Patterns of gender-specializing query reformulation

Amifa Raj*
Boise State University
Boise, USA
amifaraj@u.boisestate.edu

Bhaskar Mitra
Microsoft Research
Montréal, Canada
bmitra@microsoft.com

Nick Craswell
Microsoft
Bellevue, USA
nickcr@microsoft.com

Michael D. Ekstrand
Boise State University
Boise, USA
ekstrand@acm.org

ABSTRACT

Users of search systems often reformulate their queries by adding query terms to reflect their evolving information need or to more precisely express their information need when the system fails to surface relevant content. Analyzing these query reformulations can inform us about both system and user behavior. In this work, we study a special category of query reformulations that involve specifying demographic group attributes, such as gender, as part of the reformulated query (*e.g.*, "olympic 2021 soccer results" → "olympic 2021 *women's* soccer results"). There are many ways a query, the search results, and a demographic attribute such as gender may relate, leading us to hypothesize different causes for these reformulation patterns, such as under-representation on the original result page or based on the linguistic theory of markedness. This paper reports on an observational study of gender-specializing query reformulations—their contexts and effects—as a lens on the relationship between system results and gender, based on large-scale search log data from Bing. We find that these reformulations sometimes correct for and other times reinforce gender representation on the original result page, but typically yield better access to the ultimately-selected results. The prevalence of these reformulations—and which gender they skew towards—differ by topical context. However, we do not find evidence that either group under-representation or markedness alone adequately explains these reformulations. We hope that future research will use such reformulations as a probe for deeper investigation into gender (and other demographic) representation on the search result page.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**.

KEYWORDS

Query reformulation; Group representation; User behavior

*Work done during internship at Microsoft.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM Reference Format:
Amifa Raj, Bhaskar Mitra, Nick Craswell, and Michael D. Ekstrand. 2023. Patterns of gender-specializing query reformulation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3539618.3592034

1 INTRODUCTION

Searchers may reformulate their queries to reflect their evolving information needs or in response to less-than-relevant results from the search system to better specify the information they are looking for. A specific class of query reformulation, often referred to as *specialization* [18], involves a reformulated query on the same topic as the original query but with an expressed intent for more specific information. Specialization in query reformulation typically involves addition of one or more new terms to the original query [19, 20]. In this paper, we are interested in a particular type of query reformulation that involves specialization by adding demographic group attributes, such as gender, to the query. For example, the query "NCAA scores" may be reformulated to "NCAA *women's* scores" to clarify that the searcher is interested in the results for the women's basketball scores at the NCAA; we call these *gender-specializing query reformulations* (GSQR). Searchers reformulate their query in response to the results returned by the system which presumably did not fully meet their information need. Studying such reformulations and the search result pages (SERPs) that led to them may shed light on how that SERP and the results it contains connect to the user's information need, particularly with respect to the specifying term.

In this work, we are interested in situations where users add gender-specifying terms in their query reformulation as a lens on the relationship of system results and gender. There are many ways a query, the results, and a demographic attribute such as gender may relate. For example, the linguistic theory of *markedness* [5, 30]—that certain assumptions are assumed in describing a particular class or activity, and markers indicating demographics or other attributes are only used when deviating from the default—leads to an hypothesis that such reformulations may arise when one group dominates the SERP and a different group is desired.

We cannot assume, however, that a GSQR necessarily means that the specified gender was insufficiently represented in the original SERP, there are many reasons the searcher could reformulate their query, including by clicking on a reformulation the system suggests, or looking to entirely filter out a minority of results that don't

meet their needs. In this initial exploratory investigation of GSQR behavior, we characterize these reformulations as they appear in a real-world search log from Bing and study their contexts and effects in order to enable future research that can use such reformulations to develop deeper insight into user information need mismatch and search result representations of gender (and other demographics).

To summarize, the key contributions of our current work are:

(1) We analyze in what context GSQR occur and factors that may contribute to these reformulation patterns.
(2) We study the impact of said query reformulations on SERPs.

Finally, we conclude with a discussion of implications of our study on future research and the designe of information access systems.

2 RELATED WORK

Several research works have analyzed query reformulation to understand various associated aspects including reformulation patterns, applications, and user behavior [14, 18, 22]. Previous research identified that users often add terms (*specialization*) or remove terms (*generalization*) to modify their queries [2, 4, 17]. This behavior of query modification is found to be effective in retrieving more relevant information for users [12, 26]. Several studies worked towards identifying patterns of user query reformulation behavior, and these patterns were categorized based on search task, sequences, user intent, or semantic analysis [15, 20, 26]. Huang and Efthimiadis [15] provided a taxonomy of users' query reformulation strategies by focusing on how users write reformulations. For example, users may add, remove, substitute, or reorder to reformulate their queries. Liu and Gwizdka [19] analyzed how users reformulated queries for different search tasks and found that the type of search task can effect their reformulation behavior. Liu et al. [20] tried to understand the connection between task types, SERP of previous search and users' query reformulation behavior. Moreover they categorized query reformulation type into five groups: generalization, specialization, term substitution, repeat, and new, and further observed the effectiveness of their reformulation regarding different search tasks. Rha et al. [25] and Chen et al. [7] analyzed user behavior regarding different query reformulation techniques to identify user intention behind their reformulation. These studies observed that users may reformulate queries to find specific results, learn more about a topic, or satisfy their particular needs. This knowledge of query reformulation patterns and user intention associated with the reformulation further facilitate research on designing search engines that can better support user information need [2, 8]. In our work, we are specifically examining specialization reformulations that specify a demographic group (gender) as a lens to understand users' intent and system responses to both initial and refined queries in such settings.

3 DATA AND METHODS

Identifying GSQR in search logs. Our study focuses on query reformulation patterns that specializes the information need to specific gender groups. We adopt a narrower definition of specializing query reformulations compared to Liu et al. [19, 20]. We define a query reformulation to be specializing if the reformulated query contains all the terms in-order as in the original query and includes an additional set of contiguous terms added anywhere to the original query. We enforce a constraint that the additional terms must include one term from a list of known gender terms—*i.e.*, "woman", "man", "women", "men", "woman's", "man's", "women's", "men's", "womans", "mans", "womens", "mens", "female", "male", "male's", "female's", "males", and "females"—and optionally additional terms from a known list of prepositions—*i.e.*, "about", "against", "according to", "among", "at", "by", "except", "for", "from", "in", "like", "of", "on", "to", "with", and "without". In line with this definition, we consider reformulations like "leadership quotes" → "leadership quotes by women" and "bmi calculator" → "bmi calculator *for men*" as GSQR. Given large-scale search logs we can automatically identify instances of query reformulations matching these specified patterns. We understand the risk of considering gender as binary attribute [24]; our analyses can be further extended to non-binary gender and other demographic attributes. To consider other demographic groups beyond gender, it would be convenient if we can (semi-)automatically detect other relevant group terms. Please see Appendix A for more details.

Data. To understand user intentions behind GSQR, we analyze instances of similar reformulations in large scale search logs over the period of one year (January 1 – December 31, 2021) from Bing. We note that although specialization is a common form of reformulation [17–20], people enter a vast variety of specializing reformulation terms, and our gender terms are only a small fraction. From the search logs, we extract a sample of approximately 4.7 million pairs of consecutive queries from the same search session where the second query is a GSQR of the former. This was 3.9% of the specializing queries we considered. For both original and the reformulated queries, we extract metadata such as timestamps, entry point from which the query was submitted, web results that were displayed to the user, and a record of user clicks (if any) on those results.

Analysis methods. The focus of our study is to understand user intent behind GSQR and with that goal we did following analyses:

- To ensure that this is a recurring pattern, we consider the frequency of such reformulations in our search log data.
- We analyze how the pattern differs across query topics by using an automated text-based classifier on the original queries.
- We analyze time differences between original and reformulated queries and potential elements on the original SERP—*e.g.*, related query recommendations for intent disambiguation—that may influence the user to reformulate their queries.
- We analyze which groups are specified more often in aggregate in these reformulations by using average GloVe embeddings [23] of terms as query representation and based on the approach proposed by Bolukbasi et al. [3], we compute a genderedness measure for queries.
- We study the impact of these reformulations on the SERP.

4 RESULTS

4.1 GSQR patterns overall and by topic

Overall Pattern. As described in Section 3, each of our 4.7 million GSQR cases adds at least one term from our list of gender terms. Of those, 54% add women-related and 46% add men-related terms.

Table 1: Specialization patterns by topic. GSQR rate is compared to all-topic average. % rec. and % woman are percent of GSQR that were clicked recommendations and added woman gender terms, respectively.

Original query topic	GSQR rate	% rec.	% woman
shopping and fashion	15.7	13%	50%
health	2.0	23%	52%
sports and outdoors	2.0	7%	62%
parenting	1.7	17%	41%
animals	1.2	19%	56%
psychology	1.1	26%	54%
religion	0.9	17%	66%
art	0.8	11%	56%
literature	0.7	18%	70%
philosophy	0.7	16%	52%
history	0.6	4%	82%
photography	0.5	12%	60%
entertainment	0.4	11%	55%
other	0.4	9%	51%
science	0.3	8%	49%
cooking and food	0.3	11%	46%
politics	0.3	5%	70%
travel	0.3	9%	67%
education	0.3	7%	67%
vehicle	0.2	4%	67%
home and garden	0.2	8%	50%
technology	0.2	15%	52%
finance	0.2	7%	57%
all topics	1.0	13%	54%

The median time between the first and second query is 19 seconds. Reformulations that add men-related terms tend to happen slightly more quickly, with a median of 17 seconds, compared to 19 seconds for those adding women-related terms.

The most common methods for entering the second query are by editing at the top of the SERP (64% of our GSQR cases with median time 15 seconds) or clicking a recommended query at the top of the SERP, as in [32] (13% of our GSQR cases with median time 13 seconds). Clicking a recommended query at the top of the SERP contributes slightly higher for query reformulations specializing *for women* (14.3%) compared to *for men* (11.5%). Besides editing and clicking at the top of the SERP, other methods of GSQR may involve clicking on other query suggestions on the page or re-entering the query via other entry points (median time 60+ seconds).

So far there are some small differences between men-related and women-related GSQR, with men-related reformulation added more quickly, and women-related reformulation happening more often and having slightly higher association with recommended queries.

Query Topic Analysis. Table 1 breaks down the analysis by topical category of the original query using a topical classifier. The rate column reveals how much more prevalent GSQR is for that topic than expected, given how often we see that topic in Bing logs. The rate is particularly high in the *shopping and fashion* category and lower in categories such as *home and garden*, *technology*, and *finance*.

The next column is the percent of GSQR that came from the query recommendation feature. The rate and recommender columns are correlated (Spearman ρ 0.586, p − val = 0.003), suggesting the possibility that reformulation patterns and recommenders drive each other to some extent. It is clear that the rate is not entirely driven by the recommender, because then the % recommended column would need to explain the ratio of 15.7 for *shopping and fashion*, whereas the actual percentage is 13%, close to the overall average. The largest usage of the recommender is in the *psychology* and *health* categories.

The last column shows what percentage of GSQR are women-related, by topic. For the *shopping and fashion* topic we see a 50% split, indicating that both men and women need to specify gender in such scenarios. Women-related reformulation is lower for the parenting topic, with 41%, perhaps suggesting that finding men-related information is taking a bit more effort in this area. For the history topic, the average fraction of women-related reformulations is 82%, for example there were two instances of "on this day in history" followed by "on this day in history women".

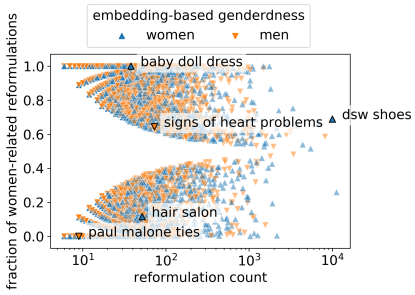

Figure 1: Scatter plot of queries that exhibit genderedness both in reformulation (y-axis) and GloVe embedding (marker). We see both types of embedding-based genderedness receiving all types of reformulation. Four example queries are highlighted.

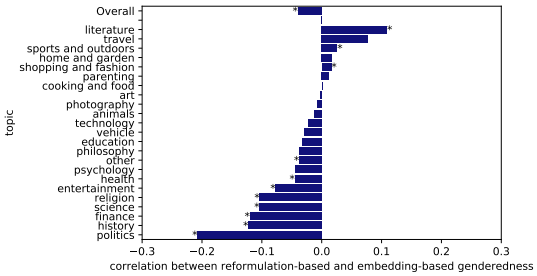

Figure 2: Spearman correlation between reformulation-based and embedding-based genderedness, with statistically significant correlation indicated by * (p-val < 0.05).

4.2 Embedding-based genderedness

We now consider embedding-based genderedness [3] of the original query, using GloVe. This allows us to study the relationship between the % women reformulation pattern and the embedding-based genderedness. For example, corresponding to the theory of markedness mentioned, we may see more men-related reformulations if the original query is associated with women (*e.g.*, "nurse" → "male nurse").

In Figure 1, each point indicates a query. Queries where the reformulation patterns were balanced between men and women were removed, according to a two-tailed binomial test (p − val < 0.05).

This gives the plot its funnel shape, since a reformulation with fewer observations (x-axis) has to be extreme on the y-axis for the binomial test to keep it. Also, queries where the embedding-based genderedness was close to 0 (within ±0.05) were removed, removing about half the queries, and keeping the most women-related queries as blue triangles and the most men-related queries as orange triangles.

The figure shows that there are many queries where the embedding-based genderedness is being corrected by reformulation, but also many where the genderedness is reinforced by the reformulation. Five queries are highlighted as examples. For "baby doll dress" and "paul malone tie", the genderedness is reinforced by the reformulation pattern, whereas for "hair salon" and "signs of heart problems", the genderedness of the query is corrected by the query pattern. We also include one query with high reformulation count "dsw shoes".

Seeing a great variety of reformulation patterns, which can contradict or reinforce original query's genderedness, raises the question of whether query genderedness and reformulation patterns are correlated overall. Figure 2 shows the overall correlation and the per-topic correlation where all the correlations are low, indicating that we do see a mix of user behaviors, as depicted in Figure 1. Over all topics ("Overall") the correlation is mildly negative, and statistically significant, meaning that an original query having a higher value on fraction of woman-related reformulations is associated with having a lower value on woman-related genderedness. The correlation is weak, but happens more often than would happen by chance. When the correlation is statistically significant ($p - val < 0.05$) the figure shows it with a *. This means that the reformulations in *literature*, *sports* and *shopping* topics are more likely to reinforce the genderedness of the original query and this pattern is present Figure 1, in the dress and ties example queries. For several topics the correlation is the other way, with reformulations tending to correct the genderedness of the original queries. Although we see several statistically significant cases of reinforcing or correcting genderedness, in all cases the associations are mild, with correlation in the range -0.2 to 0.2, so overall the user behavior is mixed, for all topics and overall.

4.3 Impact of reformulations

In our query reformulation data, we find several examples via manual inspection where the addition of gender terms surfaces more gender-specific results (*e.g.*, "ADHD symptoms" → "ADHD symptoms for women") or effectively corrects for under-representation in the original SERP (*e.g.*, "US open golf 2022" → "US open golf 2022 women"). If these reformulations are effective then we would expect behavioral search satisfaction metrics, like clickthrough rate (CTR), to improve for the reformulated query SERP in aggregate. While we cannot disclose exact CTR due to Bing disclosure limitations, we note that reformulated query SERP CTR is 2.6 times higher than original query SERP CTR. This ratio is approximately the same for both men and women as the specializing term. If we consider only queries where the reformulation was through a clarifying query recommendation, the CTR boost increases to 3.6.

CTR does not tell the whole story because uses may click on a link in the second SERP because they reformulated quickly and did not carefully examine the first SERP. We therefore examine the

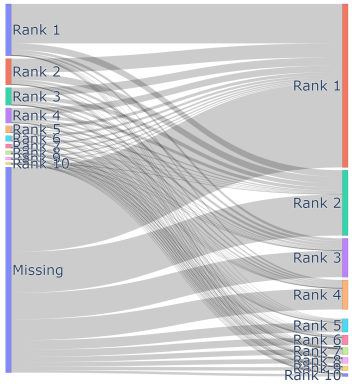

Figure 3: Visualization of the change in the rank of a result (clicked on the latter SERP) from original query SERP (left) to reformulated query SERP (right) in cases. The clicked document is often missing from the original SERP, but in several cases it was also on the original SERP, possibly near the top.

difference the query reformulation makes in the rank of the selected item. Figure 3 visualizes this analysis; each data point is GSQR event where the user clicks on a document on the reformulated query SERP. In 62% of cases, the clicked result did not appear on the original query SERP. In another 18% of cases, the clicked document appeared on the original SERP but at a lower rank, and 14% of the time they appear at the same position on both SERPs. This shows that the reformulated queries are yielding better access to the ultimately-selected results.

GSQR may also influence the relative visibility of different content sources, especially if publishers have optimized their content for particular queries. To analyze this potential phenomenon empirically, we compute the ratio of the probability of exposure for individual websites on the reformulated query SERP vs. the original SERP, which we refer hitherto as *exp-ratio*. We estimate these probabilities with the Expected Exposure technique of Diaz et al. [11] using the NDCG user behavior model. As we may expect, websites that contain group-specific content—*e.g.*, menshairstylestoday.com (exp-ratio=3.7) and menshealth.com (exp-ratio=2.3) for men and womenshealthmag.com (exp-ratio=2.8) for women—are exposed significantly more on the reformulated query SERP. Websites may also gain more exposure on the reformulated query SERP if they specifically create pages for different groups either as part of better content organization or as a search engine optimization technique. This may sometimes lead to undesirable outcomes such as underexposure of authoritative websites like mayoclinic.org (exp-ratio=0.8) and webmd.com (exp-ratio=0.7) compared to sites that are better-optimized specifically for such query patterns.

5 DISCUSSION AND CONCLUSION

In this study we find that users of search systems may reformulate their queries to look for gender-specific results, often to correct for mismatch of the their information need. Our analysis here is a first-pass approach to the problem. There can be multiple factors that can

lead users to do GSQR and by looking at the reformulated queries and how genders are represented on overall SERP—which may include images, videos, and news results—we can develop a deeper understanding of the circumstances that led to these reformulations.

For example, we identified reformulations that seek *gender-specific information* (e.g., "ADHD symptoms" → "ADHD symptoms for women") or reformulations that *intensify group representation* (e.g., "hispanic names" → "hispanic names for women") or *reinforce over-representation* (e.g., "NCAA basketball score" → "NCAA men's basketball score") where users may want to filter-out results related to another gender. We found reformulations that *correct for group under-representation* where users find fewer results about the particular gender relevant to their information need and so may explicitly reformulate the query mentioning that gender to find more results (e.g., "NCAA basketball score" → "NCAA women's basketball score"). There can be reformulation that are *influenced by other SERP elements* such as, images or videos (e.g., "hiking boots" →"hiking boots for women" because images in original query SERP are skewed towards one). And lastly, there can be *harmful reformulations* where the reformulated query contains names of people and gender identities. Many of these queries are harmful speculations about the subject's gender and often misgender the subject. We intentionally do not include any examples of such queries to avoid perpetuating further harm. Please see Appendix B for additional manual analysis of GSQRs.

Through our study, we show the importance of analyzing GSQR to get deeper insight about user behavior and systems. Trace ethnography [13] combines analysis of data traces from large-scale online systems, with ethnographic techniques to deeply understand the journeys users take in their use of the system. Query reformulations seem likely to be a useful lens to focus on such studies—applying trace ethnography to sessions that contain demographic reformulations, for instance, may produce a richer understanding of users' search goals and behavior and the system's response. As biases in group representations in search results may contribute to social harms and unfairness, more of such studies should be conducted that may shine a light on disparities in group representation and also provide insights that may be instrumental in developing measures of representational bias and representational harms [9], although there are additional challenges in measuring representational bias and fairness as we discuss in Appendix C. Such studies may be complemented by other forms of inquiry, such as lab-based user studies and online surveys, to further elicit situations where group representations are important to consider in the context of online information access. Understanding implications of how groups are represented in retrieved information may have important implications for designs of future information access systems. We hope that future research continues to engage with these questions of moral import.

Acknowledgments. This work was partially supported by the National Science Foundation under grant IIS 17-51278.

REFERENCES

[1] Arun Agrawal. 2002. Indigenous knowledge and the politics of classification. *International social science journal* 54, 173 (2002), 287–297.
[2] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. 2011. Query reformulation mining: models, patterns, and applications. *Information retrieval* 14, 3 (2011), 257–289.
[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).
[4] Peter Bruza and Simon Dennis. 1997. Query Reformulation on the Internet: Empirical Data and the Hyperindex Search Engine.. In *RIAO*, Vol. 97. Citeseer, 488–499.
[5] Mary Bucholtz. 2001. The whiteness of nerds: Superstandard English and racial markedness. *Journal of linguistic anthropology* 11, 1 (2001), 84–100.
[6] Kyla Chasalow and Karen Levy. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 77–89.
[7] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the Web Conference 2021*. 743–755.
[8] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating query reformulation behavior of search users. In *China Conference on Information Retrieval*. Springer, 39–51.
[9] Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.
[10] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
[11] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proc. CIKM*. 275–284.
[12] Susan Gauch and John B Smith. 1993. An expert system for automatic query reformation. *Journal of the American Society for Information Science* 44, 3 (1993), 124–136.
[13] R Stuart Geiger and David Ribes. 2011. Trace ethnography: Following coordination through documentary practices. In *2011 44th Hawaii international conference on system sciences*. IEEE, 1–10.
[14] Vera Hollink, Jiyin He, and Arjen de Vries. 2012. Explaining query modifications. In *European Conference on Information Retrieval*. Springer, 1–12.
[15] Jeff Huang and Efthimis N Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 77–86.
[16] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.
[17] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2009. Patterns of query reformulation during web searching. *Journal of the american society for information science and technology* 60, 7 (2009), 1358–1371.
[18] Bernard J Jansen, Mimi Zhang, and Amanda Spink. 2007. Patterns and transitions of query reformulation during web searching. *International Journal of Web Information Systems* (2007).
[19] Chang Liu and Jacek Gwizdka. 2010. Analysis of query reformulation types on different search tasks. (2010).
[20] Chang Liu, Jacek Gwizdka, Jingjing Liu, Tao Xu, and Nicholas J Belkin. 2010. Analysis and evaluation of query reformulations in different task types. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–9.
[21] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14.
[22] Bhaskar Mitra. 2015. Exploring Session Context using Distributed Representations of Queries and Reformulations. In *Proc. SIGIR*. ACM, 3–12.
[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
[24] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. 2023. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. *arXiv preprint arXiv:2301.04780* (2023).
[25] Eun Youp Rha, Wei Shi, and Nicholas J Belkin. 2017. An exploration of reasons for query reformulations. *Proceedings of the Association for Information Science and Technology* 54, 1 (2017), 337–346.
[26] Soo Young Rieh et al. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management* 42, 3 (2006), 751–768.
[27] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 315–328.
[28] Amartya Sen. 2008. The idea of justice. *Journal of human development* 9, 3 (2008), 331–342.
[29] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[30] Linda R Waugh. 1982. Marked and unmarked: A choice between unequals in semiotic structure. (1982).
[31] Pak-Hang Wong. 2012. Dao, harmony and personhood: Towards a Confucian ethics of technology. *Philosophy & technology* 25, 1 (2012), 67–86.
[32] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul Bennett, Nick Craswell, and Susan Dumais. 2020. Analyzing and Learning from User Interactions for Search Clarification. In *Proc. SIGIR*. ACM.

Appendices

A EXTENDING TO OTHER DEMOGRAPHICS

In our work, we focus primarily on analyzing query reformulations corresponding to gender-based specialization. However, similar analysis would also be meaningful considering other demographic attributes, say race and age, and intersectional identities—*e.g.*, "black women" and "elderly man". To extend our analysis beyond gender, it would be convenient if we can (semi-)automatically detect other relevant group terms. Towards that goal, we represent every specializing query reformulation in terms of a template and a keyphrase. For example, for the query reformulation "hairstyles" → "hairstyles for women over 50", we identify the template to be "hairstyles for [KEYPHRASE]" and the corresponding keyphrase as "women over 50". Obviously, this applies even to specializing query reformulations that do not strictly correspond to demographic groups—*e.g.*, for the query reformulation "durable shoes" → "durable shoes for hiking" we identify "durable shoes for [KEYPHRASE]" as the template and "hiking" as the keyphrase. Let T be the set of templates and K the set of keyphrases extracted from a search log. Furthermore, let $K_{\text{seed}} \subset K$ be a set of seed keyphrases that corresponds to demographic group attributes—*e.g.*, based on gender. Then, to identify other keyphrases that correspond to demographic group attributes in this context, we score each keyphrase $k \in K$ as follows:

$$S_k = \sum_{t \in T} p(k|t) \cdot p(t|K_{\text{seed}}) \tag{1}$$

We rank the candidate keyphrases in descending order based on this score and manually inspect the top candidates to identify the keyphrases that correspond to demographic group attributes. Using this method on a single day of search logs from a commercial web search engine and a seed set of gender-related group terms, we are able to identify the following additional group terms in the top dozen keyphrases: "kids", "girls", "boys", "teens", "seniors", and "women over 50". In addition to inspecting the keyphrases listed lower down in this ranklist, more group terms may be discoverable by employing this approach iteratively to grow the seed set of group terms and by running the analysis over a search log corresponding to a larger time period. In future work, it would be interesting to also explore other approaches, such as considering latent representations of query reformulations [22], for identifying new group terms and corresponding group-based specializing query reformulation patterns.

B MANUAL ANALYSIS

Through our manual analysis, we want to get insight on how gender is represented on SERPs and user intent behind their group-based query reformulation. By manually verifying how genders are represented on overall SERP—which may include images, videos, and news results—we aim to develop a deeper understanding of the circumstances that led to these reformulations. We manually analyze a sample of reformulations from the search log data which leads us to the following additional insights about user query reformulation behaviors in this context:

Reformulations that seek gender-specific information. We find examples where the original query SERPs do not necessarily show any gender indicators through images, videos, or snippets of web results and users are specifically looking for gender specific information. For example, when users search for "ADHD" symptoms, the original query SERP does not show any gender specific results; there is no mention of gender specific terms in the snippets of retrieved web results. Users reformulate their query for particular gender to get gender-specific ADHD symptoms. Even though the original query SERP looks gender neutral and relevant to both genders, users may want to learn more about that topic for a specific gender. In the reformulated query SERP for "ADHD symptoms for women", we notice more gender-specific results. The first result has the term "women" in it and it is the women specific section from add.org which is the official website of Attention Deficit Disorder Association. Moreover, almost all of the web results contain the term "women" and all image and video results also contain content relevant specifically to women.

Reformulations that intensify group representation or reinforce over-representation. In this scenario, users may want to filter-out results related to another gender and only want results regarding the specified gender in their reformulated query. Looking at examples, we found SERPs where results are targeted for both genders. For example, where users search for "hispanic names", original query SERP shows results that are relevant for both men and women but users want gender specific hispanic name like "hispanic names for women". In reformulated query SERP, we notice the images and videos become more gender specific. We also notice that the webpage "100 Gorgeous Hispanic Girl Names" appears in both SERPs but at 3rd position in original query SERP and is located in the 1st position in reformulated query SERP.

We found queries where the original query SERPs are slightly skewed towards one gender through it's content; particular gender is slightly more dominant in image and video representations and texts in retrieved web pages. For example, the original SERP of the query "cordless razor" has more mention of men in retrieved web pages and the images are mostly targeted to men. However, the next query is "cordless razor for men" which emphasizes on men-specific results and exclude gender-neutral results. In this scenario, reformulated query SERP shows more gender-specific results; there are more web pages on reviews of cordless razor for men. In both SERPs, the first results are from amazon, however in the reformulated query SERP, the link is directing specifically to cordless razors for men.

There are queries where original SERPs show notable differences in gender representation through their content including images, videos, news, and texts in retrieved web pages. Such unequal representation of gender in SERP can happen due to reflecting societal stereotypes or bias. For example, the SERP of the original query "NCAA basketball score" shows results mostly about men's team, games, and game scores; images, news, and videos also heavily represent men. When user reformulate their queries to "NCAA men's basketball score", they may want more results related to men and exclude any gender-neutral results. In this scenario, the reformulated query SERP does not change a lot from the original query SERP; this shows more results that are relevant to men.

Reformulations that correct for group under-representation. In this scenario, users find fewer results about the particular gender relevant to their information need and so may explicitly reformulate the query mentioning that gender to find more results. Original SERP slightly over-represent one gender through images and texts. For example, the original SERP for the query "dillards shoes" shows products related to women through images and ads and thus the results seem more relevant and targeted to women. The next query is "dillards shoes for men" which brings results suitable and relevant for men. In this scenario, reformulated SERP shows differences in image and videos than the original SERP. The top images of the original SERP has more women's shoes but in reformulated query SERP all of the images were for men. The first result of the reformulated query SERP in the men's section from the official site of Dillard's which did not appear in original query SERP.

Original SERPs show notable differences in gender representation in retrieved images, videos, news, and webpages. For example, the original SERP of the query "NCAA basketball score" shows results mostly about men's teams, games, and game scores; images, news, and videos also heavily represent men. Only one retrieved web page is about women in original query SERP. Hence, the next query is often "NCAA women's basketball score" where users can find results relevant and more specific to women. In this scenario, reformulated SERP shows visible differences than original SERP including images, videos, and news.

Reformulation that are influenced by other SERP elements. In some cases, the web results and other results on SERP may represent groups differently. For example, the original query SERP of "hiking boots" shows web pages that are relevant and suitable for both men and women. However, the images are mostly targeted for men. Even though the web pages may show gender neutral results, images are skewed towards one gender. When users reformulate their query to "hiking boots for women", they may want to see SERP which looks visually more gender-specific and relevant to their gender intent.

Harmful reformulations. During our manual analysis, we also come across few cases where the reformulated query contains names of people and gender identities. Many of these queries misgender the subject and in fact these queries are likely outcome of harmful speculations about the subject's gender. We intentionally do not include any examples of such queries to avoid perpetuating further harm.

C ON THE CHALLENGES OF MEASURING REPRESENTATIONAL BIAS AND FAIRNESS

In this study, we analyze user behavior of searchers to gain more insights about how different demographic groups are represented in search results. Our findings indicate complex interactions between group representations on SERPs and other social factors contribute to the observed query reformulation patterns. We hope that this current work serves as a step towards understanding group representations in search results, and subsequently towards developing appropriate formalization for representational bias and fairness. However, there are additional challenges in making progress towards formalizing representational bias and fairness that future work should be wary of that we discuss in this section.

Politics of classification. A particularly difficult challenge for bias and fairness analysis is associating demographic group labels with the results returned by the search system. This goes far beyond issues of developing and operationalizing appropriate taxonomies for annotation, which in itself raises challenging questions, such as how to annotate artefacts that may not include any demographic markers or includes content corresponding to multiple groups, and whether the semantics of the group labels should reflect who the content is for, about, or by. But a more fundamental issue is that any classification, and corresponding development of taxonomies, are inherently political actions [1, 10]. Even our own analysis in this study suffers from many of these issues—for example, focusing on two genders perpetuates the false and problematic framing of gender as a binary construct and even with acknowledgment of that fact continuing to conduct analysis and present results within that binary frame contributes to erasure of other gender identities. Attempts to find proxy measures that can be used to automatically assign group labels (such as gender) to results takes away the subjects' rights to self-identify and may result in their misgendering. Finally, any automatic scheme for annotating results with demographic attributes may also find harmful applications in the hands of nefarious actors and requires that as researchers we handle these questions with utmost caution and necessary sensitivity and thoughtfulness.

What is fair? Adopting a normative view of fairness in the context of information access systems is at best operationally challenging and more likely in itself problematic. For example, let us consider the normative position that a search system is responsible for appropriate representation of all groups on the search result page. Leaving aside the challenges of defining what constitutes representativeness [6] in this context, the crude framing in which we associate group labels with individual results and attempt to achieve some "fair" target distribution on the result page leads us to a fundamental question: How do we choose a fair target distribution? Let us, for example, consider the question of representation of content corresponding to men's and women's sporting events on the search result page. Is the correct target distribution that both types of content should be exposed equally? Or, should it reflect the proportion in which these sporting events occur in the real world (which due to historical gender discrimination may skew it towards a specific gender)? Or, should it reflect the proportion of online content corresponding to these events (*i.e.*, supply-side distributions which may further skew it towards a specific gender)? Or, should it model the proportion of search queries corresponding to these events (*i.e.*, demand-side distribution which may again be disproportionately skewed towards a specific gender)? In reality, the target distribution may also be influenced directly or indirectly by the commercial interests of the owner of the search system (*e.g.*, based on monetizability of content) which is even more problematic. These choices require thoughtful consideration and as system designers it is less than obvious what "fair" distribution the systems should aim for. In fact, whether fairness is the correct framing for these problems should itself be questioned [21], and we should probably also look past the cultural hegemony of our own research communities to look for alternative framings for harm reduction [27, 28, 31].

Harmful exposure. Not all exposure is desirable. In the context of representational harms, it is crucial to consider the context and the way a subject may be exposed. Examples of undesirable exposure include when searches for black identifying first names in online ad delivery systems disproportionately suggest arrest record searches [29]. Representational harms also occur when an otherwise regular search query in the context of specific demographic groups return harmful stereoptypical or sexualized caricatures—*e.g.*, the queries "asian men" *vs.* "asian women". Any measure of representational harm must therefore not only consider the amount of exposure a group receives but in what context and form.

Measurement validity. Finally, any measure of bias, unfairness, or harm must be rigorously tested for construct validity, especially for consequential validity [16]. What are the short-term and long-term consequences of optimizing search systems towards said measures? Are they likely to reduce harm or compound them, including in ways that may not be immediately obvious or easy to measure?