Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Engaging Content
Engaging People

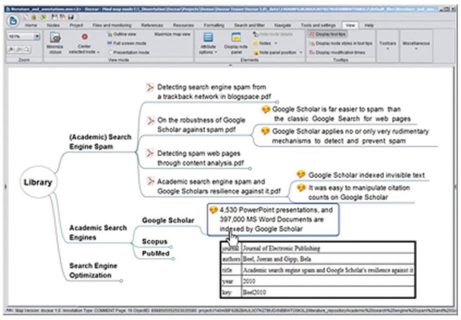
5th International Workshop on Bibliometric-enhanced Information Retrieval (BIR2017)

Apache Lucene as Content-Based-Filtering Recommender System: 3 Lessons Learned

Joeran Beel and Stefan Langer

Apache Lucene provides relevance scores for each recommendation. This information could be used, theoretically, to recommend only documents with a relevance score above a certain threshold. However, on the Web it is often reported that these scores cannot be used to compare relevancies of recommendations between different queries, or to conclude from the relevance score how relevant the search result or recommendation is overall. Our data shows a slightly different picture.

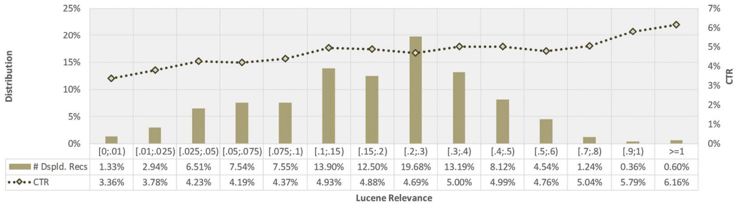
All results are based on data that we collected between May 2013, and October 2014 with Docear. Docear's recommender system delivered 418,308 recommendations to 4,674 unique users. We use click-through rate as measure for the effectiveness of delivered recommendations. Click-through rate (CTR) describes the ratio of clicked and delivered recommendations. All reported differences are statistically significant ($p < 0.05$) based on a two-tailed t-test.


Library
(Academic) Search Engine Spam
Detecting search engine spam from a trackback network in blogspace.pdf
On the robustness of Google Scholar against spam.pdf
Google Scholar is far easier to spam than the classic Google Search for web pages
Google Scholar applies no or only very rudimentary mechanisms to detect and prevent spam
Detecting spam web pages through content analysis.pdf
Academic search engine spam and Google Scholars resilience against it.pdf
Google Scholar indexed invisible text
It was easy to manipulate citation counts on Google Scholar
Academic Search Engines
Google Scholar
4,530 PowerPoint presentations, and 397,000 MS Word Documents are indexed by Google Scholar
Scopus
PubMed
Search Engine Optimization
Home
Nodes
Project
Files and monitoring
References
Resources
Formatting
Search and filter
Navigate
Tools and settings
View
Help

journal	Journal of Electronic Publishing
authors	Beel, Joeran and Gipp, Bela
title	Academic search engine spam and Google Scholar's resilience against it
year	2010
key	Beel2010

1. Lucene's relevance score allows to predict how relevant a recommendation will be for a user

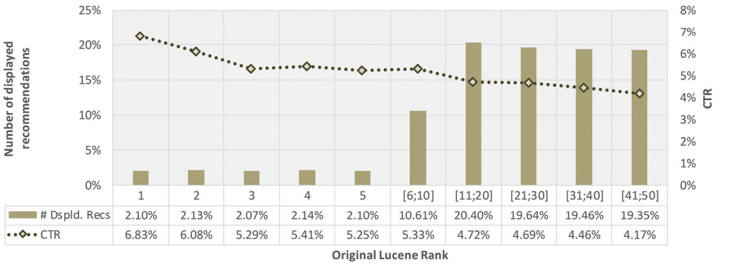
There is a notable trend: Average CTR increases, the higher Lucene relevance scores become. CTR was lowest (3.36%) for recommendations with a relevance score below 0.01, and highest (6.16%) for relevance scores of 1 and above. For recommendations with relevance scores between 0.1 and 0.8, CTR remained mostly stable around 5%. However, recommending only documents with a relevance score of 1 and above is not sensible as only a small fraction of recommendations had a relevance score of 1 and above (0.60%). Similarly, only a small fraction of recommendations had relevance scores below 0.1, so not recommending them will barely affect the overall click-through rate.



	[0;.01)	[.01;.025)	[.025;.05)	[.05;.075)	[.075;.1)	[.1;.15)	[.15;.2)	[.2;.3)	[.3;.4)	[.4;.5)	[.5;.6)	[.7;.8)	[.9;1)	>=1
# Dspld. Recs	1.33%	2.94%	6.51%	7.54%	7.55%	13.90%	12.50%	19.68%	13.19%	8.12%	4.54%	1.24%	0.36%	0.60%
CTR	3.36%	3.78%	4.23%	4.19%	4.37%	4.93%	4.88%	4.69%	5.00%	4.99%	4.76%	5.04%	5.79%	6.16%

2. It might make sense to recommend randomly 10 out of top 50 results

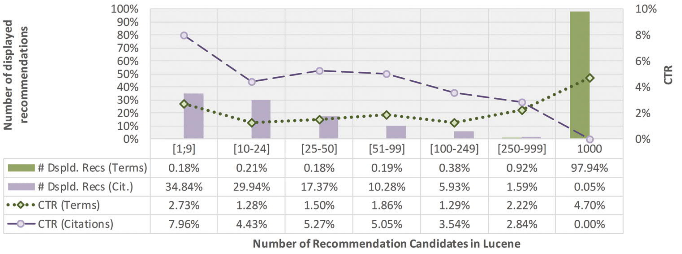
To increase diversity of recommendations, Docear's recommender system randomly chose 10 recommendations out of the top50 results returned by Lucene. This leads to lower click-through rates. Recommendations originally in Lucene's top10 results, achieved CTRs of 5.55% on average, while the top50 achieved CTRs of 4.73% on average. This means, selecting randomly 10 recommendations from the top50 candidates decreases recommendation effectiveness by around 15%, compared to showing recommendations from the top10 only.



	1	2	3	4	5	[6;10]	[11;20]	[21;30]	[31;40]	[41;50]
# Dspld. Recs	2.10%	2.13%	2.07%	2.14%	2.10%	10.61%	20.40%	19.64%	19.46%	19.35%
CTR	6.83%	6.08%	5.29%	5.41%	5.25%	5.33%	4.72%	4.69%	4.46%	4.17%

3. The number of recommendation candidates predicts the relevance of the recommendations

By default, Lucene returns 1,000 recommendations. In our system, Lucene returned 1,000 results for 91.25% of all term-based recommendations. The more recommendation candidates are available, the higher the CTR tends to be. Consequently the number of results might be a good approximation of recommendation effectiveness. If less than 1,000 results are returned it might make sense to not recommend the documents or try an alternative recommendation approach.



	[1;9]	[10-24]	[25-50]	[51-99]	[100-249]	[250-999]	1000
# Dspld. Recs (Terms)	0.18%	0.21%	0.18%	0.19%	0.38%	0.92%	97.94%
# Dspld. Recs (Cit.)	34.84%	29.94%	17.37%	10.28%	5.93%	1.59%	0.05%
CTR (Terms)	2.73%	1.28%	1.50%	1.86%	1.29%	2.22%	4.70%
CTR (Citations)	7.96%	4.43%	5.27%	5.05%	3.54%	2.84%	0.00%

This work was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD). In addition, this publication has em-anated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 13/RC/2106.