Journal of Documentation

Evaluating a threefold intervention framework for assisting researchers in literature review and manuscript preparatory tasks

Aravind Sesagiri Raamkumar, Schubert Foo, Natalie Pang,

Evaluating a threefold intervention framework for assisting researchers in literature review and manuscript preparatory tasks

Abstract

Purpose – Systems to support literature review and manuscript preparation tend to focus on only one or two of the tasks involved. This paper describes an intervention framework that redesigns a particular set of tasks, allowing for interconnectivity between the tasks and providing appropriate UI display features for each task in a prototype system.

Design/methodology/approach – A user evaluation study was conducted on the prototype system. The system supports three tasks of (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. A total of 119 researchers who had experience in authoring research papers, participated in the evaluation study. They had to select one of the provided 43 topics and execute the tasks offered by the system. Three questionnaires were provided for evaluating the tasks and system. Both quantitative and qualitative analyses were performed on the collected evaluation data.

Findings – Task redesign aspects had a positive impact in user evaluation for the second task of finding similar papers while improvement was found to be required for the first and third tasks. The task interconnectivity features seed basket and reading list were helpful for the participants in conveniently searching for papers within the system. Two of the four proposed informational display features namely information cue labels and shared co-relations, were the most preferred features of the system. Student user group found the task recommendations and the overall system to be more useful and effective than the staff group.

Originality/value – This study validates the importance of interconnected task design and novel informational display features in accentuating task-based recommendations for literature review and manuscript preparatory tasks. The potential for improvement in recommendations was shown through the task redesign exercise where new requirements for the tasks were identified. The resultant prototype system helps in bridging the gap between novices and experts in terms of literature review skills.

Keywords - *digital library, literature review, scientific paper information retrieval, scientific paper recommender system, manuscript writing, task interconnectivity, task redesign*

Paper type – Research paper

Introduction

The research lifecycle (Nicholas & Rowlands, 2011) encompasses the activities performed by researchers, ranging from identification of research opportunities to management of the overall research process. Information seeking is performed by researchers all through this lifecycle for acquiring information objects such as research topics, scientific papers, books, publication venues, collaborators, to name a few. Models of scientific information seeking such as the Ellis model (Ellis & Haugan, 1997) describe the different macro-level stages of information seeking carried out by researchers. During these stages, researchers primarily search for scientific papers for corresponding information needs since scientific papers are one of the most important and core information objects for researchers. To facilitate easier information seeking, information retrieval (IR) systems provide the necessary formal channels for querying sources where information is structurally organized. IR based systems such as academic search systems, academic databases and citation indices do not provide direct support for all scientific information seeking stages since these systems are not designed for specific LR search tasks. Instead a free-text search interface is provided in these systems for supporting ad-hoc information needs. In addition to these systems, certain reference management systems such as Mendeley (Vargas, Hristakeva, & Jack, 2016) and Docear (Beel, Langer, Genzmehr, & Nürnberger, 2013) provide paper recommendations based on researcher's paper collections. However these recommendations are mainly based on topical similarity with recently viewed papers. Researcher's task-based relevance factors have not been considered for formulating recommendations in these systems.

Research has shown that task-based IR systems are more apt for users as they address specific task requirements (Vakkari, 2001). Since Information seeking and information retrieval are intertwined in the broad context of information behavior (Wilson, 1999), the integrated information seeking and retrieval framework (Ingwersen & Järvelin, 2006) was proposed to consider task characteristics in system design. Nevertheless, there is a scarcity of such systems particularly for assisting researchers in multiple literature review (LR) and manuscript preparatory (MP) search tasks. As an addition to the free-text search engines, studies in information retrieval (IR) and recommender systems (RS) have been conducted to put forth algorithms and systems for supporting these tasks. Such user tasks include (i) building reading list for literature review (Bae, Hwang, Kim, & Faloutsos, 2014; Ekstrand et al., 2010; Jardine, 2014), (ii) finding similar papers for a given paper (Küçüktunç, Saule, Kaya, & Çatalyürek, 2015; Liang, Li, & Qian, 2011), (iii) recommending citations for particular placeholders in manuscripts (He, Kifer, Pei, Mitra, & Giles, 2011; Livne, Gokuladas, Teevan, Dumais, & Adar, 2014), (iv) recommending papers based on activity logs (Liu, Chin-Hui, & Chen, 2012; W. Yang & Lin, 2013), (v) recommending papers based on author publication history (Lee, Lee, & Kim, 2013; Sugiyama & Kan, 2013), to name a few. These approaches/techniques generate paper recommendations for the corresponding tasks.

Support for these aforementioned tasks is generally provided as part of separate systems. Support for specific LR tasks such as "building reading list" and "finding similar papers" need to be provided as constituents of a single system since these tasks are interconnected and incremental in nature. The papers from the former task are inputs to the latter task. In addition, it is to be noted that the user evaluation of the proposed techniques for task recommendations in earlier IR and RS studies, have not been performed in the context of a system where multiple tasks were supported. Therefore, the integration of these approaches towards building a task-based LR and MP assistive system remains untested. Practical implementation of these techniques in digital libraries might be a complex activity as these techniques employ disparate data pre-processing and retrieval/recommendation techniques based on preconceived requirements on the nature of the tasks.

LR search tasks have different requirements as per the stage of LR. Similar to general-purpose information seeking, researchers transition through phases such as pre-focus, problem formulation and post-focus (Vakkari, 2001) while working on an research problem. Different types of papers are required for the series of search tasks in the LR process. It is observed in previous RS studies that recommendation techniques are conceptualized for LR tasks without analysing the complexity of the tasks. There is a need to re-examine the LR tasks requirements in lieu of earlier RS studies, in order to verify whether these tasks provide the best set of papers for researchers. This redesign of LR tasks' requirements can help researchers in getting papers of different types using the required paper discovery mechanisms, in the digital libraries context. By redesign, we allude to changes in the system task corresponding to the actual user task. Task redesign is the first intervention (I1) in our current research on scientific paper retrieval/recommendation systems. Through this intervention, the intent is to improve the recommendations of scientific papers for LR tasks.

The case for a singular system supporting multiple LR tasks has been raised in the past. Scienstein (Gipp, Beel, & Hentschel, 2009) and Papyres (Naak, Hage, & Aïmeur, 2009) are systems which were designed to support LR tasks while systems such as CiteSight (Livne et al., 2014) help in supporting manuscript writing related tasks. Most of these systems are either not available for public usage due to lack of continued support or unavailability of new papers in the corpus. More importantly, they have not been evaluated with actual researchers since the intrinsic algorithms have been largely evaluated using offline mechanisms (Beel, Genzmehr, Langer, Nürnberger, & Gipp, 2013). Systems meant to support multiple LR tasks, must incorporate task interconnectivity features so that papers are managed between the tasks. These features help in simulating the connectivity in the high level information seeking stages in the LR process. Hence, such features are a prerequisite for task-based systems. Therefore, task interconnectivity was chosen as the second intervention (I2) in the current study.

With task redesign (I1) and interconnectivity (I2) as two of the three interventions, our research goal was to build a task-based assistive system to help researchers to perform three related tasks by providing recommendations for these tasks. These tasks include (i) building a reading list of research papers (Task 1), (ii) finding similar papers based on a set of papers (Task 2), and (iii) shortlisting papers from the final reading list for inclusion in manuscript based on article type (Task 3). As a part of the system development, we identified seven base features that represent the relations between a research paper and the citation network. After the task redesign (I1) exercise, we proposed a retrieval technique for Task 1 and recommendation techniques for Tasks 2 and 3, based on the seven identified features. In order to establish interconnectivity (I2) between the three tasks in the system, we implemented two paper collection features known as the seed basket and reading list. These features were meant to be managed by the users for collecting papers. The initial version of the prototype system was designed to provide a single interface for all researchers regardless of experience level, knowledge level and primary discipline.

In information systems, different categories of user interface (UI) features such as input, control, informational and personalisable features are meant to help users in both understanding and adjusting the displayed information in a convenient way (Wilson, 2011). Informational features are important in the UI as users need to be provided with the necessary information for making accurate and timely decisions on the relevance of the displayed information. Relevance is a complex notion as it functions at both system and user levels (Saracevic, 2007). Any informational display feature that can help users in making more effective and efficient relevance judgement decisions will subsequently enhance the usefulness and effectiveness of academic systems. In the assistive prototype system, four informational display features - (i) information cue labels, (ii) shared co-relations, (iii) word cloud of author-specified keywords metadata field and (iv) paper clusters were included in the task screens to highlight the unique aspects of the recommended papers. These features were meant to highlight the benefits of the task redesign exercise. The inclusion of these display features were conceived as the third intervention (I3) in the current study.

We conducted a user evaluation study with 119 researchers for evaluating the tasks and the overall system using multiple evaluation metrics. In this paper, we report the evaluation results about the impact of the three proposed interventions - task redesign (I1), task interconnectivity (I2) and informational display features (I3). Additionally, we sought to identify whether there were differences in the evaluation of student user group in comparison to academic faculty and research staff. The latter user group is collectively referred to as the staff user group in this paper.

The remaining portions of this paper are organized as follows. In section 2, the related studies are presented. The three interventions are introduced in section 3. In section 4, the prototype assistive system is introduced with details about the interface design and corpus used for the evaluation study. The findings from user evaluation study are discussed in section 5. Finally, the conclusions, limitations and future work directions are presented.

Background

Studies have been conducted for recommending papers for LR and MP tasks. Most of these studies have concentrated on a single task instead of focussing on multiple and sequential LR tasks. RS techniques have been put forth for different stages of the research lifecycle for recommending scholarly information objects. The lifecycle schematic put forth in (Nicholas & Rowlands, 2011) enlists the different stages of research. These stages are (i) identify research opportunities, (ii) find collaborators, (iii) secure support, (iv) review the literature, (v) collect research data, (vi) analyse research data, (vii) disseminate findings and (viii) manage the research process. Key RS studies conducted for the researcher tasks in some of these stages are listed in Table I. These studies are not described in detail as they are beyond the scope of this paper

Table I. RS studies for tasks in the stages of research lifecycle

In this sub-section, we will focus on studies that address multiple LR and MP tasks. One of the earliest studies catering to researcher's different LR search tasks, Mcnee (2006) proposed a theoretical model called Human Recommender Interaction (HRI). HRI connects the user dimension to tasks and recommender systems (RS) algorithms. Experience levels of users and facets of seven search tasks were linked to RS metrics. Evaluation results from the study indicated that User-based Collaborative Filtering (UBCF) algorithm provided the best results for most tasks. The main aim of the HRI approach was to map the RS algorithms to the LR tasks through RS metrics. It is observed that the approach does not make use of the characteristics of the tasks for identifying relevant papers. Secondly, the tasks in this study have not been integrated into a single system that can be useful for researchers in carrying out the LR process.

The notion of 'task' was taken as the focal point in a study where papers are recommended based on task profiles of researchers (W. Yang & Lin, 2013). Task profile of a researcher is defined as the set of papers that have been recently accessed in a digital library environment. This study can be categorized under the set of studies meant to find similar papers based on a set of papers. The proposed approach made use of textual and citation network data for formulating recommendations. This study relied on changes in the time dimension for ascertaining shifts in researcher's requirement. Even though the approach is limited to finding similar papers, it can be used at any stage of LR for researchers, for finding unknown papers.

CiteSight (Livne et al., 2014) was a system built to cater for different search needs and expectations of researchers during manuscript writing and certain LR search tasks. The system provided two types of recommendations to the user. The first type was inline recommendations where the system recommends papers for particular citation contexts in the manuscript based on the manuscript's metadata and text around the particular citation context. This recommendation approach has been studied in few earlier works as well (He et al., 2011). The second type was the global recommendations where papers that are broadly similar to the current manuscript are recommended. The system used a cache memory for this purpose. User feedback of this system indicated that participants were of the opinion that the system was able to retrieve fairly relevant research papers, closely matching their expectations. This system is helpful for researchers during manuscript writing but its usability for tasks at the start of the LR is rather unknown.

In another study meant to assist researchers with varying information needs, temporal data was used to gauge the shift in information needs of researchers (Jiang, Liu, & Gao, 2015). This approach was meant to assist researchers during the transition from contemporary to classical research topics published at different time periods. Learning-to-rank models were used on top of textual data in this study to recommend relevant citations. The proposed approach performed better than the baseline graph ranking and textual similarity approaches during evaluation. This study caters to different information needs only at a topical level and might be useful while searching for papers that cater to various related topics for a given research topic.

Based on the review of earlier studies, we identified a number of possible areas of improvement in LR and MP assistive systems. Firstly, the effectiveness of the proposed algorithms/techniques is to be validated with different datasets from computer science and other disciplines. Secondly, the user expectations from the LR and MP tasks which have been addressed in earlier studies are to be re-aligned for providing a better set of papers. Thirdly, the variety of algorithms proposed for LR tasks in disparate studies, lead to an implementation bottleneck in digital libraries where the aim is to provide recommendations for multiple LR tasks. Therefore, a common set of features are to be identified for the recommendation/retrieval generation process. In addition, there is a clear need to establish interconnectivity between the tasks in the systems so that the natural process flow in LR is retained in the systems. The 5S model (Gonçalves, Fox, Watson, & Kipp, 2004) helps practitioners in planning the design and development of generic digital libraries. Similarly, there is a need for a framework/model that facilitates the design of scientific paper recommender systems intended for supporting multiple tasks.

Proposed interventions

Before proposing the interventional mechanisms, the tasks that are addressed through the prototype system, are re-iterated. The tasks considered were (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. The two LR tasks have been handled in prior studies (Ekstrand et al., 2010; Jardine, 2014; Küçüktunç et al., 2015; Liang et al., 2011) and they can be considered as key tasks due to their importance in the LR lifecycle. The third task is a novel task meant to be useful for researchers during manuscript preparation stage. The three interventions are explained as follows.

Intervention 1 (I1): Redesigning the requirements of the identified tasks

In one of the studies conducted for proposing a framework for conceptualizing tasks in information seeking, Li & Belkin (2008) employed a facet-based approach for classifying tasks based on generic and common attributes of tasks. This faceted classification scheme can be used for understanding a given task at a detailed level. However, the scheme is not directly applicable for the current study as the facets pose issues in operationalizing. Instead, a comparatively simple and operational approach was required. Task redesign which is generally performed in the case of work tasks (Ingwersen & Järvelin, 2006), can be achieved through analysis of task complexity. The main aim of this analysis is to categorize tasks into different types using three task features – information needed, process and result (Byström & Järvelin, 1995). This analysis is similar to the business-process approach of breaking up a task into input, processing and output components (Sethi & King, 1998). The three features from the former approach correspond to the input, processing and output components of a task. These components are the most elemental components of any process vis-a-vis task. Therefore, this analysis technique was chosen for redesigning the requirements of the three tasks.

Task 1 - Building an initial reading list. In previous related studies (Bae et al., 2014; Ekstrand et al., 2010; Jardine, 2014; Wang et al., 2010), papers considered as seminal or classical or important in a particular research area, have constituted the reading list. The common characteristic of such papers is the high citation counts. Even though, it is necessary for a researcher to read important papers, such papers may not necessarily provide the overall outlook of the research area. Popular research agendas, methods and contributions can be ascertained from reading these papers. Conversely, a researcher is expected to get a holistic understanding of the research area at the start of LR. Along with popular papers, we proposed that the reading list should constitute of recently published papers and literature survey papers, covering diverse sub-topics in the given research area. The input to the task is the research topic (search keywords). Since four types of papers constitute the reading list, a different processing component was required for building the reading list. We proposed a paper ranking technique (Sesagiri Raamkumar, Foo, & Pang, 2017b) based on a novel coverage value known as the topical and peripheral coverage (TPC) (Sesagiri Raamkumar, Foo, & Pang, 2015). This value is in turn based on author-specified keywords and the citations network. The TPC value is measured by utilizing all the author-specified keywords (typically five keywords) provided in a research paper. The measurement procedure is as follows. The first step is to ascertain the keywords K provided for a paper P_i. The second step is the extraction of all papers in the corpus which have the keywords from K. The base set P_k is formed with this extracted set of papers. The third step is the extraction of the bibliographic references list $reflist_i$ and citations list $citelist_i$ of P_i. The TPC value is measured by counting the number of papers from $reflist_i$ and $citelist_i$ which are present in P_k.

Task 2 - Finding similar papers based on a set of papers. This task has been addressed by many prior studies (Küçüktunç et al., 2015; Liang et al., 2011). However, most of these studies have attempted to propose techniques for measuring similarity between two papers in the corpus. Secondly, the use of either citation network or paper textual content for similarity measurement has been predominant. There have been very few studies which have tried to use both of these data items. Secondly, when researchers execute this task in real-life settings, they intend to find similar papers based on a random set of seed papers. It is to be noted that this set of papers may be homogenous or heterogeneous in terms of inter-relationships i.e. they could be about similar or different topics. In the current study's settings, the input is a set of papers. These input papers can be about different research topics. There is no constraint on the topical similarity of the input papers. The processing component involves use of both citation network and textual content (Sesagiri Raamkumar, Foo, & Pang, 2017a). The output is a set of similar papers.

Task 3 - Shortlisting papers from the final reading list for inclusion in manuscript based on article-type choice. This task is novel in the context of tasks meant to help researchers in manuscript preparation and writing. In prior studies, techniques have been proposed for recommending citations for particular placeholders in manuscripts (He et al., 2011). These placeholders are referred to as citation contexts. However, this task was not considered for this study as it involves recommending papers from the whole corpus regardless of whether the researcher has read the paper or not. Therefore, a novel task was considered. While preparing to write a manuscript, a researcher has to decide which papers are to be cited. This scenario is taken as the third task where the aim is to recommend unique and important papers from the list of papers read during LR. For this purpose, the use of community detection algorithms (Newman, 2006) is apt as these algorithms help in forming clusters in the citation network of a given set of papers. We used the Girvan–Newman algorithm (Girvan & Newman, 2002) for this purpose. The shortlisting technique identifies the top cited papers from the clusters generated by the algorithm (Sesagiri Raamkumar, Foo, & Pang, 2016). The component-wise information for the three tasks is listed in Table II.

Table II. Task requirements redesign components

Intervention 2 (I2): Task interconnectivity

Task interconnectivity is a feature of information systems where the constituent tasks are internally connected by item collections so that retrieved items from one task becomes the input to subsequent task(s). During literature reviews, the individual search tasks are inherently interconnected as researchers alternatively broaden and narrow down their search scope (Levy & Ellis, 2006). This interconnectivity is facilitated by two entities. The first entity is the research topic and the set of related sub-topics. Researchers typically start with a broad topic and subsequently explore multiple sub-topics. The second entity is the set of papers collated during these search tasks. Reference management systems help researchers in managing papers. However, only a few reference management systems provide the option of searching the corpus of academic papers using the researcher's personal collections of papers. Docear (Beel & Langer, 2013) is a system which provides recommendations based on personal collections while Mendeley has recently launched a recommendation engine (Vargas et al., 2016). On the other hand, popular academic search systems and digital libraries provide a free-text search interface meant for finding papers for search keywords or a single input paper. Currently, the online recommendation service

F1000Prime[1] is one of the very few services where paper recommendations are provided for multiple set of papers, thereby providing support for sequential search tasks. Therefore, paper collection features are required in such systems so that researchers could execute different LR tasks without much manual effort. In order to address the aforementioned gap, we introduced two paper collections – *reading list (RL)* and *seed basket (SB)* for the three tasks in our prototype system. Reading list is the list of all papers that are read during literature review. Researcher keeps populating this list as he/she finds new papers relevant to a particular research topic. Reading list (RL) is used across all three tasks in the prototype system. Seed basket (SB) is a basket comprising of a particular set of papers. This set of papers is used to find similar papers as a part of Task 2. SB helps in connecting Task 1 to Task 2 while RL connects all the three tasks in the system.

Intervention 3 (I3): Informational display features

Informational display features in user interfaces (UI) are meant to provide adequate information to the users for making faster judgement on the usefulness and relevance of the retrieved items (Wilson, 2011). In the current study, the use of these features was specifically meant to highlight the new characteristics of the redesigned tasks (refer Table II). The absence of these features would have affected the users since there will be an increased cognitive load in recognizing the relevant documents. The display features included in the prototype system are described below.

Information cue labels. In Task 1 of the prototype system, there are four types of papers (recent, diverse, survey and popular papers) retrieved for the input research topic (refer Table I). If these types are displayed as an informational display feature, it would help the user in differentiating the retrieved papers. Information cues offer the potential to impact the user perception of retrieved results, an observation seen in past studies (Tang, 2009). The usage of cue labels is new to academic search systems although its effectiveness has been proved in other domains (Verbeke & Ward, 2006). For the prototype system, we included four types of labels aside the paper title. These labels are displayed in an automated mechanism based on the paper metadata. The labels and display intent are listed in Table III.

Table III. Information cue labels and their display intents

Word cloud of author-specified keywords. The diversity characteristic in Task 1 is meant to indicate that the reading list comprises of papers covering a wide variety of sub-topics of the input research area. If the sub-topic is displayed as cue label along with the paper title, it might not be useful for researchers who are new to the particular research area. Instead, a word cloud generated with the author-specified keywords of the retrieved papers would be more beneficial for two reasons. First, it provides a snapshot of the coverage of sub-topics within a single feature. Second, it differentiates the more frequent sub-topics from the less frequent topics with the use of font sizes (Verbeke & Ward, 2006).

Shared co-relations with seed basket papers. Current academic search systems such as Google Scholar, provide the 'Related Papers' feature in the search results to help the researchers in accessing topically-similar papers for a particular paper. In an environment

[1] F1000Prime http://f1000.com/prime

(Task 2) where similar papers are discovered for a set of papers, the same feature needs to be extended. We proposed a shared co-relations feature for this purpose. This main feature is split into two separate display features: shared co-references and shared co-citations. These features are meant to help the researchers in understanding the citation overlap between the recommended and seed basket (SB) papers.

Clusters of related papers. In Task 3, the objective was to identify unique and popular papers from the researcher's entire reading list. As described earlier, community detection algorithms were used for identifying clusters of related papers. Subsequently, the top cited papers were recommended from each cluster. In such a scenario, the researcher would be keen to know the parent clusters from which the recommended paper is shortlisted. Therefore, a corresponding informational feature was required for this purpose. The display can be both textual or in the form of a network. In the current study, we displayed the information in textual form.

The interplay between the proposed inventions for the three identified tasks is illustrated in Figure 1.

Figure 1. The three proposed interventions in the context of scientific paper recommendations

Prototype assistive system

Brief overview

The system has been named as Rec4LRW (Recommender System for Literature Review and Writing). The usage sequence of the system is as follows: A researcher runs the first task at the start of the literature review to get an initial reading list of research papers by selecting the provided input research topic (screenshot illustrated in Figure 2). In the second task, researchers select a set of seed papers (facilitated through the seed basket (SB)) from Task 1, to find similar papers (screenshots provided in Figures 3 and 4). The second task is run multiple times in a real world scenario, until the researcher is satisfied with the collected papers. The third task of the system is run at the stage of writing research manuscripts (screenshots provided in Figures 5 and 6). This task helps the researcher in identifying both important and unique papers in the final list of papers read during LR. The papers added to seed basket (SB) are part of the reading list. Before running Task 3, the user can re-run both tasks 1 and 2 for adding more papers to the reading list (RL). In Task 3 recommendations, the count of shortlisted papers is varied as per the article-type preference of the researcher. The task screens in the system have been modified to support the evaluation of the system.

Figure 2. Reading list task screen (Task 1)

Figure 3. Selecting seed papers before executing Task 2

Figure 4. Sample list of recommended papers in Task 2

Figure 5. Input options in Task 3

Figure 6. Sample list of shortlisted papers in Task 3

Corpus

An extract from the ACM Digital Library (ACMDL) is used as the dataset for the user evaluation of the Rec4LRW system. Papers from proceedings and periodicals (journals) for the period 1951 to 2011 form the dataset. The papers were shortlisted based on full text and metadata availability in the dataset, to form the sample-set/corpus for the system. The corpus contains a total of 103,739 articles and corresponding 2,320,345 references. The original data from ACM was received in the form of 4,500 XML files. Data was transferred to a MySQL database to facilitate easier storage, processing and retrieval. The references of papers were parsed using AnyStyle[2] parser for extracting article title and publication year. Apache Lucene and Mahout libraries were used for the IR and RS algorithm implementations.

User evaluation study

Purpose

A user evaluation study was conducted to determine whether researchers using the tasks provided by Rec4LRW system can be efficient and effective in conducting the corresponding real world tasks. In this context, researchers' perceptions of the system features, individual characteristics of the recommended papers and overall quality of the recommendation list were measured. A separate offline evaluation was conducted for the first task since there has been approaches proposed in the past for this task. In this paper, the findings related to the following research questions pertaining to the user evaluation study are reported.

RQ1. Do the three proposed interventions positively impact participants' experience with the tasks and the overall system?

RQ2. Do the student user group find the system functionality more useful and effective than the staff user group?

Participant recruitment

Three communication channels were used for advertising the study. Invitation mails were sent to students and staff of the authors' university. Advertisement posters were put up in notice boards across the university. Invitation mails were sent to mailing lists related to Library and Information Science (LIS) and Information Systems. The main selection criteria was the participant should have authored at least one conference or journal paper. Accordingly, a pre-screening survey was conducted to shortlist the potential participants. The study was conducted from second week of November 2015 to end of January 2016. The Rec4LRW system was made available through the internet so that the user evaluation study could be conducted in a convenient manner for the participants. Participants were permitted to perform the study from any location, with the mandatory condition of completing the study in a single sitting.

Study procedure for participants

The participants were expected to execute all the three tasks as part of the evaluation study. The reading list task (Task 1) was the first task executed by the participant. As a part of this task, the participants had to select a research topic from a list of 43 research topics. The

[2] AnyStyle parsing service http://anystyle.io/

system generated 20 recommendations for Task 1. Before running the similar papers task (Task 2), the participant had to add at least five papers into the seed basket (SB). The minimum number of papers in the SB was set to five as the intention was to highlight the ability of the system in recommending similar papers for multiple papers, a scenario which is procedurally difficult to execute manually and current systems do not provide this functionality. Subsequently, the system provided 30 recommendations for the similar papers task (Task 2). After completing the second task, the participants had to select 25 papers for the reading list (RL) so that the third task could be executed. The 25 papers along with the five papers from the SB are combined to form a total of 30 papers for the final reading list (RL). In the Task 3 screen, the participant had to select the article-type before executing the task. Subsequently, the system shortlists 'N' number of papers from the reading list (RL). The count of shortlisted paper varies with the article-type preference of the participant. The detailed study guide provided to the participants has been made available at this link[3].

Evaluation measures

The participants were required to evaluate the three tasks and the overall system. The evaluation questionnaire for each task was embedded at the bottom of the task recommendations screen. The participants had to answer the mandatory survey questions and optional subjective feedback questions as a part of the evaluation. The survey questions and the corresponding measures relevant to this paper are listed in Table IV. A five-point Likert scale was provided for measuring participant response for each question. The lowest value in the scale was 'Strong Disagree' and the highest was 'Strongly Agree'.

The task-level evaluation measures were meant to record participant perceptions on aspects related to the two interventions task redesign (I1) and task interconnectivity (I2). The impact of third intervention informational display features (I3) on user evaluation of the system has been ascertained through the qualitative feedback provided by participants. A total of seven qualitative questions were provided in the questionnaire for collecting subjective feedback of the participants regarding the three tasks and the system. Participants were requested to provide feedback about the important features and aspects that could be improved. For each task, two questions were provided: (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*. In the final questionnaire where the evaluation for both Task 3 and system was performed together, a single question was posted for eliciting feedback about the overall system. The question was – *"Please provide feedback on the system. You can mention features that you liked in the system and also other features that could be added*."

For the system-level quantitative evaluation, three constructs *Effort to use the system, Perceived system effectiveness* and *Perceived usefulness* were used. The questions for the first two constructs have been adopted from a tested RS user evaluation study (Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012) while the last construct has been adopted from the Technology Acceptance Model (TAM) (Venkatesh & Bala, 2008).

Table IV. Evaluation questions and corresponding measures

Analysis procedures

[3] Rec4LRW evaluation guide for participants http://goo.gl/dxUCuk

In the Likert scale, the values 'Agree' and 'Strongly Agree' were the two values considered for the calculation of agreement percentages in the quantitative evaluation measures. An agreement percentage above 75% was considered as an indication of higher agreement from the participants. Descriptive statistics were used to measure central tendency. Independent samples t-test was used to check the presence of statistically significant difference in the mean values of the students and staff group. Statistical significance was set at $p < .05$. Statistical analyses were done using SPSS 21.0 and R. Qualitative feedback responses for the seven feedback questions were coded using an inductive approach (Thomas, 2006) with the aim of identifying the central themes (concepts) in the participant responses. The coding was performed by one of the authors. A primary theme was identified for each comment while optional secondary themes were identified if the comments comprised of opinions about multiple aspects of the system.

Sample profile

Out of the eligible 230 participants, 138 participants signed the consent form. 119 of them completed the whole study inclusive of the three tasks in the system. The reading list task (first task) was completed by 132 participants while 121 participants completed both the first and second task. The participant demographics breakdown is based on the 132 participants who completed at least the first stage of evaluation i.e. Task 1 evaluation. The number of participants and the corresponding percentages are provided in Table V. 62 participants were graduate research students while the other 70 participants comprised of research staff, academic staff and librarians. The average research experience for students was 2 years while it was 5.6 years for staff. Majority of the participants reported that they had intermediate experience level (46.2%) and a few participants claimed they were beginners (11.4%). Most of the participants were from the engineering disciplines (65.9%) with 39% of the overall participants from the computer science discipline. Library and information studies (LIS) and electrical disciplines were also well represented with 30 participants (22.7%) each.

Table V. Evaluation questions and corresponding measures

Results

RQ1: Do the proposed three interventions positively impact user evaluation of the tasks and the overall system?

Task Redesign (I1). Figure 7 shows the agreement percentages of the 13 measures unique to the three tasks. For all the measures except *Popularity* (58.06% for students, 58.57% for staff) and *Interdisciplinarity* (56.45% for students, 61.43% for staff), students' agreement was higher than staff group in Task 1. Mental categorization of a paper as popular or interdisciplinary is subject to the experience and knowledge level of the participant. For instance, a paper could be considered popular if the citation count is relatively high or if the system displays the popular label alongside the title. However, the citation count calculation was performed internally within the dataset and the value doesn't reflect the real-time citation count of a paper. Participants were made aware of this disparity. An experienced researcher on virtue of his/her exposure to the papers of a particular research area, would be aware of the popular/seminal papers. Therefore, the researcher would be able to identify papers from different disciplines. Among the Task 1 evaluation measures, *Recency* (79.65% for students) and *Diversity* (75.81% for students) had high percentages. Task 1 findings indicate the task requirements redesign resulted in participants identifying more recent and diverse set of papers with few popular and interdisciplinary papers with the students group (70.97%) indicating that the list comprised of a fairly good mix of papers.

For Task 2, the measures *Seedbasket_Similarity* (88.71% for students, 75.41% for staff) and *Good_Spread* (87.10% for students, 73.77% for staff) indicate that participants were able to recognize the similarity of the recommended papers with the seed basket (SB) papers. Secondly, these stats indicate the ability of the recommendation technique in covering a wide variety of sub-topics. Interestingly, there was a conceivable difference between the groups for the measure *Shared_Corelations* (88.71% for students, 75.41% for staff). The difference can be partly explained by the varying expectation levels from the two groups. Experts might have been looking for papers with maximum amount of shared co-relations with SB papers. Even though, the recommendation technique for Task 2 is biased towards papers from the citation networks of SB papers, the final list could have comprised of new papers. These new papers have high textual similarity with the SB papers but without direct citation relations to the SB papers. These two scenarios could have possibly affected the evaluation for the *Shared_Corelations* measure.

For Task 3, the objective of shortlisting important papers from the reading list (RL) was met, based on the evaluation of the measure *Importance* (85.96% for students, 77.97% for staff). However, there was a perceivable tentativeness among the participants for the measure *Certainty* (70.18% for students, 61.02% for staff). *Certainty* measure indicates the participant's preference on compulsorily citing the shortlisted papers. Citation of a paper is subject to the particular citation context in the manuscript, therefore not all participants would be able to predict their citation behavior. 70% agreement from the students group can be considered to be a decent indication of trust on the recommendations. The measure *Shortlisting_Feature* (84.21% for students, 74.58% for staff) indicates that participants wanted to use this type of task on a real-time basis in their research projects.

Task Interconnectivity (I2). Among the two measures employed for recording user responses for the task interconnectivity (I2) intervention, the agreement on *Seedbasket_Usefulness* was quite high (96.77% for students, 95.08% for staff). SB helps the researcher in transitioning from an introductory task (Task 1) to a more incisive task (Task 2) where papers for specific sub-topics could be discovered. Since this feature is not available in most academic search systems and digital libraries, participants found its availability in the prototype system very useful. For the general measure *Task_Interconnectivity*, the staff group (86.44%) seem to prefer both SB and RL in the system more than the students (78.95%). The practical value of these two collection features was apparent since LR tasks require management of papers throughout the LR lifecycle. Since the usefulness of SB and RL has been well established in the current study, there is scope for including a new paper collection mechanism entitled *Publication Purpose*. This mechanism represents the actual publication intent of the researcher i.e. collecting papers for conducting LR for a conference or journal paper or even a dissertation. At any point of time, an experienced researcher would have multiple publication intents for different projects. With this proposed mechanisms, the researcher would be able to maintain multiple seed basket (SB) and reading lists (RL). We believe the inclusion of this mechanism would enable seamless interconnectivity between tasks, thereby simplifying the execution of LR tasks.

Figure 7. Agreement percentages of the evaluation measures

Informational display features (I3). The impact of the novel informational display features on the system evaluation was ascertained qualitatively from the participants' feedback data. In Table VI, the top five preferred features categories for the three tasks are listed. Two of the related categories *Information Cue Labels and Rich Metadata* were present in the top five categories for all the tasks. Information cue labels (popular, high reach, survey/review and recent) helped the participants in quickly identifying the unique papers in the recommendation lists. The usefulness of these labels was indicated in all the three tasks by the participants, thereby making them applicable for any kind of LR search tasks. The usage of cue labels could be extended beyond paper types. For instance, other types of labels such as interdisciplinary, high-impact journal/conference and altmetric score (Costas, Zahedi, & Wouters, 2015) can be added to papers so that the uniqueness of papers in the recommendation list is conveyed visually. Under the *Rich Metadata* category, the metadata fields - abstract, references count and citations count were considered beneficial by the participants. It is to be noted that references count is displayed only in certain systems. The presence of these basic metadata fields is mandatory for such systems. In the user evaluation of prior studies, this set of information was rarely displayed to the participants (Ekstrand et al., 2010; Jardine, 2014; Mcnee, 2006).

In Task 2, the *Shared Co-citations and Co-references* feature (Rank-1 (28%)) was perceived useful as it showed the relations between the recommended papers and the SB papers. This feature similar to the cue labels is novel and is not available in all current search systems. The information in this feature could be further augmented with a visual citation network that places the seed basket papers and the recommended papers as nodes so that the relations are comprehended in a contextual manner, similar to the CitNetExloperer interface (van Eck & Waltman, 2014). In Task 3, the feature of 'viewing papers in the parent cluster' of shortlisted papers was the third most preferred feature since participants found the shortlisting objective of the task and the information cue labels as the prominent favourites. Certain participants indicated that the feature helped them in discovering new papers in a serendipitous manner. In the next release of the system, the information in this feature will be displayed in the form of a network/graph.

In summary, it is clear that two of the proposed display features – information cue labels and shared co-relations were effective in accentuating the uniqueness of the recommended papers. Without these features, participants would have relied solely on the basic metadata fields. Secondly, novice researchers would have found it hard to make relevance judgement decisions on accurately ascertaining the utility of the recommended papers. It was felt that the feature of 'viewing papers in the parent cluster' would have been better received if the information was displayed in the form of citation networks, instead of a tabular representation. The word cloud feature was rarely mentioned in the participant comments. One of the possible reasons is the placement of the feature at the bottom of the screen. Secondly, the word cloud was displayed to the user only when a hyperlink was clicked. Participants skipped this hyperlink, instead they went directly to the evaluation frame in the screen. These issues will be addressed in the next release of the Rec4LRW system. The information in the 'viewing papers in the parent cluster' feature would be displayed as a citation network and the word cloud feature will be made visible to the user, by default

Table VI. Top 5 preferred feature categories

RQ2: Do the student user group find the system functionality more useful and effective than the staff user group?

Results of the independent samples t-test are presented in Table VII, along with the Cronbach alpha values of the three system constructs. For 12 of the 17 measures under the three constructs, there was no statistically significant difference between the two groups. The mean values for the five measures under the *Effort to use the System* construct were in similar range. These values indicate the same level of effort required from the participants, regardless of the group. Under the *Perceived System Effectiveness* construct, there was significant difference for the *Awareness* measure, which can be attributed to the higher experience level of the staff group. Student participants, on the other hand, are less aware of the potentials papers that are to be read during LR. There was significant difference for four out of the six measures under *Perceived Usefulness*. This finding provides a strong evidence of the usability of the system for the student user group as hypothesized, since the difference was the largest for the measures *Performance_Improvability* (M=3.91) and *Ease_Job* (M=3.96). Based on the findings from the agreement percentages and the t-test comparisons, the perception of the system appears to be consistent across the two groups for the effort involved in using the system and its effectiveness. However, the system in its current release was found more useful for graduate students in their research work.

Table VII. Independent samples t-test result

Discussion

The task redesign (I1) intervention basically entails the recommendation techniques of the three tasks. One of the expectations before the study was that the redesigned aspects of the tasks i.e. the novel characteristic(s) corresponding to the requirements of the tasks, would influence the participants more than the generic characteristics. Task 1 results indicate that *Good_Spread*, *Diversity* and *Recency* measures were well received by the student participants. The importance of diversifying recommendations has been raised in previous studies (Küçüktunç et al., 2015) and in Task 1's context, a range of papers would be beneficial for researchers at the start of LR. In addition, recent papers are an interesting case since previous studies (Bae et al., 2014; Ekstrand et al., 2010) have specifically focused on seminal papers. It is to be argued that recent papers are as important as seminal papers since researchers would want to know about the most recent research performed in a particular research area. On the flip side, three issues hampered the user experience in this task - the lack of free text search, limited coverage of the dataset (restricted to ACM DL) and fixed recommendations count. While the former two issues were implications of the study design, the participants' observation on the limited number of recommendations was an interesting case since most of the previous studies offered a maximum of 20 recommendations per task. It can be argued that users should control the number of recommendations, although there is the caveat of recommending low quality papers if the number is set very high.

Task 2 results were the most promising since the differences between the two groups were less. Similar to Task 1, *Good_Spread* and *Diversity* had high agreement percentages, thereby validating the ability of the paper discovery methods of the recommendation technique of Task 2. More importantly, the task was able to retrieve recommendations similar to most of the papers in the SB. Interestingly, many participants felt that the quality can be still improved by incorporating alternative approaches and more input dimensions. It has been observed that this task had the most number of previous studies among all the RS studies. This is due to the observation that it is a very ambiguous task as similar papers could be found on multiple grounds. Perhaps, sub-tasks could be designed for this task where each sub-task is assigned a fixed goal (e.g. find the most matching papers from the citation network of a particular paper). Notwithstanding, the incorporation of semantic textual similarity methods (Han, Kashyap, Finin, Mayfield, & Weese, 2012) in the future version of the technique can further improve the quality of recommendations. Task 3 results were promising in lieu of the task's novelty in scholarly paper recommendation studies, particularly at the stage of recommending papers during manuscript preparation. Even though, participants' responses indicated the shortlisted papers were important papers, they were still unsure about citing these papers in their manuscripts. This observation perhaps highlights the limitation of the task since the citing behavior of researchers is very much based on their personal context. Erikson & Erlandson (2014) had identified that researchers mostly cite papers for supporting their claims and for evidence purposes. Therefore, citation context based recommendations (He et al., 2011) would be more suitable for researchers. However, this task was deemed to be useful in a situation where the reading list of papers collected during the LR is a big list as finding unique and important papers from this list would be a manually complex task. The positive responses from the participants for the shortlisting feature is an encouraging sign as many participants explicitly stated the usefulness of this task if incorporated in current academic search systems. Overall, the influence of the task redesign intervention (I1) on the Rec4LRW system was positively highlighted by the participants albeit with less intensity by expert researchers (staff).

The task interconnectivity mechanisms in the Rec4LRW system basically ensure that there is connectivity established across the three tasks through the paper collections the seed basket (SB) and the reading list (RL). The study results indicate an overwhelming support for these features in terms of usability and effectiveness. The results mirror the success of the recently developed SearchAssist system (Huurdeman, Wilson, & Kamps, 2016) where the impact of search user interface (SUI) features were evaluated for different sub-tasks. It has been highlighted that current academic search systems and digital libraries do not provide the necessary interface for managing different types of LR search tasks (Du & Evans, 2011). Therefore, the positive feedback for these two features was on expected lines. Based on this study, future systems which are supposed to handle multiple LR search tasks would be best served if SB and RL are incorporated as mandatory features.

Certain novel informational display features in the Rec4LRW system were included for highlighting the redesigned task aspects, corresponding to I1. Similar to the positive responses for I2, this intervention was also well received by the participants. The effectiveness of these features was ascertained only through the subjective feedback of the participants. The information cue labels were deemed useful by the participants since these labels expedited the relevance judgment decision. The impact of labels can be attributed to Zipf's principle of least effort (Zipf, 1949) since the cognitive load on the participants was minimal. The other two most appreciated features: shared co-relations and 'view papers in parent cluster' were effective in their respective purposes. The former showed the relations between the recommended papers and the seed basket papers (in Task 2) while the latter showed why a particular paper has been shortlisted (in Task 3). These display features served as a cognitive bridge between the recommendations and the participants while highlighting the usefulness of informational features in user-interfaces (M. L. Wilson, 2011). These findings lend support to existing literature in recognizing the importance placed on UI display features since it is vital to properly highlight the recommended resources to the users.

Limitations

There are certain limitations with the proposed techniques and the user evaluation study. The retrieval technique in Task 1 takes the research topic in the form of the search keywords as input. Being the single mode of input, it could be argued that seed papers as additional input could have also been included. In the case of research topics with other alternative terms, the current study does not consider the alternative terms for the input. This scenario could have caused certain relevant papers to be missed during retrieval. In Task 2, participants were requested to include at least five papers in the seed basket. Some participants indicated that they should have been allowed to select lower number of papers in the seed basket. This is a minor limitation as our intent was to showcase the ability of Task 2 in using multiple papers for formulating recommendations. The latest papers in the dataset used in the study were published in 2011. Few participants indicated that they expected to see recently published papers. This issue could not be avoided due to the conditions in using the dataset. Similar to earlier studies, the current study also assumes that the citations of papers are treated to be equal. However, it has been shown that number of in-paper citations of a reference is critical in ascertaining the influence of the reference on the paper (Zhu, Turney, Lemire, & Vellino, 2015). In-paper citation counts have not been considered in the current study.

Conclusions

In the contemporary academic setup, the plethora of scientific information sources available to researchers, have made the task of searching relevant research papers for LR into an complex activity. Even though, scientific information seeking models (Ellis & Haugan, 1997) have highlighted the intrinsic interconnectivity in LR search tasks, the current academic systems' design is mainly suitable for meeting ad-hoc information needs. The management of papers along with the establishment of connectivity between LR tasks is left to be handled by researchers. Prior studies have proposed techniques and algorithms for recommending papers for different LR search tasks. The integrated implementation of these disparate solutions for multiple tasks in digital libraries appears to be a difficult proposition. Secondly, some of these tasks' requirements need to be changed so that better recommendation/retrieval techniques could be conceptualized to satisfy researchers. In this paper, we attempted to address these challenges and propose a prototype task-based assistive system for providing recommendations for a selected set of literature review and manuscript preparatory tasks. As a part of this system, we have incorporated three interventions to address the aforementioned issues. Using the "input-processing-output" model (Sethi & King, 1998), we redesigned the requirements of three selected tasks. Task interconnectivity was established using two paper collection features. Thirdly, we introduced novel informational display features for helping researchers in making faster relevance judgement decisions.

A user evaluation study was conducted with 119 researchers for evaluating the effectiveness and usefulness of the individual tasks and the overall system. For the task redesign intervention (I1), findings from Task 1 show that researchers found the presence of recent and a diverse set of papers covering sub-topics, to be the most explicit among all paper types. In Task 2, researchers found a good spread of recommended papers that were topically-similar to the papers in the seed basket. In Task 3, participants indicated that albeit the shortlisted papers were important papers in their reading list, they were not certain on whether those papers will be cited in their manuscripts. For the task interconnectivity intervention (I2), the usefulness and effectiveness of the reading list and seed basket features was validated with very high agreement from the participants. For the third intervention (I3), the information cue labels and the shared co-relations features from Tasks 1 and 2 were found to be most useful. From the results, it is clear that the interventions I2 and I3 had an overtly positive impact on the system evaluation while the I1 intervention worked well for Task 2.

In the system-level evaluation, the constructs *Perceived Effectiveness* and *Perceived Usefulness* differentiated the two user groups with the student group rating the system to be more useful and effective in their work. The simple UI design and features of the system was vindicated by the agreement from both groups for the construct *Effort to use the System*. From the results of the study, it is apparent that student user group prefer the system more than staff user group in its current form, they find it to be useful for their work.

Implications and future work

There are various implications of the current study. As part of theoretical implications for future task-based LR systems, the proposed threefold intervention framework serves as a base framework. Systems concentrating on isolated elements such as algorithm and display features provide incomplete user experience and insights as all these factors are responsible for providing the best expected results. In addition, the lack of system-based connectivity between tasks, fail to mirror the inherent connectivity in real-world settings. As a part of research implications, the user evaluation study results show that students prefer this type of system more than staff. It can be stated that future studies on student-staff or novice-expert comparison can be conducted as longitudinal studies to observe changes in usage patterns and perception of usefulness and other evaluation measures. The long-term objective is to evolve the prototype system into an adaptive system which can produce recommendations for different user groups with suitable variations in algorithms and user-interface features. In terms of practical implications, digital libraries and academic search systems can readily make use of the proposed recommendation techniques and the novel informational display features. The paper collection features - seed basket and reading list can be integrated with user identity in these systems so that the paper management feature is made available.

As a part of future work, a series of changes are planned to be implemented in the system. These changes are based on the feedback provided by the participants of the user evaluation study. The major changes include (i) role-based access for enforcing conditional display of certain customization/control features in the system, (ii) inclusion of more datasets in the corpus, (iii) improvement in the recommendation techniques of the second and third tasks, (iv) incorporation of grey literature articles in the recommendations lists as the explicit need and a boosting mechanism has already been identified (Raamkumar, Foo, & Pang, 2015), and (v) inclusion of relevance feedback measures. Such changes are anticipated to provide users with a better personalised experience and generate better recommendations.

References

Bae, D.-H., Hwang, S.-M., Kim, S.-W., & Faloutsos, C. (2014). On Constructing Seminal Paper Genealogy. *IEEE Transactions on Cybernetics*, *44*(1), 54–65. http://doi.org/10.1109/TCYB.2013.2246565

Beel, J., Genzmehr, M., Langer, S., Nürnberger, A., & Gipp, B. (2013). A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation - RepSys '13* (pp. 7–14). New York, New York, USA: ACM Press. http://doi.org/10.1145/2532508.2532511

Beel, J., & Langer, S. (2013). *A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research Paper Recommender Systems* (Vol. xx). Retrieved from http://docear.org/papers/A Comparison of Offline Evaluations, Online Evaluations, and User Studies.pdf

Beel, J., Langer, S., Genzmehr, M., & Nürnberger, A. (2013). Introducing Docear's Research Paper Recommender System. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Beierle, F., Tan, J., & Grunert, K. (2016). Analyzing Social Relations for Recommending Academic Conferences. *Proceedings of the 8th ACM International Workshop on Hot Topics in Planet-Scale mObile Computing and Online Social neTworking - HotPOST '16*, 37–42. http://doi.org/10.1145/2944789.2944871

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, *31*(2), 191–213. http://doi.org/10.1016/0306-4573(95)80035-R

Chen, Z., Xia, F., Jiang, H., Liu, H., & Zhang, J. (2015). Aver: Random Walk Based Academic Venue Recommendations. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 579–584). http://doi.org/10.1145/2740908.2741738

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "Altmetrics" Correlate With Citations? Extensive Comparison of Altmetric Indicators With Citations From a Multidisciplinary Perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019. http://doi.org/10.1002/asi

Du, J. T., & Evans, N. (2011). Academic Users' Information Searching on Research Topics: Characteristics of Research Tasks and Search Strategies. *The Journal of Academic Librarianship*, *37*(4), 299–306. http://doi.org/10.1016/j.acalib.2011.04.003

Ekstrand, M. D., Kannan, P., Stemper, J. A., Butler, J. T., Konstan, J. A., & Riedl, J. T. (2010). Automatically Building Research Reading Lists. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 159–166). New York, New York, USA: ACM Press. http://doi.org/10.1145/1864708.1864740

Ellis, D., & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, *53*(4), 384–403. Retrieved from http://www.emeraldinsight.com/journals.htm?articleid=864058&show=abstract

Erikson, M. G., & Erlandson, P. (2014). A taxonomy of motives to cite. *Social Studies of Science*, *44*(4), 625–637. http://doi.org/10.1177/0306312714522871

Gipp, B., Beel, J., & Hentschel, C. (2009). Scienstein : A Research Paper Recommender System. In *Scienstein : A Research Paper Recommender System* (pp. 309–315).

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*. National Acad Sciences.

Gonçalves, M. A., Fox, E. A., Watson, L. T., & Kipp, N. A. (2004). Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Transactions on Information Systems*, *22*(2), 270–312. http://doi.org/10.1145/984321.984325

Gunawardena, S. D., & Weber, R. (2009). Discovering Patterns of Collaboration for Recommendation. In *Proceedings 22th International FLAIRS Conference, FLAIRS'09. AAAI Press, Menlo Park, California.* Retrieved from http://www.aaai.org/ocs/index.php/FLAIRS/2009/paper/download/121</315\nhttp://hdl.handle.net/1860/4145

Han, L., Kashyap, A., Finin, T., Mayfield, J., & Weese, J. (2012). UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics* (pp. 16–33). Retrieved from http://ebiquity.umbc.edu/papers/select/person/tim/finin/

He, Q., Kifer, D., Pei, J., Mitra, P., & Giles, C. L. (2011). Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11* (p. 755). New York, New York, USA: ACM Press. http://doi.org/10.1145/1935826.1935926

Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C. L. (2015). A Neural Probabilistic Model for Context Based Citation Recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 2404–2410). AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=2886521.2886655

Huang, W., Wu, Z., Mitra, P., & Giles, C. L. (2014). RefSeer : A Citation Recommendation System. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on* (pp. 371–374).

Huurdeman, H. C., Wilson, M. L., & Kamps, J. (2016). Active and Passive Utility of Search Interface Features in Different Information Seeking Task Stages. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR '16* (pp. 3–12). New York, New York, USA: ACM Press. http://doi.org/10.1145/2854946.2854957

Ingwersen, P., & Järvelin, K. (2006). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer Science & Business Media. Retrieved from https://books.google.com/books?hl=en&lr=&id=zuptrfHytHMC&pgis=1

Jardine, J. G. (2014). *Automatically generating reading lists*. Retrieved from https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-848.pdf

Jiang, Z., Liu, X., & Gao, L. (2015). Chronological Citation Recommendation with Information-Need Shifting. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1291–1300). http://doi.org/10.1145/2806416.2806567

Kim, Y., Seo, J., Croft, W. B., & Smith, D. A. (2014). Automatic suggestion of phrasal-concept queries for literature search. *Information Processing and Management*, *50*(4), 568–583. http://doi.org/10.1016/j.ipm.2014.03.003

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

Küçüktunç, O., Saule, E., Kaya, K., & Çatalyürek, Ü. V. (2013). Result Diversification in Automatic Citation Recommendation, 1–4.

Küçüktunç, O., Saule, E., Kaya, K., & Çatalyürek, Ü. V. (2015). Diversifying citation recommendations. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(4), 55:1–55:21. http://doi.org/10.1145/2668106

Lee, J., Lee, K., & Kim, J. G. (2013). *Personalized Academic Research Paper Recommendation System∗*.

Levy, Y., & Ellis, T. J. (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science: International Journal of an Emerging Transdiscipline*, *9*(1), 181–212. Retrieved from http://www.inform.nu/Articles/Vol9/V9p181-212Levy99.pdf

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, *44*(6), 1822–1837. http://doi.org/10.1016/j.ipm.2008.07.005

Liang, Y., Li, Q., & Qian, T. (2011). Finding Relevant Papers Based on Citation Relations. *Web-Age Information Management*, *6897*, 403–414.

Liu, D., Chin-Hui, L., & Chen, Y.-T. (2012). Document Recommendations Based on Knowledge Flows: A Hybrid of Personalized and Group-Based Approaches. *Journal of the American Society for Information Science and Technology*, *63*(10), 2100–2117. http://doi.org/10.1002/asi

Livne, A., Gokuladas, V., Teevan, J., Dumais, S. T., & Adar, E. (2014). CiteSight : Supporting Contextual Citation Recommendation Using Differential Search. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*.

Mcnee, S. M. (2006). *Meeting User Information Needs in Recommender Systems*. Proquest. Retrieved from http://search.proquest.com/docview/305306133

Naak, A., Hage, H., & Aïmeur, E. (2009). A Multi-criteria Collaborative Filtering Approach for Research Paper Recommendation in Papyres. *E-Technologies: Innovation in an Open World*, 25–39.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–82. http://doi.org/10.1073/pnas.0601602103

Nicholas, D., & Rowlands, I. (2011). Social media use in the research workflow. *Information Services & Use*, *31*(1-2), 61–83. http://doi.org/10.3233/ISU-2011-0623

Pan, L., Dai, X., Huang, S., & Chen, J. (2015). Academic Paper Recommendation Based on Heterogeneous Graph. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, *9427*, 381–392. http://doi.org/10.1007/978-3-642-41491-6

Raamkumar, A. S., Foo, S., & Pang, N. (2015). More Than Just Black and White: A Case for Grey Literature References in Scientific Paper Information Retrieval Systems. In *Digital Libraries: Providing Quality Information* (pp. 252–257). Springer.

Saracevic, T. (2007). Relevance : A Review of the Literature and a Framework for Thinking on the Notion in Information Science . Part II : Nature and Manifestations of Relevance *. *Journal of the American Society for Information Science and Technology*, *30*(October), 1915–1933. http://doi.org/10.1002/asi

Sesagiri Raamkumar, A., Foo, S., & Pang, N. (2015). Comparison of techniques for measuring research coverage of scientific papers: A case study. In *The Tenth International Conference on Digital Information Management (ICDIM)* (pp. 132–137). http://doi.org/10.1109/ICDIM.2015.7381881

Sesagiri Raamkumar, A., Foo, S., & Pang, N. (2016). What papers should I cite from my reading list? User evaluation of a manuscript preparatory assistive task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016)* (pp. 51–62). Retrieved from http://ceur-ws.org/Vol-1610/paper7.pdf

Sesagiri Raamkumar, A., Foo, S., & Pang, N. (2017a). *Can I have more of these please? Assisting researchers in finding similar research papers from a seed basket of papers*. Manuscript submitted for publication.

Sesagiri Raamkumar, A., Foo, S., & Pang, N. (2017b). *Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems*. Manuscript submitted for publication.

Sethi, V., & King, W. R. (1998). *Organizational Transformation Through Business Process Reengineering: Applying the Lessons Learned*. Prentice Hall. Retrieved from https://books.google.com.sg/books/about/Organizational_Transformation_Through_Bu.html?id=a_bkAAAAMAAJ&pgis=1

Sie, R. L. L., van Engelen, B. J., Bitter-Rijpkema, M., & Sloep, P. B. (2014). COCOON CORE: CO-author REcommendations Based on Betweenness Centrality and Interest Similarity. In *Recommender Systems for Technology Enhanced Learning* (pp. 267–282). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4939-0530-0_13

Sugiyama, K., & Kan, M.-Y. (2013). Exploiting Potential Citation Papers in Scholarly Paper Recommendation. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13* (p. 153). New York, New York, USA: ACM Press. http://doi.org/10.1145/2467696.2467701

Tang, M.-C. (2009). A study of academic library users' decision-making process: a Lens model approach. *Journal of Documentation*, *65*(6), 938–957.

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, *27*(2), 237–246. http://doi.org/10.1177/1098214005283748

Vakkari, P. (2001). A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of Documentation*, *57*(1), 44–60. http://doi.org/10.1108/EUM0000000007075

Van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823. http://doi.org/10.1016/j.joi.2014.07.006

Vargas, S., Hristakeva, M., & Jack, K. (2016). Mendeley: Recommendations for Researchers. In *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16* (pp. 365–365). New York, New York, USA: ACM Press. http://doi.org/10.1145/2959100.2959116

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

Verbeke, W., & Ward, R. W. (2006). Consumer interest in information cues denoting quality, traceability and origin: An application of ordered probit models to beef labels. *Food Quality and Preference*, *17*(6), 453–467. http://doi.org/10.1016/j.foodqual.2005.05.010

Wang, Y., Zhai, E., Hu, J., & Chen, Z. (2010). Claper: Recommend classical papers to beginners. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)* (pp. 2777–2781). IEEE. http://doi.org/10.1109/FSKD.2010.5569227

Wilson, M. L. (2011). Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *3*(3), 1–143. http://doi.org/10.2200/S00371ED1V01Y201111ICR020

Wilson, T. (1999). Models in information behaviour research. *Journal of Documentation*.

Yang, W., & Lin, Y.-R. (2013). A task-focused literature recommender system for digital libraries. *Online Information Review*, *37*(4), 581–601.

Yang, Z., & Davison, B. D. (2012). Venue Recommendation: Submitting your Paper with Style. *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, *1*, 681–686. http://doi.org/10.1109/ICMLA.2012.127

Zarrinkalam, F., & Kahani, M. (2013). SemCiR: A citation recommendation system based on a novel semantic distance measure. *Program: Electronic Library and Information Systems*, *47*(1), 92–112. http://doi.org/10.1108/00330331311296320

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427. http://doi.org/10.1002/asi.23179

Zipf, G. (1949). Human Behavior and the Principle of Least Effort. In *Ed: Addison-Weslay*. Cambridge: MA: Addison-Wesley.

Table I. RS studies for tasks in the stages of research lifecycle

Stage/Task	Key Studies
Identify research opportunities	(Mcnee, 2006)
Find collaborators	(Gunawardena & Weber, 2009; Sie, van Engelen, Bitter-Rijpkema, & Sloep, 2014)
Review the literature	
Task of building an initial reading list at the start of LR	(Bae, Hwang, Kim, & Faloutsos, 2014; Ekstrand et al., 2010; Jardine, 2014)
Task of finding similar papers based on a single paper	(Liang, Li, & Qian, 2011; Pan, Dai, Huang, & Chen, 2015)
Task of finding similar papers based on multiple papers	(Küçüktunç, Saule, Kaya, & Çatalyürek, 2013, 2015)
Task of searching papers based on input text	(Huang, Wu, Mitra, & Giles, 2014; Kim, Seo, Croft, & Smith, 2014; Zarrinkalam & Kahani, 2013)
Disseminate Findings	
Publication Venues	(Beierle, Tan, & Grunert, 2016; Chen, Xia, Jiang, Liu, & Zhang, 2015; Yang & Davison, 2012)
Citation Contexts	(He, Kifer, Pei, Mitra, & Giles, 2011; Huang, Wu, Liang, Mitra, & Giles, 2015)

Table II. Task requirements redesign components

Task	Component	Earlier Studies	Current Study
Task 1 - Building an initial reading list	Input	Research Topics or Seed Papers	Research Topics
	Processing	Usage of citation network and paper content	Usage of citation network through author-specified keywords
	Output	Seminal or Popular papers	Recent, Diverse, Survey and Popular papers
Task 2 - Finding similar papers based on a set of papers	Input	Single paper	Multiple papers
	Processing Method	Usage of either citation network or paper content	Usage of both citation network and paper content
	Output	Similar papers	Similar papers
Task 3 - Shortlisting papers from the final reading list for inclusion in manuscript based on article-type choice	Input	Not Applicable	Full list of papers
	Processing	Not Applicable	Usage of citation network
	Output	Not Applicable	Unique and highly cited papers

Table III. Information cue labels and their display intents

Label	Display Intent
Recent	Indicate recently published papers
Popular	Indicate the paper is highly cited for the input research topic
High Reach	Indicate the paper has a high number of references in its bibliography
Survey/Review	Indicate literature survey or review papers

Table IV. Evaluation questions and corresponding measures

Measure/Construct	Question	Related Task/ System
	Task Redesign (I1)	
Popularity	The recommendation list consists of papers that appear to be popular papers for the research topic	Task 1

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	
Seedbasket_Similarity	The recommendation list consists of papers that are similar to the papers in the seed basket	Task 2
Shared_Corelations	The recommendation list consists of papers that have shared co-references and co-citations with the papers in the seed basket	
Good_Spread	The recommendation list consists of a good spread of papers for the research topic	
Importance	The shortlisted papers comprises of important papers from my reading list	Task 3
Certainty	The shortlisted list comprises of papers which I would definitely cite in my manuscript	
Task Interconnectivity (I2)		
Seedbasket_Usefulness	The feature of adding papers to the seed basket to generate similar paper recommendations is a useful feature	Task 2
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	Task 3
System Evaluation		
Effort to use the System	Construct comprising of five questions on the effort required from the participants to use the system*	System
Perceived System Effectiveness	Construct comprising of six questions on the perceptions of effectiveness of the system*	
Perceived Usefulness	Construct comprising of six questions on the perceptions of usefulness of the system*	

* - Questions listed in Table V

Table V. Demographic variables from the study

Demographic Variable	Number of Participants
Position	
Student	62 (47%)
Staff	70 (53%)
Experience Level	
Beginner	15 (11.4%)
Intermediate	61 (46.2%)
Advanced	34 (25.8%)
Expert	22 (16.7%)
Discipline Category	
Engineering & Technology	87 (65.9%)
Social Sciences	42 (31.8%)
Life Sciences & Medicine	3 (2.3%)
Discipline	
Computer Science & Information Systems	51 (38.6%)
Library and Information Studies	30 (22.7%)
Electrical & Electronic Engineering	30 (22.7%)
Communication & Media Studies	8 (6.1%)
Mechanical, Aeronautical & Manufacturing Engineering	5 (3.8%)
Biological Sciences	2 (1.5%)
Statistics & Operational Research	1 (0.8%)
Education	1 (0.8%)
Politics & International Studies	1 (0.8%)
Economics & Econometrics	1 (0.8%)
Civil & Structural Engineering	1 (0.8%)

Psychology	1 (0.8%)

Table VI. Top 5 preferred feature categories

Rank	Task 1	Task 2	Task 3
1	Information Cue Labels (41%)	Shared Co-citations & Co-references (28%)	Shortlisting Feature & Rec. Quality (24%)
2	Rich Metadata (21%)	Rec. Quality (27%)	Information Cue Labels (15%)
3	Diversity of Papers (13%)	Information Cue Labels (16%)	View Papers in Clusters (11%)
4	Rec. Quality (9%)	Seed Basket (14%)	Rich Metadata (7%)
5	Recency of Papers (4%)	Rich Metadata (9%)	Ranking of Papers (3%)

Table VII. Independent samples t-test results

Construct	α	Question [Measure]	t	M (SD)		*p*
				Students	Staff	
Effort to use the System	0.659	The system is convenient [Convenience]	0.24	3.98 (0.719)	3.95 (0.775)	0.406
		I have to invest a lot of effort in the system [Effort_Required]	-0.326	2.68 (0.909)	2.75 (1.108)	0.372
		It takes many mouse-clicks to use the system [Mouse_Clicks]	0.282	2.75 (1.138)	2.69 (1.133)	0.389
		Using the system takes little time [Little_Time]	0.009	3.54 (0.867)	3.54 (0.916)	0.496
		It takes too much time before the system provides adequate recommendations [Much_Time]	-0.536	2.67 (1.041)	2.78 (1.219)	0.296
Perceived System Effectiveness	0.807	I would recommend the system to others [Recommend]	-0.057	3.74 (0.745)	3.75 (0.921)	0.477
		Using the system is a pleasant experience [Pleasant_Experience]	0.196	3.86 (0.743)	3.83 (0.854)	0.423
		The system is useless [Useless]	-0.294	2.07 (0.884)	2.12 (0.892)	0.385
		The system makes me more aware of my choice options [Awareness]	2.433	3.96 (0.597)	3.64 (0.804)	**0.008**
		I make better choices with the system [Better_Choice]	1.426	3.72 (0.675)	3.51 (0.898)	0.078
		I can find better papers by using the system [Findability]	0.787	3.65 (0.79)	3.53 (0.897)	0.217
Perceived Usefulness	0.951	Using the system would enable me to accomplish tasks more quickly [Accomplish_Tasks]	1.665	3.89 (0.646)	3.64 (0.943)	**0.049**
		Using the system would improve my work performance [Performance_Improvability]	1.954	3.91 (0.635)	3.63 (0.908)	**0.027**
		Using the system would improve my productivity [Productivity_Improvability]	1.84	3.77 (0.756)	3.49 (0.878)	**0.034**
		Using the system would enhance my effectiveness on the work [Enhance_Effectiveness]	1.063	3.74 (0.669)	3.58 (0.932)	0.145
		Using the system would make it easier to do my job [Ease_Job]	2.253	3.96 (0.706)	3.61 (0.965)	**0.013**
		I would find the system useful in my work [Work_Usefulness]	1.399	3.96 (0.801)	3.75 (0.883)	0.082

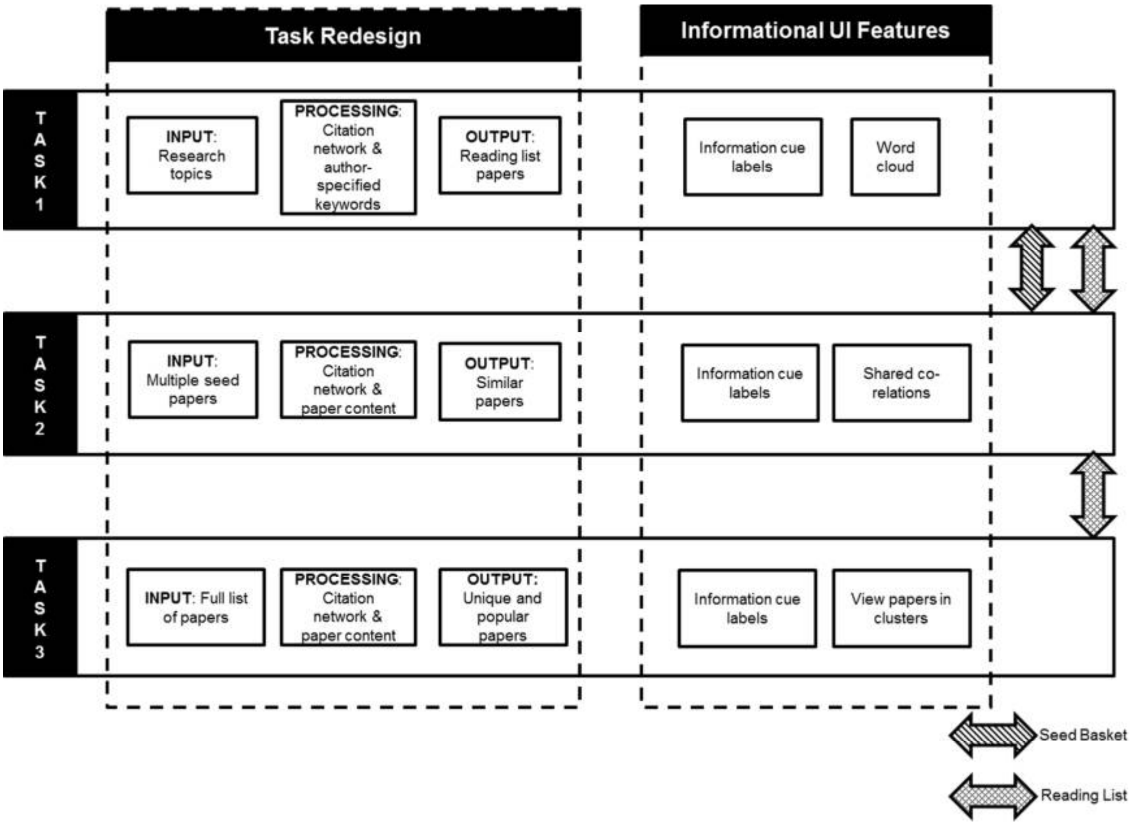

Figure 1. The three proposed interventions in the context of scientific paper recommendations

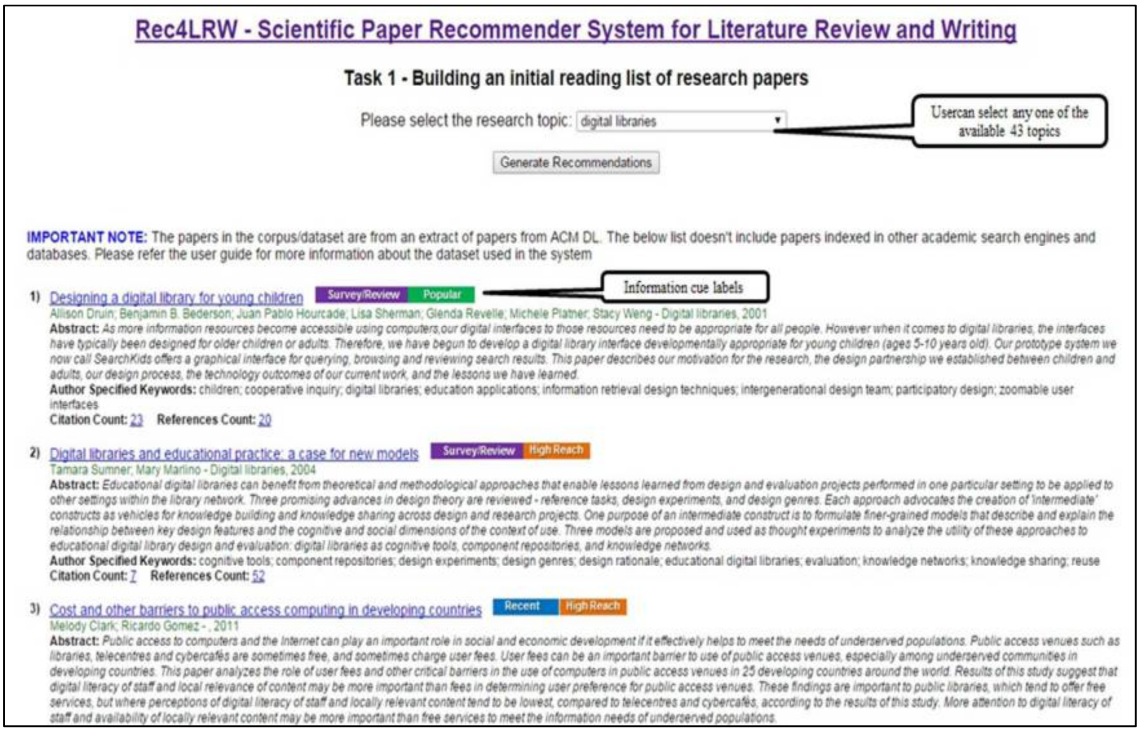

Figure 2. Reading list task screen (Task 1)

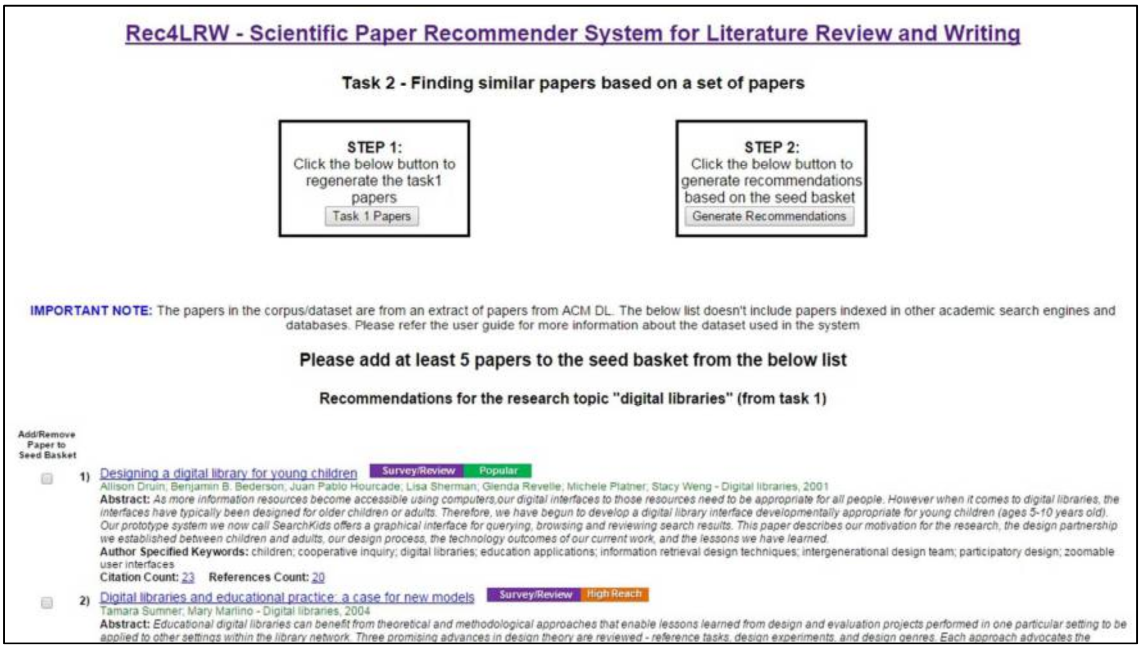

Figure 3. Selecting seed papers before executing Task 2

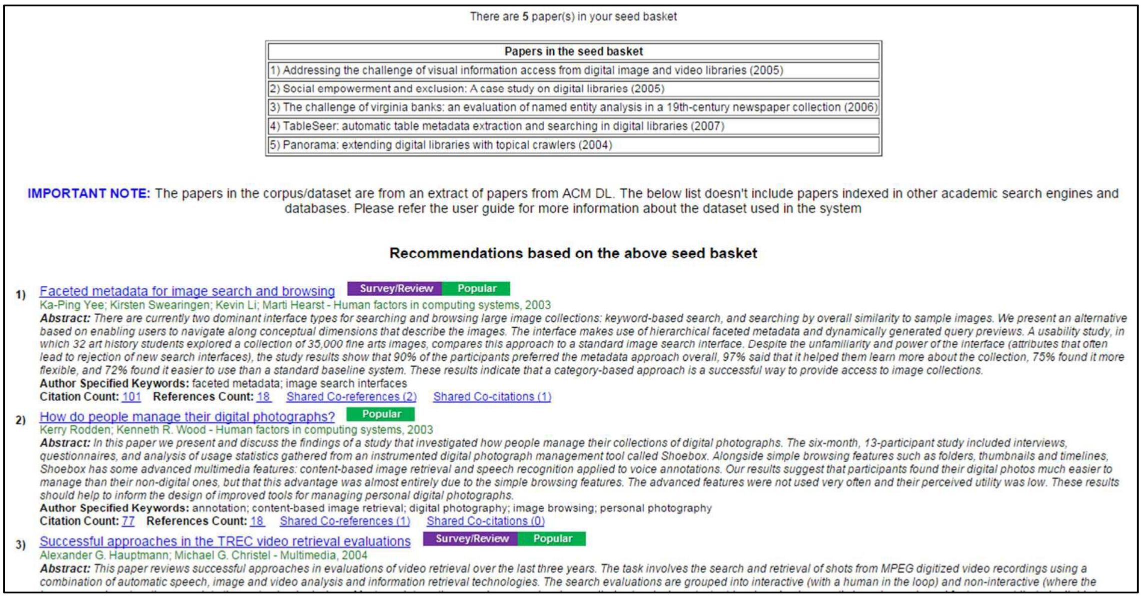

Figure 4. Sample list of recommended papers in Task 2

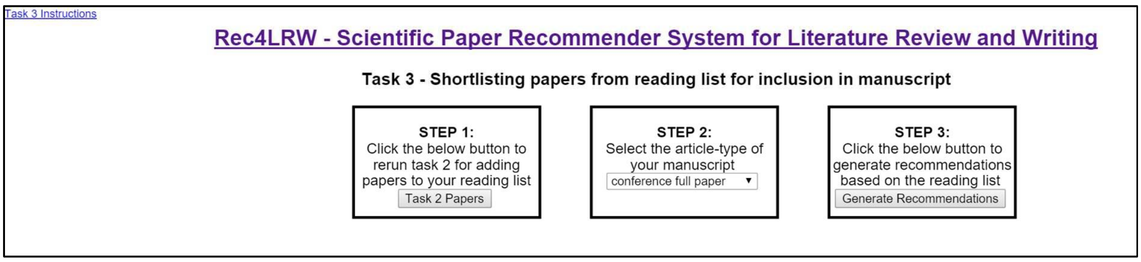

Figure 5. Input options in Task 3

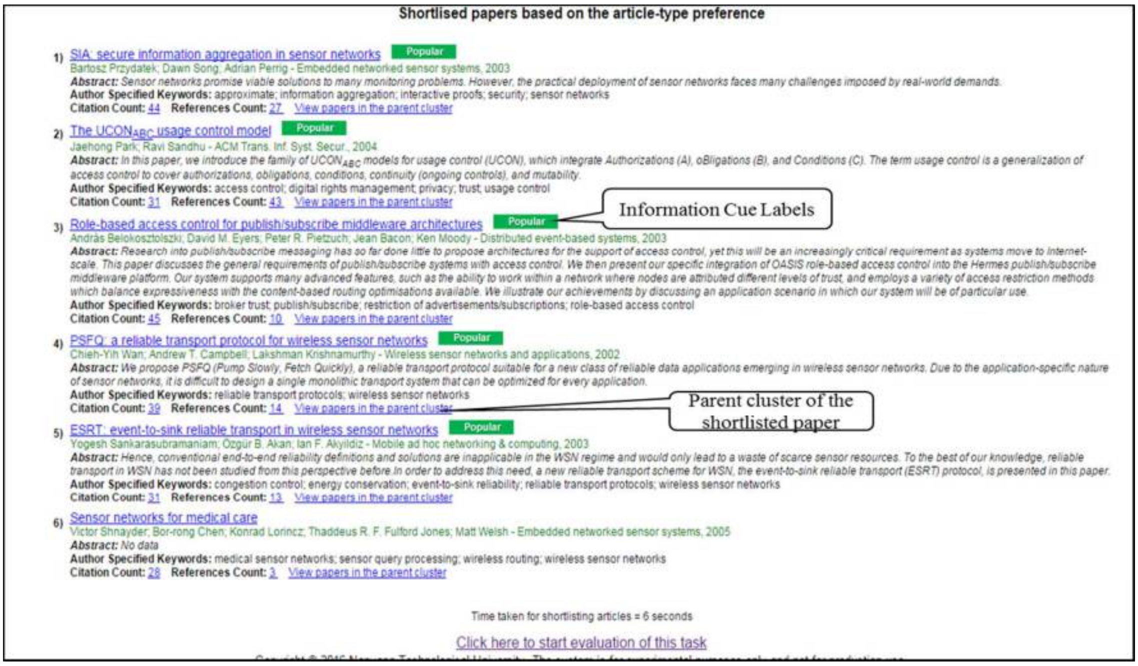

Figure 6. Sample list of shortlisted papers in Task 3

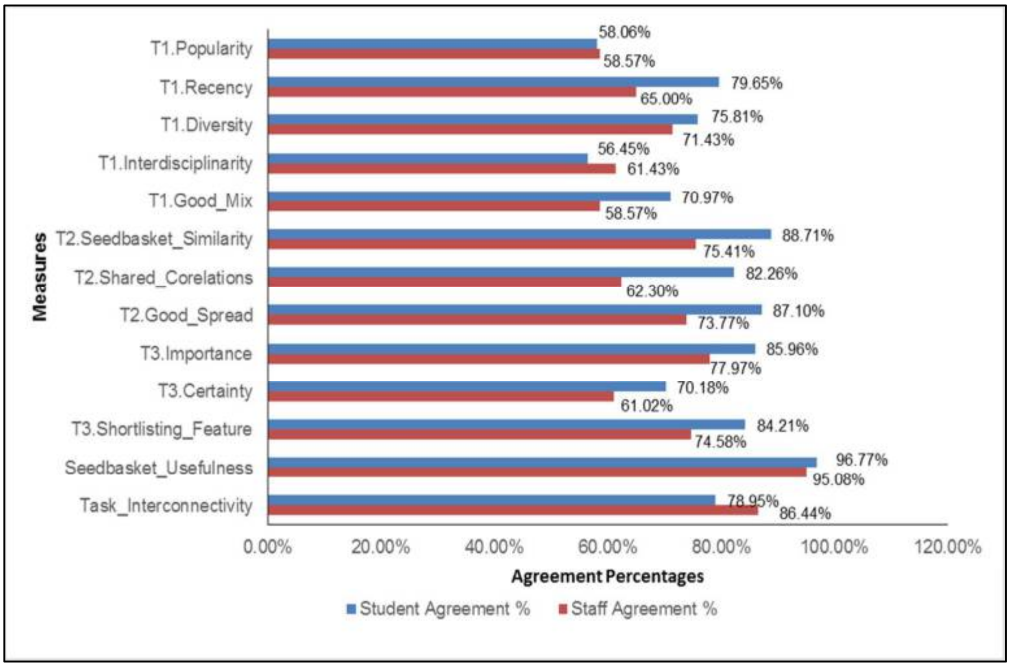

Figure 7. Agreement percentages of the evaluation measures