A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems

Joeran Beel[1] and Stefan Langer[1]

[1] Otto-von-Guericke University, Dept. of Computer Science, Magdeburg, Germany
{beel|langer}@ovgu.de

Abstract. The evaluation of recommender systems is key to the successful application of recommender systems in practice. However, recommender-systems evaluation has received too little attention in the recommender-system community, in particular in the community of *research-paper* recommender systems. In this paper, we examine and discuss the appropriateness of different evaluation methods, i.e. offline evaluations, online evaluations, and user studies, in the context of research-paper recommender systems. We implemented different content-based filtering approaches in the research-paper recommender system of *Docear*. The approaches differed by the features to utilize (terms or citations), by user model size, whether stop-words were removed, and several other factors. The evaluations show that results from offline evaluations sometimes contradict results from online evaluations and user studies. We discuss potential reasons for the non-predictive power of offline evaluations, and discuss whether results of offline evaluations might have some inherent value. In the latter case, results of offline evaluations were worth to be published, even if they contradict results of user studies and online evaluations. However, although offline evaluations theoretically might have some inherent value, we conclude that in practice, offline evaluations are probably not suitable to evaluate recommender systems, particularly in the domain of research paper recommendations. We further analyze and discuss the appropriateness of several online evaluation metrics such as click-through rate, link-through rate, and cite-through rate.

Keywords: recommender systems, evaluation, offline evaluation, online evaluation, user study

1 Introduction

Thorough evaluations are paramount to assess the effectiveness of research-paper recommender systems, and judge the value of recommendation approaches to be applied in practice or as baseline in other evaluations. The most common evaluation methods are user studies, offline evaluations, and online evaluations [1].

User studies typically measure user satisfaction through explicit ratings. Users receive recommendations generated by different recommendation approaches, rate the recommendations, and the community considers the approach with the highest average rating most effective [1]. Study participants are typically asked to quantify their overall satisfaction with the recommendations. However, they might also be asked to rate individual aspects of a recommender system, for instance, how novel or authoritative the recommendations are [2] or how suitable they are for non-experts [3]. A user study can also collect qualitative feedback, but this is rarely done in the field of (research paper) recommender systems. Therefore, we will ignore qualitative studies in this paper. We distinguish further between "lab" and "real-world" user studies. In lab studies, participants are aware that they are part of a user study, which, as well several other factors, might affect their behavior and

thereby the evaluation's results [4], [5]. In real-world studies, participants are not aware of the study and rate recommendations for their own benefit, for instance because the recommender system improves recommendations based on the ratings (i.e. relevance feedback), or ratings are required to generate recommendations (i.e. collaborative filtering).

Online evaluations originated from online advertisement and measure acceptance rates of recommendations in real-world recommender systems. Acceptance rates are typically measured by click-through rate (CTR), i.e. the ratio of clicked recommendations to displayed recommendations. For instance, if a recommender system displays 10,000 recommendations and 120 are clicked, CTR is 1.2%. Other metrics include the ratio of downloaded or bought items. Acceptance rate is typically interpreted as an implicit measure for user satisfaction. The assumption is that when a user clicks, downloads, or buys a recommended item, the user liked the recommendation. Of course, this assumption is not always reliable because users, for example, might buy a book but after reading it yet rate it negatively. If the recommender system's objective is revenue, metrics such as CTR can be explicit measures of effectiveness, namely when the operator receives money, e.g. for clicks on recommendations.

Offline evaluations typically measure the *accuracy* of a recommender system based on a ground-truth, but also *novelty* or *serendipity* of recommendations can be measured [6]. Offline evaluations were originally meant to identify a number of promising recommendation approaches [1]. These approaches should then be evaluated in detail with a user study or online evaluation to identify the most effective approaches. However, criticism has been raised on the assumption that offline evaluation could predict an algorithm's effectiveness in online evaluations or user studies. More precisely, several researchers have shown that results from offline evaluations do not necessarily correlate with results from user studies or online evaluations [7], [8]. This means that approaches that are effective in offline evaluations are not necessarily effective in real-world recommender systems. As a consequence, *McNee et al.* criticised that "the research community's dependence on offline experiments [has] created a disconnect between algorithms that score well on accuracy metrics and algorithms that users will find useful"[9]. Several more researchers voiced criticism of offline evaluations. *Jannach et al.* stated that "the results of offline [evaluations] may remain inconclusive or even misleading" and "real-world evaluations and, to some extent, lab studies represent probably the best methods to evaluate systems" [10]. *Knijnenburg et al.* reported that "the presumed link between algorithm accuracy […] and user experience […] is all but evident" [11]. Others believe that "online evaluation is the only technique able to measure the true user satisfaction" [12]. The main reason for the criticism in the literature is that offline evaluations ignore human factors, yet human factors strongly affect overall user satisfaction with recommendations. For instance, users may be dissatisfied with recommender systems, if they must wait for too long to receive recommendations [13], or the presentation is unappealing [1].

Despite the criticism, offline evaluations are the predominant evaluation method in the recommender community [14]. This is also true in the field of research-paper recommender systems, where the majority of recommendation approaches are evaluated offline, and only 34% of the approaches are evaluated with user studies and only 7% with online evaluations [15], [16]. However, online evaluations and user studies are also not without criticism. For instance, results of user studies may vary, depending on the questions [17]. *Zheng et al.* showed that CTR and relevance do not always correlate and concluded that "CTR may not be the optimal metric for online evaluation of recommender systems" and "CTR should be used with precaution" [18]. In addition, both user studies and online evaluations require significantly more time than offline evaluations, and can only be conducted by researchers who have access to a recommender system and real users, or at least some participants (e.g. students) to participate in a user study.

2 Research Objective and Methodology

In the field of research-paper recommender systems, there is no research or discussion about how to evaluate recommender systems. In addition, the existing comparisons in other recommender disciplines focus on offline evaluations and user studies [19], [20], *or* offline evaluations and online evaluations [18], but not on all three methods. Our research goal hence was to explore the adequacy of online evaluations, user studies, *and* offline evaluations. To the best of our knowledge, we are first to compare the results of all three evaluation methods, and to discuss the adequacy of the methods and metrics in detail. In addition, we are first to discuss the appropriateness of the evaluation methods in the context of research-paper recommender systems, aside from our previous paper on recommender system evaluation [21]. Compared to our previous paper, the current paper is more comprehensive, covers three instead of two evaluation methods, considers more metrics, is based on more data, and provides a deeper discussion.

To achieve the research objective, we implemented different recommendation approaches and variations in the recommender system of our literature management software Docear [22–25]. We evaluated the effectiveness of the approaches and variations with an offline evaluation, online evaluation, and user study, and compared the results.

Docear is a free and open-source literature suite, used to organize references and PDFs [24]. It has approximately 20,000 registered users and uses mind-maps to manage PDFs and references. Since 2012, Docear has been offering a recommender system for 1.8 million publically available research papers on the web [23]. Recommendations are displayed as a list of ten research papers, showing the title of the recommended papers. Clicking a recommendation opens the paper in the user's web browser. Figure 1 shows an example mind-map that shows how to manage PDFs and references in Docear. We created categories reflecting our research interests ("Academic Search Engines"), subcategories ("Google Scholar"), and sorted PDFs by category and subcategory. Docear imported annotations (comments, highlighted text, and bookmarks) made in the PDFs, and clicking a PDF icon opens the linked PDF file. Docear also extracts metadata from PDF files (e.g. title and journal name) [26], [27], and displays metadata when the mouse hovers over a PDF icon. A circle at the end of a node indicates that the node has child nodes, which are hidden ("folded"). Clicking the circle would unfold the node, i.e. make its child nodes visible again.

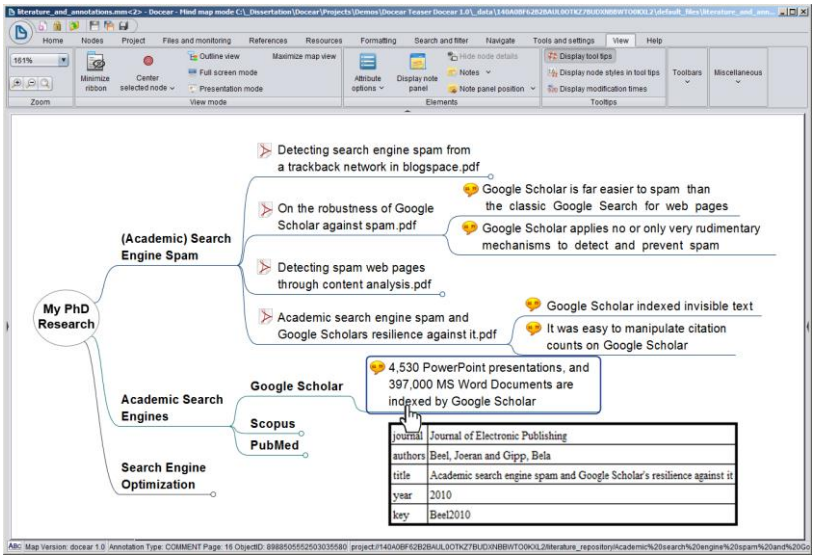

Figure 1: A screenshot of Docear, showing the management of research articles and references

Docear users, who agree to receive recommendations, automatically receive recommendations every five days upon starting the program. In addition, users can request recommendations at any time. To create recommendations, the recommender system randomly chooses one of three recommendation approaches [22], [23]. The first approach is classic content-based filtering, which utilizes terms from the users' mind-maps [28–30]. The most frequently occurring terms are extracted from the nodes, and research papers that contain these terms are recommended. For instance, if the mind-map in Figure 1 was used, terms such as *Google*, *Scholar*, *Academic*, and *Search* would be used for the user modeling because these terms occur frequently in the mind-map. Terms are weighted with TF-IDF, stop words are removed, and recommendations are generated based on cosine similarity in the vector space. The second approach utilizes citations in the same way that the first approach utilizes terms. In the example mind-map in Figure 1, the four PDF links and annotations would be interpreted as citations. The citations would also be weighted with TF-IDF and documents from the corpus that contain the same citations would be recommended. Both content-based filtering approaches are automatically assembled by a number of random variables. Details about this process are provided later in the paper. The third approach implements the stereotype concept, introduced by Rich in 1979 [31]. Based on this approach, Docear generalizes that all users are researchers, which is not strictly true since some use Docear only for its mind-mapping functionality. However, the very nature of stereotyping is to generalize, and the majority of Docear's users are researchers. To give recommendations based on the stereotype approach, we manually compiled a list of research articles that we assumed were relevant for researchers, namely articles and books about academic writing. If the stereotype approach is randomly chosen, the pre-compiled list of articles is recommended. We mainly chose the stereotype approach as a baseline and to have an approach that is fundamentally different from content based filtering. For more details on the architecture of the recommender system refer to [23].

For the offline evaluation, we considered papers that users cited in their mind-maps to be the ground-truth. For each Docear user, we created a copy of their mind-maps, and removed the paper that was most recently added to the mind-map. We then applied a randomly selected recommendation approach to the modified mind-map. Overall, we calculated 118,291 recommendation sets. To measure the accuracy of the algorithm, we analyzed whether the removed paper was within the top10 (**P@10**) or top3 (**P@3**) of the recommendation candidates. We also calculated the Mean Reciprocal Rank (**MRR**), i.e. the inverse of the rank at which the removed paper was recommended. In addition, we calculated normalized discounted cumulated gain (**nDCG**) based on the ten most recently added papers and 50 recommendation candidates. Our evaluation method is similar to other offline evaluations in the field of research-paper recommender systems, where the citations made in research papers are used as ground-truth.

We intended to conduct two user studies – one lab study and one real-world study. For the lab study, we wanted to recruit participants through our blog[1]. In our blog we asked Docear's users to start Docear, request recommendations, click each of them, and read at least the abstract of each recommended paper. Users should then rate the relevance of the recommendations from 1 to 5 stars, and if they wish, request new recommendations and continue this process for as long as they like. The study was intended to run from April to July 2014. We promoted the study in our newsletter (8,676 recipients), on Facebook (828 followers), con Twitter (551 followers), and on Docear's homepage (around 10,000 visitors per month). Despite 248 people reading the blog post, only a single user participated in the study. He rated three recommendation sets, each with ten recommendations. Ratings of a single user are not suitable to receive meaningful results. Hence, we consider this user study as a failure, and focus on results of the real-world study. The real-world study was based on ratings that users provided during their normal work with Docear. The best possible rating

[1] http://www.docear.org/2014/04/10/wanted-participants-for-a-user-study-about-docears-recommender-system/

was 5, the worst possible rating 1. Overall, 379 users rated 903 recommendation sets with 8,010 recommendations. The average rating was 2.82.

For the online evaluation, we analyzed data from Docear's recommender system, which displayed 45,208 recommendation sets with 430,893 recommendations to 4,700 users from March 2013 to August 2014. The acceptance rate was measured with the following metrics: Click-Through Rate (**CTR**) measured the ratio of clicked recommendations vs. delivered recommendations. Click-Through Rate over sets (**CTR_{Set}**) is the mean of the sets' individual CTRs. For instance, if eight out of ten recommendations had been clicked in *set I*, and two out of five recommendations in *set II*, then CTR would be $\frac{8+2}{10+5} = 66.67\%$ but CTR_{Set} would be $\frac{{}^{8}/_{10}+{}^{2}/_{5}}{2} = 60\%$. We also calculated CTR over users (**CTR_{User}**). CTR_{User} levels the effect that a few power users might have. For instance, if users A, B, and C saw 100, 200, and 1,000 recommendations, and user A clicked seven, user B 16, and user C 300 recommendations, CTR would be $\frac{7+16+300}{100+200+1000} = 24.85\%$, but CTR_{User} would be $\frac{7}{100} + \frac{16}{200} + \frac{300}{1000} = 12.36\%$, i.e. the impact of user C would be weaker. However, CTR_{User} was only calculated for two analyses (the reason is further discussed below). Link-Through Rate (**LTR**) describes the ratio of displayed recommendations against those recommendations that actually had been clicked, downloaded and linked in the user's mind-map. Annotate-Through Rate (**ATR**) describes the ratio of recommendations that were annotated, i.e. a user opened the linked PDF in a PDF viewer, created at least one annotation (bookmark, comment, or highlighted text), and imported that annotation in Docear. Cite-Through Rate (**CiTR**) describes the ratio of documents for which the user added some bibliographic data in the mind-map, which strongly indicates that the user plans to cite that document in a future research paper, assignment, or other piece of academic work.

If not otherwise stated, all reported differences are statistically significant ($p<0.05$). Significance was calculated with a two-tailed *t*-test and χ^2 test where appropriate.

3 Results

We calculated the Pearson correlation coefficient for the different evaluation metrics. Both CTR and CTR_{Set} show a strong positive correlation with ratings (r=0.78). Correlation of all other metrics, both offline and online, with user ratings is between 0.52 (CiTR) and 0.67 (nDCG). This means that CTR and CTR_{Set} are the most adequate metrics to approximate ratings. If the goal is to approximate CTR, then ratings, obviously, is the most adequate metric (r=0.78), followed by LTR (r=0.73). The other metrics have rather low correlation coefficients; the worst are nDCG (r=0.28) and MRR (r=0.30). Among the offline metrics, P@3 and P@10 correlate well with each other (r=0.92), which is to expect. MRR and nDCG also show a reasonable strong correlation with each other (r=0.71), while correlation of P@10 and MRR (r=0.56) and P@10 and nDCG (r=0.55) is rather weak.

	Approach			User Model Size						Number of Utilized Nodes						Node Selection Method				Stop Words		User Type		Labelling			Trigger Type	
	CBF (Terms)	CBF (Cit.)	Stereo-type	[1;10]	[11; 25]	[26; 100]	[101; 250]	[251; 500]	[501; 1,000]	[1;9]	[10; 49]	[50; 99]	[100; 499]	[500; 999]	1,000+	Modi-fied	Edit-ed	Creat-ed	Mov-ed	Rem-oved	Kept	Regis-tered	Anon.	Org.	Com.	No	Requ-ested	Auto
Ratings	**2.91**	2.48	2.10	2.62	2.94	**3.26**	2.92	3.00	2.94	2.56	3.17	**3.46**	3.16	3.07	2.58	2.85	2.82	3.04	**3.31**	**3.16**	2.88	**2.90**	--	2.82	**2.92**	2.88	2.83	**2.93**
CTR	**6.53%**	5.25%	4.11%	3.92%	6.27%	**7.48%**	5.40%	6.09%	4.84%	3.64%	6.62%	**7.50%**	7.08%	6.09%	6.38%	4.99%	5.38%	5.09%	**7.40%**	**6.31%**	5.94%	**5.32%**	3.86%	4.82%	4.92%	**5.39%**	**9.14%**	3.67%
CTR(Set)	**5.33%**	5.00%	4.04%	3.78%	6.27%	**7.81%**	5.60%	6.32%	4.85%	3.63%	6.62%	**7.49%**	7.18%	6.20%	6.60%	4.98%	5.57%	5.18%	**7.46%**	**6.35%**	6.01%	**5.38%**	3.83%	5.21%	5.33%	**6.46%**	**9.23%**	3.71%
CTR(Usr)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	**4.00%**	3.77%	**3.68%**	3.18%	3.57%	--	--
LTR	**2.32%**	1.89%	1.46%	1.33%	2.18%	**2.81%**	1.86%	2.03%	1.60%	1.24%	1.23%	**2.60%**	2.42%	2.39%	1.87%	2.24%	1.96%	2.33%	**2.57%**	**2.51%**	2.33%	**2.47%**	2.30%	1.75%	1.76%	**2.21%**	**3.14%**	1.33%
ATR	**1.20%**	1.15%	0.18%	0.61%	0.59%	**1.53%**	0.98%	1.18%	1.26%	0.91%	0.88%	1.18%	**1.28%**	1.38%	1.11%	0.95%	**1.38%**	1.20%	1.18%	**1.14%**	1.21%	**1.34%**	**1.34%**	1.00%	**1.23%**	0.97%	**1.32%**	1.24%
CiTR	**0.53%**	0.52%	0.02%	0.49%	0.15%	0.58%	**0.69%**	0.55%	0.24%	0.25%	0.58%	**0.72%**	0.56%	0.45%	0.44%	0.52%	**0.73%**	0.48%	0.37%	**0.58%**	0.50%	0.52%	**0.55%**	**0.59%**	0.38%	0.48%	**0.71%**	0.38%
P@3	**2.20%**	0.19%	0.00%	2.01%	1.97%	2.25%	**3.48%**	2.07%	1.85%	1.57%	1.91%	2.15%	**3.69%**	2.97%	1.85%	**2.68%**	2.40%	1.52%	2.04%	**2.71%**	2.03%	1.54%	**2.71%**	2.27%	2.03%	2.15%	--	--
P@10	**6.21%**	0.41%	0.03%	4.63%	5.82%	6.49%	**8.87%**	4.18%	2.16%	4.44%	5.81%	6.19%	**8.44%**	7.14%	5.27%	**7.08%**	7.07%	5.41%	5.81%	**7.98%**	5.17%	4.24%	**7.50%**	6.31%	6.26%	6.30%	--	--
MRR	**1.71%**	0.31%	0.04%	0.58%	0.62%	2.16%	**4.04%**	3.17%	1.20%	1.61%	1.45%	1.86%	**1.92%**	1.22%	0.45%	1.61%	1.07%	1.94%	**2.00%**	**1.73%**	1.68%	1.69%	**2.30%**	1.83%	1.60%	1.63%	--	--
nDCG	**1.37%**	0.25%	0.03%	0.90%	1.20%	1.49%	1.63%	**2.08%**	1.92%	1.09%	1.16%	1.38%	**1.54%**	1.44%	0.61%	1.17%	1.49%	**1.61%**	1.00%	**1.44%**	1.29%	**1.35%**	1.31%	1.24%	1.38%	1.33%	--	--

Figure 2: Results of the Evaluations

The user study and online evaluation both led to the same ranking of the three recommendation approaches (Figure 2): Term-based CBF performed best, i.e. CTR, CTR_{Set}, DTR, LTR, CiTR, and ratings were highest, citation-based CBF performed second best, and the stereotype approach performed worst, but still had a reasonable effectiveness. On average, LTR was around one third of

CTR. For instance, LTR for the stereotype approach was 1.46% while CTR was 4.11%. This means that one third of the recommendations that had been clicked were actually downloaded and linked in the mind-maps. ATR was around half of LTR for the CBF approaches. This means that users annotated about half of the recommendations that they downloaded. However, for the stereotype approach, ATR was only 0.18%, i.e. $\frac{1}{8}$ of LTR. Similarly, CiTR for the stereotype approach was only $\frac{1}{75}$ of LTR, while CiTR for term- and citation-based CBF was around $\frac{1}{4}$ of LTR. Apparently, stereotype recommendations were rarely annotated or cited, yet users cited every fourth content-based recommendation that they downloaded. The offline evaluation led to the same overall ranking than the online evaluation and user study. However, all four offline metrics attest that term-based CBF has significantly better effectiveness than citation based CBF (around four to ten times as effective), while user study and online evaluation only attest a slightly higher effectiveness. In addition, the effectiveness of the stereotype approach in the offline evaluation is close to zero, while user study and online evaluation show a reasonable effectiveness.

We researched not only the effectiveness of distinct recommendation approaches, but also variables such as the extent of a user's model (user model size). The user model size describes how many terms (or citations) are stored to represent the users' information needs. Whenever recommendations are requested, Docear randomly selected a user model size between 1 and 1000 terms. For term-based CBF, the highest ratings (3.26) were given for recommendations that were based on user models containing 26 to 100 terms (Figure 2). The offline metrics led to slightly different results and showed the highest effectiveness for user models containing 101 to 250 terms.

Docear's mind-maps often contain thousands of nodes. We assumed that analyzing too many nodes might introduce noise into the user models. Therefore, Docear randomly selected how many of the *x* most recently modified nodes, should be utilized for extracting terms. Based on user ratings, analyzing between 50 and 99 nodes is most effective (Figure 2). As more nodes were analyzed, the average ratings decreased. CTR, CTR_{Set}, LTR and CiTR also showed an optimal effectiveness for analyzing 50 to 99 nodes. Based on ATR, the optimal number of nodes is larger, but results were statistically not significant. The offline metrics indicate that analyzing a larger number of nodes might be sensible, namely 100 to 499 nodes.

Another variable we tested was the node modification type (Figure 2). The recommender system chose randomly, whether to utilize only nodes that were newly *created*, nodes that were *moved*, nodes that were *edited*, or nodes with any type of *modification* (created, edited, or moved). Utilizing moved nodes only, resulted in the highest ratings on average (3.31). The online metrics CTR, CTR_{Set}, and LTR as well as the offline metric MRR also have the highest effectiveness when utilizing moved nodes. Results for ATR and CiTR differ, but are statistically not significant. Based on P@N, utilizing all modified nodes is most effective, based on nDCG utilizing newly created nodes is most effective.

When the recommender system removed stop-words, the average rating was 3.16 compared to 2.88 when no stop-words were removed (Figure 2). All other metrics, except ATR, also showed higher effectiveness when stop-words were removed, but, again, results for ATR were statistically insignificant.

Docear's recommender system is open to both registered and unregistered/anonymous users[2], and we were interested whether there would be differences in the two users groups with respect to recommendation effectiveness (see also [16], [32]). CTR and CTR_{Set} show a clear difference between the two user types (Figure 2). Registered users had an average CTR of 5.32% while unregistered users had an average CTR of 3.86%. CTR_{User} is also higher for registered users (4.00%) than for anonymous users (3.77%), but the difference is not that strong. LTR and ATR also show a (slightly) higher effectiveness for registered users. The offline evaluation contradicts the findings of the

[2] Registered users have a user account assigned to their email address. For users who want to receive recommendations, but do not want to register, an anonymous user account is automatically created. These accounts have a unique random ID and are bound to a user's computer.

online evaluation: P@3, P@10, and MRR indicate that recommendations for registered users were about half as effective as for anonymous users, and nDCG showed no statistically significant difference between the user groups.

For each user, Docear randomly determined whether to display an organic label (e.g. "Free Research Papers", a commercial label (e.g. "Research Papers [Sponsored]"), or to display no label at all. For each user a fix label was randomly selected once, i.e. a particular user always saw the same label. The label had no impact on how recommendations were generated. This means, if recommendation effectiveness would differ for a particular label, then only because users would value different labels differently. In the user study, there were no significant differences for the three types of labels in terms of effectiveness: the ratings were around 2.9 on average (Figure 2). Based on CTR, CTR_{Set}, and LTR, displaying no label was most effective. In addition, commercial labels were slightly, but statistical significantly, more effective than organic labels. Based on CTR_{User}, commercial recommendations were least effective, organic labels were most effective, and 'no label' was second most effective. ATR and CiTR led to statistically not significant results, and offline metrics could not be calculated for this kind of analysis.

Two triggers in Docear lead to displaying recommendations. First, Docear displays recommendations automatically every five days when Docear starts. Second, users may explicitly request recommendations at any time. The user ratings for the two triggers are similar (Figure 2). Interestingly, the online evaluation shows a significantly higher effectiveness for requested recommendations than for automatically displayed recommendations. For instance, CTR for requested recommendations is 2.5 times higher than for automatically displayed recommendations (9.14% vs. 3.67%). An offline evaluation was not conducted because it had not been able to calculate any differences based on trigger.

4 Discussion and Conclusions

4.1 Adequacy of Online Evaluation Metrics

We used six metrics in the online evaluation, namely CTR, CTR_{Set}, CTR_{User}, LTR, ATR, and CiTR. Overall, CTR and CTR_{Set} seem to be the most adequate metrics for our scenario. They had the highest correlation with ratings, are easiest to calculate, were more often statistically significant than the other metrics and are commonly used in other research fields such as e-commerce and search engines. CTR also provided the most plausible results for the stereotype recommendations: based on CTR, the stereotype approach was reasonably effective, while the approach was ineffective based on ATR and CiTR. The result based on CTR seems more plausible since the recommendations were about academic writing and most of Docear's users should be interested in improving their writing skills. However, there is little reason for someone who is doing research in a particular research field, to annotate or even cite an article about academic writing even if the article was useful. Hence, ATR and CiTR were low, and judging stereotype recommendations based on ATR or CiTR seems inadequate to us.

However, there are other scenarios in which ATR and CiTR might be more sensible measures than CTR. For instance, imagine two algorithms called "A" and "B". Both are content-based filtering approaches but B also boosts papers published in reputable journals.[3] Most people would probably agree that algorithm B would be preferable to algorithm A. In the online evaluation, users would probably see no difference between the titles of the recommendations created with the two approaches, assuming that authors publishing in reputable journals do not formulate titles that are significantly different from titles in other journals. Consequently, recommendations of the two algorithms would appear to be similarly relevant and received similar CTR. In this example, CTR would be an inadequate measure of effectiveness and ATR and CiTR might be more appropriate.

[3] For this example we ignore the question how reputability is measured

Measuring CTR, while displaying only the title of recommendations, was criticized by some reviewers of our previous publications. The reviewers argued that titles alone would not allow thorough assessment of recommendations and CTR could therefore be misleading. In some scenarios, such as the example above with the two algorithms, one being boosted by journal reputation, this criticism could indeed apply. However, in the scenario of Docear, the results do not indicate that displaying only the title led to any problems or bias in the results. At least for the content-based recommendations, CTR correlated well with metrics such as LTR, i.e. metrics indicating that the users thoroughly investigated the recommendations.

Compared to CTR, CTR_{User} smoothed the effect of variables that strongly affected a few users. For instance, CTR_{User} was highest for organic labels, lowest for commercial labels, and mediocre for no labels – a result that one would probably expect. In contrast, CTR was highest for no label, second highest for commercial recommendations, and lowest for organic recommendations – a result that one would probably no expect. After looking at the data in detail, we found that a few users who received many recommendations (with no label) "spoiled" the results. Hence, if the objective of an evaluation was to measure overall user satisfaction, CTR_{user} was probably preferable to CTR because a few power users will not spoil the results. However, applying CTR_{user} requires more users than applying CTR since CTR_{user} requires that each user receives recommendation based on the same parameters of the variables and not per recommendation set. For instance, to calculate CTR_{user}, each user must always see the same label, user models must always be the same size for a user, or recommendations must always be based on terms *or* citations. In contrast, to calculate CTR, users may receive recommendations based on terms *and* on citations, or user models could differ in size. Consequently, to receive statistically significant results, CTR_{user} requires more users than CTR. At least for Docear, calculating CTR_{user} for variables such as user model size, number of nodes to analyze, features to utilize (terms or citations), and weighting schemes is not feasible since we would need many more users than Docear currently has.

Considering the strong correlation of CTR and ratings, the more plausible result for stereotype recommendations, and the rather low number of users being required, we conclude that CTR is the most appropriate online metric for our scenario. This is not to mean that in other scenarios other metrics might not be more sensible. Given our results and examples, we suggest that a careful justification is needed in online evaluations about which metric was chosen and why.

4.2 Online Evaluations vs. User Studies

Ratings in the user study correlated strongly with CTR (r=0.78). This indicates that explicit user satisfaction (ratings) is a good approximation of the acceptance rate of recommendations (CTR), and vice versa. Only in two cases CTR and ratings contradicted each other, namely for the impact of labels and the trigger. Based on these two analyses, it seems that ratings and CTR may contradict each other when it comes to evaluating human factors. For analyses relating to the recommendation algorithms (user model size, number of nodes to analyze, etc.), CTR and ratings always led to the same conclusions.

We argue that none of the metrics is generally more authoritative than another. Ultimately, the authority of user studies and online evaluations depends on the objective of the evaluator, and operator of the recommender system respectively. If, for instance, the operator receives a commission per click on a recommendation, CTR was to prefer over ratings. If the operator is interested in user satisfaction, ratings were to prefer over CTR. Ideally, both CTR and ratings, should be considered when making a decision about which algorithm to apply in practice or to choose as baseline, since they both have some inherent value. Even if the operator's objective was revenue, and CTR was high, low user satisfaction would not be in the interest of the operator. Otherwise users would probably ignore recommendations in the long run, and also stop clicking them. Similarly, if the objective was user satisfaction, and ratings were high, a low CTR would not be in the interest of the operator:

a low CTR means that many irrelevant recommendations are given, and if these could be filtered, user satisfaction would probably further increase. Therefore, ideally, researchers should evaluate their approaches with both online evaluation and user study. However, if researchers do not have the resources to conduct both types of evaluation, or the analysis clearly focuses on recommendation algorithms with low impact of human factors, we suggest that conducting either a user study or an online evaluation should still be considered "good practice".

4.3 Adequacy and Authority of Offline Evaluations

Our research shows only a mediocre correlation of offline evaluations with user studies and online evaluations. Sometimes, the offline evaluation could predict the effectiveness of an algorithm in the user study or online evaluation quite precisely. For instance, the offline evaluation was capable of predicting whether removing stop-words would increase the effectiveness. Also the optimal user model size and number of nodes to analyze were predicted rather accurately (though not perfectly). However, the offline evaluation remarkably failed to predict the effectiveness of citation-based and stereotype recommendations. If one had trusted the offline evaluation, one had never considered stereotype and citation-based recommendations to be a worthwhile option. The uncertain predictive power of offline evaluations, questions the often proclaimed purpose of offline evaluations, namely to identify a set of promising recommendation approaches for further analysis.

We can only speculate about why offline evaluations sometimes can predict effectiveness in user studies and online evaluations, and sometimes offline evaluations have no predictive power. One possible reason is the impact of human factors. For instance, on first glance we expected that Docear's recommendation approaches create equally relevant recommendations for both anonymous and registered users. However, the offline evaluation showed higher effectiveness for anonymous users than for registered users while we saw the opposite in the online evaluation. Although we find these results surprising, the influence of human factors *might* explain the difference: We would assume that anonymous users are more concerned about privacy than registered users[4]. Users concerned about their privacy, might worry that when they click a recommendation, some unknown, and potentially malicious website, opens. This could be the reason that anonymous users, who tend to be concerned about their privacy, click recommendations not as often as registered users, and CTR is lower on average. Nevertheless, the higher accuracy for anonymous users in the offline evaluation might still be plausible. If anonymous users tended to use Docear more intensively than registered users, the mind-maps of the anonymous users would be more comprehensive and hence more suitable for user modeling and generating recommendations, which would lead to the higher accuracy in offline evaluations. This means that although mind-maps of anonymous users might be more suitable for user modeling, the human factor "privacy concerns" causes the low effectiveness in online evaluations.

If human factors have an impact on recommendation effectiveness, we must question whether one can determine scenarios in which human factors have *no* impact. Only in these scenarios, offline evaluations would be an appropriate tool to approximate the effectiveness of recommendation approaches in online evaluations or user studies. We doubt that researchers will ever be able to reliably predict whether human factors affect the predictive power of offline evaluations. In scenarios like our analysis of registered vs. anonymous users, it is apparent that human factors may play a role, and that offline evaluations might be not appropriate. For some of our other experiments, such as whether to utilize terms or citations, we could see no plausible influence of human factors, yet offline evaluations could not predict the performance in the user study and online evaluation. Therefore, and assuming that results of offline evaluations have no inherent value, we would propose abandoning offline evaluations, as they cannot reliably fulfil their purpose. However, offline evalua-

[4] If users register, they have to reveal private information such as name and email address. If users are concerned about revealing this information, they probably tend to use Docear as anonymous user.

tions, online evaluations, and user studies typically measure different types of effectiveness. One might therefore argue that comparing the results of the three methods is like comparing apples, peaches, and oranges, and that the results of each method have some inherent value. For online evaluations and user studies, such an inherent value doubtlessly exists (see previous section), but does it exist for offline evaluations?

An inherent value would exist if those who compiled the ground-truth, better knew which items were relevant than current users who decide to click, download, or rate an item. This situation is comparable with a teacher-student situation. Teachers know which books their students should read, and although students might not like the books, or had not chosen the books themselves, the books might be the best possible choice to learn about a certain subject. Such a teacher-student situation might apply to offline evaluations.

Ground-truths inferred, for instance, from citations, theoretically could be more authoritative than online evaluations or user studies. For instance, before a researcher decides to cite a document – which would add the document to the ground-truth – the document was ideally carefully inspected and its relevance was judged according to many factors such as the publication venue, the article's citation count, or the soundness of its methodology. These characteristics usually cannot be evaluated in an online evaluation or user study. Thus, one might argue that results based on personal-collection datasets might be more authoritative than results from online evaluations and user studies where users just had a few seconds or minutes at best, to decide whether to download a paper. Assuming that offline evaluations could be more authoritative than user studies and online evaluations, the following question arises: How useful are recommendations that might objectively be most relevant to users when users do not click, read, or buy the recommended item, or when they rate the item negatively? In contrast to teachers telling their students to read a particular book, a recommender system cannot force a user to accept a recommendation. We argue that an algorithm that is not liked by users, or that achieves low CTR, can never be considered useful. Only if two algorithms performed similarly or if both approaches had at least a mediocre performance in an online evaluation or user study, an additional offline evaluation might be used to decide which of the two algorithms is more effective. However, this means that offline evaluations had to be conducted *in addition* to user studies or online evaluations, and not beforehand or as only evaluation method. Consequently, a change in the current practice of recommender systems evaluation was required.

While offline evaluations with ground-truths inferred from e.g. citations look promising on first glance, there is a fundamental problem: ground-truths are supposed to contain *all* items that are relevant for recommendation. To compile such a ground-truth, comprehensive knowledge of the domain is required. It should be apparent that most users do not have comprehensive knowledge of their domain (which is why they need a recommender system). Consequently, ground-truths are incomplete and contain only a fraction of relevant items, and perhaps even irrelevant items. If the ground-truth is inferred from citations, the problem becomes even more apparent. Many conferences and journals have space restrictions that limit the number of citations in a paper. This means that even if authors were aware of all relevant literature – which they are not – they would only cite a limited amount of relevant articles.

Citation bias enforces the imperfection of citation-based ground-truths. Authors cite papers for various reasons and these do not always relate to the paper's relevance to that author [33]. Some researchers prefer citing the most recent papers to show they are "up-to-date" in their field. Other authors tend to cite authoritative papers because they believe this makes their own paper more authoritative or because it is the popular thing to do. In other situations, researchers already know what they wish to write but require a reference to back up their claim. In this case, they tend to cite the first appropriate paper they find that supports the claim, although there may have been more fitting papers to cite. Citations may also indicate a "negative" quality assessment. For instance, in a recent literature review we cited several papers that we considered of little significance and excluded from an in-depth review [16]. These papers certainly would not be good recommendations. This

means that even if authors were aware of all relevant literature, they will not always select the most relevant literature to cite.

When incomplete or even biased datasets are used as ground-truth, recommender systems are evaluated based on how well they can calculate such an imperfect ground-truth. Recommender systems that recommend papers that are not contained in the imperfect dataset, but that might be equally relevant, would receive a poor rating. A recommender system might even recommend papers of higher relevance than those in the offline dataset, but the evaluation would give the algorithm a poor rating. In other words, if the incomplete status quo – that is, a document collection compiled by researchers who are not aware of all literature, who are restricted by space and time constraints, and who typically do biased citing – is used as ground-truth, a recommender system can never perform better than the imperfect status quo.

We consider the imperfection to be a fundamental problem. To us, it seems that the imperfection is also the most plausible reason why the offline metrics could not predict e.g. the effectiveness of citation-based and stereotype recommendations in the online evaluations and user study. As long as one cannot identify the situations in which the imperfection will affect the results, we propose that inferred ground-truths should not be used to evaluate research-paper recommender systems.

4.4 Limitations

We would like to note some limitations of our research. The offline dataset by Docear may not be considered an optimal dataset due to the large number of novice users. A repetition of our analysis on other datasets, with more advanced users, may lead to more favorable results for offline evaluations. Nevertheless, as pointed out, inferred ground-truths may never be perfect and probably always suffer from some bias, space restrictions, etc. We also focused on research-paper recommender systems. Future research should analyze the extent to which the limitations of offline datasets for research-paper recommender systems apply to other domains. This is of particular importance since ground-truths in other domains may fundamentally differ from the ground truths being used for research-paper recommender systems. For instance, we could imagine that datasets with real ratings of movies are more appropriate than some ground-truth that was inferred from e.g. citations. We also believe that the adequacy of CTR, LTR, ATR, and CiTR need more research and discussion. Although we are quite certain that CTR is the most adequate online metric for our scenario, other scenarios might require different metrics.

Please note: We make most of the data that we used for our analysis publicly available [23]. The data should allow replicating our calculations, and performing new analyses beyond the results that we presented in this paper.

References

[1] F. Ricci, L. Rokach, B. Shapira, and K. B. P., *Recommender systems handbook*. Springer, 2011, pp. 1–35.

[2] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, "Enhancing digital libraries with TechLens+," in *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, 2004, pp. 228–236.

[3] O. Küçüktunç, E. Saule, K. Kaya, and Ü. V. Çatalyürek, "Recommendation on Academic Networks using Direction Aware Citation Analysis," *arXiv preprint arXiv:1205.1143*, pp. 1–10, 2012.

[4] G. Gorrell, N. Ford, A. Madden, P. Holdridge, and B. Eaglestone, "Countering method bias in questionnaire-based user studies," *Journal of Documentation*, vol. 67, no. 3, pp. 507–524, 2011.

[5] G. Leroy, *Designing User Studies in Informatics*. Springer, 2011.

[6] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: evaluating recommender systems by coverage and serendipity," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 257–260.

[7] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl, "On the Recommending of Citations for Research Papers," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2002, pp. 116–125.

[8] A. H. Turpin and W. Hersh, “Why batch and user evaluations do not give the same results,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 225–231.

[9] S. M. McNee, N. Kapoor, and J. A. Konstan, “Don’t look stupid: avoiding pitfalls when recommending research papers,” in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 2006, pp. 171–180.

[10] D. Jannach, L. Lerche, F. Gedikli, and G. Bonnin, “What Recommenders Recommend–An Analysis of Accuracy, Popularity, and Sales Diversity Effects,” in *User Modeling, Adaptation, and Personalization*, Springer, 2013, pp. 25–37.

[11] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, “Explaining the user experience of recommender systems,” *User Modeling and User-Adapted Interaction*, vol. 22, no. 4–5, pp. 441–504, 2012.

[12] A. Said, D. Tikk, Y. Shi, M. Larson, K. Stumpf, and P. Cremonesi, “Recommender systems evaluation: A 3d benchmark,” in *ACM RecSys 2012 Workshop on Recommendation Utility Evaluation: Beyond RMSE, Dublin, Ireland*, 2012, pp. 21–23.

[13] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, “Evaluating collaborative filtering recommender systems,” *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[14] D. Jannach, M. Zanker, M. Ge, and M. Gröning, “Recommender Systems in Computer Science and Information Systems – A Landscape of Research,” in *Proceedings of the 13th International Conference, EC-Web*, 2012, pp. 76–87.

[15] J. Beel, B. Gipp, S. Langer, and C. Breitinger, “Research Paper Recommender Systems: A Literature Survey,” *International Journal on Digital Libraries*, pp. 1–34, 2015.

[16] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nürnberger, “Research Paper Recommender System Evaluation: A Quantitative Literature Survey,” in *Proceedings of the Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys) at the ACM Recommender System Conference (RecSys)*, 2013, pp. 15–22.

[17] P. Cremonesi, F. Garzotto, and R. Turrin, “Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 2, pp. 1–11, 2012.

[18] H. Zheng, D. Wang, Q. Zhang, H. Li, and T. Yang, “Do clicks measure recommendation relevancy?: an empirical user study,” in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 249–252.

[19] P. Cremonesi, F. Garzotto, S. Negro, A. V. Papadopoulos, and R. Turrin, “Looking for ‘good’ recommendations: A comparative evaluation of recommender systems,” in *Human-Computer Interaction–INTERACT 2011*, Springer, 2011, pp. 152–168.

[20] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson, “Do batch and user evaluations give the same results?,” in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 17–24.

[21] J. Beel, S. Langer, M. Genzmehr, B. Gipp, and A. Nürnberger, “A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation,” in *Proceedings of the Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys) at the ACM Recommender System Conference (RecSys)*, 2013, pp. 7–14.

[22] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberger, “Introducing Docear’s Research Paper Recommender System,” in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL’13)*, 2013, pp. 459–460.

[23] J. Beel, S. Langer, B. Gipp, and A. Nürnberger, “The Architecture and Datasets of Docear’s Research Paper Recommender System,” *D-Lib Magazine*, vol. 20, no. 11/12, 2014.

[24] J. Beel, B. Gipp, S. Langer, and M. Genzmehr, “Docear: An Academic Literature Suite for Searching, Organizing and Creating Academic Literature,” in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2011, pp. 465–466.

[25] J. Beel, B. Gipp, and C. Mueller, “SciPlore MindMapping’ - A Tool for Creating Mind Maps Combined with PDF and Reference Management,” *D-Lib Magazine*, vol. 15, no. 11, Nov. 2009.

[26] J. Beel, S. Langer, M. Genzmehr, and C. Müller, “Docears PDF Inspector: Title Extraction from PDF files,” in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL’13)*, 2013, pp. 443–444.

[27] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, “Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents,” in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries (JCDL’13)*, 2013, pp. 385–386.

[28] J. Beel, "Towards Effective Research-Paper Recommender Systems and User Modeling based on Mind Maps," *PhD Thesis. Otto-von-Guericke Universität Magdeburg*, 2015.
[29] J. Beel, S. Langer, G. M. Kapitsaki, C. Breitinger, and B. Gipp, "Exploring the Potential of User Modeling based on Mind Maps," in *Proceedings of the 23rd Conference on User Modelling, Adaptation and Personalization (UMAP).*, 2015, vol. 9146, pp. 3–17.
[30] J. Beel, S. Langer, M. Genzmehr, and B. Gipp, "Utilizing Mind-Maps for Information Retrieval and User Modelling," in *Proceedings of the 22nd Conference on User Modelling, Adaption, and Personalization (UMAP)*, 2014, vol. 8538, pp. 301–313.
[31] E. Rich, "User modeling via stereotypes," *Cognitive science*, vol. 3, no. 4, pp. 329–354, 1979.
[32] J. Beel, S. Langer, A. Nürnberger, and M. Genzmehr, "The Impact of Demographics (Age and Gender) and Other User Characteristics on Evaluating Recommender Systems," in *Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013)*, 2013, pp. 400–404.
[33] M. H. MacRoberts and B. MacRoberts, "Problems of Citation Analysis," *Scientometrics*, vol. 36, pp. 435–444, 1996.

Additional Information

Bibliographic Data

Joeran Beel and Stefan Langer. “A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems.” In Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL), edited by Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, 9316:153–168. Lecture Notes in Computer Science, 2015.

Original Article

http://link.springer.com/chapter/10.1007%2F978-3-319-24592-8_12

Public Pre-Print

http://docear.org/papers/A Comparison of Offline Evaluations...

Related Work

http://docear.org/docear/research-activities/

Academic Platforms

Authors

Joeran Beel

Stefan Langer

BibTeX

```
@INPROCEEDINGS{Beel2015x,
 author = {Joeran Beel and Stefan Langer},
 title = {A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems},
 booktitle = {Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL)},
 year = {2015},
 editor = {Kapidakis, Sarantos and Mazurek, Cezary and Werla, Marcin},
 volume = {9316},
 series = {Lecture Notes in Computer Science},
 pages = {153-168},
 doi = {10.1007/978-3-319-24592-8_12},
}
```

RefMan (RIS)

```
TY  - CHAP
T1  - A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems
A1  - Beel, Joeran
A1  - Langer, Stefan
T2  - Research and Advanced Technology for Digital Libraries
SP  - 153
EP  - 168
SN  - 3319245910
Y1  - 2015
PB  - Springer
ER  -
```

EndNote

%0 Book Section
%T A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems
%A Beel, Joeran
%A Langer, Stefan
%B Research and Advanced Technology for Digital Libraries
%P 153-168
%@ 3319245910
%D 2015
%I Springer