Bias and Fairness in Effectiveness Evaluation by Means of Network Analysis and Mixture Models

Michael Soprano, Kevin Roitero, Stefano Mizzaro
University of Udine, Italy
soprano.michael@spes.uniud.it,roitero.kevin@spes.uniud.it,mizzaro@uniud.it

ABSTRACT

Information retrieval effectiveness evaluation is often carried out by means of test collections. Many works investigated possible sources of bias in such an approach. We propose a systematic approach to identify bias and its causes, and to remove it, thus enforcing fairness in effectiveness evaluation by means of test collections.

1 INTRODUCTION

In Information Retrieval (IR) the evaluation of systems is often carried out using test collections. Different initiatives, such as TREC, FIRE, CLEF, etc. implement this setting in a competition scenario. In the well known TREC initiative, participants are provided with a collection of documents and a set of *topics*, which are representations of information needs. Each participant can submit one or more *run*, that consists in a ranked list of (usually) 1000 documents for each topic. The retrieved documents are then pooled, and expert judges provide relevance judgements for the pooled ⟨topic, document⟩ pairs. Then, an effectiveness metric (such as AP, NDCG, etc.) is computed for each ⟨run, topic⟩ pair, and the final effectiveness metric for each run is obtained by averaging its effectiveness score over the set of topics. Finally, the set of runs is ranked in descending order of effectiveness.

Different works investigated possible source of bias for this evaluation model by looking at system-topics correlations. In this work we propose to extend prior work by considering the many dimensions of the problem, and we develop a statistical model to capture the magnitude of the effects of the different dimensions.

2 BACKGROUND AND RELATED WORK

2.1 HITS Hits TREC

The output of the TREC initiative can be represented as an effectiveness matrix E as in Table 1, where each s_i is a system configuration (i.e., run), each t_j is a topic, e_{s_i,t_j} represents the effectiveness (with a metric such as AP) of the i-th system for the j-th topic, E_s and E_t represent respectively the average effectiveness of a system (with a metric such as MAP) and the average topic difficulty (with a metric such as AAP [5, 8]).

To capture the bias of such evaluation setting, Mizzaro and Robertson [5] normalise the E matrix, or more precisely each e_{s_i,t_j} in two ways: (i) by removing the system effectiveness effect, achieved by subtracting E_s from each e_{s_i,t_j}, and (ii) by removing the topic effect, achieved by subtracting E_t from each e_{s_i,t_j}. After the normalisation, Mizzaro and Robertson merge the two effectiveness matrices obtained from (i) and (ii) to form a graph in which: each link from a system to a topic expresses how a system thinks a topic is easy, and each link from a topic to a system expresses how a topic thinks a system is effective. Then, Mizzaro and Robertson compute the *hubness* and *authority* values of the systems and topics by running the HITS algorithm [3] on such graph; the hubness value for a system expresses its ability to recognise easy topics, while the hubness value for a topic expresses its ability to recognise effective systems. Results of the analysis by Mizzaro and Robertson [5], as well as by Roitero et al. [8], demonstrate that the evaluation is biased, and in particular that easy topics are better in recognising effective systems; in other words, a retrieval system to be effective needs to be effective on the easy topics.

Table 1: Effectiveness Table.

	t_1	$\cdots$	t_n	E_s
s_1	e_{s_1,t_1}	$\cdots$	e_{s_1,t_n}	$E_s(s_1)$
$\vdots$		$\ddots$		$\vdots$
s_m	e_{s_1,t_1}	$\cdots$	e_{s_m,t_n}	$E_s(s_m)$
E_t	$E_t(t_1)$	$\cdots$	$E_t(t_n)$	

2.2 HITS Hits Readersourcing

Soprano et al. [9] used the same analysis based on the HITS algorithm and described in Section 2.1 to analyse the bias present in the *Readersourcing* model [4], an alternative peer review proposal that exploits readers to assess paper quality. Due to the lack of real data, Soprano et al. run a series of simulations to produce synthetic but realistic user models that simulate readers assessing the quality of the papers. Their results show that the Readersourcing model presents some (both good and bad) bias under certain conditions derived from how the synthetic data is produced, as for example: (i) the ability of a reader to recognise good papers is independent from the fact that s/he read papers that on average get high/low judgements, and (ii) a paper is able to recognise high/low quality readers independently from its average score or from its quality.

2.3 Breaking Components Down

Breaking down the effect caused by a dimension on a complex system has been widely studied in IR. A problem of particular interest is to break down the system effectiveness score (such as AP) into the effect of systems, topics, and system components, like for example the effect of stemming, query expansion, etc. For this purpose, Ferro and Silvello [2] used Generalised Linear Mixture Models (GLMM) [6] to break down the effectiveness score of a system considering its components, Ferro and Sanderson [1] considered the effect of sub-corpora, and Zamperi et al. [10] provided a complete analysis on the topic ease and the relative effect of system configurations, corpora, and interactions between components.

IIR 2019, September 16–18, 2019, Padova, Italy

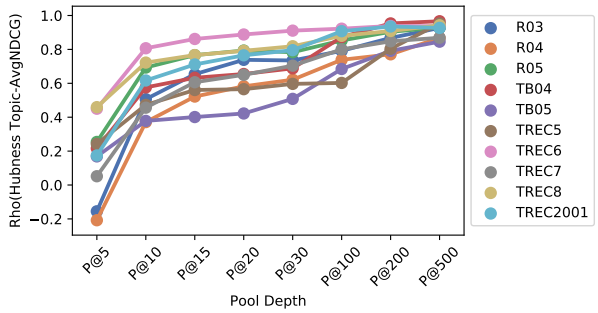

Figure 1: Effect of pool depth on the model bias.

3 EXPERIMENTS

In this paper, we propose to extend results from the related work (see Sections 2.1 and 2.2) to define a sound and engineered pipeline to find and correct the bias in the IR effectiveness evaluation setting. More in detail, we plan to investigate how the specific bias of effective systems being recognised by easy topics varies when varying the components of a test collection, such as the system population, pool depth, etc. Finally, we propose to use a GLMM as done in the related work (see Section 2.3) to compute the magnitude of effect of the various components on the bias.

3.1 Pool Effect

To investigate the effect of the different pool depths, we plan to compute, for each effectiveness metric, its value at difference cut-offs. Figure 1 shows a preliminary result: the plot shows, for the Precision metric, the different cut-off values on the x-axis and, on the y-axis, the bias value represented by the Pearson's ρ correlation between the hubness measure of systems and their average precision value; this bias represents the fact that effective systems are recognised by easy topics. As we can see from the plot, there is a clear trend suggesting that the bias grows together with the pool depth. The undesired effect that effective systems are mainly those that work well on easy topics becomes stronger when increasing pool depth.

3.2 Collection and Corpora Effect

To investigate the effect of the different collections, we plan to use different TREC collections: Robust 2003 (R03), 2004 (R04), and 2005 (R05), Terabyte 2004 (TB04) and 2005 (TB05), TREC5, TREC6, TREC7, TREC8, and TREC2001. Furthermore, we plan to break down the sub-corpora effect by considering the different corpora of the collections.

3.3 Metric Effect

We will investigate the effect of different evaluation metrics in the model bias, specifically we will consider: Precision (P), AP, Recall (R), NDCG, τ_{AP}, RBP, etc. When dealing with the metric effect, we can consider two approaches in the normalisation step: remove the average of system effectiveness and topics ease, as for example remove MAP and AAP from AP (as done by Mizzaro and Robertson [5], Roitero et al. [8], Soprano et al. [9]), or try more complex approaches; in the latter case, we can remove a score computed on a deep pool from one computed on a shallow pool (e.g., remove AP@10 from AP@1000) in order to remove the top-heaviness of a metric, or remove Precision (or Recall) from F1, to enhance the effect of precision-oriented or recall-oriented systems, and so on.

3.4 System and Topic Population Effect

Another effect we can investigate is to consider different systems and topic populations. We can consider, for example, systems ordered by effectiveness, topics ordered by difficulty, or even consider the most representative subset of systems / topics selected according to the strategy developed by Roitero et al. [7].

3.5 GLMM

Finally, we can develop a GLMM adapting the techniques used in [1, 2, 10] to study the effect that the different components described so far (see Sections 3.1–3.4) have on the bias of the model. Thus, we can define the following GLMM:

$$\text{Bias}_{ijklm} = \text{Pool}_i + \text{Collection}_j + \text{Corpora}_k + \text{System-subset}_l + \text{Topic-subset}_m + (\text{interactions}) + \text{Error}.$$

From the above equation, we can compute the *Size of Effect* index ω^2 which is an "unbiased and standardised index and estimates a parameter that is independent of sample size and quantifies the magnitude of difference between populations or the relationships between explanatory and response variables" [6]. Such index expresses the magnitude of the effect that the different components of a test collection have on the bias of the model.

4 CONCLUSIONS

Our contribution is twofold: we propose an engineered pipeline based on network analysis and mixture models that can be used to detect bias and its causes in retrieval evaluation, and we present some preliminary result. We plan to conduct the experiments described, that will allow to better understand the effect and cause of bias and fairness in the retrieval evaluation.

REFERENCES

[1] Nicola Ferro and Mark Sanderson. 2017. Sub-corpora Impact on System Effectiveness. In *Proceedings of the 40th ACM SIGIR Conference*. ACM, New York, 901–904.
[2] Nicola Ferro and Gianmaria Silvello. 2018. Toward an anatomy of IR system component performances. *JASIST* 69, 2 (2018), 187–200.
[3] Jon M. Kleinberg. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM* 46, 5 (Sept. 1999), 604–632.
[4] Stefano Mizzaro. 2003. Quality control in scholarly publishing: A new proposal. *JASIST* 54, 11 (2003), 989–1005.
[5] Stefano Mizzaro and Stephen Robertson. 2007. HITS Hits TREC: Exploring IR Evaluation Results with Network Analysis. In *Proceedings of the 30th ACM SIGIR Conference*. 479–486.
[6] Stephen Olejnik and James Algina. 2003. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods* 8, 4 (2003), 434.
[7] Kevin Roitero, J. Shane Culpepper, Mark Sanderson, Falk Scholer, and Stefano Mizzaro. 2019. Fewer topics? A million topics? Both?! On topics subsets in test collections. *Information Retrieval Journal* (2019).
[8] Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. 2017. Do Easy Topics Predict Effectiveness Better Than Difficult Topics?. In *Advances in Information Retrieval*, Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (Eds.). Springer, 605–611.
[9] Michael Soprano, Kevin Roitero, and Stefano Mizzaro. 2019. HITS Hits Readersourcing: Validating Peer Review Alternatives Using Network Analysis.. In *Proceedings of the 4th BIRNDL Workshop at the 42nd ACM SIGIR*.
[10] Fabio Zamperi, Kevin Roitero, Shane Culpepper, Oren Kurland, and Stefano Mizzaro. 2019. On Topic Difficulty in IR Evaluation: The Effect of Corpora, Systems, and System Components.. In *Proceedings of the 42nd ACM SIGIR Conference*.