Exposure Gerrymandering: Search Engine Manipulation Flying under Fairness' Radar

Tim de Jonge[1]

[1]*Radboud Universiteit, Houtlaan 4, 6525 XZ Nijmegen, The Netherlands*

Abstract

Modern society increasingly relies on Information Retrieval (IR) systems to answer various information needs. Since this impacts society in many ways, there has been a great deal of work to ensure the fairness of these systems, and to prevent societal harms. The Search Engine Manipulation Effect (SEME) is one such societal harm: voters could be influenced by means of these systems by showing biased search results. This paper introduces the notion of Exposure Gerrymandering, to illustrate how nefarious actors could create a system that appears unbiased to common fairness assessments, while substantially influencing the election at hand.

Keywords

Fairness, Information Retrieval, Search Engine Manipulation Effect

1. Introduction

Modern society increasingly relies on Information Retrieval (IR) systems to answer various information needs for a variety of needs, ranging from high impact applications like healthcare [1] and automated fact checking [2] to more everyday problems such as fashion matching [3] and music recommendation [4].

Since these IR systems impacts society in many ways, there has been a great deal of work to ensure the fairness of these systems. While it is easy to see why automated fact checking requires some care, near all IR systems have some fairness concerns. Although the targeted harm is not always named, there is a broad spectrum of IR literature mitigating societal harms [5].

One such harm is the Search Engine Manipulation Effect (SEME) [6]. This effect holds that by manipulating the results of politically loaded queries, search engines can manipulate the users votes by up to 20%, with higher outcomes for the most susceptible demographics. Follow-up research shows that this effect can be mitigated by showing warnings on skewed search results [7], and that the effect is diminished if the search engine results page provides a balanced set of results [8].

While this is promising work for those looking to mitigate the SEME, this work could be exploited by those looking to gain a political advantage. Epstein and Robertson [6] already show that the SEME can be hidden from individual users while still having a substantial effect. In this paper we introduce the notion of Exposure Gerrymandering: given strong bias to those most

FDIA '22: Future Directions in Information Retrieval, July 18–22, 2022, Lisbon, PT

tim.dejonge@ru.nl (T. d. Jonge)

CEUR Workshop Proceedings (CEUR-WS.org)

likely to change their votes, and weak and opposite bias to all other voters, one can mask this manipulation from common fairness definitions, while substantially influencing the election at hand.

2. Fairness in Information Retrieval

2.1. Existing Work

Let's look at a minimal document retrieval model, and investigate how attention can have different level of importance when we model the users as well.

Assume there is a set of documents $\mathcal{D} = \{d_i\}$, each with an already established political lean $\lambda \in [-1, 1]$; this axis does not necessarily entail the entire complexity of a political situation, but it can simply represent the axis we are investigating on at this point.

At each time t the information retrieval system gets a query $q_t \in \mathcal{Q}$, containing the information needs of a particular user x_t. In this age of datafication, the information on this user can be arbitrarily broad. For this situation, we assume that it at least holds the political preference p_t of the user, and some estimate of their tendency to be influenced by outside sources.

The system in turn orders the documents by their estimated relevance, and displays those documents it deems most relevant. There are several quantities of interest in the interaction between the user and the system in order to assess the fairness of the system.

First and foremost, we model *attention* a_{d_i}, also referred to as *exposure*. In a typical Information Retrieval system, the user is more likely to interact with a document if it is higher up in the ranking [9]. This motivates the phrasing that documents higher in the ranking get higher exposure, or more attention.

A common fairness definition is then Equity of Attention [10, 11, 12, 13]. Here, we define subgroups $G_n \subset \mathcal{D}$ which partition the set of documents; individual fairness [14] can be modeled by giving each document their own subgroup. In this paper we maintain two groups, G_+, G_-. Additionally, we require some notion of merit per document, m_{d_i}. The notion of Equity of Attention then holds that each subgroup be given attention proportional to their merit, i.e.

$$\frac{\sum_{d \in G_+} a_d}{\sum_{d \in G_+} m_d} = \frac{\sum_{d \in G_-} a_d}{\sum_{d \in G_-} m_d}$$

The concept of merit lies as of yet undefined, and thus, requires some further investigation. Natural choices for merit are $m_d = 1$ if we deem every document equally worthy without further stipulation. Alternatively, we can normalize for the amount of documents in each subgroup by choosing merit to be proportional to subgroup size ($m_d = |G_i|$): this would result in equal attention per subgroup. These options are referred to as Equality of Attention, but much more infrequently used, and for the remainder of this document, disregarded.

The main other choice is to set a document's merit equal to its relevance r_{d_i}, which is a shorthand for the desire for the user at hand to see the document d_i. Equity of Attention can then be rephrased as "give each subgroup attention to the extent that users want to see the documents in this group". This notion is libertarian in that it does not address broader societal bias either in the creation of the documents or in the users' perceptions, but particularly for marketplaces, it seems appropriate.

Although conceptually attractive, relevance seems noisy, to the point of being practically intractable. First, the relevance of a document is not directly observable outside of research contexts. This means that any assessment of real-world IR systems along this notion will have to either estimate relevance from user behaviour, or create a more labour intensive inquiry, in which users or judges have to investigate all documents and mark down which they found relevant to their query, and to which extent.

Secondly, from the first, is that this relevance is an imprecise measure. TREC, the largest applied Text Retrieval conference, mostly uses only binary relevance. Work to comprehensively define relevance in the best way to assess Information Retrieval systems in general has been ongoing for a long time, and does not seem to reach a conclusion any time soon [15, 16, 17]. To combat the noise in relevance, fairness can be assessed across multiple queries. Here, we accumulate relevance and attention across queries that the system has seen, and only compare the totals against each other, i.e. [13]

$$\frac{\sum_{q\in Q}\sum_{d\in G_+} a_d}{\sum_{q\in Q}\sum_{d\in G_+} m_d} = \frac{\sum_{q\in Q}\sum_{d\in G_-} a_d}{\sum_{q\in Q}\sum_{d\in G_-} m_d}$$

Several variations on this theme were made; the following remarks hold with regard to most approaches following these lines.

2.2. Exposure Gerrymandering

A recent salient example of a dubious IR practice is the manipulation of the US Election by Cambridge Analytica [18]. Here, Cambridge Analytica gathered a large amount of personal data of several millions of Facebook users through seemingly neutral means. Armed with this knowledge on the personal preferences of the users, they used this to influence the US Election by micro-targeting pro-Trump or anti-Hillary advertisements at those most susceptible to them; the company brass claimed that they could carry the elections by targeting only a few key voters in key districts.

This Cambridge Analytica case is an example of the Search Engine Manipulation Effect (SEME). This effect holds that biased search result rankings can shift the voting preferences of (primarily undecided) voters, meaning that a malicious actor could influence the elections through Information Retrieval System [6]. This type of campaigning is clearly attractive given the recent finding that traditional campaigning is minimally effective; except when taking unusually unpopular positions, and targeting more persuadable voters [19]. Micro-targeting allows for the taking of unpopular stances while lacking the punishment for the greater people finding out, while ensuring this information lands directly at these swing-voters.

Advertisement systems can reasonably be assessed as information retrieval systems, and there is a clear political axis to assess on. Unfortunately, in this story we can assume a malicious actor, and see what room they have within the above framing to create havoc.

The main notion this paper introduces is that of Exposure Gerrymandering. Borrowing this term from the American electoral system, the intuition is that by cleverly dividing the exposure different articles get across different users, one could create a strong influence on the users' votes towards the negative class, while the fairness criteria would not indicate a significant bias.

To see this, recall that the advertiser gets a request containing the users' political preferences, alongside with their likelihood of being influenced. If we assume here a policy where all those with polarized or fixed opinions get a small amount of political adverts for the positive class, this could easily outweigh heavy focus on more moderate voters that could be influenced. Since fairness criteria assess each query as equally weighted, adding a small amount of inconsequential bias to the majority of the population can mask heavy bias at the most sensitive points.

Some authors protect their method from this type of abuse by explicitly mentioning that the system targets utility maximisation in the sense of classical evaluation metrics. In these cases, the limited applicability is acknowledged, but a way forward still has to be found in cases where this is not guaranteed [20, 21].

3. Limitations of Algorithmic Fairness

Selbst *et al.* [21] provide an excellent overview of difficulties in abstracting the complicated notion of fairness, by indicating five traps that fair-ML work falls into. First and foremost of these is the Framing Trap, or "[The] failure to model the entire system over which a social criterion, such as fairness, will be enforced". In the case of the SEME, we can see that this is also a real risk – one needs to consider the results of showing users the rankings rather than the rankings themselves.

Important in this regard is Friedler *et al.* [22] making their distinction between several *spaces* from which one can observe a model:

- *Construct Space*: Consists of the real world, if it were ideally observable. Here, no approximations need to be made, one could simply directly observe a document's relevance
- *Observed Space*: Consist of the data-set the system has available; i.e. the measured version of Construct Space
- *Decision Space*: Consists of the output of the system

A principal reason of unfairness in modeling is then that we cannot directly observe Construct Space, and that rather, all modeling happens on noisy and biased encodings of Construct Space. To make a model fair, we have to investigate the ways in which this encoding between Observed Space and Construct Space are imperfect, and mitigate these imperfections.

This notion of Construct Space addresses the claim that data entering the system is unbiased, but this conceptualization also halts at the system decision, not taking into account how this decision would impact the relevant stakeholders. One could reasonably consider another space after Decision Space to capture outcomes for different groups after the system decision [23].

Albright [24] demonstrates that it does indeed have value to consider the larger societal structure, also after the point of computerized decision. She details the situation in the judicial system in Kentucky, where Kentucky House Bill 463 (HB463) mandated the use of a risk assessment, and standardized the possible outcomes based on this risk assessment. In exceptional cases, the judges were allowed to deviate from the risk assessment based on their own discretion. It turned out that the judges' discretion resulted in a disproportionate amount of black Americans getting higher bails than the system advised, when compared to white Americans given the same risk score. Simply designing a fair digital system is not enough, one must consider the implementation and the larger social context before being able to call the system fair in a substantial way.

These criticisms show a broad difficulty those who want to make an algorithm fair have to face: That even though an algorithm may be *prima facie* fair, or the algorithm may even appear fair through the lens of a fairness metric, that does not mean the algorithm is fair through a broader societal lens.

4. Discussion and Future Work

The concept of Exposure Gerrymandering is tough to address. Since it lies one step beyond the machine's decision, any model that attempts to diminish the effects of Exposure Gerrymandering necessarily must also model the users of the system, and the extent to which their votes would shift when faced with the biased rankings provided by the system. Further research would have to be done to investigate how large this effect can be under various circumstances, in order to adequately model Exposure Gerrymandering.

The notion of Exposure Gerrymandering could also be used to motivate a broader notion of fairness, looking beyond the system's horizon. A system impacts society in many different ways, and can appear to be fair in one regard while violating other reasonable fairness definitions. Because of this, I would advocate for a more pluralist approach to fairness, where benefits for different stakeholders are considered both at the design-step and throughout the life cycle of the system.

Taking this broader view, a non-technical solution can be found. Since Exposure Gerrymandering heavily relies on micro-targeting in order to mask its intent, reducing the degree to which advertisers can target their systems would also mitigate the effects. Dependent on the way this is implemented, this could improve user privacy, or prevent other downstream harms stemming from micro-targeting.

5. Conclusion

This paper introduces the notion of Exposure Gerrymandering: someone determined to gain political advantage can appear fair to typical fairness metrics, while achieving manipulation of the population through their IR system. Further research is needed to combat Exposure Gerrymandering, and to investigate whether there is a broader fairness approach that can capture this type of challenge.

References

[1] A. Hashavit, H. Wang, R. Lin, T. Stern, S. Kraus, Understanding and mitigating bias in online health search, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing

Machinery, New York, NY, USA, 2021, p. 265–274. URL: https://doi.org/10.1145/3404835.3462930. doi:10.1145/3404835.3462930.

[2] N. Vo, K. Lee, Learning from fact-checkers: Analysis and generation of fact-checking language, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 335–344. URL: https://doi.org/10.1145/3331184.3331248. doi:10.1145/3331184.3331248.

[3] X. Yang, X. He, X. Wang, Y. Ma, F. Feng, M. Wang, T.-S. Chua, Interpretable fashion matching with rich attributes, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 775–784. URL: https://doi.org/10.1145/3331184.3331242. doi:10.1145/3331184.3331242.

[4] A. B. Melchiorre, N. Rekabsaz, E. Parada-Cabaleiro, S. Brandl, O. Lesota, M. Schedl, Investigating gender fairness of recommendation algorithms in the music domain, Information Processing & Management 58 (2021) 102666.

[5] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5454–5476. URL: https://aclanthology.org/2020.acl-main.485. doi:10.18653/v1/2020.acl-main.485.

[6] R. Epstein, R. E. Robertson, The search engine manipulation effect (seme) and its possible impact on the outcomes of elections, Proceedings of the National Academy of Sciences 112 (2015) E4512–E4521.

[7] R. Epstein, R. E. Robertson, D. Lazer, C. Wilson, Suppressing the search engine manipulation effect (seme), Proceedings of the ACM on Human-Computer Interaction 1 (2017) 1–22.

[8] T. Draws, N. Tintarev, U. Gadiraju, A. Bozzon, B. Timmermans, This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 295–305.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, G. Gay, Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search, ACM Transactions on Information Systems (TOIS) 25 (2007) 7–es.

[10] A. Singh, T. Joachims, Fairness of exposure in rankings, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2219–2228.

[11] M. Morik, A. Singh, J. Hong, T. Joachims, Controlling fairness and bias in dynamic learning-to-rank, in: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 429–438.

[12] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, B. Carterette, Evaluating stochastic rankings with expected exposure, in: Proceedings of the 29th ACM international conference on information & knowledge management, 2020, pp. 275–284.

[13] A. J. Biega, K. P. Gummadi, G. Weikum, Equity of attention: Amortizing individual fairness in rankings, in: The 41st international acm sigir conference on research & development in information retrieval, 2018, pp. 405–414.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

[15] E. M. Voorhees, I. Soboroff, J. Lin, Can old trec collections reliably evaluate modern neural retrieval models?, arXiv preprint arXiv:2201.11086 (2022).

[16] E. Sormunen, Liberal relevance criteria of trec- counting on negligible documents?, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 324–330.

[17] D. E. Losada, J. Parapar, A. Barreiro, When to stop making relevance judgments? a study of stopping methods for building information retrieval test collections, Journal of the Association for Information Science and Technology 70 (2019) 49–60.

[18] H. Berghel, Malice domestic: The cambridge analytica dystopia, Computer 51 (2018) 84–89.

[19] J. L. Kalla, D. E. Broockman, The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments, American Political Science Review 112 (2018) 148–166.

[20] S. Fazelpour, Z. C. Lipton, Algorithmic fairness from a non-ideal perspective, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 57–63.

[21] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 59–68.

[22] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, On the (im) possibility of fairness, arXiv preprint arXiv:1609.07236 (2016).

[23] H. Weerts, L. Royakkers, M. Pechenizkiy, Does the end justify the means? on the moral justification of fairness-aware machine learning, arXiv preprint arXiv:2202.08536 (2022).

[24] A. Albright, If you give a judge a risk score: evidence from kentucky bail decisions, Harvard John M. Olin Fellow's Discussion Paper 85 (2019) 16.