Multi-method Evaluation in Scientific Paper Recommender Systems*

Aravind Sesagiri Raamkumar
Wee Kim Wee School of Communication and Information
Nanyang Technological University
Singapore
aravind002@ntu.edu.sg

Schubert Foo
Wee Kim Wee School of Communication and Information
Nanyang Technological University
Singapore
sfoo@ntu.edu.sg

ABSTRACT

Recommendation techniques in scientific paper recommender systems (SPRS) have been generally evaluated in an offline setting, without much user involvement. Nonetheless, user relevance of recommended papers is equally important as system relevance. In this paper, we present a scientific paper recommender system (SPRS) prototype which was subject to both offline and user evaluations. The lessons learnt from the evaluation studies are described. In addition, the challenges and open questions for multi-method evaluation in SPRS are presented.

KEYWORDS

Scientific paper recommender systems, multi-method evaluation, user evaluation, research paper recommender systems

ACM Reference Format:
Aravind S. Raamkumar and Schubert Foo. 2018. Multi-method Evaluation in Scientific Paper Recommender Systems. In *UMAP'18 Adjunct: 26th Conference on User Modeling, Adaptation and Personalization Adjunct, July 8–11, 2018, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3213586.3226215

1 INTRODUCTION

For the different stages in the of scientific research lifecycle, Recommender System (RS) techniques have been conceptualized to recommend information objects such as publication venues and collaborators in addition to the standard scientific papers [6]. In particular, scientific paper recommender systems (SPRS) research has been an active area of research. SPRS techniques utilize data from sources such as the citation network, paper metadata, full-text and system log files for generating recommendations. As per [1], offline evaluations are more prevalent in this SPRS area, accounting to about 69% of all studies. User-based and online evaluations seem to be uncommon due to the complexity and uncertainty factors.

UMAP'18 Adjunct, July 8–11, 2018, Singapore.

ISBN 978-1-4503-5784-5/18/07...$15.00
https://doi.org/10.1145/3213586.3226215

Offline evaluations are comparatively convenient to conduct as users are not involved. It is observed that large-scale user evaluations have been conducted mainly as part of doctoral dissertations [4, 6, 9]. We developed a task-based SPRS prototype called Rec4LRW [10] for helping researchers with literature review and manuscript preparatory tasks. Our focus was more on conducting user evaluation studies as recommendations was just one of the multiple aspects of this system. In the next section, we introduce the Rec4LRW prototype along with its features and information about the evaluation studies. The lessons learned from the evaluation studies are described at the end of the section. The challenges for multi-method evaluation in SPRS are put forth in the final section.

2 REC4LRW SYSTEM

The Rec4LRW system [10] was developed to assist researchers in two main literature review search tasks and one manuscript preparatory task. The three tasks are *(i) building an initial reading list of research papers, (ii) finding similar papers based on a set of papers,* and *(iii) shortlisting papers from the final reading list for inclusion in manuscript based on article-type choice.* The recommendation techniques for these tasks are based on a combination of graph ranking algorithms, IR ranking functions, collaborative filtering and community detection algorithms. The system was built as a prototype to showcase not only the task recommendations but also the task interconnectivity features and novel UI display features. A sample screenshot from the first task of Rec4LRW is provided in Figure 1. A snapshot of the data from ACM digital library was used as the corpus of the system.

2.1 Evaluation Studies

The offline evaluation of these three tasks was challenging due to the requirements of the tasks i.e. there were no previous studies conducted for the identified requirements. Secondly, we tried building gold standard lists for the tasks by seeking help from topical experts, but the outcome was not encouraging due to expert unavailability and uncertain nature of heuristics for selecting papers. Hence, the standard IR/RS evaluation metrics could not be used for the study. However, we proceeded with performing offline evaluation for the first task as its task input was the same from previous studies [3]. We used the *rank aggregation* [2] evaluation methodology for benchmarking the proposed technique with other relevant techniques. Offline evaluation was not conducted for the second and third tasks due to the novel requirements of the tasks.

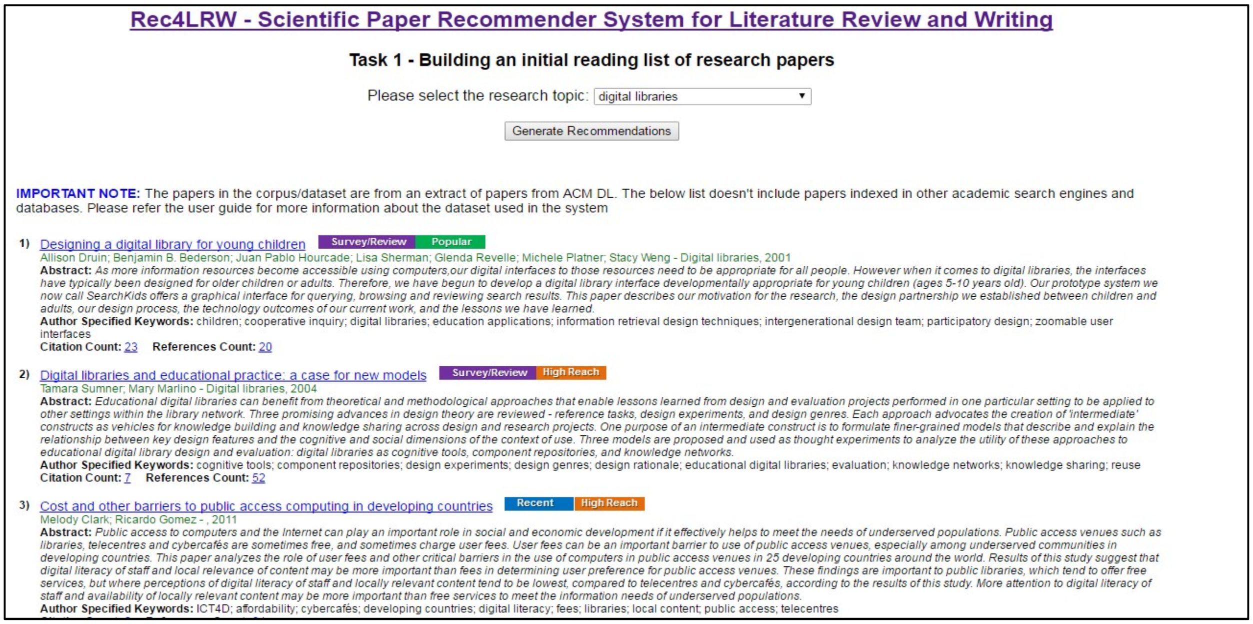

Figure 1. Reading list task screen (Task 1) in Rec4LRW system

A large-scale user evaluation study of the Rec4LRW system was conducted with 119 researchers who had experience in conducting research and writing research papers. These researchers were divided into two groups of staff (53%) and students (47%) for analysis purpose. The purpose of the user evaluation study was to determine whether researchers using the tasks provided by Rec4LRW system, can be efficient and effective in conducting the corresponding LR tasks. Researchers' perceptions of the individual characteristics of the recommended papers, overall quality of the recommendation list and system features were measured.

The specific instructions for the participants of the user evaluation study were as follows. In Task 1, participants had to select a research topic from a list of 43 research topics in the task screen. On selection of a topic, the system provided 30 recommendations. Before executing Task 2, the participant had to add at least five papers from Task 1 into the seed basket (SB)[1]. Subsequently, the system provided 30 recommendations for this task. For Task 3, the participants were requested to add at least 30 papers in the personalized reading list (RL)[2]. The participant had to then select the article-type and run the task so that the system could retrieve the shortlisted papers.

In Table 1, the evaluation goals and the corresponding measurement methods of this study are listed. The quantitative evaluation measures and constructs used in this study are listed in Table 2. These measures facilitated collection of user responses for three aspects – recommendations, UI and system.

Table 1. Rec4LRW Evaluation Goals and Measurement Methods

Evaluation Goals	Measurement Methods
Ascertain the agreement percentages of the evaluation measures for the three tasks and the overall system and identify whether the values are above a preset threshold criterion of 75%	Percentages comparison, Independent samples t-test
Test the hypothesis that students benefit more from the recommendation tasks/system in comparison to staff	
Measure the correlation between the measures and build a regression model with *Good_List* as the dependent variable	Spearman correlation coefficient, Multiple linear regression, Paired samples t-test
Track the change in user perceptions between the three tasks. This is similar to the first evaluation goal since the agreement percentages will be used for the analysis	Percentages comparison
Compare the pre-study and post-study variables for understanding whether the target participants are benefitted from the tasks	Percentages and crosstab comparison
Identify the top most preferred and critical aspects of the task recommendations and the system using the subjective feedback of the participants	Qualitative descriptive coding [8]

[1] Seed basket (SB) is a task interconnectivity feature in the system to connect Task 1 to Task 2.
[2] Personalized reading list (RL) is a task interconnectivity feature for collecting all the papers from Tasks 1 and 2, which the participants find to be relevant for their literature review

Most of these measures were conceptualized based on the specific task requirements. The system constructs *Effort to use the System* and *Perceived System Effectiveness* were adopted from a user experience RS study [5]. The third system construct *Perceived Usefulness* was adopted from the TAM model [12]. Five-point Likert scale was provided for measuring participant response for survey-type questions in the questionnaires. Subjective feedback was collected using two questions *(i) From the displayed information, what features did you like the most?* and *(ii) Please provide your personal feedback about the execution of this task.* The responses were collected using three questionnaires at different stages of the evaluation.

Table 2. Rec4LRW User Evaluation Measures and Constructs

Measure	Description
Relevance*	The recommendation list is relevant to the research topic
Usefulness*	The recommendation list is useful for reading at the start of your literature review
Good_List*	This is a good recommendation list, at an overall level
Popularity+	The recommendation list consists of papers that appear to be popular papers for the research topic
Recency+	The recommendation list consists of a decent quantity of recent papers
Diversity+	The recommendation list consists of papers from different sub-topics
Interdisciplinarity+	The recommendation list consists of interdisciplinary papers
Good_Mix+	The recommendation list consists of a good mix of diverse, recent, popular and literature survey papers
Good_Spread+	The recommendation list consists of a good spread of papers for the research topic
Familiarity+	The papers in the recommendation list appear familiar to you
Novelty+	The papers in the recommendation list are unknown to you
Serendipity+	The recommendation list consists of some unexpected papers that you were not expecting to see
Expansion_Required+	There is a need to further expand this recommendation list
User_Satisfaction+	Your satisfaction level for this recommendation list
Seedbasket_Usefulness^	The feature of adding papers to the seed basket to generate similar paper recommendations is a useful feature
Seedbasket_Similarity^	The recommendation list consists of papers that are similar to the papers in the seed basket
Shared_Corelations^	The recommendation list consists of papers that have shared co-references and co-citations with the papers in the seed basket
Task_Interconnectivity~	I would like to see the feature of managing reading list and seed basket papers between the three tasks in academic search systems and databases
Importance~	The shortlisted papers comprise of important papers from my reading list
Certainty~	The shortlisted list comprises of papers which I would definitely cite in my manuscript
Shortlisting_Feature~	I would like to see the feature of shortlisting papers from reading list based on article-type preference, in academic search systems and databases
Effort to use the System	System construct comprising of five questions on the effort required from the participants to use the system
Perceived System Effectiveness	System construct comprising of six questions on the perceptions of effectiveness of the system
Perceived Usefulness	System construct comprising of six questions on the perceptions of usefulness of the system

Note: X - Common to all tasks, X+ - Specific to Tasks 1 and 2, X^ - Specific to Task 2, X~ - Specific to Task 3*

2.2 Lessons Learned from the Evaluation Studies

Through the user evaluation study conducted with researchers, it was convincingly established that students preferred the task recommendations and the overall system. 82% of the students felt that they would be accomplish their tasks more quickly with the system. On the other hand, staff participants found the system to be useful albeit less effective (for instance, 60.38% of staff participants felt that the system would enhance their effectiveness). The incorporation of open-ended questions in the evaluation questionnaires was most beneficial since many participants gave thoughtful feedback about different aspects of the system. In retrospect, qualitative feedback from the participants yielded the most useful and important findings from the evaluation study. With the voluminous feedback data (n=109), we were able to put forth a conceptual framework [11] to guide future SPRS studies from a multidisciplinary viewpoint.

From a quantitative evaluation viewpoint, the regression model testing yielded interesting results. To the best of our knowledge, this was the first study where regression testing was used in a SPRS user evaluation study. With *Good_List* as the dependent variable, we tried to identify the statistically significant predictors. The predictors for Task 1 were *Recency, Novelty, Serendipity, Usefulness* and *User_Satisfaction.* For Task 2, the predictors were *Seedbasket_Similarity* and *Usefulness* while for Task 3, the predictors were *Relevance, Usefulness* and *Certainty*. An example interpretation of these results is as follows. For Task 1, a user might find the recommendation list to be a good list if there are adequate number of recent, novel and unanticipated papers that are useful for the task at hand. Two observations can be made on these predictors from the three tasks. First, they are mostly specific to nature of the recommendation task and *Usefulness* was the only evaluation measure which was common predictor in the three tasks. The regression testing helped us in better understanding the expectations of users and we intend to focus more on these paper types in our future studies. We are of the view that regression testing should be performed after user evaluation studies of SPRS studies, particularly when the recommendations are supposed to satisfy multiple requirements. However, validating these regression models in multiple studies with different participant demographics, is important if causation is to be established. In our study, we used paired samples t-test for validation by using the same dataset, due to certain constraints.

3 CHALLENGES

3.1 Standardized Datasets and Ground Truth

There is a lack of standardized datasets for conducting research studies. Different versions of the datasets from CiteSeer, Microsoft Academic Graph (MAG) and Association of Computation linguistics (ACL) are majorly used by studies. A common version of a dataset is very rarely used across studies, thereby affecting cumulative research to a certain extent. There is no specific TREC track where common datasets could be shared. Proprietary data could be one of the concerns affecting this area. The other major issue is the unavailability of gold standard lists to perform accuracy and relevancy checks of proposed techniques. It is often observed that these lists are not made public, with a exception of few studies [4]. Regardless of the availability of the gold standard lists, a related question is *How dependable are the gold standard lists in SPRS evaluation since relevance is largely dependent on user perspective?*

3.2 Combination of Evaluation Methods

During the evaluation of the Rec4LRW system, only one of the three tasks was subject to offline evaluation. Since there were different variants of the proposed recommendation technique for the first task, the offline evaluation helped in selecting the best performing technique. This technique was then chosen for implementation in the Rec4LRW system. Hence, offline evaluation helped in selecting a recommendation technique which was subsequently evaluated by the users. In situation where there is feasibility to conduct multi-method evaluation, the question is – *Should the evaluations be conducted in a parallel or serial manner?*

3.3 Considerations for Usability Studies

Usability can be defined as a measure of system use in terms of many dimensions such as effectiveness, efficiency, learnability, safety and enjoyability [7]. Usability studies are generally conducted with participants being closely observed. UI bugs and overall user experience are best measures through such studies. In the case of SPRS, we feel usability studies can be conducted at a stage when the developed system is close to production readiness i.e. usability testing could be the final evaluation method. A valid question in this context would be *What type of data should be collected during usability testing in SPRS evaluation?*

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office.

REFERENCES

[1] Beel, J., Genzmehr, M., Langer, S., Nürnberger, A. and Gipp, B. 2013. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation - RepSys '13* (New York, New York, USA, 2013), 7–14.

[2] Dwork, C., Kumar, R., Naor, M. and Sivakumar, D. 2001. Rank aggregation methods for the web. *Proceedings of the 10th international conference on World Wide Web* (Hong Kong, Hong Kong, 2001), 613–622.

[3] Ekstrand, M.D., Kannan, P., Stemper, J.A., Butler, J.T., Konstan, J.A. and Riedl, J.T. 2010. Automatically Building Research Reading Lists. *Proceedings of the fourth ACM conference on Recommender systems* (New York, New York, USA, 2010), 159–166.

[4] Jardine, J.G. 2014. *Automatically generating reading lists.* University of Cambridge.

[5] Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H. and Newell, C. 2012. Explaining the user experience of recommender systems. *User Modelling and User-Adapted Interaction.* 22, 4–5 (2012), 441–504. DOI:https://doi.org/10.1007/s11257-011-9118-4.

[6] Mcnee, S.M. 2006. *Meeting User Information Needs in Recommender Systems.* University of Minnesota.

[7] Rogers, Y., Sharp, H. and Preece, J. 2011. *Interaction Design: Beyond Human-Computer Interaction.* Wiley.

[8] Saldana, J. 2009. *The Coding Manual for Qualitative Researchers.* SAGE.

[9] Sesagiri Raamkumar, A. 2018. *A task-based scientific paper recommender system for literature review and manuscript preparation.* Nanyang Technological University.

[10] Sesagiri Raamkumar, A., Foo, S. and Pang, N. 2017. Evaluating a threefold intervention framework for assisting researchers in literature review and manuscript preparatory tasks. *Journal of Documentation.* 73, 3 (May 2017), JD-06-2016-0072. DOI:https://doi.org/10.1108/JD-06-2016-0072.

[11] Sesagiri Raamkumar, A., Foo, S. and Pang, N. 2016. Proposing a Scientific Paper Retrieval and Recommender Framework. *Proceedings of International Conference on Asia-Pacific Digital Libraries, ICADL 2016* (Tsukuba, Japan, 2016).

[12] Venkatesh, V. and Bala, H. 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences.* 39, 2 (May 2008), 273–315. DOI:https://doi.org/10.1111/j.1540-5915.2008.00192.x.