Evan Carnevale

Professor Albluwi

CSC 110

March 13th, 2016

Great Algorithms Part II

Have you ever looked for your missing favorite shirt that you wore last week in order to make sure it gets properly washed? As a human, the task is simple: there is a dirty pile of laundry that needs to be sorted in order to find your favorite shirt. One would simply go through the dirty pile until they found their shirt. But what would you do if your wardrobe was the size of Nordstrom’s and the dirty pile of laundry was considerably larger? The task of finding your favorite shirt would certainly be more difficult and time consuming. Similarly, this relates to search engine indexing and how it finds a particular search query (“the needle”) in the “world’s largest haystack”, the World Wide Web. The search engine index algorithm is broken into two phases depending on the user’s query. In a matter of seconds, the search engine matches and ranks the results from the endless web based on their relevance to the query. Thus the importance of this algorithm is evident based on timing and relevance when displaying information to the user. Historically, indexing has its roots in a very basic form but its effectiveness has come a long way for the method to be used in applications that are used every day, millions of times a day. The search engine index algorithm is crucial and efficient in the display of the most relevant information based upon a user’s query, which will continue to be successfully utilized by millions of people, millions of times a day now and in the future.

The index algorithm is considered one of the oldest useful ideas in computer science, but its history is relatively short in regards to search engines. As noted by archaeologists, indexing is almost as old as writing itself. These archaeologists found a 5000-year-old Babylonian temple library that cataloged its cuneiform tablets by subject. In 1994, companies such as Infoseek and Lycos were some of the earliest commercial offerings that utilized the index algorithm but in 1995, AltaVista launched its search engine, which became known as the “king of search engines” at that time. Their algorithm, like many other search engine index algorithms, successfully processed a vast amount of information available by providing quality and relevant results in the fastest way possible. Since the initial development, many different companies have improved the index algorithm including the three that dominate the market today: Google, Yahoo and Microsoft’s Bing. Google especially is atop of the market because their main goal, according to Ben Gomes the Vice-President of Engineering is “to make improvements to search that just answer the user’s need and gets them to the exact answer faster and faster” (Google). Their algorithm attempts to implement a seamless connection between the user’s thoughts, their information needs and the search results that they find.

The index algorithm is a fundamental element behind the structure of any effective search engine. Early on and with the explosion of new information available on the web, it was getting increasingly hard to find the exact content that the user wanted. By using this algorithm, user’s have available to them a vast amount of quality information in a matter of seconds based on their search query. By working similar to the index of a book, the web index of a query for a search engine is broken up into two phases: matching and ranking. The user will first enter a query, which is treated as a word list. The matching phase will match all the pages that mention the query, which is typically a very large number of cases. This presents a problem because there is so much information from just the query that needs to be narrowed down to the select handful of results that will provide the most useful information the user is looking for. The information on each page is also treated as a word list where each position corresponds to a page and the order of words. By comparing these positions, phrase queries can narrow down and eliminate many of the thousands of hits based on the location and order of words from the query. More importantly, this is where the second phase, ranking, begins. Ranking takes the narrowed down results from the query and puts them in order of relevance. In turn, this provides the most useful information hits in order of most useful to less useful for the user. Also, by using nearness location of phrase queries, search engines are able to improve their rankings just based upon whether or not they are located near the next word in the query. Finally, the index algorithm uses a “metaword trick” which narrows the ranking process down even further by looking at the titles of pages based on the query. Pages that tend to have the query in their title are more likely to contain relevant information than pages that just mention the query in the body. A simple example of how the index algorithm works would be to assume that the World Wide Web only consisted of three pages. Page one has the words “the cat sat on the mat”, page two has the words “the dog stood on the mat” and page three has the words “the cat stood while the dog sat”. For each word on each page, there is a specific location representing the page and the number where that word resides on the page. For example, “the” occurs on all the following positions: 1-1, 1-5, 2-1, 2-5, 3-1. If the query were “the cat”, the search engine would use nearness to see that “cat” appears on 1-2 and 3-2, which both match the query because the second number of each position increments by one in comparison to the positions for “the”. This means that those particular words are located next to each other. The phrase in the query determines how each page will be ranked because relevance is based on how suitable or useful a given page is. By using the index based on the query, word location helps solve the issue of ranking.

Two applications that use the index algorithm are search engines and books. These both relate because they operate similarly to one another, although search engines tend to be much faster than humans when it comes to searching the whole World Wide Web. The index of a book is a list of words or phrases that have associated locations where useful material relating to that heading can be found within the book. For example, a book on science could have an index entry “physics 168, 180”. This indicates that the pages 168 and 180 have the word and information about “physics” on their pages. Search engines work the same way, only on a grander scale. Google for example, allows the user to type in a query which then displays all the pages, ranked in order from most to least relevant, based on the particular query. The effectiveness of their algorithm solely relies not on the number of times the query was matched but based on their ranking system. On September 26th, 2013 Google released one of the most significant enhancements to the search engine algorithm to date, Hummingbird. According to analytics expert Neil Patel, “Hummingbird gives Google a precise and fast platform where search users can easily find what they’re looking for when they type a given keyword in the search engine”. Google stepped away from the importance of keywords and instead focused on the user. Patel also states that, “This new algorithm makes use of over 200 ranking factors to determine the relevance and quality score of a particular site” (Neil Patel). By updating their ranking system, Google began to address the questions the users may have in the best way possible rather than focusing on the ranking of the query.

The search engine index algorithm definitely is important and relevant to society, especially at the rate that information is transmitted today. Prior to the 1990s, search engines were mainly ineffective in the sense that it took a long time to search the web for exactly the information that is needed. Search engines have improved with time as many companies at the same time have improved their index algorithm. Companies such as Google, Yahoo and Microsoft have put many others out of business because their algorithms provide the most efficient and most relevant results based on the query. Frankly speaking, their algorithms are simply better than those of other companies in terms of the display of relevant information. Immediately, the effects are seen; Information at hand today comes rapidly and is at the fingertips of anyone with access to a search engine whether it is via smartphone, computer or tablet. The irony of this effect is that it has instilled impatience with users who have to wait more than just a few seconds to get the new information that they wish to learn about! In the long term, the effect of the search engine index algorithm will bring relevant and current information to users faster and faster. Vice-President of Engineering at Google Ben Gomes stated it perfectly when he said that Google’s main goal is to create a “seamless connection” between the user’s thoughts and the results of the new information they searched for.

The search engine index algorithm will definitely have effects on the future. The overall effectiveness of search engines could result in a loss of detail for users and their ability to effectively scan through documents to pick up information. People are already becoming accustomed to the rapid access of information, but this may be a double-edged sword because search engines are in a sense “trimming the fat” off of new information and providing exact answers to user queries. By using a query but leaving a semi-broad scope within the search engine allows the user to sort through information themselves and thus extract the information based on their particular search. Search engines are in place to eliminate the thought behind users having to locate the information by hand. Although there could be a loss of intellect on the users end by just being given an answer, the benefits of this algorithm are extremely clear. Users are provided new information about a subject in less than a second. The algorithm allows the opportunity for users to read about a historical event from 1960 and then research the effects of texting and driving in less than a minute. The only way the algorithm may improve is in the ranking system, which each company individually hires researchers working to improve their index algorithms each day. According to a study conducted by Professor Thorsten Joachims, searchers typically have a strong bias when it comes to ranking. Joachims used a search engine and swapped the order of the first two search hits. Despite its relevance to the query, users still clicked on the top entry 34%of the time and on the second hit 12%of the time (Nielsen). This bias strongly shows that the user is gullible but also, it shows to companies like Google just how important the ranking system is in displaying the correct, relevant information to the user. This indicates that the future of the search engine index algorithm is bright and will always be relevant as the process of transmitting information only becomes more rapid.

# Bibliography

Google. The Evolution of Search. 27 November 2011. 13 March 2016 <https://www.youtube.com/watch?v=mTBShTwCnD4>.

Neil Patel. The Ultimate Google Algorithm Cheat Sheet. 17 February 2015. 13 March 2016 <http://neilpatel.com/2015/02/17/the-ultimate-google-algorithm-cheat-sheet/>.

Nielsen, Jakob. The Power of Defaults. 26 September 2005. 13 March 2016 <https://www.nngroup.com/articles/the-power-of-defaults/>.