Google clasifica las búsquedas en unas 25 temáticas de primer nivel

34. Google classifies its queries by subject matter, and it has developed more than two dozen “level-one” classifications. *Id.* at 7029:2-16 (J. Baker) (discussing PSDX11 at 17). Users tend to use a GSE for a short period of time to search for a particular topic and then allow time to pass before using a GSE to search for a different topic. Put differently, users do not typically search multiple different subject matters during an unbroken time period. *Id.* at 8419:9-15 (Israel) (discussing DXD29 at 25) (78% of users searched within only one vertical in a short period). Yet,

# ¿Cuántas categorías hay en Trends? ¿Y en la API de NLP de Google?

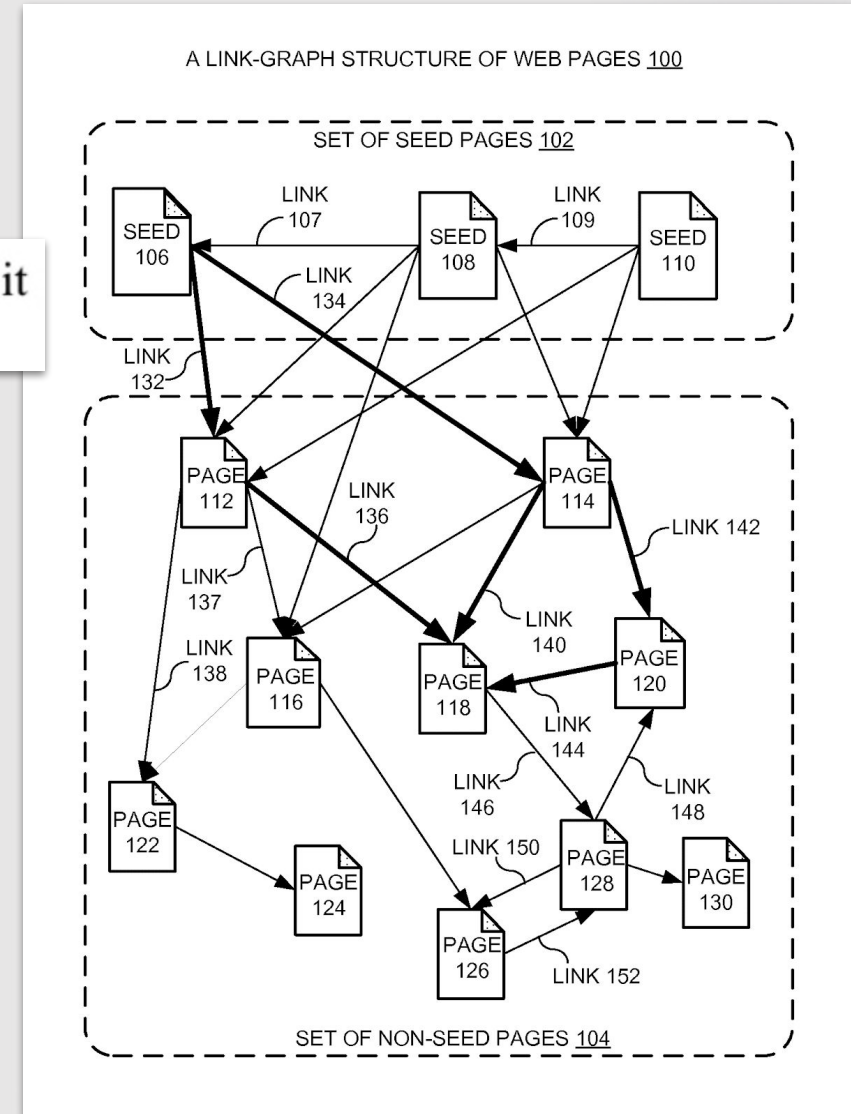


# PageRank, según distancia a la “página semilla” más cercana (Nearest Seed)

**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.

<https://patents.google.com/patent/US9165040>

<https://useo.es/patente-anti-spam-google/>



useo Cercanía a “High craps pages”, URLs para las que Google predice muchos clics orgánicos

- `onsiteProminence` (`type: integer()`, `default: nil`) – Onsite prominence measures the importance of the document within its site. It is computed by propagating simulated traffic from the homepage and high craps click pages. It is a 13-bit int.

Autocompletion results for "craps" Press RETURN for full-text search

|                                                                               |        |
|-------------------------------------------------------------------------------|--------|
| GoogleApi.ContentWarehouse.V1.Model.QualityNavboostCrapsCrapsData             | MODULE |
| GoogleApi.ContentWarehouse.V1.Model.QualityNavboostCrapsCrapsDevice           | MODULE |
| GoogleApi.ContentWarehouse.V1.Model.QualityNavboostCrapsFeatureCrapsData      | MODULE |
| GoogleApi.ContentWarehouse.V1.Model.QualityNavboostCrapsCrapsClickSignals     | MODULE |
| GoogleApi.ContentWarehouse.V1.Model.QualityNavboostCrapsStatsWithWeightsProto | MODULE |

### 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.
-  (popularity) signal that uses Chrome data.

focused, if it has > 50% of the URLs with watch pages (with smoothing prior). ariane/4045246

- `metadata` (type: `GoogleApi.ContentWarehouse.V1.Model.QualityNsrNsrDataMetadata.t`, default: `nil`) -
- `chromeInTotal` (type: `number()`, default: `nil`) - Site-level Chrome views.
- `chardVariance` (type: `number()`, default: `nil`) -
- `chardScoreVariance` (type:

# GoogleApi.ContentWarehouse.V1.Model.QualitySitemapScoringSignals

Scoring signals for **computing the sitelink score.**

## Attributes

- `annotations` (type: `list(String.t)`, default: `nil`) -
- `chromeTransCount` (type: `String.t`, default: `nil`) -
- `chromeTransProb` (type: `number()`, default: `nil`) -
- `chromeWeight` (type: `number()`, default: `nil`) -
- `country` (type: `list(String.t)`, default: `nil`) -
- `countryConfidence` (type: `list(number())`, default: `nil`) -
- `impressions` (type: `String.t`, default: `nil`) -
- `langConfidence` (type: `list(number())`, default: `nil`) -
- `language` (type: `list(String.t)`, default: `nil`) -
- `localCountryIdentifier` (type: `list(String.t)`, default: `nil`) -
- `longClicks` (type: `String.t`, default: `nil`) -
- `longCtr` (type: `number()`, default: `nil`) -
- `navboostScore` (type: `number()`, default: `nil`) -
- `navmenuScore` (type: `number()`, default: `nil`) -
- `pagerank` (type: `integer()`, default: `nil`) -
- `recentLongCtr` (type: `number()`, default: `nil`) -
- `targetCdocLanguages` (type: `list(integer())`, default: `nil`) -
- `titleScore` (type: `number()`, default: `nil`) -

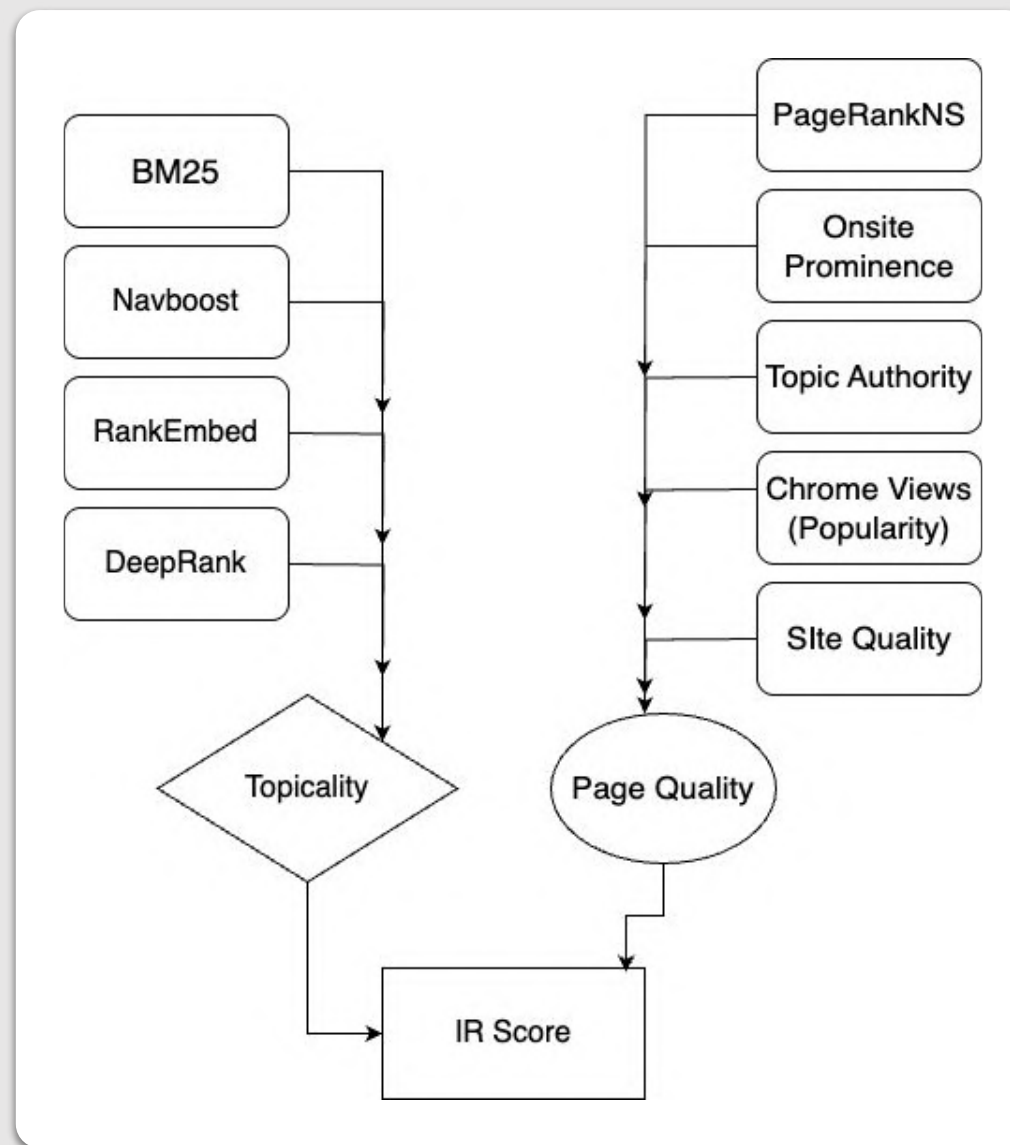
## Site Engagement

☐ Show WebUI (chrome:// and chrome-untrusted://) pages.

| Origin                           | Base  | Bonus | Total▼ |       |
|----------------------------------|-------|-------|--------|-------|
| https://mail.google.com/         | 68.55 | 0     | 68.55  | _____ |
| https://s_____.monday.com/       | 50.6  | 0     | 50.6   | _____ |
| _____                            | 45.86 | 0     | 45.86  | _____ |
| https://www.google.com/          | 42.79 | 0     | 42.79  | _____ |
| https://docs.google.com/         | 31.54 | 0     | 31.54  | _____ |
| https://teams.microsoft.com/     | 27.72 | 0     | 27.72  | _____ |
| https://lookerstudio.google.com/ | 15.95 | 0     | 15.95  | _____ |
| https://calendar.google.com/     | 9.87  | 0     | 9.87   | _____ |
| _____                            | 7.95  | 0     | 7.95   | _____ |
| https://search.google.com/       | 6.42  | 0     | 6.42   | _____ |
| https://drive.google.com/        | 5.65  | 0     | 5.65   | _____ |
| https://wordpress.org/           | 5.51  | 0     | 5.51   | _____ |
| _____                            | 4.88  | 0     | 4.88   | _____ |
| https://meet.google.com/         | 4.41  | 0     | 4.41   | _____ |
| https://accounts.google.com/     | 3.71  | 0     | 3.71   | _____ |



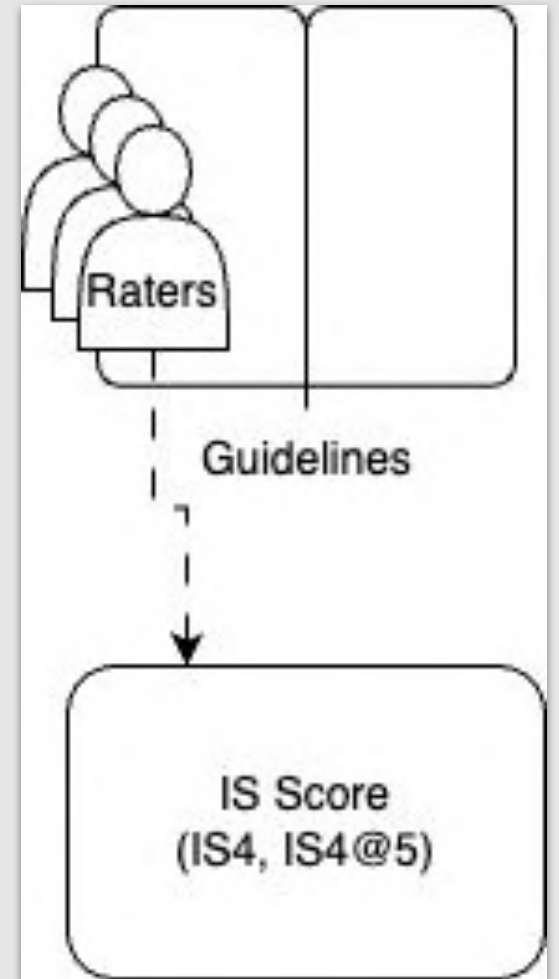
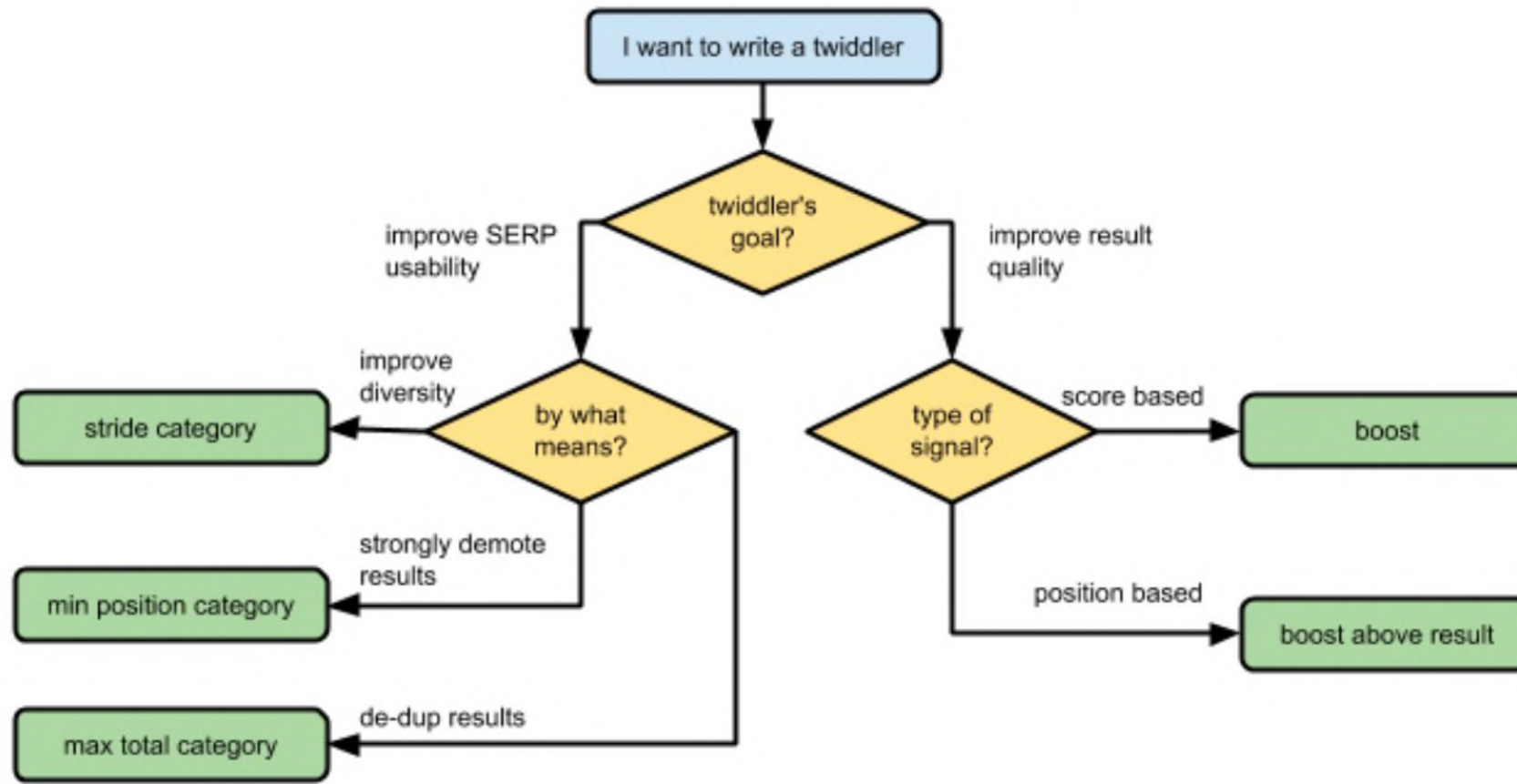
# Recapitulamos



# ¿Más? Sí, Twiddlers, el papel de los raters y muchas cosas más...

7/3/2018

Twiddler Quick Start Guide - Superroot



# Juan González Villa

Fundador y director, Agencia  
USEO

 <https://useo.es/juan-gonzalez-villa-seo/>

 [juan@useo.es](mailto:juan@useo.es)

## 10 Links Azules

