IR-Cologne at TREC 2017 OpenSearch Track: Rerunning Popularity Ranking Experiments in a Living Lab

Narges Tavakolpoursaleh[1], Mandy Neumann[2], and Philipp Schaer[2]

GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany[1]
narges.tavakolpoursaleh@gesis.org
TH Köln (University of Applied Sciences), Cologne, Germany[2]
mandy.neumann@th-koeln.de
philipp.schaer@th-koeln.de

1 Introduction

In this paper, we will present our work on popularity-based relevance ranking within the SSOAR open access repository system where we reused a popularity data-driven ranking approach. We applied the same normalization method as last year, namely the Characteristic Scores and Scale Method (CSS). Our main focus was to test if we could reproduce the results of last year's track. We, therefore, see this work not as a sole engineering task to produce the best possible popularity ranking method for scientific literature but as a test bed for reproducibility experiments in the domain of living labs.

The TREC 2016 OpenSearch track was focused on ad-hoc search for scientific literature and three scientific search engines and document repositories were part of this living lab-centered evaluation campaign: (1) CiteSeerX, (2) Microsoft Academic Search, and (3) SSOAR - Social Science Open Access Repository. From these three only SSOAR remained in this year's OpenSearch track. The first author of this paper is responsible for the implementation of the living lab infrastructure and the LL4IR API that is necessary to include an online system into the OpenSearch evaluation campaign. This work is based on her Master's thesis at University of Bonn [8]. Details of the implementation are described in the two overview papers of the OpenSearch track [1,3].

2 Method

In previous LL4IR and OpenSearch campaigns [5,7] we experimented with different kinds of popularity data as a reranking factor. For our ranking method, we consider the number of views and the number of downloads in SSOAR as a popularity factor. Table 1 represented a sample of these data for documents in our corpus with a different year of publication and date of availability in SSOAR. We discovered different biases in the data that introduced flaws into the ranking when using them in a straight-forward manner, like the biases raw

Table 1: Sample of usage data for document in SSOAR

#downloads	#views	available-form	publication-year	site_docid
372	1147	2012-08-29	1998	document449
42	131	2015-12-01	2009	document45488
687	481	2012-08-29	2011	document29377
25	138	2014-04-14	2004	document38204
465	557	2012-08-29	1998	document1909

data introduces into the ranking due to e.g. different publication years, dates of availability or the lack of a common scale to compare the different usage data with each other.

We implemented a method to normalize the usage data to remove biases and to enable some kind of comparability. Our method was based on a procedure called the Characteristic Scores and Scales method (CSS) described by Glänzel [2]. The CSS method is used to find characteristic partitions for citation distributions. For more details about the method see our last year's paper [6]. In short: The method extracts classes of papers based on their citations interpreted as "poorly cited", "fairly cited", "remarkably cited", or "outstandingly cited". These partitions can then be used to normalize different kinds of usage data distributions. Plassmeier et al. [4] showed that these classes are applicable to other usage data sets like the number of record views or the number of loans at local libraries.

In this year we didn't change our ranking approach but reran the experiment with the new queries for OpenSearch 2017. Our aim was to get a direct comparison of two identical systems in the same living lab setting. We would like to compare the different usage patterns and see if the overall results of the two experiments are the same to check on the reproducibility of these kinds of experiments.

In table 2 we can see the results of our popularity data crawling for SSOAR compared to those from last year. We analyzed approx. 18 000 documents with an average of 287 downloads and 416 record views per document in SSOAR. Although the absolute number of candidate documents in our corpus is lower compared to last year, the per document popularity data is richer. Our 2017 corpus statistics listed in the table 3 shows small derivations after we dropped all documents that have null values for one of their attributes i.e. download numbers, view numbers, date of availability or year of publication.

3 Results

We applied the CSS normalization method on the usage data of our whole corpus. In figure 1 we can see the unprocessed view and download counts for four different time slices in the data. The plots show the counts for the in-system

Table 2: Comparison of corpus statistics on the popularity data gathered from SSOAR for the years 2016 and 2017.

	2016		2017	
	downloads	views	downloads	views
docs total	24 760	24 760	18 003	18 003
docs w. usage data	21 523	24 724	15 344	18 003
max	504 720	21 788	38 818	24 450
sum	6 549 674	9 822 049	5 175 322	7 492 304
avg	264.51	396.66	287.47	416.17

Table 3: Filtered statistics on the popularity data gathered from SSOAR collection for TREC 2017.

	downloads	views
docs total	17 732	17 732
avg	289.93	421.16
std	660.44	548.52
min	0	2
25%	37	170
50%	196	350
75%	340	517
max	38 818	24 450

publication dates for the year 2009, 2011, 2013, and 2015. We see the same skewed popularity data as in last year's experiment. In the CSS normalized data, these large differences in the counts are slightly decreased and therefore allow for a better integration of the ranking formula and a better comparison. We can still see the biases introduced by the different publication years that are present in the raw data distributions. In another approach, we applied the CSS normalization method on the usage data for each publication year individually. Figure 2 confirms that in consequence the curves come together significantly and could possibly provide a better comparison. However, in our ranking approach, we applied the CSS normalization on the undivided collection since we want to compare the result with the result of the previous year.

SSOAR shared 1165 search terms and a list of approx. 18 000 candidate documents for the Open Search track in TREC 2017. These numbers increased in comparison to TREC 2016 in its last round where 1127 search terms (including 118 terms naming browsing categories instead of ad-hoc search terms) were distributed to the participants of the Open Search track. During the experimental phases in 2016 and 2017, SSOAR reported 16 730 and 33 583 feedback records, respectively. The number of reported clicks on candidate documents were 329 for 2016 and 475 clicks for 2017.

Table 4 describes the result of our ranking method in OpenSearch 2017 and 2016. Last year the number of clicks in the result is not reported. We (Gesis)

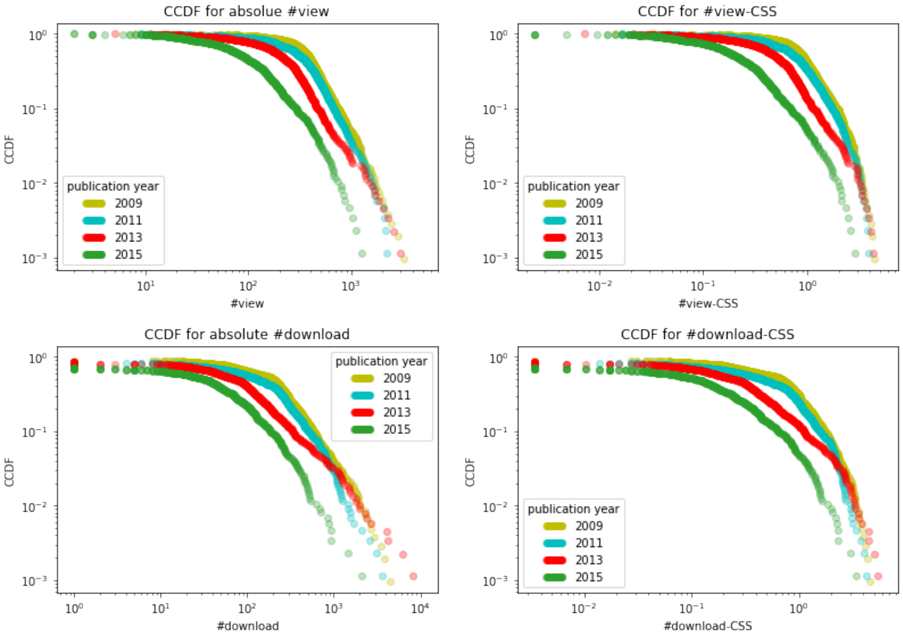

Fig. 1: Plot of raw number of the view and download (left) and CSS normalization (right) rates for four different time slices in the usage data of SSOAR.

Table 4: Result from TREC OpenSearch 2016 and 2017 for Team GESIS / IR-Cologne

	Win	Ties	Losses	Outcome	Impression	Clicks
IR-Cologne 2017	9	2	6	0.60	3700	31
Gesis 2016 round#1	1	460	0	1.0	460	NG
Gesis 2016 round#2	1	96	0	1.0	97	NG

won the competition by having just one win click more than the base system. In 2017, however, we had more wins and losses. Still, our system performed better than the other ranking systems, Webis with an outcome of 0.462 and ICTNET with an outcome of 0.400.

Figure 3 shows a comparison between our experimental ranking scores in TREC 2016 and TREC 2017 for one sample search term. While the retrieval system remained stable for the two years, the computed ranking score differ a lot. The figure shows only those documents that were included in both systems in 2016 and 2017. They values in the figure were sorted according to the 2017 ranking score to show the corresponding ranking score for each value. The documents on the top of the list in 2017 have much higher ranking score in comparison to 2016, but the items in the tail have almost the same ranking score in both years.

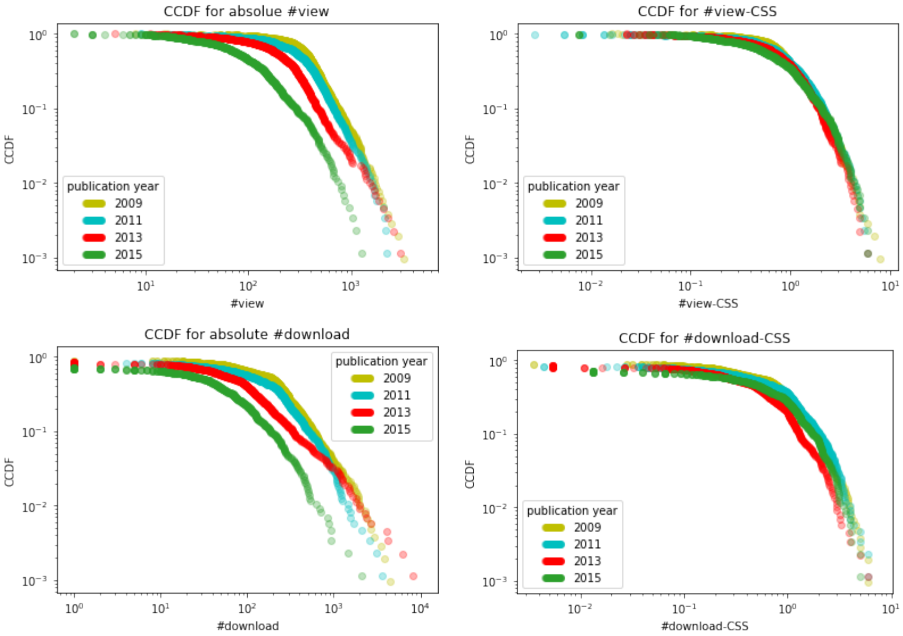

Fig. 2: Plot of raw view and download (left) and individual CSS normalization (right) rates for three different time slices in the usage data of SSOAR.

The reason for this diversity of ranking scores in two years may lie in new documents in the candidate list and also in the sharp growth of the number of views and download for some documents in TREC 2017. The impact on the resulting ranking is huge as a Kendall's tau value of −0.239 was observed for these two rankings. This was an observation we already had during last year's TREC: The popularity values introduce a high fluctuation of the resulting rankings.

Acknowledgement

We would like to thank the whole SSOAR team that supported us to implement the LL4IR API into the productive system and opening up the system for academic research.

References

1. Balog, K., Schuth, A., Tavakolpoursaleh, N., Schaer, P., Chuang, P.Y., Wu, J., Giles, C.L.: Overview of the trec 2016 open search track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016. vol. Special Publication 500-321. National Institute of Standards and Technology (NIST) (2016), http://trec.nist.gov/pubs/trec25/trec2016.html

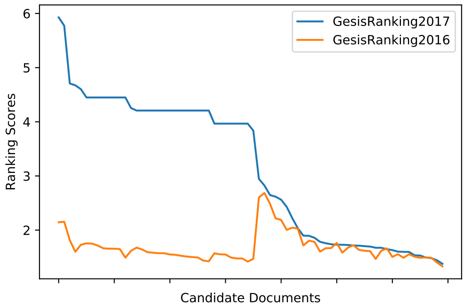

Fig. 3: Comparison plot of the experimental ranking scores for a sample query in 2016 and 2017.

2. Glänzel, W.: Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. Journal of Informetrics 1(1), 92–102 (Jan 2007)
3. Jagerman, R., de Rijke, M., Krisztian Balog, P.S., Schaible, J., Tavakolpoursaleh, N.: Overview of trec opensearch 2017. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017. National Institute of Standards and Technology (NIST) (2017)
4. Plassmeier, K., Borst, T., Behnert, C., Lewandowski, D.: Evaluating popularity data for relevance ranking in library information systems. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community. p. 125. American Society for Information Science (2015), `http://dl.acm.org/citation.cfm?id=2857195`
5. Schaer, P., Tavakolpoursaleh, N.: Historical clicks for product search: Gesis at clef ll4ir 2015. In: Cappellato, L., Ferro, N., Jones, G.J.F., SanJuan, E. (eds.) Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. CEUR Workshop Proceedings, vol. 1391. CEUR-WS.org (2015), `http://ceur-ws.org/Vol-1391/26-CR.pdf`
6. Schaer, P., Tavakolpoursaleh, N.: Ideas for a standard ll4ir extension - living labs from a system operator's perspective. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. CEUR Workshop Proceedings, vol. 1609, pp. 591–592. CEUR-WS.org (2016), `http://ceur-ws.org/Vol-1609/16090591.pdf`
7. Schaer, P., Tavakolpoursaleh, N.: Popularity ranking for scientific literature using the characteristic scores and scale method. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016. vol. Special Publication 500-321. National Institute of Standards and Technology (NIST) (2016), `http://trec.nist.gov/pubs/trec25/papers/THKoeln-GESIS-O.pdf`
8. Tavakolpoursaleh, N.: A Living Lab Evaluation Envirtonment for Academic Document Repositories. Master's thesis, University of Bonn, Germany (2016)