Project 5: Data Bias in Web Results

Gilbert Yap
December 6, 2020
EC601 Product Design in Electrical and Computer Engineering

At some point in the 21st century, one of the default answers to difficult questions became "You should Google it". However, being the world's first choice for information retrieval comes at a price. Society has entered an age where it is not just concerned with the preservation and transmission of data, but also with its integrity and credibility. Two major features of modern websites are their search functionality and the related "autocomplete" feature. As websites and search engines' userbase exponentially increased, automated learning-based systems were deployed to help improve these two features to accommodate the influx of data. Unfortunately, both of these features have opportunities to spread misinformation through the data presented to users. The order in which information is displayed in search functions has an effect on what people consider to be "important" or popular. Even before conducting a search, autocomplete gives users an impression of what other people are looking for, but misleading suggestions can perpetuate false information or harmful ideas. In order to prevent sensationalism, websites need to be aware of the biases and potentially problematic data that they serve to their users.

The Pew Research Center stated in 2019 that as many as 1 in 5 Americans get their news from Twitter [3]. Social media has continued to grow out of its early usage as a way to connect friends and family, and into a platform for any person or organization to keep followers up to date on their products, services, or general thoughts. Users want to unify their feeds and sources of news so that they have one fewer website to visit or one fewer password to remember. Built-in search features in sites such as Facebook and Twitter provide users with a way to directly search for desired information across their website, but they need be careful about the way that data is presented. [4] states that users tend to click links near the top of the search results and place more importance on them if it matches the criteria of their query. Often these links are sorted in specific ways relating to popularity, though there are cases where websites will implement a unique sorting algorithm that takes into account user clicks and time spent on a specific link. This seems like a good idea at first, since websites believe that links that more users go to must be more important or more relevant.

Kulshrestha [4] conducted an experiment during the 2015 Democratic and Republican National Conventions with the goal of observing bias created by Twitter's search result ranking system. She found that while the content of all resulting tweets from a search like "Bernie Sanders" had a slight Democratic-bias, the ranking system placed even more tweets with Democratic-bias near the top of the search results, giving the impression of a heavily Democratic-bias. This may seem like an obvious result when searching for a Democratic candidate, but in fact a similar phenomenon occurred for searches for "Donald Trump". The overall bias of the returned tweets for "Donald Trump" was neutral (this can be misleading since a neutral bias score can mean there were an equal amount of biased tweets in both directions that lead to a neutral average), which could be hard to believe for some users. However, the same phenomenon did not occur in the search results for "Ted Cruz" and "Marco Rubio", both of which had an small Republican-leaning bias among all returned tweets, but the top search results seemed more heavily Republican-biased. Switching the method of sorting from Twitter's default to the "Most Favorited" or "Most Liked" methods barely changed the bias scores. The conclusion that was drawn was that the most popular tweets dictated the bias of the search results, even if the total pool of returned tweets did not support same political leaning.

While it is true that 80% of tweets come from the top 10% of Twitter users [3], the ranking of search results can make it appear that the majority of users are supporting one perspective. This can be an issue for users who are looking for neutral sources of information. Contradictory information can be a source of contention for users who identify with a specific poltical party. If you are a Republican but your search for a Republican candidate is filled with Democratic-biased content, you may feel like your political party is not being represented properly. If your search result confirms your own bias (logged-in users may receive further ranking bias than users without accounts) then your social media feed may start to resemble an echo chamber [1], where the diversity of ideas shrinks and users' feeds become increasingly one-sided. Kulshrestha calls this effect

"ranking bias" and proposes her methodology to be used on other social media websites or search engines to determine their search result bias. Since the method only relies on public information, third-party investigators can audit the search functionality of other websites using her same methods.

In some instances, a complete search query is not even needed for data to be presented. The concept of "auotcomplete" is beneficial not only to the user, but to the developer as well [2]; search suggestions reduce the number of unique searches performed on the search engine, making it easier to serve users who are looking for the same information. Since so many people use search engines like Google or Bing, these companies employ methods to have the autocomplete suggestions learn from the most popular queries. Unfortunately, several issues have arisen from this. Search suggestions can be problematic or harmful to users in the form of misinformation, slander, unintended adult content, promotion of illegal or harmful behavior, etc [2]. For example, the organization UN Women [5] held an ad campaign to raise awareness of real Google autocomplete results in 2013 related to women. Examples include "women shouldn't [have rights]", "women cannot [drive]", and "women should [stay at home]" (the [] portions represent a suggestion). It is unclear if these became top suggestions due to manipulation of the autocomplete system or if they were searched organically. Regardless, these suggestions perpetuate harmful stereotypes about women and serve as proof that search suggestions needs to be moderated in some capacity, which can be a challenge in of itself.

The issue with moderation of search results is that there is no hard rule as to what is deemed appropriate or inappropriate. For example, the query "where are guns" could be a user about to research gun violence ("where are guns the most used") or it could be a user looking to purchase a gun ("where are guns available"). If purchasing a gun is illegal in the country that the user resides in, the engine may block both of these suggestions unintentionally. Some moderation methods which Google has employed is a mix of filtering and blocking [2] to prevent the display of problematic or controversial search terms.

As of December 5th 2020, a started search for "is donald t" shows most suggestion for Donald Trump, but a search for "is hilary c" does not give any suggestions at all. This appears to be to prevent the dissemination of false information regarding Hilary Clinton that arose during her campaign in 2016. This could be viewed by supporters of Hilary Clinton as Google favoring Donald Trump by allowing his searches to be displayed, while Hilary Clinton's are removed. Google's response to instances of this type of blocking [2] was to assure users that they are still free to conduct their search, albeit without search suggestions. The users most harmed by this would be users with physical disabilities that makes typing difficult, as it could prevent them from easily searching for certain information. Search engines like Google have a moral obligation to prevent harm to its users, but compromise could be at the expense at some of their users.

For some, learning that there exists bias in everyday online dependencies like Google is akin to learning that some doctors receive kickbacks from pharmaceutical companies. When the internet made its way into the average person's home, people trusted that it would be used for good. Like all tools it can also be used maliciously if left without rules and guidelines. With as much as a fifth of the American population using a website like Twitter to get news information, social media companies need to be more careful about the bias caused by ranking of their data when presented to users. Companies like Google need to additionally understand how features like autocomplete can influence the thoughts and perceptions of users before they can even begin their web searches. As a society, our handling of the ever-growing heap of online data can influence the progression or regression of cultural ideas. Groups like UN Women demonstrate that although many perceived issues related to gender to be "resolved", sexism still exists and is reinforced by modern day technology. Luckily, there are many researchers like Kulshrestha who are investigating these websites and are presenting their information to the public to help curb misinformation and bias from unintentional or malicious uses of web data.

References

[1] Matteo Cinelli et al. *Echo Chambers on Social Media: A comparative analysis*. 2020. arXiv: 2004.09603 [physics.soc-ph].

[2] Timothy J. Hazen et al. *On the Social and Technical Challenges of Web Search Autosuggestion Moderation*. 2020. arXiv: 2007.05039 [cs.CY].

[3] Adam Hughes and Stefan Wojcik. *10 facts about Americans and Twitter*. Sept. 2020. URL: https://www.pewresearch.org/fact-tank/2019/08/02/10-facts-about-americans-and-twitter/.

[4] Juhi Kulshrestha et al. "Quantifying Search Bias". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Feb. 2017). DOI: 10.1145/2998181.2998321. URL: http://dx.doi.org/10.1145/2998181.2998321.

[5] *UN Women ad series reveals widespread sexism*. Oct. 2013. URL: https://www.unwomen.org/en/news/stories/2013/10/women-should-ads.