FROM PAGE RANK TO RANKBRAIN

Renée Ridgway

The concept of Page Rank has its basis in the Scientific Citation Index (SCI), a form of academic hierarchy that has now been grafted as a conceptual paradigm for the way we find information and how that information is prioritised for us. The eponymous Page Rank algorithm was developed in 1998 and is basically a popularity contest based on votes. A link coming from a node with a high rank has more value than a link coming from a node with low rank. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

Secret recipes

Presently, 'keyword search' is still the way Google Search organises the internet by crawling and indexing,1 which determines the importance of a website based on the words it contains, how often other sites link to it, and dozens of other measures. With Google Search the emphasis is to keep the attention of the user and to have them click on the higher rankings, effortlessly. However as Gillespie points out, the exact workings are opaque and vary for diverse users, "the criteria and code of algorithms are generally obscured—but not equally or from everyone" (Gillespie 185). Based on users' histories, location and search terms, the searcher is 'personalised' through a set of criteria.2 Not only are the creators of content of web pages kept in check by search engines, but the tracking of different factors, or signals, determine the ranking of an individual page. Mostly through reverse engineering, a whole 'Search Engine Optimisation' (SEO) industry has developed around 'gaming' the algorithm to figure out its recipe or signals.

Signals

During the past 18 years, Google has constantly tweaked their proprietary algorithm, containing around 200 ingredients or 'signals' in the recipe.3 "Signals are typically factors that are tied to content, such as the words on a page, the links pointing at a page, whether a page is on a secure server and so on. They can also be tied to a user, such as where a searcher is located or their search and browsing history."4 Links, content, keyword density, words in bold, duplicate content, domain registration duration and outbound link quality are some other examples of factors, or 'clues'. One of the major changes in 2010 to the core algorithm of Page Rank was the 'Caffeine' update, which enabled an improvement in the gathering of information or indexing, instead of just sorting. 'Panda' was an update that was implemented in 2011 that downranks sites, which are considered lower quality, enabling higher quality pages to rise. In April 2012 Google launched the 'Penguin' update that attempts to catch sites, and now devalues spam instead of demoting (adjusting the rank) of the entire site. As of September 30, 2016, it updates in real time as part of the core algorithm.5

Analogous to the components of engine that has had it parts replaced, where Penguin and Panda might be the oil filter and gas pump respectively, the launch of 'Hummingbird' in August 2013 was Google's largest overhaul since 2001. With the introduction of a brand new engine the emphasis has shifted to the contextual — it's less now about the keyword and more about the intention behind it — the semantic capabilities are what are at stake. Whereas previously certain keywords were the focus, at the moment the other words in the sentence and their meaning are accentuated. Within this field of 'semantic search' the 'relationality linking search queries and web documents'6 is reflected with the 'Knowledge Graph,'7 along with 'conversational search' that incorporates voice activated enquiries.

If Hummingbird is the new Google engine from 2013, the latest replacement part is then 'RankBrain'. Launched around early 2015 it ostensibly 'interprets' what people are searching for, even though they may have not entered the exact keywords. 'RankBrain' is rumoured to be the third most important signal, after links and content (words) and infers the use of a keyword by applying synonyms or stemming lists.8 The complexity level of the queries has gone up, resulting in an improvement of indexing web documents. User's queries have also changed and are now not only keywords but also multi-words, phrases and sentences that could be deemed 'long-tail' queries. These need to be translated to a certain respect, from 'ambiguous to specific' or 'uncommon to common,' in order to be processed and analysed.9 This reciprocal adaptability between the users and interface has been verified by previous research. Therefore it is probable that Google assigns these complex queries to groups with similar interests in order to 'collaboratively filter' them.10

Machine learning

"Algorithms are not always neutral. They're built by humans, and used by humans, and our biases rub off on the technology. Code can discriminate."11

As of June 2016 'RankBrain' is being implemented for every Google Search query and the SEO industry speculates it's summarising the page's content. The murmur is that the algorithm is adapting, or 'learning' as it were from people's mistakes and its surroundings. According to Google the algorithm learns offline, being fed historical batched searches from which it makes predictions. "And algorithms are made and remade in every instance of their use because every click, every query, changes the tool incrementally" (Gillespie 173). This cycle is constantly repeated and if the predictions are correct, the latest versions of 'RankBrain' go live.12

Previously there were not computers powerful or fast enough, or the data sets were too small to carry out this type of testing. Nowadays the computation is distributed over many machines, enabling the pace of the research to quicken. This progress in technology facilitates a constellation or coming together of different capabilities from various sources, through models and parameters. Eventually the subject, or learner, in this case the algorithm, is able to predict, through repetition. Where is the human curator in all of this? "There is a case to be made that the working logics of these algorithms not only shape user practices, but also lead users to internalize their norms and priorities" (Gillespie 187). The question then is to what extent is there human adaption to algorithms in this filtering or curation process, how much do algorithms affect human learning and whether not only discrimination but also agency can be contagious.13

README

'From Page Rank to Rank Brain' is an essay that attempts to 'decloak' as well as 'update' public knowledge about Google a.k.a. Alphabet's ranking algorithm. This text has then been altered through 3 'translation' processes.

Drawing on Constant's collection of scripts,1 the first translation used 'encryptionlinessha1.py' that 'provides the ultimate reduction (although at the expense of human as well as machine legibility) by encrypting every line of your text as a 128-bit hash value. Each hash value can of course be reversed again if you try to match it with every single line of every single text existing.'2 The second translation uses a little python script called the 'The Synonymizer' that corrupts your writing style by swapping out words in your text with randomized synonyms from WordNet.3 With the third translation, the text was first read with the 'text to speech' voice of 'Alex' and saved as an audio file, then uploaded to 'gentle', a robust yet lenient 'forced aligner' built on Kaldi.4 Forced aligners are computer programs that take media files and their transcripts and return extremely precise timing information for each word (and phoneme) in the media. How does it work? "As in all of these Machine Learning cases, you have to follow the data."5

1.https://gitlab.constantvzw.org/machineresearch/reduction/tree/master/filters (Proximus NV » OVH SAS)
2.In cryptography, SHA-1 (Secure Hash Algorithm 1) is a cryptographic hash function designed by the United States National Security Agency and is a U.S. Federal Information Processing Standard published by the United States NIST in 1993. SHA-1 produces a 160-bit (20-byte) hash value known as a message digest. A SHA-1 hash value is typically rendered as a hexadecimal number, 40 digits long.
3."Note:it may also corrupt the meaning of your text which replaces 'choice words' with synomyms." WordNet: http://wordnet.princeton.edu/. (Proximus NV » Hurricane Electric, Inc. » Princeton University) Thanks 2 Dave Young
4.http://lowerquality.com/gentle/ (Proximus NV » Level 3 Communications, Inc. » Advania hf. » Thor Data Center ehf)
5.In this case, it's the CALLHOME corpus, which is 120 unscripted 30-minute telephone conversations between native speakers of English in the 1990s.https://catalog.ldc.upenn.edu/LDC97S42. (Proximus NV » Cogent Communications » University of Pennsylvania) Thanks 2 Robert M. Ochshorn

Works Cited

Feuz, Martin; Fuller, Matthew; Stalder, Felix. "Personal Web Searching in the age of Semantic Capitalism: Diagnosing the Mechanics of Personalisation". First Monday, peer-reviewed journal on the internet. Volume 16, Number 2-7, February 2011. Web. http://firstmonday.org/article/view/3344/2766 (Proximus NV » Tinet Spa » Cogent Communications » University of Illinois at Chicago) Gillespie, Tarleton. "The Relevance of Algorithms". Media Technologies, ed. Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot. Cambridge, MA: MIT Press, 2014, pp. 167-193. Print.
Page, Lawrence and Brin, Sergey. The Anatomy of a Large-Scale Hypertextual Web Search Engine (1999). Web. http://infolab.stanford.edu/ (Proximus NV » Hurricane Electric, Inc. » Stanford University)~backrub/google.html

1 Since 2013, Google.com is the most visited website in the world, according to Alexa. "Google processes over 40,000 search queries every second which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide." In 1999, it took Google one month to crawl and build an index of about 50 million pages. In 2012, the same task was accomplished in less than one minute. 16% to 20% of queries that get asked every day have never been asked before. Every query has to travel on average 1,500 miles to a data centre and back to return the answer to the user. A single Google query uses 1,000 computers in 0.2 seconds to retrieve an answer. http://www.internetlivestats.com/google-search-statistics/. (Proximus NV » Level 3 Communications, Inc. » Colo4, LLC » PrivateSystems Networks)
2 No space here to elaborate, please see Personalisation as Currency: http://www.aprja.net/?p=2531 (Proximus NV » NORDUnet » RIPE Network Coordination Centre)
3 Google usually describes that is has around 200 major ranking signals, yet there have been discussions of 1000 or even 10000 sub-signals. http://searchengineland.com/bing-10000-ranking-signals-google-55473 (Proximus NV » Tinet Spa » EGIHosting)
4 http://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440 (Proximus NV » Tinet Spa » EGIHosting)
5 "Some sites want to do this because they've purchased links, a violation of Google's policies, and may suffer a penalty if they can't get the links removed. Other sites may want to remove links gained from participating in bad link networks or for other reasons." http://searchengineland.com/google-penguin-doesnt-penalize-bad-links-259981 (Proximus NV » Tinet Spa » EGIHosting)
6 According to David Amerland, author of Google Semantic Search. http://searchengineland.com/hummingbird-has-the-industry-flapping-its-wings-in-excitement-reactions-from-seo-experts-on-googles-new-algorithm-173030 (Proximus NV » Tinet Spa » EGIHosting)
7 Knowledge Graph was launched in 2012 and combines 'semantic search' information added to search results so that users do not query further. However this has lead to a decrease of page views on Wikipedia of different languages. https://en.wikipedia.org/wiki/KnowledgeGraph (Proximus NV » RIPE Network Coordination Centre » Telia Company AB » Wikimedia Foundation, Inc.)
8 In regard to information retrieval, 'stemming' is when words are reduced to their 'stem' or root form. "Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation". https://en.wikipedia.org/wiki/Stemming (Proximus NV » RIPE Network Coordination Centre » Telia Company AB » Wikimedia Foundation, Inc.)
9 http://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440
10 http://firstmonday.org/article/view/3344/2766
11 Victoria Turk. http://motherboard.vice.com/enuk/read/when-algorithms-are-sexist (Proximus NV » Belgacom International Carrier Services SA » Amazon.com, Inc.)
12 http://searchengineland.com/faq-all-about-the-new-google-rankbrain-algorithm-234440
13 During the writing of my PhD I use Google Search for

my research and have allowed myself to be personalized on my Apple computer without installing plugins, etc. that would attempt to prevent it.