Auditing Search Engines for Differential Satisfaction Across Demographics

Rishabh Mehrotra[*]
University College London
R.Mehrotra@cs.ucl.ac.uk

Ashton Anderson
Microsoft Research
ashton@microsoft.com

Fernando Diaz
Microsoft Research
fdiaz@microsoft.com

Amit Sharma
Microsoft Research
amshar@microsoft.com

Hanna Wallach
Microsoft Research
wallach@microsoft.com

Emine Yilmaz
University College London
emine.yilmaz@ucl.ac.uk

ABSTRACT

Many online services, such as search engines, social media platforms, and digital marketplaces, are advertised as being available to any user, regardless of their age, gender, or other demographic factors. However, there are growing concerns that these services may systematically underserve some groups of users. In this work, we present a framework for internally auditing such services for differences in user satisfaction across demographic groups, using search engines as a case study. We first explain the pitfalls of naively comparing the behavioral metrics that are commonly used to evaluate search engines. We then propose three methods for measuring latent differences in user satisfaction from observed differences in evaluation metrics. To develop these methods, we drew on ideas from the causal inference and multilevel modeling literature. Our framework is broadly applicable to other online services, and provides general insight into interpreting their evaluation metrics.

1. INTRODUCTION

Modern search engines are complex, relying heavily on machine learning methods to optimize performance. Although machine learning can address many challenges in web search, there is also increasing evidence that suggests that these methods may systematically and inconspicuously underserve some groups of users [7, 3]. From a social perspective, this is troubling. Search engines are a modern analog of libraries and should therefore provide equal access to information, irrespective of users' demographic factors [1]. Even beyond ethical arguments, there are practical reasons to provide equal access. From a business perspective, equal access helps search engines attract a large and diverse population of users. From a public-relations perspective, service providers and the decisions made by their services are under increasing scrutiny by journalists [11] and civil-rights enforcement [9, 4] for seemingly unfair behavior.

One way to assess whether a search engine provides equal access is to look for differences in user satisfaction across demographic groups. If users from one group are consistently less satisfied than users from another, then these users are likely not being provided with equal search experiences. However, measuring differences in satisfaction is non-trivial. One demographic group may issue very different queries than another. Or, two groups may issue similar queries, but with different intents. Any differences in aggregate evaluation metrics will therefore reflect these contextual differences, as well as any differences in user satisfaction. Moreover, search engines are often evaluated using metrics based on behavioral signals, such as the number of clicks or time spent on a page. Because these signals may themselves be systematically influenced by demographics, we cannot interpret metrics based on them as being direct reflections of user satisfaction. For example, if younger users read more slowly than older users, then a metric based on the time spent on a page will, on average, be higher for younger users, regardless of their level of satisfaction.

In this paper, we propose three methods for measuring latent differences in user satisfaction from observed differences in evaluation metrics. All three methods are internal auditing methods—i.e., they use internal system information. Internal auditing methods [13, e.g.,] differ from external auditing methods [2; 18; 10; 15, e.g.,], which rely only on publicly available information.

Our first two methods aim to disentangle user satisfaction from other demographic-specific variation; if we can recover an estimate of user satisfaction for each metric and demographic group pairing, then we can compare these estimates across groups. For our third method, we take a different approach. Instead of estimating user satisfaction and then comparing these estimates, we estimate the latent differences directly. Because we are not interested in absolute levels of satisfaction, this is a more direct way to achieve our goal.

We used all three methods to audit Bing—a major

[*] Work conducted at Microsoft Research.

search engine—focusing specifically on age and gender. Overall, we found no difference in satisfaction between male and female users, but we did find that older users appear to be slightly more satisfied than younger users.

2. DATA AND METRICS

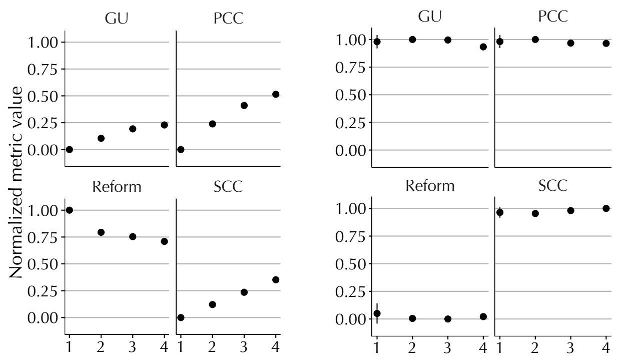
We selected a random subset of desktop and laptop users of Bing from the English-speaking US market, and focused on their log data from a two week period during February, 2016. We removed spam using standard bot-filtering methods, and discarded all queries that were not manually entered. By performing these preprocessing steps, we could be sure that any observed differences in evaluation metrics were not due to differences in devices, languages, countries, or query input methods.

We enriched these data with user demographics, focusing on self-reported age and (binary) gender information obtained during account registration. We discarded data from any users older than 74, and binned the remaining users according to generational boundaries: (1) younger than 18 (post-millennial), (2) 18–34 (millennial), (3) 35–54 (generation X), and (4) 55–74 (baby boomers).[1] To validate each user's self report, we predicted their age and gender from their search history, following the approach of Bi et al. [6]. We then compared their predicted age and gender to their self report. If our prediction did not match their self report, we discarded their data. Approximately 51% of the remaining users were male. In contrast, the distribution of users across the four age groups is much less even, with the younger age groups containing substantially fewer users (<1% and 13% for post-millennial and millennial, respectively) than each of the older age groups (41% and 45% for generation X and baby boomers).

Finally, we labeled the remaining queries with topic information, using the web classifier of Bennett et al. [5].

After these steps, we were left with 32 million search impressions, involving 16 million distinct queries. (A search impression is a unique view of a results page presented to a user in response to a query.) These queries were issued by 4 million users over 17 million sessions.

We considered four different evaluation metrics, each intended to operationalize user satisfaction: *graded utility*, a model-based metric that provides a four-level estimate of user satisfaction based on search outcome and user effort [17]; *reformulation rate*, the fraction of queries that were followed by another, reformulated, query [16]; *page click count*, the total number of clicks made by a user on a results page, thought to reflect their level of engagement; and *successful click count*, the number of clicks with dwell times longer than 30 seconds [8]. For graded utility, page click count, and successful click count, higher values mean higher satisfaction; for reformulation rate, the relationship is reversed.



(a) (b)

Figure 1: Query-averaged values for each metric and age group pairing. (a) All data. (b) Context-matched data. Error bars (one standard error) are present in all plots, but are mostly so small that they cannot be seen.

3. DEMOGRAPHIC DIFFERENCES

In this section, we describe observed differences between the demographic groups. We focus on the types of queries issued by users and on evaluation metrics.

First, we found that users from different demographic groups issued different types of queries. A higher proportion of female users (28%) issued navigational queries than male users (26%). Although similar proportions of male and female users (∼17%) issued head queries, slightly more male users issued tail queries. (The top 20% and bottom 30% of queries are called head and tail queries, respectively.) Based on these differences alone, we would expect male users to exhibit lower values of the evaluation metrics described in the previous section. We also found that a higher proportion of older users (30%) issued naviational queries than younger users (13%), while younger users (39%) were more likely to issue tail queries than older users (30%). The queries issued by the youngest age group were most similar to the second-youngest age group. We observed the same pattern for the other age groups, suggesting that users who are close in age are more likely to issue similar queries than users whose ages are further apart.

Next, we compared the evaluation metrics described in section 2 across demographic groups, without controlling for any confounding demographic-specific variation. For each metric and demographic group pairing (e.g., graded utility and millennial), we computed the average metric value for each query issued by that group (by averaging over impressions) and then averaged these values. By computing query-averaged values, we ensured that our results were not dominated by the most popular queries. Finally, we normalized the query-averaged values to lie between zero and one: we

[1] http://www.pewresearch.org/fact-tank/2016/04/25/millennials-overtake-baby-boomers/

identified the minimum and maximum values for each metric over the groups, subtracted the corresponding minimum off of each value, and then divided each result by the corresponding maximum minus the minimum.

We provide the normalized query-averaged values for each metric and age group pairing in figure 1a. The metrics all follow the same trend: older users have better values (lower for reformulation rate, higher for the other metrics) than younger users. However, as described in section 1, we cannot conclude that this trend means that older users are genuinely more satisfied than younger users; these differences may be due to other demographic-specific variation. For example, users from different age groups issued different types of queries, so this trend may simply reflect this contextual difference.

In contrast, we found very little variation across genders; however, again, this is not conclusive evidence.

4. CONTEXT MATCHING

In this section, we present our first method for disentangling user satisfaction from other demographic-specific variation. This method recovers an estimate of user satisfaction for each metric and demographic group pairing by controlling for two confounding contextual differences: the query itself and the intent of the user.

We drew on well-established ideas from the causal inference literature to develop a matching method similar to those used in medicine and the social sciences [20]. Specifically, for each demographic factor (i.e., age or gender), we made sure that the impressions from that factor's groups were as close to identical as possible.

To do this, we first restricted the data to navigational queries because they are generally less ambiguous than informational queries [21]. We then retained only those queries with at least ten impressions from each demographic group. To control for the intent of the user, we followed the approach of Radlinski et al. [19]. For each query, we identified the search result with the most final successful clicks. (A final successful click is a successful click—i.e., a click with a dwell time longer than 30 seconds—that terminates the query.) We then discarded any impression whose final successful click was not on that result. Finally, to be certain that the users had the same choices available to them when making those clicks, we kept only those impressions that were of the same results page (up to the first eight results). After these steps, we were left with 1.2 million impressions, involving 19,000 queries, issued by 617,000 users.

In figure 1b, we provide normalized query-averaged values for each metric and age group pairing, computed using the context-matched data. There is much less variation across age groups than in figure 1a (all data). This suggests that the trend described in section 3 is unlikely to be due to differences in user satisfaction.

Again, we found very little variation across genders.

5. MULTILEVEL MODELING

In this section, we present our second method for disentangling user satisfaction from other demographic-specific variation. Like our context-matching method, this method recovers an estimate of user satisfaction by controlling for confounding contextual differences; however, it only controls for characteristics of the query itself and not for the intent of the user. We were therefore able to use this method without restricting the data.

We drew on the multilevel modeling literature [14] to develop a new statistical model for the effect of query difficulty on evaluation metrics, controlling for the topic of the query and demographics of the user who issued the query. We then used this model to examine the effects of age and gender on each of the four evaluation metrics described in 2, for fixed query difficulties and topics. Because the model does not control for the intent of the user, these effects may be due to differences in intent, as well as any differences in user satisfaction.

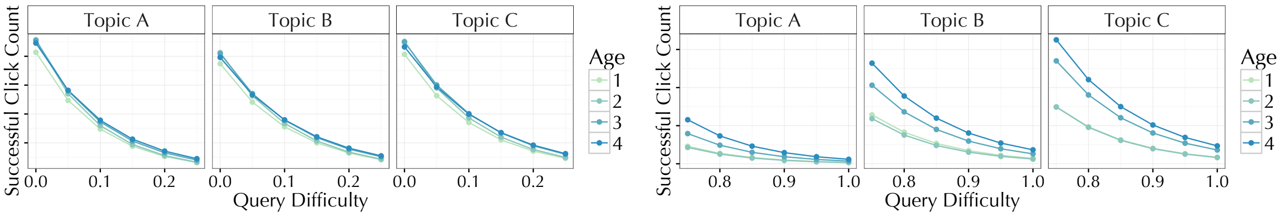
The model operationalizes the following intuition: We expect that queries with different difficulties will lead to different values of the evaluation metrics described in section 2. We also expect that queries about different topics will lead to different values, as will queries issued by users with different demographics. The model uses two levels to capture this intuition: the first level accounts for differences across age, gender, and topic combinations; the second level models these differences.

Letting Y_i denote the value of one of the metrics described in section 2 (i.e., graded utility, reformulation rate, page click count, or successful click count) for the i^{th} impression in our data set, the model assumes that

$$\mathbb{E}[Y_i] = f^{-1}(\alpha_{a_i g_i t_i} + \beta_{a_i g_i t_i} X_i), \tag{1}$$

where $f(\cdot)$ is a link function; a_i and g_i are the age and gender of the i^{th} impression's user; and t_i and X_i are the topic and difficulty of the i^{th} impression's query. We can interpret $\alpha_{a_i g_i t_i}$ and $\beta_{a_i g_i t_i}$ as the intercept and slope, repectively, of Y_i with respect to X_i—the model has a different intercept and slope for each age, gender, and topic combination. At the second level, the model further assumes that each intercept $\alpha_{a_i g_i t_i}$ is a linear combination of age, gender, and topic indicator variables, as well as a corresponding interaction term. Finally, the model assumes that the coefficients at the second level are Gaussian distributed around zero.

To estimate the difficulty of each query, we sorted the queries issued by each demographic group according to their graded utility values. We then averaged each query's percentile positions in these lists to obtain an estimate of its difficulty that is uncorrelated with the demographics of the users who issued it. Most methods for estimating the difficulty of a query are based on behavioral signals, such as the reformulation rate or the time spent on a page [12, 22]. Because behavioral



(a)

(b)

Figure 2: Successful click count according to our model for various query difficulties. (a) Easiest. (b) Hardest.

signals may themselves be systematically influenced by demographics, we were unable to use these methods.

We used a random sample of 1.4 million impressions to fit a different version of the model for each evaluation metric. Because graded utility ranges from negative one to positive one, we used a Gaussian model with an identity link function; because reformulation rate ranges from zero to one, we used a binomial model with a logit link function; and, because page click count and successful click count are both non-negative integers, we used a Poisson model with a log link function.

Again, we found that gender had little effect on any of the metrics, while age had an effect on all four. For each topic and age group pairing, we used each metric's version of the model (with g_i arbitrarily fixed to male) to predict the values of that metric for query difficulties between zero and one in increments of 0.05. In figure 2, we depict these values for successful click count; we show only the six easiest (2a) and six hardest (2b) query difficulties. These plots indicate that older users have slightly higher values than younger users, especially for more difficult queries. Again, we cannot conclude that this means that older users are genuinely more satisfied than younger users; these differences may be due to unmodeled demographic-specific variation.

6. ESTIMATING DIFFERENCES

In this section, we present our third method. This method estimates latent differences in user satisfaction across demographic groups directly. Specifically, it considers randomly selected pairs of impressions (for the same query, issued by users from different demographic groups) and uses a high-precision algorithm to estimate which impression led to greater user satisfaction. Then, using these labels, it models differences in satisfaction.

We restricted the data to only those queries that were issued by users from at least three demographic groups and that had at least ten impressions. We then randomly selected 10% (~62,000) of these queries. For each query, we randomly selected 10,000 pairs of impressions, resulting in a total of 2.7 billion pairs. Finally, for each pair, we compared the impressions' values of the evaluation metrics and labeled one of the impressions as leading to greater user satisfaction if there was a difference so large that it was unlikely to explained by anything other than a genuine difference in user satisfaction.

```
**if** $\mathrm{RR}_i < \mathrm{RR}_j$ **return** $+1$
**if** $\mathrm{RR}_i > \mathrm{RR}_j$ **return** $-1$
**if** $\mathrm{GU}_i - \mathrm{GU}_j > 0.4$ **return** $+1$
**if** $\mathrm{GU}_j - \mathrm{GU}_i > 0.4$ **return** $-1$
**if** $\mathrm{SCC}_i - \mathrm{SCC}_j > 2$ **return** $+1$
**if** $\mathrm{SCC}_j - \mathrm{SCC}_i > 2$ **return** $-1$
**if** $\mathrm{GU}_i - \mathrm{GU}_j > 0.2$ and $\mathrm{SCC}_i - \mathrm{SCC}_j > 1$ **return** $+1$
**if** $\mathrm{GU}_j - \mathrm{GU}_i > 0.2$ and $\mathrm{SCC}_j - \mathrm{SCC}_i > 1$ **return** $-1$
**else** **return** $0$
```

Figure 3: Algorithm for labeling a pair of impressions.

We provide the algorithm that we used to compare the impressions' metric values in figure 3. We obtained the thresholds using the model described in section 5; however, we omit a full discussion to conserve space.

We used a single-level model to estimate latent differences in satisfaction across demographic groups. This model is similar to the one described in section 5, but does not include query-specific terms. Letting $S_i - S_j$ denote the latent difference in user satisfaction between the i^{th} and j^{th} impressions, the model assumes that

$$P(S_i - S_j > 0) = f^{-1}(\mu_0 + \gamma_{a_i} + \gamma_{a_j} + \gamma_{g_i} + \gamma_{g_j} + \gamma_{a_i \times g_i \times a_j \times g_j}), \quad (2)$$

where $f(\cdot)$ is a logit link function and $a_i \times g_i \times a_j \times g_j$ denotes an interaction term. The model also assumes that the coefficients are Gaussian distributed around zero.

We fit the model using pairs of impressions from different demographic groups, labeled as either $+1$ or -1 via the algorithm in figure 3. Again, we found that gender had little effect. For each age group pairing, we therefore used the model (with g_i and g_j arbitrarily fixed to male and female) to predict $P(S_i - S_j > 0)$. We visualize the probabilities for each pairing in figure 4.

This figure suggests that older users are more satisfied than younger users, with the difference increasing for users whose ages are further apart. However, because the probabilities are close to 0.5, the difference is relatively small for each age group pairing. These results are consistent with the trends described in sections 3 and 5; though, again, we note that these differences

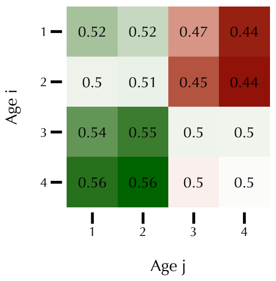


Figure 4: $P(S_i - S_j > 0)$ for each age group pairing. Standard errors (not shown) are between 0.001–0.004.

may be due to other demographic-specific variation.

7. DISCUSSION

Internally auditing search engines for equal access is much more complicated than comparing evaluation metrics for demographically binned search impressions. In this paper, we addressed this challenge by proposing three methods for measuring latent differences in user satisfaction from observed differences in evaluation metrics. We then used these methods to audit Bing, focusing specifically on age and gender. Overall, we found no difference in satisfaction between male and female users, but we did find that older users appear to be slightly more satisfied than younger users. By using three different methods, with complementary strengths, we can be confident that any trends detected by all three methods are genuine, though we cannot conclude that they are due to differences in user satisfaction, as opposed to other demographic-specific variation. We hypothesize that we would be able to attribute such trends to unmodeled differences between demographic groups if we were to see the same trends when using our three methods to audit an independently developed search engine.

We conclude that there is a need for deeper investigations into observed differences in evaluation metrics across demographic groups, as well as a need for new metrics that are not confounded with demographics.

8. REFERENCES

[1] *Code of Ethics for Librarians and other Information Workers.* International Federation of Library Associations and Institutions, 2012.

[2] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing black-box models by obscuring features. arXiv:1602.07043, 2016.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias, 2016.

[4] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

[5] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. In *WWW*, 2010.

[6] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: Social data meets search queries. In *WWW*, 2013.

[7] T. Bolukbasi. Quantifying and reducing stereotypes in word embeddings. *CoRR*, 2016.

[8] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: a comparison to eye tracking. In *SIGIR*, 2009.

[9] Cecliai, Smith, and D. Patel. Big data: A report on algorithmic systems, opportunity, and civil rights. Technical report, Executive Office of the President of the United States, 2016.

[10] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence. In *IEEE Symposium on Security and Privacy*, 2016.

[11] N. Diakopoulos. Algorithmic accountability. *Digital Journalism*, 3(3):398–415, 2015.

[12] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR*, 2010.

[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, 2015.

[14] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press, 2006.

[15] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv:1610.02413*, 2016.

[16] A. Hassan. Beyond clicks: query reformulation as a predictor of search satisfaction. In *CIKM*, 2013.

[17] J. Jiang and A. Hassan. Understanding and predicting graded search satisfaction. In *WSDM*, 2015.

[18] M. Lecuyer, R. Spahn, Y. Spiliopolous, A. Chaintreau, R. Geambasu, and D. Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *CCS*, 2015.

[19] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *WWW*, 2010.

[20] D. B. Rubin. *Matched sampling for causal effects.* Cambridge University Press, 2006.

[21] Y. Wang and E. Agichtein. Query ambiguity revisited: Clickthrough measures for distinguishing informational and ambiguous queries. In *HLT*, 2010.

[22] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *CIKM*, 2009.