Ranking Robustness in Adversarial Retrieval Settings

Gregory Goren
gregory.goren@campus.technion.ac.il
Technion — Israel Institute of Technology

ABSTRACT

The adversarial retrieval setting over the Web entails many challenges. Some of these are due to, essentially, ranking competitions between document authors who modify them in response to rankings induced for queries of interest so as to rank-promote the documents. One unwarranted consequence is that rankings can rapidly change due to small, almost indiscernible changes, of documents. In recent work, we addressed the issue of ranking robustness under (adversarial) document manipulations for feature-based learning-to-rank approaches. We presented a formalism of different notions of ranking robustness that gave rise to a few theoretical findings. Empirical evaluation provided support to these findings. Our next goals are to devise learning-to-rank techniques for optimizing relevance and robustness simultaneously, study the connections between ranking robustness and fairness, and to devise additional testbeds for evaluating ranking robustness.

ACM Reference Format:
Gregory Goren. 2019. Ranking Robustness in Adversarial Retrieval Settings. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3331184.3331414

1 INTRODUCTION

In the Web retrieval setting, many document authors are "ranking-incentivized". That is, they are interested in having their documents highly ranked for some queries by search engines. To this end, they often respond to rankings by introducing modifications to their documents (a.k.a., search engine optimization [7]). These modifications could be almost indiscernible by users, yet cause changes to rankings. We argue that this is an unwarranted result of the ranking competition that takes place between document authors. A case in point, users can loose faith in the search system if they observe these rapid ranking changes.

One of our main goals in this proposal is to address the issue of ranking robustness in the face of adversarial document manipulations. While the robustness (stability) of classification algorithms, specifically, with respect to adversarial manipulations [2–5, 9], has attracted much research attention, we are not aware of prior work on addressing ranking robustness.

In recent work [6], we described the first formalism, to the best of our knowledge, of different notions of ranking robustness with respect to (adversarial) document manipulations. We focused on feature-based learning-to-rank approaches [8].

Our main research goal is devising learning-to-rank approaches that are optimized for both relevance and robustness. This is an intriguing challenge since robustness is with respect to temporal changes of rankings based on responses of document authors to induced rankings. While in our recent work [6] we demonstrated how robustness can be improved via regularization mechanisms, here we are interested in devising loss functions that are directly optimized for both relevance and robustness.

Our next order of business is studying the connections between ranking robustness and ranking fairness [1, 10]. A case in point, recent work demonstrated the merits of introducing randomization to ranking functions so as to increase fairness [1]. However, such randomization hurts ranking robustness. Our goal is to study whether this tradeoff is inherent to the connection between robustness and fairness, and how it can be controlled.

Additional goal on our research agenda is developing datasets that will allow to study and evaluate ranking robustness. For example, to evaluate a newly proposed ranking approach that presumably promotes robustness, one has to use this ranking function in a "live" system where authors respond to rankings induced by this function.

Acknowledgments. We thank the reviewers for their comments. This research is based upon joint work with Oren Kurland, Moshe Tennenholtz and Fiana Raiber. The work is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 740435).

REFERENCES

[1] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proc. of SIGIR*. 405–414.
[2] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2015. Analysis of classifiers' robustness to adversarial perturbations. *CoRR* abs/1502.02590 (2015).
[3] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2016. Robustness of classifiers: from adversarial to random noise. In *Proc. of NIPS*. 1624–1632.
[4] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2017. The Robustness of Deep Networks: A Geometrical Perspective. *IEEE Signal Processing Magazine* 34, 6 (2017), 50–62.
[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *Proc. of ICLR*.
[6] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2018. Ranking Robustness Under Adversarial Document Manipulations. In *Proc. of SIGIR*. 395–404.
[7] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web Spam Taxonomy. In *Proc. of AIRWeb 2005*. 39–47.
[8] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval.* Springer. I–XVII, 1–285 pages.
[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proc. of ICLR*.
[10] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proc. of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 22.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331414