The Digital Divide in Search Query Composition Patterns

Amelia R. Malone

Department of Computer Science
University of Maryland
College Park, Maryland 20742
amalone2@terpmail.umd.edu

ABSTRACT

Web search is an essential task for internet users to find information for daily activities; however, little research addresses the differentiated usage patterns and search query composition patterns for different demographics of search users. This paper hopes to continue to broaden the understanding of the digital divide to include search query composition patterns associated with different demographics of users. By understanding the search query composition patterns, particularly with searches for educational and employment opportunities, we hope to create a framework for understanding and studying algorithmic discrimination based on usage patterns. As a first step, we conducted an exploratory online search activity in which participants entered three unique search queries for each of five prompts and then answered demographic questions and internet usage questions. By correlating micro-trends from the search queries to the demographics of the participants, we hope to provide a foundation to understand possible demographic differences in search queries.

1. INTRODUCTION

Web search is an important internet task that helps users find relevant information about topics they find important. Many search topics can directly impact users' economic and personal wellbeing, such as health information, educational advice, and job searches. Although search is a necessary task for virtually all internet users, researchers have found differences between the usage patterns of different demographics of users. Researchers have also found that differences in search formulation based on different demographic groups can be used to infer demographic characteristics of users of unknown demographics [1]. This differentiated use and the ability to infer demographic information about users is a possible path in which to study digital discrimination. Research has previously found this discrimination by studying the differences in ads selected for display while using White or African American identifying names in search queries [2]. Researchers have also uncovered differences in the ads served to profiles identified with certain demographics of users [3].

Relatedly, the *digital divide*, a term for unequal access to computing resources, has been documented in a variety of situations, such as computers and internet connections. Even when these differences in access are addressed, the digital divide continues to remain in the different usage patterns of different groups, such as older users spending less time online even when they have comparable access to computers as younger users [4]. These remaining differences are known as the *second-level digital divide* [5].

Previous research establishes a connection between different search usage patterns with different demographics of users, and previous research also establishes a *second-level digital divide*. However, to date, we have found no research linking the differing usage patterns of search users to algorithmic discrimination or the potentially different outcomes that come from differentiated usage.

Our research attempts to identify trends in the different searches constructed by different demographic groups that could be considered differentiated usage and thus potential patterns for differentiating and studying the search experience for different users. To address this question, we conducted an exploratory internet-based activity and survey of 80 participants. We asked participants to create three different search queries in response to five search prompts about job or education related information, and answer demographic and technology usage questions in order to understand more about their background. We were most interested in understanding income or wealth differences, with some secondary interest in other demographics.

We focused our research on search queries because of their widespread use by a large cross-section of the population and their importance in helping people access information they deem necessary. We chose to study search queries for educational and employment information because they are topics that span different demographics.

We found some micro-trends which included men being more likely to search for a specific programming language than women, the majority searches for all prompts included many of the words of the prompt, and one of the prompts included a high number of sentence or question responses.

Future work could build upon these micro-trends by testing if there is a connection between these differentiated search patterns and potentially discriminatory search results or advertisements.

2. RELATED WORK

In this section, we discuss prior research in three related research areas: the digital divide, search engine personalization, and online discrimination via search engines.

2.1 The Digital Divide

The digital divide, or unequal access to technology, is an important force in shaping the experience of many technology users [6]. The digital divide is known to include differences in access to computers or the internet, as well as skill differences in users. Differences in usage patterns between those with access to computing resources remain even when access issues are resolved, so the digital divide is more complex than simply computer access. For example, the frequency of web search is different amongst users of different demographics and is more closely related to educational background than demographic characteristics such as gender [7]. Differences in technical skill and online abilities were also found to be related to socio-economic status, and this online ability was also determined to be related to differentiated use of technology. Researchers believe that users are more comfortable using the internet to complete certain tasks when those users have more technological skill [5].

There are also cultural differences that separate users, such as age and race. These cultural differences in turn have been found to influence the amount of time spent using technology and in what way that time is spent [8]. Differentiated usage also includes the *knowledge gap,* which can influence what people search online and what information they are able to find through those searches [4].

There has been significant effort to mitigate the physical digital divide through policy initiatives, such as ensuring computers in lower-income schools, but less effort has been made to diminish the digital divide that comes from this usage divide, or *second-level digital divide*. This is thought to be because this secondary digital divide is seen as a perpetuation of previous divides, such as differences in literacy rates showing up in a new context [9].

2.2 Search Engine Personalization

Search Engine personalization is an important usability feature of modern search engines. Personalization can change the search results for a user based on their location and demographic information, which can help the user more easily locate pertinent information. Personalization on Search engines, such as Google, has been found to account for on average 11.7% of differences in search results; however, this personalization was also only found to occur on that search engine when users were logged into their accounts [10]. Although personalization algorithms used by major search companies remain opaque, researchers have been able to successfully visualize users' personalization of search results [11]. Due to the prevalence of search personalization, there is concern by some that this personalization will give people search results that make them less likely to see a diverse set of opinions [12].

Researchers have also found differences in search formulation for different demographic groups, which can then be used to infer demographic characteristics of users of unknown demographics, such as guessing gender with 80% accuracy [1].

2.3 Discrimination in Web Search

Search engines have also been studied in the context of algorithmic discrimination.

Researchers found differences in Google ad delivery based on whether the search term was a black or white identifying name, with black identifying names being more likely to show advertisements related to finding arrest records than white identifying names, regardless of the actual criminal record for each of the names [2].

Researchers have also found discriminatory results relating to Google's relationship between user behaviors, user profiles, ads, and ad settings. The gender of the profile searching the web influenced the ads being served to the user, such that women were served ads for high paying jobs less frequently [3].

It is not currently well understood how search personalization contributes to discrimination and bias with differentiated results and ads.

3. METHODOLOGY

We conducted an online activity in order to answer our research questions. As a part of the online activity, participants were given five prompts for which they had to compose three unique search queries each. The prompts were designed to ask participants about education and job searches they, or someone they imagine, would enter.

Our institution's Internal Review Board (IRB) approved the study. We now discuss our recruitment process, procedure and analysis, and limitations of our results.

3.1 Recruitment

We recruited participants from Amazon Mechanical Turk (AMT) and Craigslist during February and March 2016, and successfully collected answers from 80 participants, 37 from AMT and 43 from Craigslist. We limited AMT participants to those in the United States and with a 90%+ approval rating for their work. We were not able to choose the demographics of the AMT respondents. We submitted Craigslist postings to the most trafficked Craigslist sites in the United States, including New York City, San Francisco, Washington, D.C., and Seattle, and posted on the Craigslist board dedicated to recruiting volunteers on those city sites [13]. Participants were asked to respond to the ad via email in order to receive a demographic survey. If respondents met the qualifications (Over 18, living in the US, and fluent in English) and added to the diversity of the participant pool, they were invited to complete the search activity. Our AMT participants received $1 in compensation for their time via the AMT platform, and our Craigslist participants were compensated with a $3 Amazon gift card via email.

3.2 Procedure

After completing a consent form, participants were shown an attention-check passage adapted from prior work, which included answering a text question and completing a simple math problem [14]. Due to a high number (N=8 of 22) of Craigslist participants failing the attention check, we changed it to be more explicit, and fewer Craigslist participants (N=0 of 29) failed the revised version. Only one of 44 participants on AMT failed the original reading check; this high pass rate could be due to the relative frequency of reading or quality checks in AMT tasks, with AMT workers knowing that their work can be rejected should they fail.

After the attention check, participants were asked to create three searches for each of five different situations. The prompts were as follows:

- Pretend you were interested in learning to code in an online setting. What are three different searches you might type into a search engine, such as Google or Bing? Searches may include more than one word.
- If you were interested in getting a degree from an online setting, what are three different searches you would put into your search engine, such as Bing or Google? Searches may include multiple words.
- Pretend that you wanted to get a GED. What are three different searches you would put into your search engine to help you learn about getting a GED? Searches may include multiple words.
- Pretend you were interested in getting a job, with your current qualifications, and were entering searches into your search engine, such as Google or Bing. What are three different searches you would put into a search engine? Searches may include more than one word.
- Imagine that you have a GED and are interested in finding a job in the healthcare field. What are three different searches you would do to search for educational opportunities? Searches may include multiple words.

We chose to use employment- and education-related prompts because potential discrimination in these search topics is important to understand, and income differences may be particularly salient in these topics.

Participant answers were rejected if they incorrectly answered the reading check questions or if they disregarded directions by writing the same search query twice for the same prompt.

Finally, participants completed a demographic survey consisting of questions about their gender, ethnicity, and household income. The demographic survey also probed in which industry their job was, their current employment status, and their mother's highest level of educational attainment, which has been suggested as a strong indicator, or even a proxy, of socio-economic status [15]. They were also asked to rate their knowledge of internet terms on a 5-point Likert scale from "No Knowledge" to "Expert Knowledge", in order to measure their web-use skill in the context of the general population. These terms have been shown to be a better proxy for online skills than the number of hours participants spend online [16]. We also asked participants about their knowledge and use of online education and job search sites, taken from the ten most popular online learning sites, the ten most popular online colleges, and the ten most trafficked employment sites [17] [18] [19]. For the questions asking about their search engine of choice, web browser of choice, web-use skill term knowledge, and knowledge and usage of education and employment sites, the answers were displayed in a random order.

3.3 Analysis

We analyzed the search queries using an open-coding process on MAXQDA 12 software, done by one researcher [20]. The codes were applied to each search individually. We analyzed the majority of the results through frequency counts. Since each participant wrote three queries per prompt, we note in the results which results are in the context of the total number of queries, and which are in the context of the number of users with that particular search query characteristic. If a participant had at least one of their three prompts with that characteristic, we counted the participant as having that search query characteristic.

In order to more easily see trends between different demographic groups, such as women and men, or participants with a household income under $30,000 and participants with a household income over $30,000, we calculated the term frequency-inverse document frequency (tf-idf) of the searches for each prompt individually, as well as each prompt divided by gender, household annual income under or over $30,000, and those with and without a college degree. In order to calculate the tf-idf of the searches, each search was considered its own document and the set of all searches for a particular prompt was used for all total documents.

3.4 Limitations

Our methodology has several limitations related to using surveys and conducting research over Internet platforms. Participants may not have given complete or correct answers, so wherever possible, self-reported answers were cross-checked with data collected by the survey platform, such as longitude and latitude points and zip code. Participants may also have participated more than once. To mitigate this possibility, the directions clearly stated that participants would only be compensated once for completing the study, longitude and latitude points were checked to ensure that each response was from a relatively unique geographic area, and Craigslist respondents were asked to complete a qualifying survey to catch potential participants responding more than once before they completed the online activity. Participants on AMT were discouraged from completing the activity multiple times through quality checks on the Qualtrics platform and by each participant entering their unique AMT identifying number.

Participants could also respond to survey questions without reading the directions and/or questions, so an attention check was used to exclude participants who did not or could not read and understand the directions.

We piloted the search prompts to ensure that the wording was not biasing, and to test the topics participants would be asked about. Admittedly, the chosen topics may not be optimal for finding income differences amongst different users. Our sample size may be too small for finding and understanding these particular differences and are too small to make statistical claims about our results.

Because we used online recruitment, our sample may not representative of the general population. When prompting users to make search queries on a given topic, it is inevitable that word choice in the prompt will influence participants' answers at least somewhat. We attempted to mitigate this word choice bias by asking participants to write three searches.

4. RESULTS

In this section, we discuss the results of our study. First, we will give an overview of the demographics and technology usage of all participants, and then we give a demographic and technology comparison between the Craigslist participants and the AMT participants. We then go on to discuss micro-trends found in the answers to the educational prompts and the employment prompts. We then discuss trends that occurred across the prompts.

4.1 Participants

We sent the link with the Craigslist demographic survey to 72 potential participants. We received 75 responses to our Craigslist demographic survey, three of which were participants repeating the survey. After rejecting potential participants who took the survey more than once, potential participants younger than 18, and potential participants outside of the United States, we asked 51 participants to complete the study. Forty-nine participants went on to complete the search activity and two of the participants failed to start the search activity. For the AMT participants, but all but one of the AMT participants who started the activity and demographic survey completed the entire activity and survey. We had a total of 44 AMT participants complete the survey and activity, 43 of whom correctly passed the reading test.

In total, we had 80 participants successfully complete both the demographic survey and the internet activity, by answering all of the questions, answering the reading checks correctly, and giving three distinct searches for each prompt. Of those 80 participants, 40 were men and 40 were women. Our participants ranged in age from 18-60, with the majority of our participants (N=56) between the ages of 23 and 44. All of our participants had at least a high school degree or the equivalent, with four having only a high school degree, about a quarter finishing some college (N=22), the majority completing an Associate's or Bachelor's degree (N=43), and a smaller number completing a Graduate or Professional degree (N=10). Our participants were also asked about their mother's highest level of education completed, and about a quarter of them responded with some high school or a high school degree (N=4, N=14), about a quarter responded with come college (N=18), the majority responded with a college degree (N=37), and a handful had mothers with a Graduate or Professional degree (N=6).

Our participants included ten with an annual household income of under $15,000 per year, eight with a household annual income of $15,001-$30,000, 35 with a household income of $31,000-$60,000, 18 with a household income of $61,000-$100,000, and nine with an annual household income of over $100,000 per year. For employment status, the majority (N=58) were employed, roughly an equal number were students or unemployed (N=10, N=8), and a smaller number identified as other, to which some indicated freelancing or disability (N=4). A majority of our participants identified their race or ethnicity as Caucasian (N=49), with fewer participants identifying themselves as Asian, Black or African American, and the fewest identifying themselves as Hispanic (N=14, N=11, N=4). Two declined to identify their race or ethnicity.

4.1.1 Participant Technology Usage

In order to gain a better understanding of the participants' technology usage, we asked them their level of knowledge of six web-use terms on a 5-point Likert scale [16]. For the high-level understanding terms, "advanced search" and "PDF", participants answered with an average of 3.34 and 3.38 respectively. For the medium-level understanding terms "spyware" and "wiki," participants answered with an average of 2.76 and 3.19 respectively, and for the low-level understanding terms "cache" and "phishing," participants responded with an average of 2.64 and 2.80, respectively. The trend of the averages decreasing as the technology terms refer to lower level processes is in keeping with the trend for the general population, as reported by Hargittai et al [16].

Almost all of our participants used Google as their most frequent search engine of choice (N=74). Roughly half of our participants reported spending five or fewer hours per day online (N=37). A number of our participants had used for-profit online education sites, the most common of which was the University of Phoenix (N=16, N=12).

4.1.2 Characteristics of Craigslist and AMT Participants

We noticed that the demographic characteristics of the Craigslist (N=37) and AMT (N=43) participants were different. Our Craigslist participants were predominantly women, with 24 women and 13 men, while our AMT participants were predominantly male, with 27 men and 16 women. The annual household income of our Craigslist participants was more predominantly between $30,001-$100,000, with 29 participants identifying themselves as in that range and eight participants being below $30,000 or above $100,000. AMT workers were more likely to be in the lowest or highest income brackets, with 18 participants having an annual household income below $30,000 or above $100,000, and 25 having an annual household income between $30,001-$100,000.

4.2 Educational Prompts

4.2.1 Coding prompt

This prompt asked participants to search as if they "were interested in learning to code in an online setting."

Five participants mentioned computer-programming languages in their search, the majority of whom were men (N=4). Even though the prompt did not explicitly ask participants to search for coding classes, many participants did search for variants of "course," "class," or "tutorial," a majority of whom were women (N=31, N=19). Three of the participants searched for a specific online website to learn coding, "lynda.com" and "codeacademy," while one participant searched for a "coding MOOC."

Participants who had annual household incomes under $30,000 were just as likely to have included "free" in their search query as those in higher household-income brackets (N=4, N=4).

Three participants who passed the reading tests provided seemingly off-topic search queries such as "deep fryer," "Presidential Primary Election Dates," and "Homes for rent Baltimore." This is the highest number of off-topic search queries of any of the prompts, perhaps suggesting that the participants failed to understand or misinterpreted what "learning to code" meant. This may have been a cultural issue, as these participants entered search queries more related to the prompt for all the other prompts.

Perhaps in a sign that the prompt heavily influenced the search results that participants entered, many participants entered the exact prompt, "Learn to code," and a smaller, but still noticeable number put "learn to code" as a part of their search query (N=16, N=9)

4.2.2 Online Degree Prompt

The next prompt asked participants to write searches they would enter if they "were interested in getting a degree from an online setting."

Only two participants, both women, named specific schools in their search, "Phoenix University", "Walden University online courses", and "Kaplin unniversity online", all for-profit online schools, and one non-profit school "Bccc online courses". Both participants reported using the for-profit schools they mentioned in their searches. In order to find online education sites, it was more common for participants to search for the "best" or "top" online degree programs (N=22), with no discernable demographic difference between the subset of participants who chose to search this way and the larger sample.

Cost or value was a consideration for six participants, four of whom had annual incomes of $30,001-$60,000. A handful of participants searched for "accredited" educational institutions (N=7), but we found no discernable difference between this subset and the larger sample in our collected demographics.

Similarly to the "coding" prompt, many entered searches of words that were already in the prompt.

4.2.3 GED Prompt

The participants were then asked to write searches they would enter if they "wanted to get a GED." A higher percentage of the searches relative to other prompts were questions or phrases. This surprised us, and we do not have a strong theory as for why participants responded to this with longer searches, given the similar wording of the prompts. There is a possibility it could be due to the shortness of the word "GED", so participants wrote more, but this would need to be studied more for verification. Nearly half of our participants formulated searches as a question or phrase (N=39). Of all entered searches for this prompt, 25% were sentences (N=60/240), as opposed to 5% for the online degree prompt (N=14/280). Five participants put in their local cities, only two participants mentioned specific GED programs.

4.3 Employment Prompts

4.3.1 Job with Current Qualifications Prompt

The participants were then asked to write searches they would enter if that participant "were interested in getting a job with your current qualifications." This prompt generated more unique searches because of the link to the participant, instead of the participant imagining or pretending to be another person. As such, it was harder to find commonalities amongst the searches.

We did, however, find that the majority of searches were either directly for a specific job the participant was looking for, or for a job-posting site. Only 1.4% of searches were posed as questions (N=4/280). A handful of participants, a majority of whom were female (N=9, N=6) mentioned the specific job posting websites Craigslist (N=3), Monster (N=3), Indeed (N=2), Yahoo Jobs (N=1), Coroflot (N=1), and Career Builder (N=1), in a total of eleven searches.

4.3.2 GED Job in Healthcare Prompt

Our last prompt asked users what they would search if they, "have a GED and are interested in finding a job in the healthcare field." Many of our participants associated this with nursing, by searching for nursing schools or nursing assistant positions, while only one of our participants associated a doctor with this hypothetical job (N=10, N=1). Participants also searched for their specific location in finding employment in 4.5% of the searches, but no demographic trends emerged for participants including location in their search (N=6, N=11/240). Few of the searches were articulated as a question, or formulated in a sentence (N=6/240).

4.4 Tf-idf of Searches

We also computed the tf-idf of the searches for each prompt as a whole and for each prompt divided into two subgroups based on gender, those with a household income under and over $30,000, and participants with and without a college degree. Due to the small number of participants, dividing the participants into more than two groups would not have yielded a large enough sample in the subgroup to have meaningful results. Even when only dividing the searches into two groups to run tf-idf, differences between the words that initially seemed promising were often slight or nonexistent differences between the groups the searches themselves were analyzed again. For example, in the coding prompt, participants with a household annual income of less than $30,000 had a tf-idf of 3.4 for the word 'free' while participants with an income above $30,000 had a tf-idf of 4.44 for 'free'. However, the same number of participants in both groups used 'free' in a search, but one participant in the lower-income group used it twice.

5. FUTURE WORK

Given the limitations of our study, alternate research methods may be better suited for further study in this research area. A diary study, in which participants' searches are recorded over the course of weeks or months, could yield searching patterns that are more reflective of the topics of users' searches and their realistic usage patterns. This might have more external validity for different demographics of users. This methodology would also lessen the tendency of participants to rush through composing searches and remove the potential for participants to put part of a prompt as their search query. However, a diary study may be impacted by the knowledge gap, and it would be difficult to understand how user searches were impacted by the user's previous knowledge.

Another approach could be to study the clients of career centers or other organizations, such as public libraries, that assist people in educational or employment opportunities. By analyzing the searches made at these job centers, differences in search behaviors based on demographic characteristics and socioeconomic status might become evident.

Finally, future work could include feeding the results of this study to a controlled search engine to look for patterns amongst the ads or websites served. This would necessitate stronger statistical support for different search characteristics between demographic groups, and then recreating those characteristics for the search engine to respond to.

6. SUMMARY

Search is an important internet task. Given the capability of search engines to categorize users' demographic characteristics based on their usage patterns, and the noted potential for algorithmic discrimination based on search input and personalization, there is the potential for search usage to create deliberate or inadvertent online discrimination. By asking participants to search based on prompts, we were able to understand more about differences in searching patterns amongst different users that might potentially contribute to discrimination. We were also able to find interesting micro-trends, but given the small sample size are unsure if these would hold in a larger sample size.

7. ACKNOWLEDGMENTS

Our many thanks to Elissa Redmiles and Dana Purcell for their input and assistance.

8. REFERENCES

[1] Weber I, and Castillo, C. The demographics of web search. In Proceedings of the 33th annual international ACM SIGIR conference on Research and development in information retrieval, 2010.

[2] Sweeney, L. Discrimination in Online Ad Delivery. Communications of the ACM, 2013, Vol. 56 No. 5, 44-54.

[3] Datta A, Tschantz MC, Datta A. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. arXiv. 2014;cs.CR.

[4] Hargittai, E. & Hinnant, A. (2008). Digital Inequality: Differences in Young Adults' Use of the Internet. *Communication Research*. 35**(5)**: 602-621.

[5] Hargittai, E., & Hsieh, Y. P. (2013). Digital Inequality. In W. H. Dutton (Ed.), The Oxford Handbook of Internet Studies (pp.129-150). Oxford, UK: Oxford University Press.

[6] Van Dijk, J. (2006). Digital divide research, achievements and shortcomings. Poetics, 34, 221–235.

[7] Goel S., Hofman J. M., and Sirer M. I., Who Does What on the Web: A Large-scale Study of Browsing Behavior. ICWSM, 2012.

[8] Jackson, L., Barbatis, G., von Eye, A., Biocca, F., Zhao, Y. and Fitzgerald, H. 2003. Internet use in low-income families: implications for the digital divide. IT&Society, 1(5): 141–165.

[9] Van Dijk, J. & Hacker, L. (2003) Digital divide as a complex and dynamic phenomenon. Information Society, 19, 315–326.

[10] Hannak, A., Sapiezynski, P., Kakhki, A. M., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. Measuring personalization of web search. In WWW, 2013.

[11] Dillahunt, T., Brooks, C., Gulati, S. (2015). Detecting and visualization filter bubbles in Google and Bing. *CHI '15*

Extended Abstracts, Apr 18-23, 2015, Seoul, Republic of Korea.

[12] Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. London: Viking/Penguin Press.

[13] "Craigslist Posting Service." [Online]. Available: https://www.craigslistpostingservice.net/craigslist-most-trafficked-cities/

[14] Egelman S, Péer E. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). CHI. 2015:2873–82.

[15] Hargittai, E. & Shaw A. (2015). Mind the Skills Gap: The Role of Internet Know-How and Gender in Differentiated Contributions to Wikipedia. *Information, Communication and Society*.

[16] Hargittai, E. & Hsieh, Y. P. (2012). Succinct survey measures of web-use skills. Social Science Computer Review.

[17] "Best College Reviews." [Online]. Available: http://www.bestcollegereviews.org/50-top-online-learning-sites/

[18] "The Best Schools." [Online]. Available: http://www.thebestschools.org/features/popular-online-colleges/

[19] "The eBiz MBA." [Online]. Available: http://www.ebizmba.com/articles/job-websites

[20] http://www.maxqda.com/