Erschienen in: Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation / Alejandro Bellogin (Hrsg.). - New York : ACM, 2013. - S. 7-14. - ISBN 978-1-4503-2465-6

A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation

Joeran Beel
Docear
Magdeburg, Germany
beel@docear.org

Stefan Langer
Docear
Magdeburg, Germany
langer@docear.org

Bela Gipp
University of California
Berkeley, USA
gipp@berkeley.edu

Marcel Genzmehr
Docear
Magdeburg, Germany
genzmehr@docear.org

Andreas Nürnberger
Otto-von-Guericke University
Magdeburg, Germany
andreas.nuernberger@ovgu.de

ABSTRACT
Offline evaluations are the most common evaluation method for research paper recommender systems. However, no thorough discussion on the appropriateness of offline evaluations has taken place, despite some voiced criticism. We conducted a study in which we evaluated various recommendation approaches with both offline and online evaluations. We found that results of offline and online evaluations often contradict each other. We discuss this finding in detail and conclude that offline evaluations may be inappropriate for evaluating research paper recommender systems, in many settings.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering.*

General Terms
Management, Documentation

Keywords
Research paper recommender systems, evaluation, offline evaluation, click-through rate, online evaluation, comparative study

1. INTRODUCTION
In the past 14 years, more than 170 research articles were published and in 2013 alone, an estimated 30 new research articles are expected to appear in this field (Figure 1) [8]. The more recommendation approaches are proposed, the more important their evaluation becomes to determine the best performing approaches and their individual strengths and weaknesses. Determining the 'best' recommender system is not trivial and there are three main evaluation methods, namely user studies, online evaluations, and offline evaluations to measure recommender systems quality [19].

In user studies, users explicitly rate recommendations generated by different algorithms and the algorithm with the highest average rating is considered the best algorithm [19]. User studies typically ask their participants to quantify their overall satisfaction with the recommendations. A user study may also ask of participants to rate a single aspect of a recommender system, for instance, how novel or authoritative the recommended research papers are, or how suitable they are for non-experts [1,2,17]. Alternatively, a user study can collect qualitative feedback, but because this approach is rarely used for recommender system evaluations [8], we will not address it further. It is important to note that user studies measure user satisfaction at the time of recommendation. They do not measure the accuracy of a recommender system because users do not know, at the time of the rating, whether a given recommendation really was the most relevant.

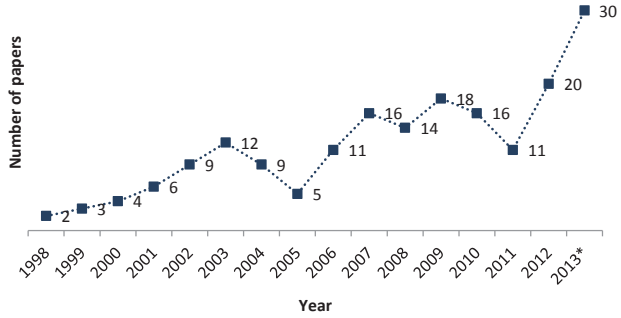


Figure 1: Published papers per year [8]

In online evaluations, recommendations are shown to real users of the system during their session [19]. Users do not rate recommendations but the recommender system observes how often a user accepts a recommendation. Acceptance is most commonly measured by click-through rate (CTR), i.e. the ratio of clicked recommendations[1]. For instance, if a system displays 10,000 recommendations and 120 are clicked, the CTR is 1.2%. To

[1] Aside from clicks, other user behavior can be monitored, for example, the number of times recommendations were downloaded, printed, cited, etc.

compare two algorithms, recommendations are created using each algorithm and the CTR of the algorithms is compared (A/B test). Aside from user studies, online evaluations *implicitly* measure user satisfaction, and can directly be used to estimate revenue if recommender systems apply a pay-per-click scheme.

Offline evaluations use pre-compiled offline datasets from which some information has been removed. Subsequently, the recommender algorithms are analyzed on their ability to recommend the missing information. There are three types of offline datasets, which we define as (1) true-offline-datasets, (2) user-offline-dataset, and (3) expert-offline-datasets.

'*True-offline-datasets*' originated in the field of collaborative filtering where users explicitly rate items (e.g. movies) [18]. True-offline-datasets contain a list of users and their ratings of items. To evaluate a recommender system, some ratings are removed, and the recommender system creates recommendations based on the information remaining. The more of the removed ratings the recommender predicts correctly, the better the algorithm. The assumption behind this method is that if a recommender can accurately predicted some known ratings, it should also reliably predict other, unknown, ratings.

In the field of research paper recommender systems, users typically do not rate research articles. Consequently, there are no true-offline-datasets. To overcome this problem, implicit ratings commonly are inferred from user actions (e.g. citing, downloading, or tagging a paper). For instance, if a user writes a research paper and cites other articles, the citations are interpreted as positive votes of the cited articles [3]. To evaluate a recommender system, the articles a user has cited are removed from his authored paper. Then, recommendations are generated (e.g. based on the text in the authored paper) and the more of the missing citations are recommended the more accurate the recommender is. Instead of papers and citations, any other document collections may be utilized. For instance, if users manage research articles using a reference management software, such as JabRef or Zotero, some (or all) of the users' articles could be removed and recommendations could be created using the remaining information. We call this type of dataset '*user-offline-dataset*' because it is inferred from the users' decision whether to cite, tag, store, etc. an article.

The third type of datasets, which we call '*expert-offline-datasets*', are those created by experts. Examples of such datasets include TREC or the MeSH classification. In these datasets, papers are typically classified by human experts according to information needs. In MeSH, for instance, terms from a controlled vocabulary (representative of the information needs) are assigned to papers. Papers with the same MeSH terms are considered highly similar. For an evaluation, the information need of the user must be determined and the more of the papers satisfying the information need are recommended, the better.

In contrast to user studies and online evaluations, offline evaluations measure the *accuracy* of a recommender system. Offline datasets are considered a ground-truth that represents the ideal set of papers to be recommended. For instance, in the previous example, we assumed that the articles an author cites are those articles to be best recommended. Thus, the fewer of the author-cited articles are predicted by the recommender system, the less accurate it is. To measure accuracy, precision at position *n* (P@n) is typically used to express how many of the relevant articles were recommended within the top *n* results of the recommender. Other common evaluation metrics include recall, F-measure, mean reciprocal rank (MRR) and normalized discounted cumulative gain (NDCG). Only MRR and NDCG take into account the position of recommendations in the generated recommendation list. For a comprehensive overview of offline evaluations including evaluation metrics and potential problems refer to [4,18,19].

Typically, offline evaluations are meant to identify the most promising recommendation approaches [5,6,19]. These most promising approaches should then be evaluated in more detail with a user study or an online evaluation to identify the best approaches. However, we found that most approaches are *only* evaluated with offline evaluations [8], rendering the results one-sided. In addition, some arguments have been voiced that offline-evaluations are not adequate to evaluate recommender systems [6,9,17]. Research indicates that offline evaluations and user studies sometimes contradict each other [7,8,13]. This means, algorithms that performed well in offline evaluations did not always perform well in user studies. This is a serious problem. If offline evaluations could not reliably predict an algorithm's performance and hence cannot fulfill their purpose in a user study or an online evaluation, the question arises what they are good for.

2. RESEARCH OBJECTIVE & METHODOLOGY

Initial comparisons of user studies and offline evaluations, in the field of research paper recommender systems, were mostly not very sound because user studies contained relatively few participants [8]. The studies also did not examine whether the results of offline evaluations and online evaluations correlated. The limited discussion that does exist focused on recommender systems in general [9–13] that were not developed for research paper recommendations in particular.

Therefore, we conducted a comprehensive evaluation of a set of algorithms using (a) an offline evaluation and (b) an online evaluation. Results of the two methods were compared to determine whether and when results of the two evaluation methods contradicted each other. Subsequently, we discuss differences and validity of evaluation methods focusing on *research paper* recommender systems. The goal was to identify which of the evaluation methods were most authoritative, or, if some methods are unsuitable in general. By 'authoritative', we mean which evaluation method one should trust when results of different methods contradict each other.

We performed both evaluations using the literature management software Docear, a desktop software for Windows, Linux, and MacOS, which we developed [14]. Docear manages electronic literature and references in mind maps and offers a recommender system for research papers. Weekly, or upon user request, the recommender system retrieves a set of research papers from Docear's server and recommends them to users. Typically, one set contains ten recommendations[2]. When a user clicks on a recommendation, this is recorded. For information on Docear's

[2] For some recommendation approaches coverage was low. In these cases, less than ten recommendations were made.

recommendation approaches refer to [15]. For information on Docear's users refer to [21].

With the consent of Docear's recommendation service users, Docear has access to:

- the mind maps containing text, citations, and links to papers on the users' hard drives
- the papers downloaded as a PDF
- the annotations created directly in the PDF files, i.e. comments, highlighted text, and bookmarks
- the BibTeX references created for further use in, e.g., Microsoft Word or LaTeX.

Docear's recommender system selects randomly out of several factors to create an algorithm to generate recommendations. Among the factors that are randomly chosen are stop-word removal (on/off); feature type (citations or text); number of mind maps to analyze (only the current mind map or the past *x* created/edited/ maps); feature weighting scheme (TF only; several TF-IDF variations); and many additional factors. This way, one randomly assembled algorithm might utilize terms in the currently edited mind map with the terms being weighted by TF only, while another algorithm utilizes citations made in the last 5 created mind maps, with the citations weighted by TF-IDF (i.e. CCIDF [16]).

For the online evaluation, 57,050 recommendations were delivered to 1,311 users. The primary evaluation metric was click-through rate (**CTR**). We also calculated mean average precision (**MAP**) over the recommendation sets. This means, for each set of recommendations (typically ten), the average precision was calculated. For example, when two of ten recommendations were clicked, average precision was 0.2. Subsequently, the mean was calculated over all recommendation sets of a particular algorithm.

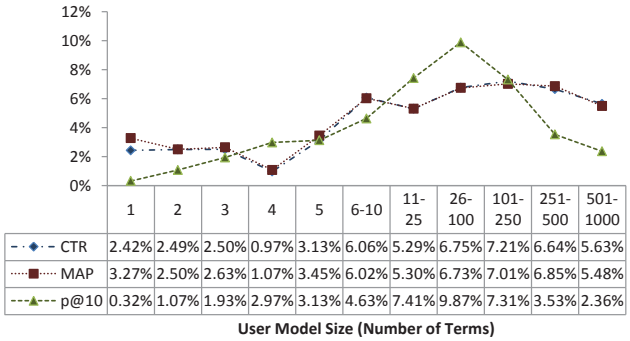
For the offline evaluation, we removed the paper that was last downloaded from a user's document collection together with all mind maps and mind map nodes that were created by the user after downloading the paper. An algorithm was randomly assembled to generate ten recommendations and we measured the recommender's precision, i.e. whether the recommendations contained the removed paper. Performance was measured as precision[3] at rank ten (**P@10**). The evaluation was based on 5,021 mind maps created by 1,491 users.

3. RESULTS

MAP and CTR coincide highly for all evaluated algorithms. Whether terms or citations (Figure 7), the weighting schemes (Figure 4), user model size (Figure 2), or the number of analyzed nodes (Figure 5), both CTR and MAP never contradicted each other[4]. Since MAP is based on CTR this finding is not a surprise, however, to our knowledge this has not been shown empirically before. This finding also implies that there is no need to report both metrics in future papers. While reporting either metric is sufficient, CTR is probably preferable, since CTR makes use of more information than MAP and thus variations in the results should be lower.

In some cases, the offline evaluation had predictive power for an algorithm's performance in Docear. Docear randomly selects the user model size, i.e. the number of terms that represent the user's interests. In the offline evaluation, precision increased with the number of terms a user model contained – up until 26-100 terms (Figure 2). For larger user models, precision decreased. CTR also increased the more terms a user model contained and decreased once the user model became too large; although for CTR, the maximum was achieved for 101-250 terms. Figure 2 clearly shows a high correlation between offline and online evaluation. Even the absolute results are comparable in many cases. For instance, for a user model size of 101-250 terms CTR was 7.21% and P@10 was 7.31%.

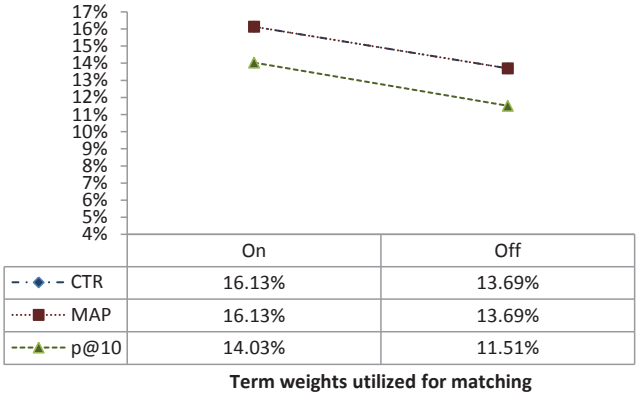
Docear randomly selects whether to keep term weights for the matching process of user models and recommendation candidates. This means, after terms with the highest weight have been stored in the user model, Docear uses the terms either with or without their weight to find papers that contain the same terms. Again, the offline evaluation satisfactorily predicts results of the online evaluation (Figure 3).



	1	2	3	4	5	6-10	11-25	26-100	101-250	251-500	501-1000
CTR	2.42%	2.49%	2.50%	0.97%	3.13%	6.06%	5.29%	6.75%	7.21%	6.64%	5.63%
MAP	3.27%	2.50%	2.63%	1.07%	3.45%	6.02%	5.30%	6.73%	7.01%	6.85%	5.48%
p@10	0.32%	1.07%	1.93%	2.97%	3.13%	4.63%	7.41%	9.87%	7.31%	3.53%	2.36%

User Model Size (Number of Terms)

Figure 2: Impact of user model size (number of terms)



	On	Off
CTR	16.13%	13.69%
MAP	16.13%	13.69%
p@10	14.03%	11.51%

Term weights utilized for matching

Figure 3: Storing term weights

To calculate term weights, the offline evaluation is not as predictive. On the one hand, both offline and online evaluations show that a TF-IDF weighting based on a user's mind maps performs best (Figure 4). On the other hand, the offline evaluation shows a little but significantly better performance for a TF-only

[3] In addition to precision, we also measured mean reciprocal rank (MRR) and normalized discounted cumulative gain (NDCG). However, results did not differ notably from precision. Hence, due to space restrictions, these results are omitted.

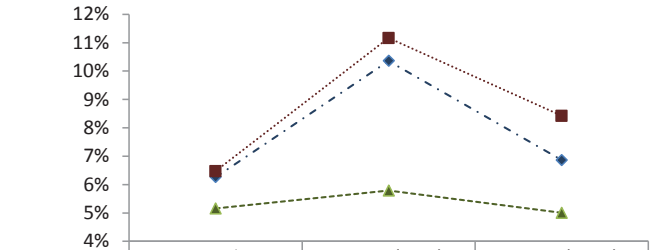
[4] It could still be possible that MAP over users will differ.

weight than for a TF-IDF measure based on the entire corpus. The online evaluation contradicts the offline evaluation, with TF-IDF on the corpus performing better than TF-only.

In many other situations, offline evaluations could not predict an algorithm's performance in practice, i.e. in a real-word system measured with CTR. We will only present some of the results due to space restrictions.

Docear randomly selects how many of the most recently *x* created, edited, or moved nodes in a mind map are utilized (from each of the *x* nodes, the contained terms are used to construct the user model). The offline evaluation predicts that performance is best for analyzing the 50-99 most recently edited, created, or moved nodes (Figure 5). If more nodes were analyzed, precision strongly decreased. However, in practice analyzing the most recent 500-1000 nodes achieved the highest CTR (9.21%) and adding more nodes only slightly decreased CTR.

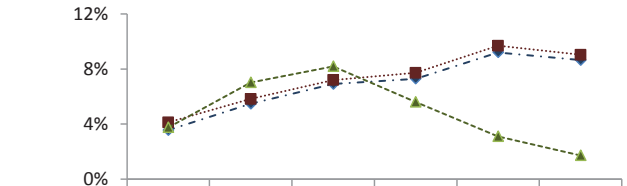
In some cases, offline evaluations contradicted results obtained by the online evaluation. The offline evaluation predicted that analyzing only edited nodes would achieve the best performance with a precision of 5.59% while analyzing moved nodes would only achieve a precision of 4.08% (Figure 6). In practice, results did not coincide. Analyzing moved nodes resulted in the best performance, with a CTR of 10.06% compared to CTR of 6.28% for analyzing edited nodes.



	TF Only	TF-IDF (MM)	TF-IDF (Corp)
CTR	6.26%	10.36%	6.86%
MAP	6.47%	11.17%	8.42%
p@10	5.16%	5.79%	5.00%

Weighting Schemes

Figure 4: Feature weighting

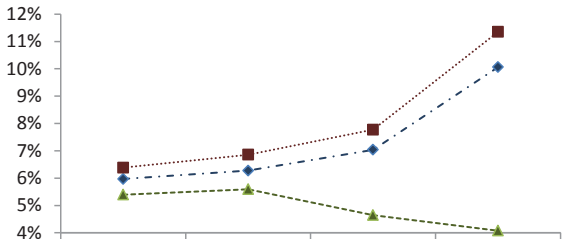


	<=9	10-49	50-99	100-499	500-999	>=1000
CTR	3.56%	5.48%	6.91%	7.29%	9.21%	8.65%
MAP	4.13%	5.83%	7.20%	7.72%	9.70%	9.04%
p@10	3.76%	7.03%	8.20%	5.61%	3.10%	1.72%

# of nodes being analyzed

Figure 5: Number of utilized nodes

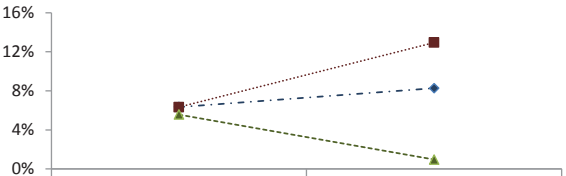
For the fundamental question of whether to utilize terms or citations for generating recommendations, results also differed. For term-based recommendations, Docear extracted the most frequent terms from the user's mind maps and recommended those papers that frequently contain the extracted terms. For citation-based recommendations, Docear extracted the most frequent citations from the users' mind maps, and those research papers that frequently contain the same citations were recommended (comparable to CCIDF [16]). For term-based recommendations, the offline evaluation predicted a precision of 5.57%, which was quite accurate – the actual CTR was 6.34% (Figure 7). For citation-based recommendations, however, the offline evaluation predicted a disappointing result of 0.96%. In practice, the citation-based approach had a CTR of 8.27% and thus even outperformed the text-based approach. If one had relied on offline evaluations, one probably had not considered trying citations in practice because they performed so poorly in the offline evaluation.



	All	Edited	Created	Moved
CTR	5.97%	6.28%	7.04%	10.06%
MAP	6.38%	6.86%	7.77%	11.35%
p@10	5.39%	5.59%	4.65%	4.08%

Node Selection Method

Figure 6: Node selection method



	Terms Only	Citations Only
CTR	6.34%	8.27%
MAP	6.33%	12.95%
p@10	5.57%	0.96%

Utilized Features

Figure 7: Terms vs. citations

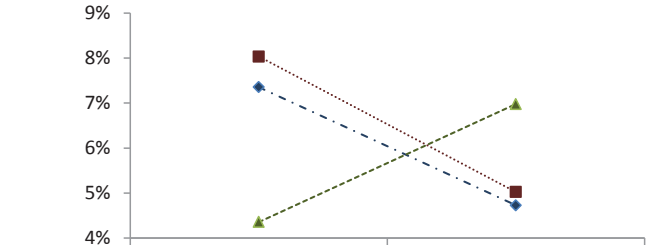
Docear is open to both registered and unregistered users[5]. The offline evaluation predicted that recommendations for anonymous users would achieve higher performance (P@10=6.98%) than for registered users (4.36%). This is interesting in itself. Apparently,

[5] Registered users have a user account tied to their email address. All mind maps created by users who wish to receive recommendations are uploaded to Docear's server, where they are analyzed. 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 all mind maps of these users are uploaded to Docear's server.

there must be significant differences in the mind maps created by anonymous and registered users. However, in practice, registered users achieve significantly higher CTR (7.35%) compared to anonymous users (4.73%) (Figure 8). This again shows that offline evaluations could not predict true system performance.

3.1 Discussion

It is commonly assumed that once the most promising algorithms have been determined via offline evaluations they should be evaluated in detail with user studies or online evaluations. However, our research showed that offline evaluations could *not* reliably predict an algorithm's click-through performance in practice. Instead, offline evaluations were only sometimes able to predict CTR, which leads us to formulate three questions.



	Registered	Anonymous
CTR	7.35%	4.73%
MAP	8.03%	5.02%
p@10	4.36%	6.98%

Figure 8: User types

1. Why do offline evaluations only sometimes accurately predict performance in real-world systems?

We see two possible answers why offline evaluations may not (always) have predictive power.

The first reason, which is also discussed in the literature [17], may be the ignorance of human factors. Offline evaluations can only evaluate the accuracy of recommender systems. However, there are further factors – the human factors – influencing whether users are satisfied with recommender systems and click recommendations or rate them positively. Users may be dissatisfied with accurate recommender systems, if they must wait for too long to receive recommendations [18], the presentation is unappealing [19], labeling of recommendations is suboptimal, or recommendations are given for commercial reasons [20][6]. User satisfaction may also differ by demographics – older users tend to be more satisfied with recommendations than younger users [21].

An influence of human factors seems likely for some of our experiments, especially when comparing registered vs. unregistered users. We would assume that unregistered users are more concerned about privacy and tend to refuse an analysis of their private mind maps. Hence, we would expect unregistered users to have lower acceptance rates of recommendations, i.e. lower CTRs. The analysis of CTR indicates that this assumption is valid. Anonymous users had lower CTRs than registered users. However, the offline evaluation predicted the contrary, namely that *registered* users had lower CTRs. It seems plausible to us that the offline evaluation was wrong because it could not consider the human factors, which might be quite strong in this particular experiment. In other experiments, e.g., when we compared user model sizes, offline evaluations had some predictive power. This may be the case, because the influence of human factors was the same for different user model sizes or not relevant at all.

The second reason why offline-evaluations may not always have predictive power relates to the imperfection of offline-datasets. Offline-datasets represents a ground-truth that contains all and only those papers relevant for recommendations. To compile a valid ground-truth, users would have to be aware of the entire literature in their field. Consequently, one must conclude that user-offline-datasets are incomplete, containing only a fraction of all relevant documents and maybe even some papers of low or no relevance.

In case of user-offline-datasets based on citations, this problem becomes even more apparent. Many papers contain only few references because of space limitations. As such, citations do not make an ideal ground-truth because the dataset will never contain all relevant papers. This means, even if authors were aware of all relevant literature – which they are not – they would only add a limited amount of the relevant articles to their document collection (e.g. by citing them).

When incomplete datasets are used as ground-truth, recommender systems are evaluated based on how well they can calculate an incomplete ground-truth. Recommender systems that were recommending other but equally relevant papers, which happened to not be contained in the incomplete offline dataset, would receive a poor rating. A recommender system might even recommend papers of higher relevance than those in the offline dataset, but the offline evaluation would also 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 and are restricted by space and time constraints – is used as ground-truth, a recommender system can never perform better than the imperfect status quo.

The inherent citation bias further enforces unsuitability of citations for use in offline evaluations. Authors cite papers for various reasons, and these do not always relate to the paper's relevance to that author [22–24]. Some researchers prefer citing the most recent papers to show they are "up-to-date" in their field, even if the cited papers are not the most relevant. Other authors tend to cite authoritative papers because they believe this makes their paper more authoritative, or because it is the popular thing to do. In other situations, researchers already have in mind what they wish to write but require a reference to back up their claim. In this case, they tend to cite the first-best paper they find that supports the claim, although there may have been more fitting papers to cite. This means, even if authors were aware of all relevant literature in their field, they will not always select the most relevant literature to cite.

[6] Identical recommendations, which were labeled once as organic and once as commercial, influenced user satisfaction ratings despite having equal relevance.

This again leads to incomplete and biased document collections, which results in suboptimal evaluations.[7]

The argument of incomplete and biased offline-datasets may explain why offline evaluations only sometimes have predictive power. A correlation between offline and online evaluations would occur when a suboptimal dataset had the same effect on all evaluated algorithms. If the suboptimal dataset had different effects on two algorithms, the offline evaluation would deliver different results than an online evaluation.

2. Is it possible to identify the situations where offline evaluations have predictive power?

If one could identify the situations in which human factors have the same or no impact on two algorithms, offline evaluations could be purposefully applied in these situations. In scenarios like our analysis of registered vs. anonymous users, it is apparent that human factors may play an important role and offline evaluations should not be used. For some of our other experiments, such as whether to utilize terms or citations, we can see no plausible influence of human factors (but still results did not correlate). We doubt that researchers will ever be able to determine reliably in advance whether human factors play such an important role that offline evaluations would not have predictive power. Agreeing with this assumption, and assuming that the sole purpose of offline evaluations was to predict CTR or the relevance ratings of users, the only solution was to abandon offline evaluations entirely.

The same conclusion applies for the argument regarding incomplete datasets. Retrospectively, one may be able to explain why an offline evaluation could not predict the performance in practice, due to incompleteness of the dataset. However, we doubt that there could ever be a way to determine in advance if an offline dataset is incomplete or when suboptimal datasets have the same negative effects on two algorithms. Therefore, if one accepts that offline datasets inferred from users' data are incomplete and maybe even biased, and that one cannot determine to what extent datasets are incomplete and biased, the conclusion can only be to avoid offline evaluations when evaluating research paper recommender systems.

3. Is it problematic, that offline evaluations do not (always) have predictive power?

Theoretically, it could be that results of offline evaluations have some inherent value, and it might make sense to apply an algorithm in practice, or use it as a baseline, if it performed well in an offline evaluation although it received low CTR or user ratings. This scenario requires that users who compile the offline dataset have a better knowledge of document relevance than those users the recommendations are shown to.

In the case of expert-datasets, one might argue that topical experts can better judge research paper quality and relevance than average users and hence know better what is relevant to users than the users themselves. Therefore, evaluations using expert-datasets might have some inherent value and might be more authoritative than results obtained from online evaluations or user studies. For instance, if experts were asked to compile an introductory reading list on recommender systems for bachelor students, they could probably better select the most relevant documents than the bachelor students themselves could. Even if users were not particularly satisfied with the recommendations, and rated them poorly, the recommendations would still have the highest level of topical relevance to users.

However, such an expert-created list for bachelor students may not be suitable for PhD students who wanted to investigate the topic of recommender systems in more depth. Thus, another expert list would be needed for the PhD students; another for senior researchers; another for foreign language students, etc. Overall, there would be an almost infinite number of lists required to cater to all types of user backgrounds and information needs. Such a comprehensive dataset does not exist and probably never will.

When today's expert-datasets are used for evaluations, an evaluation focuses only on one very specific use-case that neglects the variety of uses-cases in real applications. This is particularly true for expert datasets like MeSH, because these controlled vocabularies were not created for evaluating research paper recommender systems. MeSH was created to index and facilitate search for medical publications. Articles with the same MeSH tags are classified as highly similar. If a recommender system is to be evaluated for its suitability to recommend highly similar papers, MeSH could be an appropriate ground-truth. However, users have diverse information needs and identifying topically highly similar papers is only one of them.

Therefore, we conclude that expert-datasets might be a more authoritative evaluation method than online evaluations and user studies in theory, but for real-world applications there will likely never be adequate expert-datasets to cover the variety of user needs. Hence, we suggest that expert-offline-datasets should not be used for offline evaluations of recommender systems, except if it was highly plausible that the expert-dataset fully represents the real-world use-cases of the recommender system.

User-offline-datasets do not suffer the problem of overspecialization and should typically represent a large variety of use-cases. However, the question arises why users who contribute to the offline dataset should know better what is good for the current users than the current users themselves? The answer may be that the users from whom the dataset was inferred, had more time to compile their collections than the users the recommendations are given to. In online evaluations and user studies, researchers must make decisions using information such as title and abstract to decide whether a recommendation is relevant. In contrast, before a researcher decides to cite a document, the document was carefully inspected – in the ideal case – 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. In conclusion, a paper contained in a researcher's document collection, or cited by a researcher, may thus be a stronger indicator of relevance than if the recommendation was solely clicked or rated positively in an online evaluation or user study.

[7] In the case of a citation recommender, one may argue that biased citation-based evaluations are acceptable because others would also like to cite the same papers. However, there is still the problem that a citation list is incomplete and a recommender system is punished in an offline evaluation when it recommends other possibly more relevant papers, which the user may even have preferred to cite.

There is a plausible example in which results from a user-offline-evaluation may be more authoritative than results from an online evaluation or user study. Imagine two algorithms A and B. Both are the same content-based filtering approaches but B additionally boosts papers that were published in reputable journals[8]. In an online study, users would judge the relevance of recommendations using the titles of the recommendations (if a title is interesting, users click the recommendation). We assume that authors publishing in reputable journals do not formulate titles that are significantly different from titles in other journals. Because users would see no difference between the recommendations by looking at the titles, all recommendations would appear as similarly relevant and received similar CTR. In contrast, offline-datasets likely contained more papers from reputable journals because these papers are likely of higher quality than articles published in less known journals. As a result, in an offline evaluation algorithm B would show better performance than algorithm A. Most likely, people would agree that algorithm B indeed should be preferred over algorithm A in a practical setting and hence the offline evaluation would have identified the best algorithm while the online evaluation did not.

Following the argument that offline evaluations are based on more thorough assessments than online evaluations, one might conclude that evaluations using user-offline-datasets indeed might be more authoritative than online evaluations or user studies. However, this was only one example for a very specific scenario, and one important question remains: How useful are recommendations that objectively might be most relevant to users when users do not accept, i.e. click or positively rate, them? The problem is that in contrast to a teacher telling his students to read a particular paper, 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 a very low CTR can never be considered useful. Only if two algorithms performed alike in an online evaluation (or user study) an additional offline evaluation might be used to decide which of the two algorithms should be used in the real-world system, or as baseline for future evaluations. However, in this case, results from offline evaluations were only valuable in combination with results from online evaluations or user studies. In addition, the criticism of ignoring human factors and incomplete datasets still applies. As such, it could not be known for sure if the results of the offline evaluation are truly correct.

4. SUMMARY & OUTLOOK

Our results cast doubt on the meaningfulness of offline evaluations. We showed that offline evaluations could often not predict CTR, and we identified two possible reasons.

The first reason for the lacking predictive power of offline evaluations is the ignorance of human factors. These factors may strongly influence whether users are satisfied with recommendations, regardless of the recommendation's relevance. We argue that it probably will never be possible to determine when and how influential human factors are in practice. Thus, it is impossible to determine when offline evaluations have predictive power and when they do not. Assuming that the only purpose of offline evaluations is to predict results in real-world settings, the plausible consequence is to abandon offline evaluations entirely.

The second reason why (user-) offline-datasets may not always have predictive power is due to their incompleteness. This is attributable to insufficient user knowledge of the literature, or biases arising in the citation behavior of some researchers. Our results led to the conclusion that sometimes incomplete and biased datasets may have the same negative effects on different algorithms. In other situations, they have different effects on different algorithms, which is why offline evaluations could only sometimes predict results of online evaluations. Since we see no way of knowing when negative effects of incomplete datasets would be the same for two algorithms, we concluded that user-offline-datasets are not suitable for predicting the performance of recommender systems in practice.

However, we also argued that offline evaluations might have some inherent value and it may make sense to apply algorithms in real-world systems if they performed well in offline evaluations but poorly in online evaluations or user studies. The underlying assumption is that users who contributed to the offline dataset know better than users receiving recommendations, which papers are relevant for certain information needs. Theoretically, this could be the case for datasets compiled by experts but we argued that expert-datasets are overspecialized and not practically feasible, and thus unsuitable for evaluations of recommender systems. Evaluations based on user-offline-datasets could have some value to determine which algorithms are best if the algorithms performed consistent in online evaluations and user studies. However, this also means that offline evaluations alone are of little value.

Our study represents a first step in the direction of deciding whether and when offline-evaluations should be used. Future research should clarify with more certainty whether offline-evaluations are indeed unsuitable for evaluating research paper recommender systems. We cannot exclude with certainty that we did not miss an important argument, or that there may be a way to determine the situations in which offline evaluations do have predictive power. In addition, the offline dataset by Docear might not be considered an optimal dataset due to the large numbers of novice users. A repetition of our analysis on other datasets may possibly lead to more favorable results for offline evaluations. It might also make sense to repeat our study with more offline-metrics such as recall, or NDCG, and additionally conduct a large-scale user study. It might also be argued that CTR is not an ideal evaluation measure and it should not be considered the goal of an offline-evaluation to predict CTR. CTR only measures how interesting a title appears to a user. Measuring instead how often users actually *cite* the recommended paper may be a more appropriate measure, and offline-evaluations likely correlate more with this measure.

In addition, it should be researched to what extent the limitations of offline datasets for research paper recommender systems apply to other domains and 'true-offline-datasets'. True-offline-datasets are not relevant for research paper recommender systems but for many other recommender systems. They contain ratings of real users and we could imagine that they possibly represent a near-perfect ground truth. Results of true-offline evaluations would not contradict results from online evaluations. Although, there is also doubt on how reliable user rating are [25].

[8] We ignore the question of how reputation is measured

In summary, we require a more thorough investigation of the usefulness of offline evaluations and more sound empirical evidence before we can abandon offline evaluations entirely. Meanwhile, we would suggest treating results of offline evaluations with skepticism.

5. REFERENCES

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

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

[3] 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," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, New Orleans, Louisiana, USA: ACM, 2002, pp. 116–125.

[4] J.A. Konstan and J. Riedl, "Recommender systems: from algorithms to user experience," *User Modeling and User-Adapted Interaction*, 2012, pp. 1–23.

[5] 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, 2012, pp. 441–504.

[6] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," *The Journal of Machine Learning Research*, vol. 10, 2009.

[7] G. Karypis, "Evaluation of item-based top-n recommendation algorithms," *Proceedings of the tenth international conference on Information and knowledge management*, ACM, 2001, pp. 247–254.

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

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

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

[11] D. Jannach, L. Lerche, F. Gedikli, and G. Bonnin, "What Recommenders Recommend–An Analysis of Accuracy, Popularity, and Sales Diversity Effects," *User Modeling, Adaptation, and Personalization*, Springer, 2013.

[12] G. Shani and A. Gunawardana, "Evaluating recommendation systems," *Recommender systems handbook*, Springer, 2011, pp. 257–297.

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

[14] J. Beel, B. Gipp, S. Langer, and M. Genzmehr, "Docear: An Academic Literature Suite for Searching, Organizing and Creating Academic Literature," *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, ACM, 2011, pp. 465–466.

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

[16] K.D. Bollacker, S. Lawrence, and C.L. Giles, "CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications," *Proceedings of the 2nd international conference on Autonomous agents*, ACM, 1998, pp. 116–123.

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

[18] 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, 2004, pp. 5–53.

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

[20] J. Beel, S. Langer, and M. Genzmehr, "Sponsored vs. Organic (Research Paper) Recommendations and the Impact of Labeling," *Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013)*, T. Aalberg, M. Dobreva, C. Papatheodorou, G. Tsakonas, and C. Farrugia, eds., Valletta, Malta: 2013, pp. 395–399.

[21] J. Beel, S. Langer, A. Nürnberger, and M. Genzmehr, "The Impact of Demographics (Age and Gender) and Other User Characteristics on Evaluating Recommender Systems," *Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013)*, T. Aalberg, M. Dobreva, C. Papatheodorou, G. Tsakonas, and C. Farrugia, eds., Valletta, Malta: Springer, 2013, pp. 400–404.

[22] T.A. Brooks, "Private acts and public objects: an investigation of citer motivations," *Journal of the American Society for Information Science*, vol. 36, 1985, pp. 223–229.

[23] M. Liu, "Progress in documentation the complexities of citation practice: a review of citation studies," *Journal of Documentation*, vol. 49, 1993, pp. 370–408.

[24] M.H. MacRoberts and B. MacRoberts, "Problems of Citation Analysis," *Scientometrics*, vol. 36, 1996, pp. 435–444.

[25] X. Amatriain, J. Pujol, and N. Oliver, "I like it... i like it not: Evaluating user ratings noise in recommender systems," *User Modeling, Adaptation, and Personalization*, 2009, pp. 247–258.