Name: Avani Gupta

Roll number: 2019121004

Assignment - 2

Study Google architecture paper, look for more recent information on Google search engine architecture. List down the top 10 improvements.

There were many imporvements after the google was launched. They introduced NLP, Machine Learning, Deep Learning approches afterwards. Many goolge servies like Google assistant are an add-on. Here are significant improvements google did :-

1. Broad Core Algorithm Update – March 2019 Core Update

It aimed to provide better results for the search queries

- It automatically down ranked sites even though it wasn't the sites' fault
- You could do nothing to 'fix' the sites to increase their rankings again

Google announced the update via Twitter on March 12th, 2018

Google also mentioned that the update was to assist pages which were under-rewarded but had great usage for the audience and encouraged such sites to continue to create quality content.

2. Panda

Launched: Feb 24, 2011

Rollouts: ~monthly

Goal: De-rank sites with low-quality content

Google Panda is an algorithm used to assign a content quality score to webpages and down-rank sites with low-quality, spammy, or thin content. Initially, Panda was a filter rather than a part of Google's core algorithm, but in January 2016, it was officially incorporated into the ranking algo. While this doesn't mean that Panda is now applied to search results in real time, it does indicate that both getting filtered by and recovering from Panda now happens faster than before.

3. Penguin

Launched: April,24,2012

Rollouts: May 25, 2012; Oct 5, 2012; May 22, 2013; Oct 4, 2013; Oct 17, 2014; September 27, 2016; October 6, 2016; real-time since

Goal: De-rank sites with spammy, manipulative link profiles

Google Penguin aims to identify and down-rank sites with unnatural link profiles, deemed to be spamming the search results by using manipulative link tactics. Since late 2016, Penguin is part of Google's core ranking algo and operates in real time, which means that penalties are now applied faster, and recovery also takes less time.

4. Pirate

Launched: Aug 2012

Rollouts: Oct 2014

Goal: De-rank sites with copyright infringement reports

Google's Pirate Update was designed to prevent sites that have received numerous copyright infringement reports from ranking well in Google search. The majority of sites affected are relatively big and well-known websites that made pirated content (such as movies, music, or books) available to visitors for free, particularly torrent sites. That said, it still isn't in Google's power to follow through with the numerous new sites with pirated content that emerge literally every day.

5. Hummingbird

Launched: August 22, 2013

Rollouts: —

Goal: Produce more relevant search results by better understanding the meaning behind queries

Google Hummingbird is a major algorithm change that has to do with interpreting search queries, (particularly longer, conversational searches) and providing search results that match searcher intent, rather than individual keywords within the query.

While keywords within the query continue to be important, Hummingbird adds more strength to the meaning behind the query as a whole. The use of synonyms has also been optimized with Hummingbird; instead of listing results with the exact keyword match, Google shows more theme-related results in the SERPs that do not necessarily have the keywords from the query in their content.

6. Pigeon

Launched: July 24, 2014 (US)

Rollouts: December 22, 2014 (UK, Canada, Australia)

Goal: Provide high quality, relevant local search results

Google Pigeon (currently affecting searches in English only) dramatically altered the results Google returns for queries in which the searcher's location plays a part. According to Google, Pigeon created closer ties between the local algorithm and the core algorithm, meaning that the same SEO factors are now being used to rank local and non-local Google results. This update also uses location and distance as a key factor in ranking the results.

Pigeon led to a significant (at least 50%) decline in the number of queries local packs are returned for, gave a ranking boost to local directory sites, and connected Google Web search and Google Map search in a more cohesive way.

7. Mobile Friendly Update

Launched: April 21, 2015

Rollouts: —

Goal: Give mobile friendly pages a ranking boost in mobile SERPs, and de-rank pages that aren't optimized for mobile

Google's Mobile Friendly Update (aka Mobilegeddon) is meant to ensure that pages optimized for mobile devices rank at the top of mobile search, and subsequently, down-rank pages that are not mobile friendly. Desktop searches have not been affected by the update.

Mobile friendliness is a page-level factor, meaning that one page of your site can be deemed mobile friendly and up-ranked, while the rest might fail the test.

8. RankBrain

Launched: October 26, 2015 (possibly earlier)

Rollouts: —

Goal: Deliver better search results based on relevance & machine learning

RankBrain is a machine learning system that helps Google better decipher the meaning behind queries, and serve best-matching search results in response to those queries.

While there is a query processing component in RankBrain, there also is a ranking component to it (when RankBrain was first announced, Google called it the third most important ranking factor). Presumably, RankBrain can somehow summarize what a page is about, evaluate the relevancy of search results, and teach itself to get even better at it with time.

The common understanding is that RankBrain, in part, relies on the traditional SEO factors (links, on-page optimization, etc.), but also looks at other factors that are query-specific. Then, it identifies the relevance

features on the pages in the index, and arranges the results respectively in SERPs.

9. Possum

Launched: September 1, 2016

Rollouts: —

Goal: Deliver better, more diverse results based on the searcher's location and the business' address

The Possum update is the name for a number of recent changes in Google's local ranking filter. After Possum, Google returns more varied results depending on the physical location of the searcher (the closer you are to a certain business physically, the more likely you'll see it among local results) and the phrasing of the query (even close variations now produce different results). Somewhat paradoxically, Possum also gave a boost to businesses that are outside the physical city area. (Previously, if your business wasn't physically located in the city you targeted, it was hardly ever included into the local pack; now this isn't the case anymore.) Additionally, businesses that share an address with another business of a similar kind may now be de-ranked in the search results.

10. Fred

Launched: March 8, 2017

Rollouts: —

Goal: Filter out low quality search results whose sole purpose is generating ad and affiliate revenue

The latest of Google's confirmed updates, Fred got its name from Google's Gary Illyes, who jokingly suggested that all updates be named "Fred". Google confirmed the update took place, but refused to discuss the specifics of it, saying simply that the sites that Fred targets are the ones that violate Google's webmaster guidelines. However, the studies of affected sites show that the vast majority of them are content sites (mostly blogs) with low-quality articles on a wide variety of topics that appear to be created mostly for the purpose of generating ad or affiliate revenue.

2)Explore the concept of barrels and describe how different indexing units can co-exist.

Barrels are memory locations present in secondary memory where partially indexed pages are (maybe stored by docID in case of forward Index)stored. The indexer distributes the hits(which record the word, position in document, an approximation of font) size, and capitalization into a set of "barrels", creating a partially sorted forward index.

Each barrel holds a range of wordID’s. If a document contains words that fall into a particular barrel, the docID is recorded into the barrel, followed by a list of wordID’s with hitlists which correspond to those words. The inverted index consists of the same barrels as the forward index, except that they have been processed by the sorter. For every valid wordID, the lexicon contains a pointer into the barrel that wordID falls into. It points to a doclist of docID’s together with their corresponding hit lists. This doclist represents all the occurrences of that word in all documents. We chose a compromise between these options, keeping two sets of inverted barrels -- one set for hit lists which include title or anchor hits and another set for all hit lists. This way, we check the first set of barrels first and if there are not enough matches within those barrels we check the larger ones. After each document is parsed, it is encoded into a

number of barrels. Every word is converted into a wordID by using an in-memory hash table -- the lexicon.

New additions to the lexicon hash table are logged to a file. Once the words are converted into wordID's, their occurrences in the current document are translated into hit lists and are written into the forward barrels. In order to generate the inverted index, the sorter takes each of the forward barrels and sorts it by wordID to produce an inverted barrel for title and anchor hits and a full text inverted barrels.

Different Indexing units co-exist in barrels. Parallel operation is performed on barrels as described above. The main difficulty with parallelization of the indexing phase is that the lexicon needs to be shared. Writing a log of all the extra words that are not in a base lexicon helps. That way multiple indexers can run in parallel and then the small log file of extra words can be processed by one final indexer.