Research Paper Recommender System Evaluation: A Quantitative Literature Survey

Joeran Beel
Docear, Magdeburg
Germany
beel@docear.org

Stefan Langer
Docear, Magdeburg
Germany
langer@docear.org

Marcel Genzmehr
Docear, Magdeburg
Germany
genzmehr@docear.org

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

Corinna Breitinger
Univ. of California, Berkeley, USA
breitinger@berkeley.edu

Andreas Nürnberger
OvGU, FIN, ITI
Magdeburg, Germany
andreas.nuernberger@ovgu.de

ABSTRACT

Over 80 approaches for academic literature recommendation exist today. The approaches were introduced and evaluated in more than 170 research articles, as well as patents, presentations and blogs. We reviewed these approaches and found most evaluations to contain major shortcomings. Of the approaches proposed, 21% were not evaluated. Among the evaluated approaches, 19% were not evaluated against a baseline. Of the user studies performed, 60% had 15 or fewer participants or did not report on the number of participants. Information on runtime and coverage was rarely provided. Due to these and several other shortcomings described in this paper, we conclude that it is currently not possible to determine which recommendation approaches for academic literature are the most promising. However, there is little value in the existence of more than 80 approaches if the best performing approaches are unknown.

Categories and Subject Descriptors

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

General Terms

Measurement, Algorithms, Performance, Experimentation

Keywords

Research paper recommender systems, evaluation, comparative study, recommender systems, survey

1. INTRODUCTION

Recommender systems for research papers are becoming increasingly popular. In the past 14 years, over 170 research articles, patents, web pages, etc. were published in this field. Interpolating from the numbers of published articles in this year, we estimate 30 new publications to appear in 2013 (Figure 1). Recommender systems for research articles are useful applications, which for instance help researchers keep track of their research field. The more recommendation approaches are proposed, the more important their evaluation becomes to determine the best approaches and their individual strengths and weaknesses.

Evaluating recommender systems requires a definition of what constitutes a good recommender system, and how this should be measured. There is mostly consensus on what makes a good recommender system and on the methods to evaluate recommender systems [1,11,62]. However, at least in related research fields, authors often do not adhere to evaluation standards. For instance, three quarters of evaluations published in the *User Modeling and User-Adapted Interaction* (UMAI) journal were statistically not significant, and often had serious shortcomings in their evaluations [2]. These results raise the question whether researchers in the field of research paper recommender systems might ignore evaluation standards in the same way as authors of the UMAI journal.

In the remainder of this paper, we describe the main features, which contribute to a 'good', i.e. a high quality, recommender system, and the methods used to evaluate recommender systems. We then present our research objective and methodology, and conclude with the results and a discussion.

1.1 Features of Recommender System Quality

1.1.1 Accuracy

The first factor that contributes to a good recommender is its accuracy, i.e. its capacity to satisfy the individual user's information need [62]. Information needs vary among users due to different background and knowledge [3], preferences and goals [4], and contexts [108]. One user may be interested in the most *recent* research papers on mind mapping, while another may be interested in the *first* publication introducing recommender systems, or the most *popular* medical research on lung cancer, but only in a given *language,* etc. Items that satisfy the information needs are "relevant" to the user [62]. Accordingly, a good recommender system is one that recommends (the most) relevant items. To do so, a recommender system must first identify its users' information needs and then identify the items that satisfy those needs. How well a recommender system performs at this task is reflected by its accuracy: the more relevant, and the less irrelevant items it recommends, the more accurate it is.

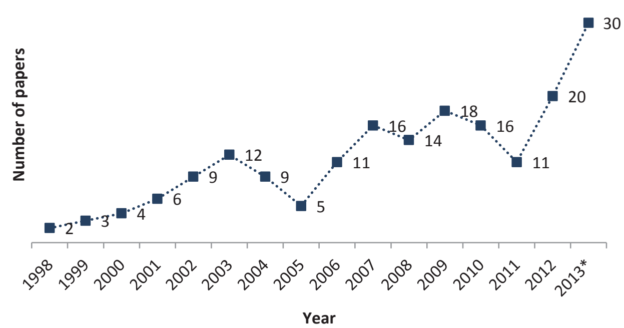

Figure 1: Published papers per year[1]

A prerequisite to achieve high accuracy is high coverage of the available items [5]. Coverage describes how many papers of those in the recommender's database may be recommended with the recommendation approach. For text-based approaches, coverage is usually 100%. For many citation-based approaches, coverage is usually significantly lower, because only a fraction of all documents is cited, and can hence be recommended [58].

[1] Based on the papers we reviewed for this article. Numbers for 2013 were estimated by interpolating from the number of articles published until our survey was conducted (late April 2013).

RepSys '13, October 12 2013, Hong Kong, China. Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2465-6/13/10…$15.00.
http://dx.doi.org/10.1145/2532508.2532512

1.1.2 User Satisfaction

The second factor that contributes to a good recommender system is its ability to provide "satisfaction" to the user [6]. At first glance, one may assume that an accurate recommender system, i.e. one that recommends the most relevant items, satisfies the user. However, many additional factors influence user satisfaction. One of these factors is serendipity [9,60]. If milk was recommended to a customer in a supermarket, this could be a very accurate recommendation, but not a satisfying one [60]. Milk is an obvious product to buy in a supermarket. Therefore, most customers would be more satisfied with more diverse recommendations (that still should be accurate to some extent). Users may also be dissatisfied with accurate recommender systems, if they must wait for too long to receive recommendations [62], the presentation is unappealing [11], labeling of recommendations is suboptimal, or recommendations are given for commercial reasons [7][2]. User satisfaction may also differ by demographics – older users tend to be more satisfied with recommendations than younger users [8]. In addition, costs can play a role. Typically, recommender systems are free but some systems charge users a fee or are only available as part of subscription packages. One example is the reference manager Mendeley, which offers its recommender system *Mendeley Suggest* only to its premium users. The time a user must invest before receiving recommendations may also influence user satisfaction. Some systems expect users to specify their interests manually. In other systems, users' interests are inferred automatically, which significantly reduces the user's required time commitment. The mentioned factors are only a small selection. There are many more factors influencing whether a user is satisfied with a recommender system [9,11].

1.1.3 Satisfaction of the Recommendation Provider

The third factor contributing to a good recommender system is its ability to satisfy the recommendation provider. Typically, it is assumed that providers of recommender systems are satisfied when their users are satisfied, but this is not always the case. One interest of the providers is keeping costs low, where costs may be measured in terms of labor, disk storage, memory, CPU power, and traffic [11]. As such, a good recommender system may also be defined as one that can be developed, operated, and maintained at a low cost. Other providers, e.g. publishers, may have the goal of generating a profit from the recommender system [61]. With this goal, a publisher would prefer to recommend items with higher profit margins even if user satisfaction was not that high. A news-website might have the goal of keeping their readers as long as possible on their website [61]; in which case, a recommender would preferably suggest longer articles even if shorter articles might result in higher user satisfaction.

In most situations, there will be a tradeoff between the three factors. For instance, clustering strongly reduce runtimes, and hence costs, but also decreases accuracy [10]; and when the primary goal is to generate revenue, user satisfaction may suffer. Of course, user satisfaction should never be too low because then users might ignore the recommendations completely.

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

1.2 Evaluating Methods

Knowing the three features contributing to a good recommender system – recommendation accuracy, user satisfaction, and provider satisfaction – leads to the question how these three features are to be quantified and compared. Aspects related to time and money, such as runtime, costs, and revenue, can easily be measured and are thus not covered in detail in the remainder of this paper. To measure a recommender's accuracy and to gauge user satisfaction three evaluation methods are commonly used: user studies, online evaluations, and offline evaluations [11][3].

In user studies, users explicitly rate recommendations generated with different algorithms and the algorithm with the highest average rating is judged the best algorithm [11]. In online evaluations, recommendations are shown to users as they use the real-world system [11]. Users do not rate recommendations; rather, the system observes how often users accept a recommendation. Acceptance is typically measured by click-through rate (CTR), i.e. the ratio of clicked recommendations[4]. To compare two algorithms, recommendations are created using each algorithm and then CTR of the algorithms are compared (A/B test). Offline evaluations use pre-compiled offline datasets from which some information is removed for the evaluation. Subsequently, the recommender algorithms are analyzed on their ability to recommend the removed information.

Which of the three evaluation methods is most suitable is still under debate. Typically, offline evaluations are considered suitable to pre-select a set of promising algorithms, which are subsequently evaluated in online evaluations or by a user study [11]. However, there is serious criticism of offline evaluations [60–65,106,111].

1.3 Further Considerations

Another important factor in evaluating recommender systems is the baseline against which an algorithm is compared. Knowing that a certain algorithm has a CTR of e.g. 8% is not useful if the CTRs of alternative approaches are unknown. Therefore, novel approaches should be compared against a baseline representative of the state-of-the-art. Only then is it possible to quantify whether a novel approach is better than the state-of-the-art and by what margin.

Additionally, a statistically significant number of participants is crucial to user study validity, as well as sufficient information on algorithm complexity and runtime, the use of representative datasets, and several other factors [11]. Only if all these factors are considered, will an evaluation produce valid results that allow identifying the best recommendation approaches. Of course, it is also important that researchers publish all relevant details about their evaluation and their approaches to allow others to verify the validity of the conducted evaluations and to implement the approaches.

[3] We ignore provider's satisfaction in the remainder since this type of satisfaction should usually relate to numbers that are easy to measure, e.g., revenue or costs.

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

2. RESEARCH OBJECTIVE & METHODOLOGY

The research objective we pursued was to examine the validity of evaluations performed for existing research paper recommender systems. In reviewing the literature, we assess how suitable existing evaluations are for identifying the most promising research paper recommender systems.

To achieve this objective, we conducted a quantitative analysis of the status quo. We seek to answer the following questions.

1. To what extent do authors perform user studies, online evaluations, and offline evaluations? (see Section 3.1)
2. How many participants do user studies have? (see Section 3.2)
3. Against which baselines are approaches compared? (Section 3.3)
4. Do authors provide information about algorithm's runtime and computational complexity? (Section 3.4)
5. Which metrics are used for algorithm evaluation, and do different metrics provide similar rankings of the algorithms? (Section 3.5)
6. Which datasets are used for offline evaluations (Section 3.6)
7. Are results comparable among different evaluations based on different datasets? (Section 3.7)
8. How consistent are online and offline evaluations? Do they provide the same, or at least similar, rankings of the evaluated approaches? (Section 3.8)
9. Do authors provide sufficient information to re-implement their algorithms or replicate their experiments? (Section 3.9)

To identify the status quo, we reviewed 176 papers, including a few patents, presentations, blogs, and websites on 89 research paper recommendation approaches[5] [14–56,58,59,66–100,102–110]. We distinguish between *papers* and *approaches* because often one approach is presented or evaluated in several papers. For instance, there are three papers on the recommender system *Papyres* and all cover different aspects of the same system [12,13,74]. Therefore, we count *Papyres* as one recommendation approach. To cite an approach, for which more than one paper exists, we subjectively selected the most representative paper. For our analysis, we also 'combined' the content of all papers relating to one approach. If an approach was once evaluated using an online evaluation, and in another paper using an offline evaluation, we say that the approach was evaluated with both online and offline evaluations. Space restrictions keep us from providing an exhaustive bibliography of the 176 papers reviewed, so that we only cite the 89 approaches, i.e. one representative paper for each approach.

Papers were retrieved using Google Scholar, the ACM Digital Library and Springer Link by searching for `[paper | article | citation] [recommender | recommendation] [system | systems]` and downloading all articles that had relevance for research paper recommendations[6]. In a second step, the bibliography of each article was examined. When an entry in the bibliography pointed to an article not yet downloaded, the cited article was also downloaded and inspected for relevant entries in its bibliography.

3. RESULTS

3.1 Evaluation Methods

19 approaches (21%) were not evaluated [14–26], or were evaluated using system-unique or uncommon and convoluted methods [27–31,93]. In the remaining analysis, these 19 approaches are ignored. Of the remaining 70 approaches, 48 approaches (69%), were evaluated using an offline evaluation [32–52,54,58,59,74,78,80,83,86,88–92,94–100,102–107,109], 24 approaches (34%) with a user study [66–74,76,77,79,81,82,87,102–108,110], five approaches (7%) were evaluated in real-world systems with an online evaluation [53–56,68] and two approaches (3%) were evaluated using a qualitative user study [84,85] (Table 1)[7].

Interesting in this context is the low number of online evaluations (7%) and the prevalence of offline evaluations (69%). Despite active experimentation in the field of research papers recommender systems, we observed that many researchers have no access to real-world systems to evaluate their approaches and researchers who do, often do not use them. For instance, C. Lee Giles and his co-authors, who are some of the largest contributors in the field [57–59,94,96,99,100], could have conducted online experiments with their academic search engine CiteSeer. However, they chose primarily to use offline evaluations. The reason for this may be that offline evaluations are more convenient than conducting online evaluations or user studies. Results are available within minutes or hours and not within days or weeks as is the case for online evaluations and user studies. However, as stated, offline-evaluations are subject to various criticisms [60–65,106,111].

Table 1: Evaluation methods[7]

Offline	User Study	Online	Qualitative
48	24	5	2
69%	34%	7%	3%

3.2 Number of Participants in User Studies

Four of the 24 user-studies (17%) were conducted with less than five participants [66,67,102,104]. Another four studies had five to ten participants [77,79,103,110]. Three studies had 11-15 participants [68,81,87], and another four studies had 16-50 participants [69–71,105]. Only six studies (25%), were conducted with more than 50 participants [72–74,106–108]. Three studies failed to mention the number of participants [75,76,82] (Table 2). Given these findings, we conclude that most user studies were not large enough to arrive at meaningful conclusions on algorithm quality.

Table 2: Number of participants in user studies

	Number of Participants					
	n/a	<5	5-10	11-15	16-50	>50
Absolute	3	4	4	3	4	6
Relative	13%	17%	17%	13%	17%	25%

[5] We use the term 'approach' not only for distinct recommendation concepts like content based or collaborative filtering, but also for minor variations in recommendation algorithms.

[6] The relevance judgment was done manually by using the title and if in doubt consulting the abstract.

[7] Some approaches were evaluated with several methods at the same time. Therefore, percentages do not add up to 100.

3.3 Baselines

Thirteen of the evaluated approaches (19%) were not evaluated against a baseline (Table 3) [77–88,102]. The evaluations' usefulness is low because knowing that in certain circumstances an algorithm has a certain CTR allows no conclusion on how it compares against other algorithms. Another 50 approaches (71%) were evaluated against trivial baselines, such as simple content-based filtering without any sophisticated adjustments. These trivial baselines do not represent the state-of-the-art and are not helpful for deciding which of the 89 approaches are most promising. This is in particular true, since different approaches were not evaluated against the *same* simple baselines. Even for a simple content-based approach, there are many variables such as whether stop-words are filtered, if and which stemmer is applied, from which document section (title, abstract, etc.) the text is extracted, etc. This means, almost all approaches were compared against different baselines.

Only seven authors (10%) evaluated their approaches against state-of-the-art approaches proposed by other researchers in the field. Only these seven evaluations allowed drawing some conclusions on which approaches may perform best. The authors, however, compared the seven approaches only against some state-of-the-art approaches. It remains unclear how they would have performed against the remaining state-of-the-art approaches[8].

Table 3: Baselines

	No Baseline	Simple Baseline	St.of the Art Bsln.
Absolute	13	50	7
Relative	19%	71%	10%

3.4 Runtimes & Computational Complexity

Only eight approaches (11%) provided information on runtime. Runtime information, however, is crucial. In one comparison, the runtimes of two approaches differed by factor 600 [100]. For many developers, an algorithm requiring 600 times more CPU power than another would probably not be an option. While this example is extreme, it frequently occurred that runtimes differed by factor five or more, which can also affect the decisions on algorithm selection.

Computational complexity was reported by even fewer evaluations. Computational complexity may be less relevant for researchers but highly relevant for providers of recommender systems. It is important for estimating the long-term suitability of an algorithm. An algorithm may perform well for a few users but it might not scale well. Hence, algorithms with, for example, exponentially increasing complexity most likely will not be applicable in practice.

3.5 Use of Offline Evaluation Metrics

Out of the 48 offline evaluations, 33 approaches (69%) were evaluated with *precision* (Table 4). Recall was used for eleven approaches (23%), F-measure for six approaches (13%) and NDCG for six approaches. Seven approaches (15%) were evaluated using other measures [88–91,97,98,105]. Overall, results of the different measures highly correlated – that is algorithms, which performed well using precision also performed well using, for instance, NDCG.

Table 4: Evaluation measures[7]

	Precision	Recall	F-Measure	NDCG	MRR	Other
Absolute	33	11	6	6	4	7
Relative	69%	23%	13%	13%	8%	15%

3.6 Use of Datasets

Researchers used different datasets to conduct their offline evaluations (Table 5). Fourteen approaches (29%) were evaluated using data from CiteSeer and five approaches (10%) were evaluated using papers from ACM. Other data sources included CiteULike (10%), DBLP (8%) and a variety of others, many not publicly available (52%). Even when data originated from the same sources, this did not guarantee that the same datasets were used. For instance, fourteen approaches used data from CiteSeer but no single '*CiteSeer dataset*' exists. Authors collected CiteSeer data at different times and pruned datasets differently. Some authors removed documents with less than two citations from the corpus [92], others with less than three citations [107], and others with less than four citations [93]. One study removed all papers with less than ten and more than 100 citations and all papers citing less than 15 and more than 50 papers [94]. Of the original dataset of 1,345,249 papers, only 81,508 remained, about 6%. The question arises how representative results can be based on such a pruned dataset.

Table 5: Data sources

	CiteSeer	ACM	CiteULike	DBLP	Others
Absolute	14	5	5	4	25
Relative	29%	10%	10%	8%	52%

In conclusion, it is safe to say that no two studies performed by different authors, used the same dataset. This raises the question to what extent results based of different datasets are comparable?

3.7 Universality of Offline Datasets

Seven approaches were evaluated on different offline datasets [95–100,110].

The analysis of these seven evaluations confirms a well-known finding: results from one dataset do not allow any conclusions on the *absolute* performance achievable in another dataset. For instance, an algorithm, which achieved a recall of 4% on an IEEE dataset, achieved a recall of 12% on an ACM dataset [110].

However, the analysis also showed that the *relative* performance among different algorithms remained quite stable over different datasets. Algorithms performing well on one dataset (compared to some baselines) also performed well on other datasets (compared to the same baselines). Dataset combinations included CiteSeer and some posts from various blogs [97], CiteSeer and Web-kd [98], CiteSeer and CiteULike [100], CiteSeer and Eachmovie [99], and IEEE, ACM and ScienceDirect [110]. Only in one study results differed notably, however, the absolute ranking of the algorithms remained stable [100] (see Table 6). In this paper, the proposed approach (CTM) performed best on two datasets with a MRR of 0.529 and 0.467 respectively. Three of the four baselines performed similarly on the CiteSeer dataset (all with a MRR between 0.238 and 0.288). However, for the CiteULike dataset the TM approach performed four times as well as CRM. This means, if TM had been compared with CRM, rankings would have been similar on the

[8] It is interesting to note that in all published papers with an evaluation against a baseline, at least one of the proposed approaches performed better than the baseline(s). It never occurred that a paper reported on a non-effective approach. This invited a search for possible explanations. First, authors may intentionally select baselines such that their approaches appear favorable. Second, the simple baselines used in most evaluations achieve relatively unrefined results, so that any alternative easily performs better. Third, authors do not report their failures, which ties in with the fourth point, which is that journals and conferences typically do not accept publications that report on failures.

CiteSeer dataset but different on the CiteULike dataset. As mentioned, for all other reviewed evaluations no such variations in the rankings were observed.

Table 6: MRR of different recommendation approaches on CiteSeer and CiteULike datasets

		Dataset	
Rank	Approach	CiteSeer	CiteULike
1	CTM	0.529	0.467
2	TM	0.288	0.285
3	cite-LDA	0.285	0.143
4	CRM	0.238	0.072
5	link-LDA	0.028	0.013

Overall, a sample size of seven is small, but it gives at least some indication that the impact of the chosen dataset is rather low. This finding is interesting because in other fields it has been observed that different datasets lead to different results [101].

3.8 Consistency of Offline Evaluations and User Studies

Six approaches were evaluated using an offline evaluation in addition to a user study [102–107]. Of these six evaluations, one did not compare its approach against any baseline [102]. The remaining five evaluations reported non-uniform results. In two cases, results from the offline evaluations were similar to results of the user studies [103,105]. However, the user studies had only five and 19 participants respectively. As such, results should be interpreted with some skepticism. Three other studies reported that results of the offline evaluations contradicted the results of the user studies [104,106,107]. Two of these studies had more than 100 participants; the other study only had two participants. The findings indicate that results from user studies and offline evaluation do not necessarily correlate, which could question the validity of offline evaluations in general [111].

Interestingly, the three studies with the most participants were all conducted by the authors of TechLens [105–107], who are also the only authors in the field of research paper recommender systems discussing the potential shortcomings of offline evaluations [108]. It seems that other researchers in this field are not aware of problems associated with offline evaluations although there has been quite a discussion.

3.9 Sparse Information on Algorithms

Many authors provided sparse information on the exact workings of their proposed approaches. Hence, replication of their evaluations, or re-implementing their approaches, for example, to use them as a baseline, is hardly possible. For instance, one set of authors stated they had created content-based user models based on a user's documents. From which document section (title, abstract, keywords, body, etc.) the text was taken was not explained. However, taking text from titles, abstracts or the body makes a significant difference [109,110].

4. SUMMARY & OUTLOOK

The review of 176 publications has shown that no consensus exists on how to evaluate and compare research paper recommender approaches. This leads to the unsatisfying situation that despite the many evaluations, the individual strengths and weaknesses of the proposed approaches remain largely unknown. Out of 89 reviewed approaches, 21% were not evaluated. Of the evaluated approaches, 19% were not evaluated against a baseline. Almost all evaluations that compared against a baseline, compared against trivial baselines. Only 10% of the reviewed approaches were compared against at least one state-of-the-art approach.

In addition, runtime information was only provided for 11% of the approaches, despite this information being crucial for assessing algorithm practicability. In one case, runtimes differed by factor 600. Details on the proposed algorithms were often sparse, which makes a re-implementation difficult in many cases. Only five approaches (7%) were evaluated using online evaluations. The majority of authors conducted offline evaluations (69%). The most frequent sources for retrieving offline datasets were CiteSeer (29%), ACM (10%), and CiteULike (10%). However, the majority (52%) of evaluations were conducted using other datasets and even the datasets from CiteSeer, ACM, and CiteULike differed, since they were all fetched at different times and pruned differently. Because of the different datasets used, individual study outcomes are not comparable. Of the approaches evaluated with a user study (34%), the majority (58%) of these studies had less than 16 participants. In addition, user studies sometimes contradicted results of offline evaluations. These observations question the validity of offline evaluations, and demand further research.

Given the circumstances, an identification of the most promising approaches for recommending research papers is not possible, and neither is a replication for most evaluations. We consider this a major problem for the advancement of research paper recommender systems. Researchers cannot evaluate their novel approaches against a state-of-the-art baseline because no state-of-the-art baseline exists. Similarly, providers of academic services, who wish to implement a recommender system, have no chance of knowing which of the 89 approaches they should implement.

We suggest the following three points of action to ensure that the best research paper recommender approaches can be determined:

1. Discuss the suitability of offline evaluations for evaluating research paper recommender systems (we started this already with the preliminary conclusion that offline evaluations are unsuitable in many cases for evaluating research paper recommender systems [111]).
2. Re-evaluate existing approaches, ideally in real-world systems with suitable baselines, sufficient study participants, and with information on runtimes and computational complexity.
3. Develop a framework including the most promising approaches, so other researchers can easily compare their novel approaches against the state-of-the-art.

If these actions are not taken, researchers will continue to evaluate their approaches without comparable results, and although many more approaches would exist, it would be unknown which are most promising for practical application, or against which to compare new approaches.

5. REFERENCES

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

[2] D.N. Chin, "Empirical evaluation of user models and user-adapted systems," *User modeling and user-adapted interaction*, vol. 11, 2001, pp. 181–194.

[3] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," *The adaptive web*, 2007, pp. 3–53.

[4] S. Lam, D. Frankowski, and J. Riedl, "Do you trust your recommendations? An exploration of security and privacy issues in recommender systems," *Emerging Trends in Information and Communication Security*, 2006, pp. 14–29.

[5] N. Good, J.B. Schafer, J.A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl, "Combining collaborative filtering with personal agents for better recommendations," *Proceedings of the National Conference on Artificial Intelligence*, JOHN WILEY & SONS LTD, 1999, pp. 439–446.

[6] C.D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England, 2009.

[7] 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.

[8] 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.

[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] R. Burke, "Hybrid recommender systems with case-based components," *Advances in Case-Based Reasoning*, Springer, 2004, pp. 91–105.

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

[12] A. Naak, "Papyres : un système de gestion et de recommandation d'articles de recherche," Université de Montréal, 2009.

[13] A. Naak, H. Hage, and E. Aιmeur, "Papyres: A Research Paper Management System," *Proceedings of the 10th E-Commerce Technology Conference on Enterprise Computing, E-Commerce and E-Services*, IEEE, 2008, pp. 201–208.

[14] C. Bancu, M. Dagadita, M. Dascalu, C. Dobre, S. Trausan-Matu, and A.M.F. Florea, "ARSYS-Article Recommender System," *Proceedings of the 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2012, pp. 349–355.

[15] 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.

[16] S.C. Cazella and L.O.C. Alvares, "Combining Data Mining Technique and Users' Relevance Opinion to Build an Efficient Recommender System," *Revista Tecnologia da Informação, UCB*, vol. 4, 2005.

[17] P. Chirawatkul, "Structured Peer-to-Peer Search to build a Bibliographic Paper Recommendation System," Saarland University, 2006.

[18] L. Fernández, J.A. Sánchez, and A. Garcιa, "Mibiblio: personal spaces in a digital library universe," *Proceedings of the fifth ACM conference on Digital libraries*, ACM, 2000, pp. 232–233.

[19] G. Geisler, D. McArthur, and S. Giersch, "Developing recommendation services for a digital library with uncertain and changing data," *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, ACM, 2001, pp. 199–200.

[20] B. Gipp, J. Beel, and C. Hentschel, "Scienstein: A Research Paper Recommender System," *Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09)*, Virudhunagar (India): IEEE, 2009, pp. 309–315.

[21] A. Nakagawa and T. Ito, "An implementation of a knowledge recommendation system based on similarity among users' profiles," *Proceedings of the 41st SICE Annual Conference*, IEEE, 2002, pp. 326–327.

[22] H.-E. Popa, V. Negru, D. Pop, and I. Muscalagiu, "DL-AgentRecom-A multi-agent based recommendation system for scientific documents," *Proceedings of the 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, IEEE, 2008, pp. 320–324.

[23] L.M. Rocha, "TalkMine: a soft computing approach to adaptive knowledge recommendation," *Studies in fuzziness and soft computing*, vol. 75, 2001, pp. 89–116.

[24] J. Vassileva, "Supporting peer-to-peer user communities," *Proceedings of the Conference On the Move to Meaningful Internet Systems*, Springer, 2002, pp. 230–247.

[25] A.S. Vivacqua, J. Oliveira, and J.M. de Souza, "i-ProSE: inferring user profiles in a scientific context," *The Computer Journal*, vol. 52, 2009, pp. 789–798.

[26] Q. Yang, S. Zhang, and B. Feng, "Research on Personalized Recommendation System of Scientific and Technological Periodical Based on Automatic Summarization," *Proceedings of the 1st International Symposium on Information Technologies and Applications in Education*, IEEE, 2007, pp. 34–39.

[27] C. Hess, *Trust-based recommendations in multi-layer networks*, IOS Press, 2008.

[28] N.F. Matsatsinis, K. Lakiotaki, and P. Delia, "A system based on multiple criteria analysis for scientific paper recommendation," *Proceedings of the 11th Panhellenic Conference on Informatics*, 2007, pp. 135–149.

[29] S. Watanabe, T. Ito, T. Ozono, and T. Shintani, "A paper recommendation mechanism for the research support system papits," *Proceedings of the International Workshop on Data Engineering Issues in E-Commerce*, IEEE, 2005, pp. 71–80.

[30] S.-S. Weng and H.-L. Chang, "Using ontology network analysis for research document recommendation," *Expert Systems with Applications*, vol. 34, 2008, pp. 1857–1869.

[31] S.-Y. Yang and C.-L. Hsu, "A New Ontology-Supported and Hybrid Recommending Information System for Scholars," *Proceedings of the 13th International Conference on Network-Based Information Systems (NBiS)*, IEEE, 2010, pp. 379–384.

[32] H. Avancini, L. Candela, and U. Straccia, "Recommenders in a personalized, collaborative digital library environment," *Journal of Intelligent Information Systems*, vol. 28, 2007, pp. 253–283.

[33] N. Agarwal, E. Haque, H. Liu, and L. Parsons, "A subspace clustering framework for research group collaboration," *International Journal of Information Technology and Web Engineering*, vol. 1, 2006, pp. 35–58.

[34] A. Arnold and W.W. Cohen, "Information extraction as link prediction: Using curated citation networks to improve gene detection," *Proceedings of the 4th International Conference on Wireless Algorithms, Systems, and Applications*, Springer, 2009, pp. 541–550.

[35] M. Baez, D. Mirylenka, and C. Parra, "Understanding and supporting search for scholarly knowledge," *Proceeding of the 7th European Computer Science Summit*, 2011, pp. 1–8.

[36] S. Bethard and D. Jurafsky, "Who should I cite: learning literature search models from citation behavior," *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, 2010, pp. 609–618.

[37] F. Ferrara, N. Pudota, and C. Tasso, "A Keyphrase-Based Paper Recommender System," *Proceedins of the IRCDL'11*, Springer, 2011, pp. 14–25.

[38] M. Gori and A. Pucci, "Research paper recommender systems: A random-walk based approach," *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE Computer Society Washington, DC, USA, 2006, pp. 778–781.

[39] S.-Y. Hwang, W.-C. Hsiung, and W.-S. Yang, "A prototype WWW literature recommendation system for digital libraries," *Online Information Review*, vol. 27, 2003, pp. 169–182.

[40] Y. Huang, "Combining Social Networks and Content for Recommendation in a Literature Digital Library," National Sun Yat-Sen University, Taiwan, 2007.

[41] K. Jack, "Mahout Becomes a Researcher: Large Scale Recommendations at Mendeley," *Presentation at Big Data Week Conferences*, 2012.

[42] 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.

[43] J. Lin and W.J. Wilbur, "PubMed Related Articles: a Probabilistic Topic-based Model for Content Similarity," *BMC Bioinformatics*, vol. 8, 2007, pp. 423–436.

[44] G.H. Martın, S. Schockaert, C. Cornelis, and H. Naessens, “Metadata impact on research paper similarity,” *14th European Conference on Digital Libraries*, Springer, 2010, pp. 457–460.
[45] S. Pohl, F. Radlinski, and T. Joachims, “Recommending Related Papers Based on Digital Library Access Records,” *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, Vancouver, BC, Canada: ACM, 2007, pp. 417–418.
[46] T. Strohman, W.B. Croft, and D. Jensen, “Recommending citations for academic papers,” *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2007, pp. 705–706.
[47] K. Sugiyama and M.-Y. Kan, “Scholarly paper recommendation via user’s recent research interests,” *Proceedings of the 10th annual joint conference on Digital libraries*, ACM, 2010, pp. 29–38.
[48] H. Wu, Y. Hua, B. Li, and Y. Pei, “Enhancing citation recommendation with various evidences,” *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, IEEE, 2012, pp. 1160–1165.
[49] H. Xia, J. Li, J. Tang, and M.-F. Moens, “Plink-LDA: using link as prior information in topic modeling,” *Proceedings of the Conference on Database Systems for Advanced Applications (DASFAA)*, Springer, 2012, pp. 213–227.
[50] P. Yin, M. Zhang, and X. Li, “Recommending scientific literatures in a collaborative tagging environment,” *Proceedings of the 10th international conference on Asian digital libraries*, Springer, 2007, pp. 478–481.
[51] F. Zarrinkalam and M. Kahani, “SemCiR,” *Program: electronic library and information systems*, vol. 47, 2013, pp. 92–112.
[52] F. Zarrinkalam and M. Kahani, “A multi-criteria hybrid citation recommendation system based on linked data,” *Proceedings of the 2nd International eConference on Computer and Knowledge Engineering*, IEEE, 2012, pp. 283–288.
[53] 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.
[54] T. Bogers and A. van den Bosch, “Recommending scientific articles using citeulike,” *Proceedings of the 2008 ACM conference on Recommender systems*, ACM New York, NY, USA, 2008, pp. 287–290.
[55] M. Mönnich and M. Spiering, “Adding value to the library catalog by implementing a recommendation system,” *D-Lib Magazine*, vol. 14, 2008, pp. 4–11.
[56] S.E. Middleton, N.R. Shadbolt, and D.C. De Roure, “Ontological user profiling in recommender systems,” *ACM Transactions on Information Systems (TOIS)*, vol. 22, 2004, pp. 54–88.
[57] Q. He, D. Kifer, J. Pei, P. Mitra, and C.L. Giles, “Citation recommendation without author supervision,” *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 2011, pp. 755–764.
[58] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, “Context-aware citation recommendation,” *Proceedings of the 19th international conference on World wide web*, ACM, 2010, pp. 421–430.
[59] L. Rokach, P. Mitra, S. Kataria, W. Huang, and L. Giles, “A Supervised Learning Method for Context-Aware Citation Recommendation in a Large Corpus,” *Proceedings of the Large-Scale and Disributed Systems for Information Retrieval Workshop (LSDS-IR)*, 2013, pp. 17–22.
[60] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, 2005, pp. 734–749.
[61] A. Gunawardana and G. Shani, “A survey of accuracy evaluation metrics of recommendation tasks,” *The Journal of Machine Learning Research*, vol. 10, 2009, pp. 2935–2962.
[62] 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.
[63] 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.
[64] 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, pp. 25–37.
[65] 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.
[66] P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, “A framework for tag-based research paper recommender system: an IR approach,” *Proceedings of the 24th International Conference on Advanced Information Networking and Applications (WAINA),*, IEEE, 2010, pp. 103–108.
[67] A. Woodruff, R. Gossweiler, J. Pitkow, E.H. Chi, and S.K. Card, “Enhancing a digital book with a reading recommender,” *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2000, pp. 153–160.
[68] J. Bollen and H. Van de Sompel, “An architecture for the aggregation and analysis of scholarly usage data,” *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, ACM, 2006, pp. 298–307.
[69] B. Gipp and J. Beel, “Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis,” *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI’09)*, B. Larsen and J. Leta, eds., Rio de Janeiro (Brazil): International Society for Scientometrics and Informetrics, 2009, pp. 571–575.
[70] K. Uchiyama, H. Nanba, A. Aizawa, and T. Sagara, “OSUSUME: cross-lingual recommender system for research papers,” *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, ACM, 2011, pp. 39–42.
[71] Y. Wang, E. Zhai, J. Hu, and Z. Chen, “Claper: Recommend classical papers to beginners,” *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, 2010, pp. 2777–2781.
[72] U. Farooq, C.H. Ganoe, J.M. Carroll, I.G. Councill, and C. Lee Giles, “Design and evaluation of awareness mechanisms in CiteSeer,” *Information Processing & Management*, vol. 44, 2008, pp. 596–612.
[73] S. Kang and Y. Cho, “A novel personalized paper search system,” *Proceedings of the International Conference on Intelligent Computing*, Springer, 2006, pp. 1257–1262.
[74] A. Naak, H. Hage, and E. Aımeur, “A multi-criteria collaborative filtering approach for research paper recommendation in papyres,” *Proceedings of the 4th International Conference MCETECH*, Springer, 2009, pp. 25–39.
[75] S. Huang, G.R. Xue, B.Y. Zhang, Z. Chen, Y. Yu, and W.Y. Ma, “Tssp: A reinforcement algorithm to find related papers,” *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, 2004, pp. 117–123.
[76] Y. Jiang, A. Jia, Y. Feng, and D. Zhao, “Recommending academic papers via users’ reading purposes,” *Proceedings of the sixth ACM conference on Recommender systems*, ACM, 2012, pp. 241–244.
[77] K. Hong, H. Jeon, and C. Jeon, “UserProfile-based personalized research paper recommendation system,” *Proceedings of the 8th International Conference on Computing and Networking Technology*, IEEE, 2012, pp. 134–138.
[78] J. He, J.-Y. Nie, Y. Lu, and W.X. Zhao, “Position-Aligned translation model for citation recommendation,” *Proceedings of the 19th international conference on String Processing and Information Retrieval*, Springer, 2012, pp. 251–263.
[79] A. Kodakateri Pudhiyaveetil, S. Gauch, H. Luong, and J. Eno, “Conceptual recommender system for CiteSeerX,” *Proceedings of the third ACM conference on Recommender systems*, ACM, 2009, pp. 241–244.
[80] Y. Lu, J. He, D. Shan, and H. Yan, “Recommending citations with translation model,” *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM, 2011, pp. 2017–2020.
[81] J.M. Morales-del-Castillo, E. Peis, and E. Herrera-Viedma, “A filtering and recommender system prototype for scholarly users of

digital libraries," *Proceedings of the Second World Summit on the Knowledge Society*, Springer, 2009, pp. 108–117.

[82] G. Mishra, "Optimised Research Paper Recommender System Using Social Tagging," *International Journal of Engineering Research and Applications*, vol. 2, 2012, pp. 1503–1507.

[83] C. Pan and W. Li, "Research paper recommendation with topic analysis," *Proceedings of the International Conference on Computer Design and Applications (ICCDA)*, IEEE, 2010, pp. 264–268.

[84] K. Stock, A. Robertson, F. Reitsma, T. Stojanovic, M. Bishr, D. Medyckyj-Scott, and J. Ortmann, "eScience for Sea Science: A Semantic Scientific Knowledge Infrastructure for Marine Scientists," *Proceedings of the 5th IEEE International Conference on e-Science*, IEEE, 2009, pp. 110–117.

[85] K. Stock, V. Karasova, A. Robertson, G. Roger, M. Small, M. Bishr, J. Ortmann, T. Stojanovic, F. Reitsma, L. Korczynski, B. Brodaric, and Z. Gardner, "Finding Science with Science: Evaluating a Domain and Scientific Ontology User Interface for the Discovery of Scientific Resources," *Transactions in GIS*, vol. 1, 2013, pp. 1–28.

[86] M. Zhang, W. Wang, and X. Li, "A Paper Recommender for Scientific Literatures Based on Semantic Concept Similarity," *Proceedings of the International Conference on Asian Digital Libraries*, 2008, pp. 359–362.

[87] Z. Zhang and L. Li, "A research paper recommender system based on spreading activation model," *Proceedings of the 2nd International Conference on Information Science and Engineering (ICISE)*, IEEE, 2010, pp. 928–931.

[88] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences of the United States of America*, National Acad Sciences, 2004, pp. 5220–5227.

[89] N. Ratprasartporn and G. Ozsoyoglu, "Finding related papers in literature digital libraries," *Proceedings of the 11th European Conference on Digital Libraries*, Springer, 2007, pp. 271–284.

[90] A. Vellino, "A comparison between usage-based and citation-based methods for recommending scholarly research articles," *Proceedings of the American Society for Information Science and Technology*, Wiley Online Library, 2010, pp. 1–2.

[91] C. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang, "CARES: a ranking-oriented CADAL recommender system," *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, ACM, 2009, pp. 203–212.

[92] R. Dong, L. Tokarchuk, and A. Ma, "Digging Friendship: Paper Recommendation in Social Network," *Proceedings of Networking & Electronic Commerce Research Conference (NAEC 2009)*, 2009, pp. 21–28.

[93] A. Daud and A.H. Muhammad Akramand Rajpar Shaikh, "Scientific Reference Mining using Semantic Information through Topic Modeling," *Research Journal of Engineering & Technology*, vol. 28, 2009, pp. 253–262.

[94] C. Caragea, A. Silvescu, P. Mitra, and C.L. Giles, "Can't See the Forest for the Trees? A Citation Recommendation System," *iConference 2013 Proceedings*, 2013, pp. 849–851.

[95] N. Lao and W.W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine learning*, vol. 81, 2010, pp. 53–67.

[96] D. Zhou, S. Zhu, K. Yu, X. Song, B.L. Tseng, H. Zha, and C.L. Giles, "Learning multiple graphs for document recommendations," *Proceedings of the 17th international conference on World Wide Web*, ACM, 2008, pp. 141–150.

[97] R.M. Nallapati, A. Ahmed, E.P. Xing, and W.W. Cohen, "Joint latent topic models for text and citations," *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2008, pp. 542–550.

[98] S. Kataria, P. Mitra, and S. Bhatia, "Utilizing context in generative bayesian models for linked corpus," *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010, pp. 1340–1345.

[99] D.M. Pennock, E. Horvitz, S. Lawrence, and C.L. Giles, "Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach," *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 2000, pp. 473–480.

[100] W. Huang, S. Kataria, C. Caragea, P. Mitra, C.L. Giles, and L. Rokach, "Recommending citations: translating papers into references," *Proceedings of the 21st ACM international conference on Information and knowledge management*, ACM, 2012, pp. 1910–1914.

[101] 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.

[102] X. Tang and Q. Zeng, "Keyword clustering for user interest profiling refinement within paper recommender systems," *Journal of Systems and Software*, vol. 85, 2012, pp. 87–101.

[103] Y. Liang, Q. Li, and T. Qian, "Finding relevant papers based on citation relations," *Proceedings of the 12th international conference on Web-age information management*, Springer, 2011, pp. 403–414.

[104] Z. Huang, W. Chung, T.H. Ong, and H. Chen, "A graph-based recommender system for digital library," *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, ACM, 2002, pp. 65–73.

[105] M.D. Ekstrand, P. Kannan, J.A. Stemper, J.T. Butler, J.A. Konstan, and J.T. Riedl, "Automatically building research reading lists," *Proceedings of the fourth ACM conference on Recommender systems*, ACM, 2010, pp. 159–166.

[106] 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.

[107] 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.

[108] 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.

[109] K. Jack, "Mendeley: Recommendation Systems for Academic Literature," *Presentation at Technical University of Graz (TUG)*, 2012.

[110] C. Nascimento, A.H. Laender, A.S. da Silva, and M.A. Gonçalves, "A source independent framework for research paper recommendation," *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, ACM, 2011, pp. 297–306.

[111] 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," *Proceedings of the Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys) at the ACM Recommender System Conference (RecSys)*, 2013.