Title	Proposing a Scientific Paper Retrieval and Recommender Framework(Main article)
Author(s)	Sesagiri Raamkumar, Aravind; Foo, Schubert; Pang, Natalie
Citation	Sesagiri Raamkumar, A., Foo, S., & Pang, N. (2016). Proposing a Scientific Paper Retrieval and Recommender Framework. Lecture Notes in Computer Science, 10075, 92-97.
Date	2016
URL	http://hdl.handle.net/10220/41670
Rights	© 2016 Springer International Publishing AG. This is the author created version of a work that has been peer reviewed and accepted for publication by International Conference on Asian Digital Libraries, Lecture Notes in Computer Science, Springer. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [http://dx.doi.org/10.1007/978-3-319-49304-6_12].

Proposing a Scientific Paper Retrieval and Recommender Framework

Aravind Sesagiri Raamkumar, Schubert Foo, Natalie Pang

Wee Kim Wee School of Communication and Information,
Nanyang Technological University, Singapore
{aravind002, sfoo, nlspang}@ntu.edu.sg

Abstract. In this paper, we propose a framework that combines aspects of user role modeling and user-interface features with retrieval and recommender systems components. The framework is based on emergent themes identified from participants feedback in a user evaluation study conducted with a prototype assistive system. 119 researchers participated in the study for evaluating the prototype system that provides recommendations for two literature review and one manuscript writing tasks.

Keywords: scientific paper recommender systems, scientific paper retrieval systems, literature review, manuscript writing, user roles, personalization

1 Introduction

Special purpose information retrieval (IR) and recommender systems (RS) implementations have been devised for providing relevant research papers to different LR search and manuscript writing (MW) tasks [1, 2]. Two issues are observed in such implementations: First, the applications are piecemeal approaches thereby forcing the researcher to depend on multiple systems to complete important LR search tasks. Second, there are a wide variety of algorithms and data items used in these studies, making it a difficult proposition for a contextual integration of services. With the aim of addressing these issues, we selected two key LR search tasks and one MW task for developing a system called Rec4LRW [3]. The recommendation techniques of the tasks are based on a set of features that capture the important characteristics of the research paper and the constituent bibliographic references and citations. Along with the traditional metadata fields displayed with the recommended papers, new informational display features were introduced in the system to help the user in making faster and efficient decisions on the relevance and usefulness of retrieved/recommended papers. Using a quantitative and qualitative approach, an evaluation study was carried out with the system. A total of 119 university student and staff participants who had experience in writing research papers participated in the study.

In this paper, we first present the emergent themes derived from the feedback comments of the participants. Secondly, these themes are further utilized for conceptualizing a specialized framework called scientific paper retrieval and recommender

adfa, p. 1, 2011.

Framework (SPRRF). This framework is meant to guide our future studies with the Rec4LRW system and also to help researchers and developers in better designing systems meant for recommending papers. SPRRF integrates elements from user modeling, IR/RS, search user interfaces (SUI) and exploratory search; therefore most of the contextual entities related to a task are reinforced to complement each other.

2 Prototype System and User Evaluation Study

The three tasks offered by the Rec4LRW system are (i) building a 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 preference of the user. In the task screens of the system, new informational display features are included for helping researchers in understanding the uniqueness of the recommended papers. For all the three tasks, information cue labels depicting the paper-type of the recommended paper are displayed. The four labels used are *popular, high reach, survey/review and recent*. An extract from the ACM Digital Library (ACM DL) for the period 1951 to 2011, is used as the corpus of the system. The sample set for the evaluation study was formed by extracting papers with full text and metadata availability in the extract. The final corpus contained a total of 103,739 articles.

A user evaluation study was conducted to determine the usefulness and efficiency levels of the three recommendation tasks and the overall system. An online prescreening survey was conducted to screen the potential participants. A user guide[1] with the necessary instructions was provided to the participants at the start of the study. The evaluation questionnaire in each task was accommodated at the bottom of the screen. The participants had to answer the survey questions and subjective feedback questions as a part of the evaluation. Participants' subjective feedback responses were coded by the corresponding author using an inductive coding style. The aim of the coding exercise was identifying the central themes from the comments of the 119 participants.

3 Emergent Themes from Participants Feedback Data

3.1 Distinct User Groups and Information Cues

Information Systems (IS) across different domains provide content based on the specific role of the user. The role can determine both the display features and the content to be displayed to the user. In industrial and corporate IS, these roles are utilized to enforce security settings simulating the hierarchy of employees. In academic digital libraries, these roles have not been considered extensively even though attempts have been made to classify users based on varying experience levels [1]. This type of classification can be challenged in relation to the task. Conversely, research papers can be

[1] Rec4LRW user guide http://goo.gl/dxUCuk

classified on content-oriented aspects such as quality of research, extent of contribution, article-type and parent discipline. From the participants' feedback, the existence of two user groups was inherently visible. One group required control features in the UI for sorting the recommendations and viewing the articles through topical facets. These participants also gave preferences on the algorithm for retrieving papers as researchers tend to follow distinctive paths to arrive at the required papers. The other group of users was largely satisfied with both the recommendations quality and the ranked display of papers. They were not interested in manipulating the display for achieving alternative rankings. Secondly, they trusted the background algorithms used for the recommendations.

The utility of information cues in positively impacting users' perceptions has been underlined in earlier studies [4]. Rec4LRW's unique informational display features such as the information cue labels enabled the participants to better understand the recommended papers. Apart from the four cue labels from the current design of the Rec4LRW system, more labels indicating the interdisciplinary and article-type aspects of the recommended papers can be introduced. Cue labels appear to be a most promising feature for inclusion of such systems as most participants found them to be most useful.

3.2 Two Types of Serendipity and Algorithms

Serendipitous discovery of research papers is a challenging problem as it is complex to model the interestingness of particular unread papers to researcher's current interests. This problem has been handled before in earlier studies [5]. The approaches from such studies are to be classified under the *forced serendipity* category as the resultant recommendations are based on corresponding models. The alternate way of serendipitously encountering research papers is based on purely un-modelled scenarios. For instance, the 'View Papers in the Parent Cluster' feature in the Rec4LRW system helped participants in noticing papers which they have not read earlier. In addition, it can be stated that *natural serendipity* can be facilitated by incorporating more transparency in the recommendation process.

The recommendation and retrieval algorithms proposed in earlier studies have been predominantly static and fixed. The obvious advantage of fixed algorithms is the validity and reproducibility. Nonetheless, factors such as relevance feedback-based changes and choice of algorithms are to be considered for future systems. These two factors contribute to the fluidity level in algorithms. In the case of the first factor, user's actions and choices dictate future recommendations. For the second factor, users expect a list of appropriate algorithms to be presented to them. Some participants in the study suggested heuristics to identify papers for Task 1 and 2. Providing a list of algorithms is expensive in terms of computational capability as these algorithms need to be optimized for superior performance. Nevertheless, user satisfaction will probably improve with algorithmic independence.

3.3 Inclusion of Control Features and Bibliometric Data

In digital libraries, the importance of control features in UI cannot be overstated as these systems serve as an entry point to the large corpuses of papers. Even though, algorithms help in ranking the top most relevant papers for a user's search requirement, not all users would want to select the papers from the ranked list. During the user evaluation study, it was noticed that many users felt handicapped by the absence of control features such as sorting and advanced search features in the Rec4LRW system. Informational display features in RS mostly do not represent an extensive set of bibliometric data. In traditional digital libraries, the inclusion of this data has become commonplace as users rely on these metrics for relevance judgment. However, in the case of previous RS, only simple metrics such as the citation count and reference count were included. In the user study, participants explicitly stated the need to include metrics such as impact factor and h-index along with the other metadata. The main challenge for including these metrics in the user interface is the computing overhead for calculating these values for all the papers in the corpus. Further exacerbating this issue, most of the prototype systems use different datasets, thereby re-use of metrics data is not a viable option.

3.4 Diversification of Corpus and Task Interconnectivity

The evaluation of algorithms in most of the prior studies has been restricted to datasets from certain disciplines such as computer science and related disciplines. Even though there is large level of uniformity in hard and soft sciences on the approaches followed for scientific information seeking, not much is known about the differences in relevance heuristics for LR tasks. Therefore, future studies should include papers from "far-apart" disciplines for the evaluation. In systems where multiple search tasks are supported, task interconnectivity mechanism is an essential component. With this component, certain redundant user actions can be avoided. In the user study, a good number of participants appreciated the utility of seed basket and reading list towards management of the paper across the three tasks.

4 The Framework

There are three high-level components in the *Scientific Paper Retrieval and Recommender Framework (SPRRF)* as shown in Figure 1.

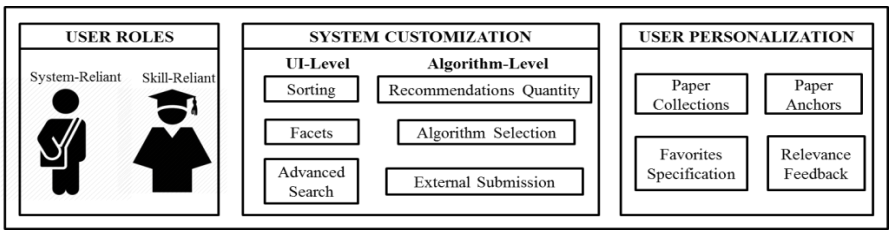

Fig. 1. Scientific Paper Retrieval and Recommender Framework (SPRRF)

4.1 User Roles

Our proposed classification of user-role is based on the levels of system customization and user personalization preferred by the user. The two proposed roles are (i) skill-reliant and (ii) system-reliant users. Skill-reliant user role caters to users who prefer to customize the UI and system-level features to a high extent. These users prefer to have sorting, advanced search and filtering options to sieve through the recommended results. They prefer to have control over algorithm logic and the required quantity of output papers. On the other hand, System-reliant role caters to users who prefer to trust the system in its default settings i.e. fixed algorithm logic, low levels of customization and non-personalized options. Research students mostly fall under this category since they preferred the system to a higher extent [3]. The characteristics of these user roles are influenced by the other two components in SPRRF. Assignment of the role to a new user to the system can be made with a simple selection from the user during the first visit. Accordingly, a user could change the role setting during future visits.

4.2 System Customization

Under this component, there are two sub-components. They are UI and algorithm customizations. The *UI Customization* sub-component involves control features. Although, there are different types of control features, three main features are considered adequate. These are (i) sort options, (ii) topical facets and (iii) advanced search options. Sort options provide alternative schemes such as sorting by publication date, citation count and textual similarity. Topical facets are hyperlinks provided in the navigation pane of the results page. The author-specified keywords from research papers are ideal candidates for topical facets. Advanced search options include more text boxes for executing field-specific search queries which can be combined using Boolean operators. These UI customization features in specialized RS will help in simulating a familiar experience for users who have been using traditional digital libraries. The second level of system customization - *Algorithmic Customization*, is related to the retrieval/recommender algorithm. There are three customization features. These are (i) setting the recommendations count, (ii) selecting the algorithm and (iii) submission of external papers through Bibtex files. In all the previous studies and the current study, the recommendations count has been fixed by the researchers based on different rationale. Nevertheless, users will be benefited with this flexible option of setting recommendations count. On a down side, papers with very low relevance scores could be retrieved if the recommendations count is set high. Certain tasks such as the Task 1 in the current study provide scope for choosing from different algorithms. These algorithms use different rules and information paths for identifying the candidate papers. Hence, the available algorithms could be provided as choices to users for selection. The third feature is the 'upload' option for loading Bibtex files so that similar papers could be found based on the citations in the Bibtex files.

4.3 User Personalization

The extent of user personalization applicable for scientific paper recommendations is limited in comparison with other domains such as e-commerce, films and music. Through the SPRRF, a different perspective of personalization is presented with four features. These are (i) paper collections, (ii) favorites specification, (iii) paper anchors and (iv) relevance feedback. The seed basket and reading list which are already available in the Rec4LRW system are apt paper collection features for enforcing explicit personalization at task level. Anchoring or pinning certain papers in the seed basket or reading list, is the second feature meant for exerting strong influence on recommendations. This helps in acquiring highly personalized results. Alternatively, different weights could be set to the seed papers so that recommendations could be formulated accordingly. User specification of favorites among authors, conferences and journals is the third personalization feature for manipulating recommendations. This feature is set at the user profile level, thereby making these favorites global for all the recommendation tasks carried out by the user. Relevance feedback based re-orientation of recommendations is the fourth feature of user personalization that can really benefit researchers in training the system to their individual tastes.

5 Conclusion

In this paper, we have proposed a specialized framework meant to cater for future studies in scholarly search tasks. A detailed version of the framework with proofs for the emergent themes has been made available as a technical report [6].

References

1. McNee, S.M., Kapoor, N., Konstan, J.A.: Don't Look Stupid : Avoiding Pitfalls when Recommending Research Papers. In: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. pp. 171–180 (2006).
2. Küçüktunç, O., Saule, E., Kaya, K., Çatalyürek, Ü. V: TheAdvisor : A Webservice for Academic Recommendation. In: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. pp. 433–434 (2013).
3. Sesagiri Raamkumar, A., Foo, S., Pang, N.: Making Literature Review and Manuscript Writing Tasks Easier for Novice Researchers through Rec4LRW System. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16. pp. 229–230. ACM Press, New York, New York, USA (2016).
4. Tang, M.-C.: A study of academic library users' decision-making process: a Lens model approach. J. Doc. 65, 938–957 (2009).
5. Sugiyama, K., Kan, M.-Y.: Serendipitous Recommendation for Scholarly Papers. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. pp. 307–310 (2011).
6. Sesagiri Raamkumar, A., Foo, S., Pang, N.: A Framework for Scientific Paper Retrieval and Recommender Systems. (2016). http://arxiv.org/abs/1609.01415.