Brock Cooper
u0821793
3/8/2015

Part 3: Algorithm Research - PageRank

PageRank is the original Google search algorithm that was named after Larry Page, one of the founders of Google. Larry Page and Sergey Brin developed the algorithm at Stanford University in 1996. PageRank works by counting the number of links and judging their quality to determine how important a website is. The basic idea is that more important websites are likely to receive more links from other websites. [1]

Google utilizes a web crawler to systematically browse the web. A web crawler starts with a list of URLs, visits the URLs, identifies, those hyperlinks and recursively repeats the process [2]. As the web crawler is visiting each website a PageRank is also assigned to the page. Every website is given a PageRank score between 0 and 10 on an exponential scale. The web pages with the highest score have the greatest number of links to their pages. USA.gov, Twitter, and Adobe Reader download are some examples of websites with a PageRank of 10 [3].

The PageRank of one page also depends on the PageRank of the pages pointing to that page. The tricky part is that we don't know what the PageRank of those pages are until we calculate the rank of the pages pointing to those pages. Pages also can point to each other so the calculation gets to be very complex or even impossible to imagine. PageRank is calculated on an iterative basis, meaning that you can calculate the PageRank of one page without knowing the final value of its links and each time you run the calculation you are only closer to finding out the final value of the page's PageRank.

Google has since added many countless aspects of their ranking algorithm to counteract SEO tactics such as Google Bombs or Googlewashing, which artificially trick Google's PageRank into thinking it was a highly ranked page to appear at the top of search results [5].

Google has many competitors that use a variety of different page ranking algorithms. Since PageRank is patented, other search engines don't use the exact calculation and process of PageRank but they do use similar assumptions and ideas to build their ranking to produce similar results. The HITS algorthim used by Twitter and Ask.com rates pages with two scores: 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 [7]. The TrustRank algorithm used at Yahoo makes a point to identify and separate spam links from useful links by first having an expert manually examine a set of seed links and having the crawler seeking out similar types of links as the ones examined by the expert [8]. These various techniques have likely been implemented in similar fashion at Google as well.

References:

[1] http://web.archive.org/web/20090424093934/http://www.google.com/press/funfacts.html
[2] http://en.wikipedia.org/wiki/Web_crawler
[3] http://www.bruceclay.com/blog/what-is-pagerank/
[4] http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm
[5] http://en.wikipedia.org/wiki/Google_bomb
[6] https://www.youtube.com/watch?v=KyCYyoGusqs
[7] http://en.wikipedia.org/wiki/HITS_algorithm
[8] http://en.wikipedia.org/wiki/TrustRank