Measuring Gender Stereotype Reinforcement in Information Retrieval Systems*

Discussion paper

Alessandro Fabris[1], Alberto Purpura[1], Gianmaria Silvello[1] and Gian Antonio Susto[1]

[1]*Department of Information Engineering, University of Padua, Padua, Italy*

Abstract

Can we measure the tendency of an Information Retrieval (IR) system to reinforce gender stereotypes in its users? In this abstract, we define the construct of Gender Stereotype Reinforcement (GSR) in the context of IR and propose a measure for it based on Word Embeddings. We briefly discuss the validity of our measure and summarize our experiments on different families of IR systems.

Keywords

Fairness, Gender Stereotypes, Information Retrieval, Search Engines, Word Embeddings

1. Introduction

Search Engines (SEs) increasingly act as the gatekeepers of information. Their role in information access is undisputed, with a user base exceeding 90% of all people connected to the internet [1]. SEs inevitably influence users, helping them map concepts and link entities across queries and documents. For this reason, they can play an important role in countering or reinforcing stereotypical associations [2].

Stereotypes are generalised beliefs about groups of individuals, held widely in a population of interest. They arise from a co-occurrence of features, such as membership to a group and display of certain traits and roles. The extent to which an individual believes a stereotypical trait to be common in a given group is often measured through an association test between groups and traits [3].

Male and female are highly salient categories in human cognition, available from an early age for stereotypical associations.[2] As a result, western societies maintain a wide range of gender stereotypes, relating e.g. to professions, career, competence, care, predisposition for science and mathematics. The same stereotypes are also found in artifacts and technology produced by the same societies. For example, the search results of popular image SEs were found to contain gender-stereotypical associations [5] and to influence users' cognition accordingly [6]. Only recently, novel approaches to measure gender bias in text-based SEs have been proposed [4, 7].

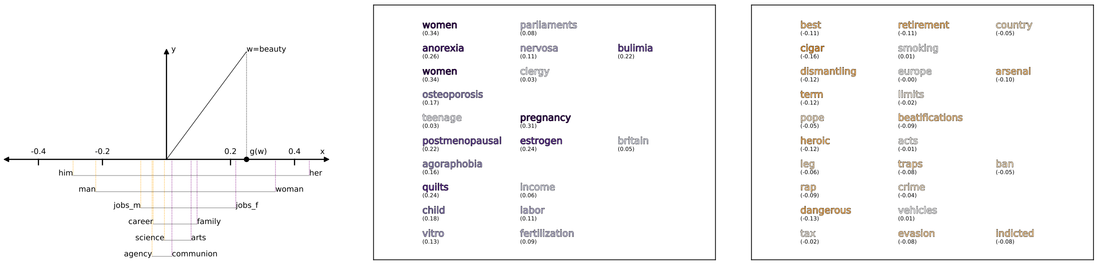
IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

fabrisal@dei.unipd.it (A. Fabris); purpuraa@dei.unipd.it (A. Purpura); silvello@dei.unipd.it (G. Silvello); gianantonio.susto@unipd.it (G. A. Susto)

CEUR Workshop Proceedings (CEUR-WS.org)

[2] The present binary framing of gender is a consequence of this fact and a clear limitation of our work. This is a common weakness for work in this space; addressing it is far from trivial.

* Extended abstract of Fabris et al. [4].



(a) Gender direction

(b) “Female” queries

(c) “Male” queries

Figure 1: Adapted from [4]. **Left**. Gender projection $g(w)$ for different words and concepts related to known gender stereotypes. Four stereotypical dychotomies are depicted in Figure 1a, which relate to gender concentrations in professions (jobs_m vs jobs_f), work-related choices (career vs family), predisposition for subjects (science vs arts), competence and warmth (agency vs communion). For each of these concepts, Figure 1a depicts the average projection of a set of words that have been proposed as representative of the concept. The projection of stereotypically female concepts is always (statistically significantly) lower than the projection of their male counterpart. **Middle and right**. Gendered queries from the Robust04 collection [8] according to gender projection $g(\cdot)$. The text is printed with color-coded gradient where strongly “masculine” words are orange, strongly “feminine” words are purple, neutral words are white. Indeed the words in these queries relate to gender either intrinsically (women), biologically (menopause) or stereotypically (e.g. child relates to family, quilts to steretypical occupations, heroic to agency, dangerous is contrary to communion).

In this work we provide an overview of the Gender Stereotype Reinforcement (GSR) construct and measure of Fabris et al. [4].

2. GSR: Construct, Measure and Validity

In the context of IR, we define GSR as a SE’s tendency to reinforce (or counter) gender-stereotypical associations in its users. Direct measurement of this construct would require impractical longitudinal user studies of counterfactual nature. Fabris et al. [4] propose a computational approach, based on Word Embeddings (WEs).

Indeed, WEs have been found to reliably encode several gender stereotypes [9, 10], typically along a single direction of the embedded space. More precisely, Bolukbasi et al. [9] show how to isolate a problematic direction, called *gender subspace*, where gender-related concepts are clustered in accordance with gender stereotypes. To illustrate this concept, Figure 1a depicts the gender direction w_g of Word2vec embeddings [11] along the x axis. A sample word, $w =$ beauty, is projected onto the gender direction, where it is closer to intrinsically female words (her, woman) and to stereotypically female concepts than to their male counterparts.

Let us indicate by $g(w) = (w \cdot w_g)/(|w||w_g|)$ the function associating a word w with its normalized scalar projection on the gender direction w_g. By extension, $g(\cdot)$ maps a query q_i into the average projection of its words $g(q_i)$. Let us call $g(x)$ the *genderedness* of x. Moreover, we apply function $g(\cdot)$ to the ranked list of documents $\mathcal{L}_i$ returned by an IR system s in response to query q_i. We indicate it by $g(\mathcal{L}_i)$ and define it as the average projection of words in ranked documents d_k, weighted according to the rank of each document in $\mathcal{L}_i$. In symbols $g(\mathcal{L}_i) = \sum_{d_k \in \mathcal{L}_i} w_k \cdot g(d_k)$, with w_k computed according to a *DCG-like* logarithmic discount [12].[3] Given a set of N queries $\mathcal{Q}$ and a collection of documents $\mathcal{D}$ available for retrieval, we define the GSR of an IR system s over $(\mathcal{Q}, \mathcal{D})$ in terms of the correlation between the genderedness of queries in $g(q_i), q_i \in \mathcal{Q}$ and the genderedness of ranked lists of documents $g(\mathcal{L}_i)$ produced in response. More precisely, GSR is defined as

$$m_s(\mathcal{Q}, \mathcal{D}) = \frac{1}{\sigma^2_{g(q)}} \frac{1}{N} \sum_{i=1}^{N} (g(q_i) - \mu_q)(g(\mathcal{L}_i) - \mu_{\mathcal{L}}), \tag{1}$$

where $\mu_q, \mu_{\mathcal{L}}$ represent the average genderedness of queries and ranked lists, while $\sigma^2_{g(q)}$ is a scaling factor to go from correlation to slope coefficient. Informally, Equation 1 captures the agreement between the language of queries and documents along stereotypically gendered lines, induced by an IR system s.

A thorough assessment of the suitability of this equation to measure the GSR construct is an important and complex endeavour undertaken in [4]. Here we show the precision of the projection function $g(\cdot)$ in finding interesting queries for the study of gender stereotypes in SEs. Figure 1b (1c) shows the ten queries with lowest (highest) genderedness $g(q)$ in the Robust04 collection [8], which are the most associated with women (men) according to $g(q)$. Indeed these are *gendered* queries, ranging from intrinsically gendered (mentioning `women`), to biologically gendered (mentioning `menopause`), to stereotypically gendered (with `quilts` and `child` among words in stereotypically female queries, `dangerous` and `heroic` in stereotypically male queries).

3. Experiments and Discussion

Our experiments on the Robut04 collection [8], omitted here for brevity, compare IR ranking algorithms from different families. We consider lexical models (e.g. BM25 - [13]), semantic models (e.g. w2v add - [14]) and neural architectures (e.g. MatchPyramid - [15]). We find that semantic models, based on biased WEs, are most prone to reinforcing gender stereotypes, while neural systems based on the same word representations can mitigate this effect. Indeed neural models exhibit low GSR, comparable to that of lexical systems such as BM25. Moreover, we test the reliability of these conclusions by measuring GSR according to two different sets of WEs (Word2Vec [11] and fastText [16]), finding strong agreement between the two. Finally, we assessed the impact of debiasing WEs [9] on downstream IR tasks. By measuring system performance and GSR both before and after debiasing, we find this approach to be superficial and insufficient to reduce the tendency of an IR system to reinforce gender stereotypes.

[3] To be precise, $g(\mathcal{L}_i)$ should be query-dependent [4]. Here we neglect this aspect.

Acknowledgments

Part of this work was supported by MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).

References

[1] K. Purcell, J. Brenner, L. Rainie, Search engine use 2012, 2012. URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/media/Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf.

[2] S. U. Noble, Algorithms of oppression: How search engines reinforce racism, NYU Press, 2018.

[3] A. G. Greenwald, D. E. McGhee, J. L. Schwartz, Measuring individual differences in implicit cognition: the implicit association test., Journal of personality and social psychology 74 (1998) 1464–1480.

[4] A. Fabris, A. Purpura, G. Silvello, G. A. Susto, Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms, Information Processing & Management 57 (2020) 102377.

[5] J. Otterbacher, J. Bates, P. Clough, Competent men and warm women: Gender stereotypes and backlash in image search results, in: Proc. of CHI 2017, ACM, 2017, p. 6620–6631.

[6] M. Kay, C. Matuszek, S. A. Munson, Unequal representation and gender stereotypes in image search results for occupations, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 3819–3828.

[7] N. Rekabsaz, M. Schedl, Do neural ranking models intensify gender bias?, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2065–2068.

[8] D. Harman, The darpa tipster project, SIGIR Forum 26 (1992) 26–28.

[9] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Advances in neural information processing systems, 2016, pp. 4349–4357.

[10] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[12] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems 20 (2002) 422–446.

[13] S. E. Robertson, U. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval (FnTIR) 3 (2009) 333–389.

[14] I. Vulić, M.-F. Moens, Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings, in: Proc. of SIGIR 2015, ACM, 2015, p. 363–372.

[15] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, X. Cheng, Text matching as image recognition, in: Proc. of AAAI 2016, AAAI Press, 2016, p. 2793–2799.

[16] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, ACL, 2017, pp. 427–431.