Information Processing and Management 53 (2017) 577–594

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems

Aravind Sesagiri Raamkumar*, Schubert Foo, Natalie Pang

*Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore*

ARTICLE INFO

*Article history:*
Received 1 May 2016
Revised 22 December 2016
Accepted 23 December 2016

*Keywords:*
Reading list
Literature review
Digital libraries
Scientific paper information retrieval
Author-specified keywords
Scientific paper recommender systems

ABSTRACT

An initial reading list is prepared by researchers at the start of literature review for getting an overview of the research performed in a particular area. Prior studies have taken the approach of merely recommending seminal or popular papers to aid researchers in such a task. In this paper, we present an alternative technique called the AKR (Author-specified Keywords based Retrieval) technique for providing popular, recent, survey and a diverse set of papers as a part of the initial reading list. The AKR technique is based on a novel coverage value that has its calculation centered on author-specified keywords. We performed an offline evaluation experiment with four variants of the AKR technique along with three state-of-the-art approaches involving collaborative filtering and graph ranking algorithms. Findings show that the Hyperlink-Induced Topic Search (HITS) enhanced variant of the AKR technique performs better than other techniques, satisfying most requirements for a reading list. A user evaluation study was conducted with 132 researchers to gauge user interest on the proposed technique using 14 evaluation measures. Results show that (i) students group are more satisfied with the recommended papers than staff group, (ii) popularity measure is strongly correlated with the output quality measures and (iii) the measures familiarity, usefulness and 'agreeability on a good list' were found to be strong predictors for user satisfaction. The AKR technique provides scope for extension in future information retrieval (IR) and content-based recommender systems (RS) studies.



## 1. Introduction

The activities in scientific information seeking (Ellis, Cox, & Hall, 1993) differ from general purpose information seeking on account of variegated information paths and relevance judgement criteria for finding the suitable information objects (research papers). The information needs of researchers keep changing as per the stage in the scientific publication lifecycle. The corresponding search tasks are inherently complex, uncertain and multifaceted since searching is performed in multiple information sources that entail different search queries (Du & Evans, 2011). The apparent differences between novices and experts in terms of the required skills has been observed (Karlsson et al., 2012). Reasons such as lack of depth in the skills of novices, particularly doctoral students (Spezi, 2016), low system skills (Bullock, 2013) and lack of confidence in handling ICT (Markauskaite, 2007) have been identified. The nuances in handling the features provided by academic search systems

* Corresponding author.
*E-mail addresses:* aravind002@ntu.edu.sg (A. Sesagiri Raamkumar), sfoo@ntu.edu.sg (S. Foo), nlspang@ntu.edu.sg (N. Pang).

http://dx.doi.org/10.1016/j.ipm.2016.12.006
0306-4573/

have been highlighted with marked differences between novices and experts in multiple studies (Brand-Gruwel, Wopereis, & Vermetten, 2005; Tabatabai & Shore, 2005; Yoo & Mosa, 2015) where the experts' ability in carefully formulating a problem before conducting search is highlighted as a key difference. Complex search techniques such as zooming, concertina and intersections-finding are required for identifying the relevant research papers (Levy & Ellis, 2006; Ridley, 2012). These search techniques are executed along with citation chaining/searching procedures (Bates, 1989; White & Griffith, 1981), for exploring the citation network of research papers. Hence, academic information searching is inherently different from general purpose information searching, thereby differentiating the overall design of academic databases and search engines. The design along with the retrieval/recommendation techniques is expected to cater to researchers of varying experience levels.

Unlike traditional search engines, academic search systems rank results (research papers) largely based on citation count as it has been proven to provide better results than mere topical matching (Lawrence, Lee Giles, & Bollacker, 1999). However, some studies have commented on the insufficiency of solely relying on citation count for gauging a paper's contribution to a particular research area (Lehmann, Lautrup, & Jackson, 2003). It is worth noting that the ranking criteria of papers are subject to the user's task and the underlying information need. Complementing the free-text search engines, task-based Information Retrieval (IR) and Recommender Systems (RS) have been designed to provide results based on contextual factors (Ricci, Rokach, & Shapira, 2011), closer to the user's task. IR and RS research has provided the necessary algorithms and techniques to build information systems and digital libraries. A practitioner could utilize these techniques as a black box and build an information system on top of it. CiteSeer is one such digital library that incorporates both IR and RS techniques for retrieving and recommending papers, however this system is meant to be useful for ad-hoc search needs and not specific literature review (LR) tasks.

Two important LR tasks are building a reading list of research papers for initial reading and finding similar papers based on a set of papers. At the start of LR, a reading list of research papers is essential for researchers who are venturing into new research areas. A research paper's relative position in the citation network is the main criteria towards its selection in the reading list. This relative position or importance has been perceived as paper seminality and popularity in earlier studies (Bae, Hwang, Kim, & Faloutsos, 2014; Ekstrand et al., 2010; Jardine, 2014; Wang, Zhai, Hu, & Chen, 2010). Seminal papers can be regarded as popular papers in terms of higher citation counts. These seminal/popular papers of a particular field help the researcher in gaining a limited understanding due to two reasons. Firstly, such papers are relatively dated, thereby limiting the scope of knowledge to particular time periods. Secondly, the papers are restricted to particular sub-topic(s) in the given area. Most research areas are broad and many finer sub-topics emerge as research progresses in the particular area. Consequently, seminality is one of the required characteristics of a reading list. Diversity is one of the other characteristics since diverse set of papers are required for covering different sub-topics in the research area. The case for citation diversification has been raised in an earlier study (Küçüktunç, Saule, Kaya, & Çatalyürek, 2015). Recency is another important characteristic as recently published papers represent the latest research performed in the area. Recent papers help in comparing arguments, methodologies and results with earlier studies (Leedy & Ormrod, 2005). Literature survey/review papers are also important inclusions in the reading list as they provide an overview of research in the given area. Survey papers encompass most recent and seminal papers, thereby indirectly addressing two of the other characteristics of reading list (Jesson, Matheson, & Lacey, 2011).

These four characteristics are essential as the expectations of a reading list can be different for researchers based on varying levels of expertise, primary discipline and type of research (academic, applied or translational). We address the identified four characteristics with a novel technique in the current study. The research objective of the current study is to propose a retrieval technique for generating reading list, meant for use at the start of a researcher's literature review on a given research area. This reading list is mainly meant to help researchers in getting a holistic overview of the given research area and it is to be used at the start of the literature review, as addressed by earlier studies (Ekstrand et al., 2010; Jardine, 2014). After reviewing the papers in the reading list, researchers can carry forward with the subsequent search of literature on specific sub-topics, for identifying research gaps and formulating research problems (Levy & Ellis, 2006).

In this paper, the requirements of a comprehensive reading list are first put forth. For addressing the identified requirements, we conceptualize a novel research coverage value known as Topical and Peripheral Coverage (TPC) which is based on author-specified keywords from research papers. Similar to previously proposed approaches (Bae et al., 2014; Ekstrand et al., 2010), this value measurement technique also relies on citation networks of papers. The generation of the citation networks are guided by the author-specified keywords from research papers. At a conceptual level, the value is aimed at identifying diverse papers for research topics along with peripheral papers in the case of inter-disciplinary topics. Finally, a corresponding retrieval technique called as the AKR (Author-specified Keywords based Retrieval) technique that makes use of TPC value in ranking top 20 papers for a reading list, is subsequently proposed. The AKR technique is devised based on feasibility for practical implementation in digital libraries, both in IR and RS contexts.

Offline and user evaluations were conducted to evaluate the proposed AKR technique. A dataset from the ACM Digital Library (ACMDL) was used for the experiments. In the offline evaluation experiment, two basic variants of AKR techniques and two other variants where HITS score is used to boost the TPC value, were benchmarked against three approaches involving collaborative filtering and graph ranking algorithms from earlier studies. The HITS enhanced variant of AKR technique provided the best results satisfying the most requirements of a reading list. This technique was implemented as one of the three tasks in the Rec4LRW system, which is meant for assisting researchers in literature review and manuscript preparatory tasks. A user evaluation study was conducted with 132 researchers. 43 research topics were provided for selection in the task. Data for 14 evaluation measures were collected through a survey questionnaire.

The paper is organized as follows. In Section 2, the related studies are presented. Section 3 is dedicated to reading list generation technique. Requirements of a reading list along with the conceptual model of Topical and Peripheral Coverage (TPC) value and the AKR technique are described. The offline evaluation results are presented in Section 4. In Section 5, the Rec4LRW system and its features are introduced. The findings from user evaluation study are discussed in Section 6. The limitations of this research work are listed in Section 7. Finally, Section 8 presents the conclusions, limitations and future work directions.

## 2. Related work

The introduction of citation indexing systems (Giles, Bollacker, & Lawrence, 1998) paved the way for improving the mechanism of searching research papers. Thereafter, the citation count of a research paper has largely remained the default choice for ranking research papers in search engines. This scenario often leads to a phenomenon called as 'Matthew Effect' (Merton, 1968) where highly cited papers consistently emerge at the top of the search results. The ranking algorithm in one of the most popular and contemporary academic search engines, Google Scholar gives increased preference to highly cited papers (Beel & Gipp, 2009). Most of the current academic search engines and databases mainly offer free-text search functionality, thereby being applicable for ad-hoc search tasks during different stages of literature review. Therefore, the usage of topical similarity (for e.g. TF-IDF (Salton & Buckley, 1988)) and citation count as ranking measures can be considered adequate for such systems. The usability of these systems is much left to the searching skills of the users, particularly in aspects such as selection of search keywords, usage of advanced search and filter options, and the subsequent relevance judgement of the retrieved papers.

In order to simplify the step where search keywords are selected, there have been systems proposed to accept draft text as input instead of search keywords. A combination of translation and LDA-based topic models are used for this purpose in the Refseer system (Huang, Wu, Mitra, & Giles, 2014). Alternatively, phrasal concepts have been used in another study where the input draft text is used to recommend concepts extracted from candidate papers (Kim, Seo, Croft, & Smith, 2014). This approach is aimed at simplifying the task of retrieving relevant papers by modifying the user queries.

Since free-text search engines are not designed for specific tasks, the case of task-based IR and RS systems gained prominence (Vakkari, 2001). Task-based IR systems are applicable in scenarios where the user tasks can be clearly defined. Most of the task-based approaches help in identifying seminal papers from the corpus. Graph ranking algorithms such as PageRank (Page, Brin, Motwami, & Winograd, 1999) and Hyperlink-induced Topical Search (HITS) (Kleinberg, 1999) have been used since the in-links and out-links of papers in citation graphs provide scope for calculating the popularity of papers (Liu & Lin, 2007; Waltman & Yan, 2014). For improving the results, these algorithms have been enhanced by incorporating textual (Jardine & Teufel, 2014) and temporal data (Singh, Shubhankar, & Pudi, 2011). In a recent study conducted with the Microsoft Academic Graph dataset, certain assumptions are made on the goodness of papers. The intuition is that "good authors write good papers, and good papers are not only cited by many others, but also published in good journals or conferences" (Feng et al., 2016). Mutual reinforcement between authors, venues and papers has been used to rank papers, after stages of propagation and refinement within this tripartite network. This technique provides scope for implementation in digital library environments.

Only in the recent past, studies with the explicit objective of generating "reading lists" of important/seminal research papers have been conducted. The advent of Collaborative Filtering (CF) algorithms (Goldberg, Nichols, Oki, & Terry, 1992) have benefitted this area of research. Basic versions of CF algorithms have been found to produce better results than Content-based (CB) filtering methods for multiple LR search tasks (Mcnee, 2006). Continuing the trend, CF techniques weighted with scores from graph ranking algorithms PageRank, HITS and Stochastic Approach for Link-Structure Analysis (SALSA) provided even better results (Ekstrand et al., 2010) than traditional CF algorithms.

The identification of seminal papers in the citation network is subject to the visible and latent characteristics exhibited by such papers. In another study, two unique characteristics behind the citing behaviour of classical (seminal) papers were used to conceptualize an approach for identifying seminal papers (Wang et al., 2010). These two characteristics are (i) the nature of seminal papers being consistently downloaded for reading and (ii) the habit of researchers in citing references from a classical paper. A paper ranking score is calculated by operationalizing the two characteristics.

Novelty of a research paper towards the particular research area is the focus of a study where the citation network is leveraged to identify critical papers (Chen et al., 2011). A methodology called as Citation Authority Diffusion (CAD) is proposed to identify important papers from Google Scholar for particular research topics. This approach is reliant on set of input research papers for identifying the relevant papers. Re-ranking of retrieved papers from portals such as ACM DL is the focus of a study which acknowledges the deficiency of current academic search engines in not giving preference to popular papers (Amolochitis, Christou, Tan, & Prasad, 2013). The proposed re-ranking heuristic takes into account the following data (i) the position of query terms in the papers, (ii) annual distribution of the citation counts and (iii) papers' index terms from the ACM classification system.

Bae et al. (2014) proposed an algorithm based on Random Walk with Restart (RWR) to measure the seminality score of papers. The method incorporates inter-paper similarity scores in addition. However, this approach is aimed at building a genealogy of seminal research papers so that researchers are able to track the progress of particular trends. Another recent study employed a hybrid approach by combining topic models with PageRank. The proposed ThemedPageRank algorithm has been evaluated to provide better results than Google Scholar (Jardine, 2014). This approach also takes the age of re-

search papers into account. Graph centrality measures such as betweenness centrality in graphs have also been found to be useful in ascertaining seminality of papers, due to their ability of capturing latent information present in citation graphs (Runelöv, 2015).

Most of the earlier approaches (Bae et al., 2014; Ekstrand et al., 2010) are based on the citation network as it is one of the main structures that can help in ascertaining the relative position of a paper with respect to other papers and the corresponding research topic(s). Novel citation distribution models (Yang & Han, 2015) can be used to create new citation indices that help in identifying the top papers by considering the depth and breadth of citation distributions across spatial and temporal dimensions. Graph ranking and other hybrid algorithms have been used to ascertain the paper importance.

There are certain shortcomings in these studies. Firstly, it is important to note that the utility of these methods have not been tested in the interdisciplinary research sphere. Secondly, there are certain limitations such as lack of semantic relations between citations (Hjørland, 2013) in the prior studies surveyed in this paper. Semantic similarity data could potentially augment the existing citation similarity data which is primarily calculated based on co-occurrences. . Thirdly, these approaches assume that seminality of a research paper is the main characteristic for inclusion in a reading list. When other characteristics related to recent, survey and diverse set of papers are considered for a reading list, new techniques are required.

## 3. Reading list in literature review

In this section, the requirements of a reading list are first put forth. Second, the Topical and Peripheral Coverage (TPC) value measurement technique is described along with the conceptual model. Finally, the retrieval technique for building the reading list is outlined.

### 3.1. Requirements

A reading list is defined as "*a list of sources (recommended by a teacher or university lecturer) which provide additional or background information on a subject being studied*" (Collins, 2016). In the context of literature review (LR), there have been no formal definitions set for this collection of papers. In previous studies (Bae et al., 2014; Ekstrand et al., 2010; Jardine, 2014; Wang et al., 2010), papers considered as seminal, classical or important in a particular research area, have constituted the reading list. The common aspect of these papers is the high citation counts. Even though, it is necessary for a researcher to read important papers, such papers may not 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. The search activities performed by researchers for building a reading list, typically follows the analytical strategy (Fidel, 2012) since the information need is vague. The scientific information seeking model (Ellis et al., 1993) indicates the exploratory nature of researchers' search tasks during the initial stages of information seeking. In the *starting* phase of this IS model, researchers read papers suggested by experts/supervisors in addition to literature survey/review papers and other secondary resources. Researchers make use of the initial set of papers in this pre-focus stage so that they could zoom into the sub-topics. These sub-topics are used in subsequent directed search sessions where the exact problem is formulated. Therefore, a variety of papers are expected by researchers in this first LR task. Accordingly, we propose four set of requirements for a reading list.

*Requirement 1 (R1):* The reading list should contain popular papers. As noted earlier, popular papers are the only constituents of reading lists in earlier studies. These papers have a seminal status in the particular research area. Generally, these papers have very high citation counts, indicating their popularity. Therefore, an adequate quantity of popular papers is required for a reading list.

*Requirement 2 (R2):* The reading list should contain survey papers. Literature survey papers provide an overview of the existing research performed in a research area. Researchers generally read survey papers as a starting point in their literature review (Ellis & Haugan, 1997). These papers provide synthesis of prior studies along with the problem areas and research opportunities. Ideally, a reading list of about 20 papers should contain at least one or two survey papers. The prescribed count of survey papers in a reading list is subjective as there is no empirical data about the researchers' expectations. It would be convenient for researchers, if the system provides a feature where the user could set the number of papers.

*Requirement 3 (R3):* The reading list should contain recent papers. Prior studies and current search systems give lesser preference to recent papers. Recently published papers provide information about the latest research performed in a research area. Additionally, recent papers potentially cite important papers. Researchers can look at such bibliographic references to discover other interesting papers.

*Requirement 4 (R4):* A reading list should contain papers from sub-topics of the main research area. Research areas comprise of sub-topics which span out to become self-contained research areas in themselves over a period of time. Additionally, there are interdisciplinary research areas which are subject to different research methods from corresponding disciplines. Therefore, diversity is a necessary characteristic for a reading list. A diversified reading list is meant to provide a bird's eye of the whole research landscape, particularly for broad research topics.

Based on the above requirements, a reading list is defined *as "a list with an agreeable mix of popular, recent and survey papers covering diverse sub-topics in the particular research area"*. In the current study, the number of papers for each require-

ment is not fixed as a separate study needs to be conducted with researchers across disciplines for ascertaining the paper count for each requirement.

### 3.2. Topical and peripheral coverage (TPC)

With the objective of meeting the requirements of a reading list, we proposed two novel coverage value measurement techniques based on author-specified keywords (AK) from research papers (Sesagiri Raamkumar, Foo, & Pang, 2015). They are Topical Coverage (TC) and Topical and Peripheral Coverage (TPC). Prior approaches in research coverage measurement are based solely on the citations network (Bae et al., 2014; Mcnee, 2006). Graph ranking techniques (Ekstrand et al., 2010; Jardine & Teufel, 2014) have been used to ascertain the importance of papers in the network. The usefulness of these methods has not been tested in the context of finding interdisciplinary research papers. These methods do not make use of author-specified keywords from research papers to build the citations network. Author-specified keywords represent the central topics addressed in research papers, directly specified by the authors. Publication houses generally allow authors to include a maximum of five keywords per paper. These keywords are primarily useful in classifying the papers. More importantly, there is an opportunity to identify inter-disciplinary research papers using these keywords as authors tend to use terms from their primary discipline's vocabulary. From the two proposed coverage values, Topical and Peripheral Coverage (TPC) value was found to provide better results during our initial pilot tests. The measurement technique of TPC is re-introduced as follows.

The Topical and Peripheral Coverage (TPC) value is measured by utilizing all the author-specified keywords provided in a research paper. The first step is to identify the keywords $K$ provided for a paper $P_i$, followed by extracting all the papers in the corpus which have the keywords in $K$. This extracted set of papers forms the base set $P_k$. In the next step, extraction of the bibliographic references list $reflist_i$ and citations list $citelist_i$ of $P_i$ is performed. The final coverage (TPC) value is measured by counting the number of papers from $reflist_i$ and $citelist_i$ that are present in $P_k$. The TPC value can also be multiplied with other values such as PageRank or hub/authority score from HITS so that its effectiveness can be increased. The TPC value helps in building a reading list that satisfy the four requirements. Since the value calculation is based on author-specified keywords, both broad and narrow topics are covered in the base set $P_k$, thereby directly addressing R4. On top of this setup, papers with either high number of bibliographic references or citations tend to get higher TPC scores. Hence, the other three requirements are addressed. R1 papers have high citation counts while R2 and R3 papers are expected to have high number of bibliographic references. However, R3 recent papers with low number of references will get low scores. The conceptual model for TPC is provided below.

$$\text{For each } P_i (1 \le i \le n), reflist_i : \left\{ P_j \middle| P_i \text{ references } P_j \right\} \tag{1}$$

$$\text{For each } P_i (1 \le i \le n), citlist_i : \left\{ P_j \middle| P_i \text{ is cited by } P_j \right\} \tag{2}$$

$$K : \text{list of all author specified keywords in } P_n \tag{3}$$

$$P^k : \text{list of all papers in the corpus having any } k \text{ in } K \text{ as author specified keyword} \tag{4}$$

$$TPC \text{ for } P_i : \left| reflist_i \in P^k \right| + \left| citlist_i \in P^k \right| \tag{5}$$

### 3.3. Retrieval technique for building reading list

The retrieval technique for this task is operationalized by keeping TPC value as the main ranking entity. Since the technique is conceptually based on author-specified keywords, we label it as AKR technique where AKR stands for Author-specified Keywords based Retrieval. The input to this task is the research topic. Earlier studies have used different types of input into their respective techniques. Ekstrand et al. (2010) used a set of seed papers as input. Wang et al. (2010) trained their system with a set of papers for classification purposes. Bae et al. (2014) used a set of seminal papers as input. For the current task, we consider the naturalistic scenario of using research topic as the input. The research topic is expressed as search keywords by users in academic search systems. The AKR technique is split into two stages.

#### 3.3.1. Stage 1 - content-based filtering

The objective of this stage is to construct a representative list of papers related to the input research topic. Three metadata fields article title, article abstract and author-specified keywords are used for the text matching. These fields are merged to form a single field so that the similarity matching efficiency could be maximized. The Okapi BM25 similarity score (Jones, Walker, & Robertson, 2000) is used for the similarity matching. The top 200 matching papers are retrieved to form set $S$. The number of papers in $S$ is set to 200 so that most of the closely matching documents are in consideration for the final ranking in Stage 2. In earlier studies (Strohman, Croft, & Jensen, 2007), the top 100 papers have been retrieved during the evaluation. In this study, we wanted to have a bigger base set so that most of the relevant research papers are retrieved.

**Table 1**
Techniques used in offline evaluation experiment.

| Order | Abbr. | Technique Description |
|---|---|---|
| A | AKRv1 | Basic AKR technique with weights $W_{CC} = 0.25$, $W_{RC} = 0.25$, $W_{CO} = 0.5$ |
| B | AKRv2 | Basic AKR technique with weights $W_{CC} = 0.1$, $W_{RC} = 0.1$, $W_{CO} = 0.8$ |
| C | HAKRv1 | HITS enhanced AKR technique boosted with weights $W_{CC} = 0.25$, $W_{RC} = 0.25$, $W_{CO} = 0.5$ |
| D | HAKRv2 | HITS enhanced AKR technique boosted with weights $W_{CC} = 0.1$, $W_{RC} = 0.1$, $W_{CO} = 0.8$ |
| E | CFHITS | IBCF technique boosted with HITS |
| F | CFPR | IBCF technique boosted with PageRank |
| G | PR | PageRank technique |

*3.3.2. Stage 2 - ranking the final list of papers*

The 200 papers from *S* are ranked based on ranking scheme which is majorly based on the TPC value. The objective of this stage is to shortlist the top 20 papers. We have set the count of top papers as 20 since users rarely go beyond the top 2 or 3 search results pages in general-purpose search sessions (Van Deursen & Van Dijk, 2009). However, there is a caveat associated with this setting. In the case of academic search, users might be more patient to browse through the search results until they identify the required number of relevant papers. A composite rank $CR_P$ is used for ranking papers from *S*. In order to give importance to citation and bibliographic reference counts albeit to a lesser level, the composite rank is based on three values – citation count, reference count and TPC value. Weights are used to set the importance of these values. The sum of these weights should add up to 1. The inclusion of these weights is to vary the preference given to the values. All the three values are normalized before the $CR_P$ is calculated. The formula for $CR_P$ is provided in the equation below.

$$CR_P = \left(\left(\frac{CC_p - \min_S CC}{\max_S CC - \min_S CC}\right) * W_{CC}\right) + \left(\left(\frac{RC_p - \min_S RC}{\max_S RC - \min_S RC}\right) * W_{RC}\right) + \left(\left(\frac{TPC_p - \min_S TPC}{\max_S TPC - \min_S TPC}\right) * W_{TPC}\right) \quad (6)$$

where $CC_P$ is the citation count of paper *P* from set *S*. min*CC* is the minimum citation count value from *S*. max*CC* is the maximum citation count value from *S*. $W_{CC}$ is the importance weight for citation count. $RC_P$ is the references count of paper *P* from *S*. min*RC* is the minimum references count value from *S*. max*RC* is the maximum references count value from *S*. $W_{RC}$ is importance weight for reference count. $TPC_P$ is the TPC value of paper *P* from *S*. min*TPC* is the minimum TPC value from *S*. max*TPC* is the maximum TPC value from S. $W_{CO}$ is the importance weight for TPC value.

## 4. Offline evaluation experiment

An offline evaluation experiment was conducted for comparing the proposed AKR technique with approaches from earlier studies. Offline evaluation is one of the most frequently adopted forms of evaluation in studies of this domain (Beel, Genzmehr, Langer, Nürnberger, & Gipp, 2013) since it is relatively easier to perform in terms of cost and resources. Offline evaluation is objectively performed at a level of abstraction where multiple techniques are compared using a set of evaluation metrics and the whole activity generally doesn't involve human participants. Some studies (Jardine, 2014) have been conducted in the past where real-time systems such as Google Scholar, have been benchmarked against proposed recommendation techniques. This scenario is applicable if the task objective is similar on both sides. In the current study, the task's objective is to build a reading list of papers for initial reading. Systems such as Google Scholar, Scopus or CiteSeer do not claim to provide a reading list of papers for the search keywords. Therefore, conducting an offline evaluation experiment by benchmarking with such systems was not considered.

From earlier studies, item-based collaborative filtering (IBCF) boosted with graph ranking algorithms such as PageRank and HITS, have provided better results than content-based filtering techniques (Ekstrand et al., 2010). Both these approaches are considered for the experiment. The third approach considered is the traditional PageRank technique. The proposed AKR technique is evaluated in four variants. In the first set of variants, the basic AKR technique is implemented with two different weights combinations. The first variant AKRv1 has the weights set ($W_{CC} = 0.25$, $W_{RC} = 0.25$, $W_{CO} = 0.5$) where TPC is given 50% weightage while citation count and reference count share the remaining 50% weightage. Second variant AKRv2 has the weights set ($W_{CC} = 0.1$, $W_{RC} = 0.1$, $W_{CO} = 0.8$) where TPC is given a dominant 80% importance while citation count and reference count share the remaining 20% weightage. The second set of AKR variants were introduced after considering the findings from an earlier case study (Sesagiri Raamkumar et al., 2015). It was observed that TPC's effectiveness in meeting the requirements of a reading list will be enhanced if it is boosted with HITS score. Therefore, the third AKR variant is called the HAKRv1 technique (HITS enhanced AKR) with the same weights set as AKRv1 and the fourth variant is HAKRv2 with the same weights set as AKRv2. The seven techniques used for the offline evaluation are listed in Table 1.

### *4.1. Dataset and technical details*

An extract from the ACM Digital Library (ACMDL) is used as the dataset for the offline and user evaluations. Papers from proceedings and periodicals (journals) for the period 1951 to 2011 form the dataset. The papers were shortlisted based on

**Table 2**
Aggregated ranks generated using CE algorithm with spearman footrule distance.

| Paper type (Requirement) | Optimal aggregated ranks | | | | | | | Min. Obj. function score |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| Recent papers (R1) | B | A | C | D | E | F | G | 10.66 |
| Popular papers (R2) | F | E | C | D | G | A | B | 11.89 |
| Literature survey papers (R3) | C | G | D | A | E | F | B | 13.38 |
| Diverse papers (R4) | C | D | G | A | B | F | E | 12.15 |

*Note*. The optimal ranked list is displayed for the paper type in each row. For instance, the AKR variants occupy all the top ranks for the requirement R1.The letters corresponding to the seven techniques are listed in Table 1.

full text and metadata availability in the dataset, to form the sample set/corpus for the system. The sample set contains a total of 103,739 articles and corresponding 2320,345 references. The original data from ACM was received in the form of 4500 XML files. Data was transferred to a MySQL database to facilitate easier storage, processing and retrieval. The references of papers were parsed using AnyStyle parser (AnyStyle, 2015) for extracting article title and publication year. Apache Lucene and Mahout libraries were used for the IR and RS algorithm implementations.

### 4.2. Experiment details

As the intent was to perform the offline evaluation with a wide range of research topics, it was decided that top 200 author-specified keywords from the dataset, would be identified as the research topics. This identification is based on the number of papers that use the keywords. 14 keywords were removed as these keywords were extensions or plurals of the base topic (for e.g., digital libraries – digital library, social networking – social network), thereby a total of 186 keywords were used for the experiment. The experiment was performed in three steps. The first and third steps are common for all the seven techniques. In the first step, top 200 papers were retrieved using the BM25 similarity algorithm (refer stage 1 in Section 3.3.1). In the second step, the top 20 papers were identified using the seven techniques. Basically, the papers were ranked based on the descending order of the scores. A total of 744 reading lists were generated as a result. In the third step, the evaluation related metrics are calculated.

For requirements R1, R2 and R3, the number of relevant papers are enumerated. For requirement R4 (diverse papers), a different approach is followed. In earlier RS studies, measurement of novelty and diversity metrics is based on distance between the recommended items (Castells, Vargas, & Wang, 2009). Longer distance between two recommended items indicates high diversity while shorter distance indicates low diversity. In the current study's domain, since there are references and citations networks for the papers, subgraph properties could be used to infer the level of diversity. If *G(V,E)* is a graph built with references and citations of papers for a particular topic, *G1(V1,E1)* is a subgraph built with just the final 20 papers from the evaluated retrieval technique. The number of edges *E1* from *G1* is an indication of level of diversity in the final list of papers. If there is more number of edges, it means there are many inter-referencing/inter-citing connections between the papers, thus implying a less diverse list of papers and vice-verse for high diversity.

Traditional evaluation metrics such as Precision, Recall and DCG (Discounted Cumulative Gain) were not used for the evaluation as the intent was to identify the approach which best satisfies the four requirements. The evaluation rhetoric is based on the following intuition. For each of the four requirements, the number of *X* papers is first identified in the final top 20 list (for instance, *X* being recent, popular or survey paper). The paper count is compared across the seven techniques. Ranks are assigned to the technique based on the highest counts. This step is repeated for all the 186 topics. At this stage, we have ranked lists for 186 topics for each requirement. Rank aggregation approaches (Dwork, Kumar, Naor, & Sivakumar, 2001) have been proposed to identify the optimal ranked list of items over multiple lists. In the current study, we employ the RankAggreg library (Pihur & Datta, 2009) to identify the optimal ranked lists for the four requirements. Specifically, Cross-Entropy Monte-Carlo algorithm (CE) with two distance measures Spearman footrule distance and Kendall's tau distance is used for the rank aggregation. Based on the aggregated ranks for the four requirements, the best technique which satisfies the most requirements is finally selected.

### 4.3. Findings

The ordering for the seven techniques is provided in Table 1. The optimal aggregated ranked lists and minimum objective scores for the two distance measures are displayed in Tables 2 and 3. The lowest score (minimum objective function score) was attained by the CE algorithm run with Kendall distance. The comparison between spearman distance, mean ranks and samples of raw data is provided in Fig. 1–4. The figures help in comparing the optimal rank lists with the mean (average) ranking method, for each paper type. In the X axis, the techniques are listed in the optimal order so that the distinct ranks can be observed. In the Y axis, the ranks are displayed. From the figures, the drawback of using a simple averaging method in rank aggregation can be noticed since certain techniques share the same ranks.

**Table 3**
Aggregated ranks generated using CE algorithm with kendall distance.

| Paper type (Requirement) | Optimal aggregated ranks | | | | | | | Min. obj. function score |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| Recent papers (R1) | A | B | C | D | E | F | G | 6.30 |
| Popular papers (R2) | F | C | E | D | G | A | B | 7.31 |
| Literature survey papers (R3) | C | D | G | E | A | F | B | 8.10 |
| Diverse papers (R4) | C | D | G | A | E | B | F | 7.37 |

*Note*. The optimal ranked list is displayed for the paper type in each row. For instance, the AKR variants occupy all the top ranks for the requirement R1.The letters corresponding to the seven techniques are listed in Table 1.

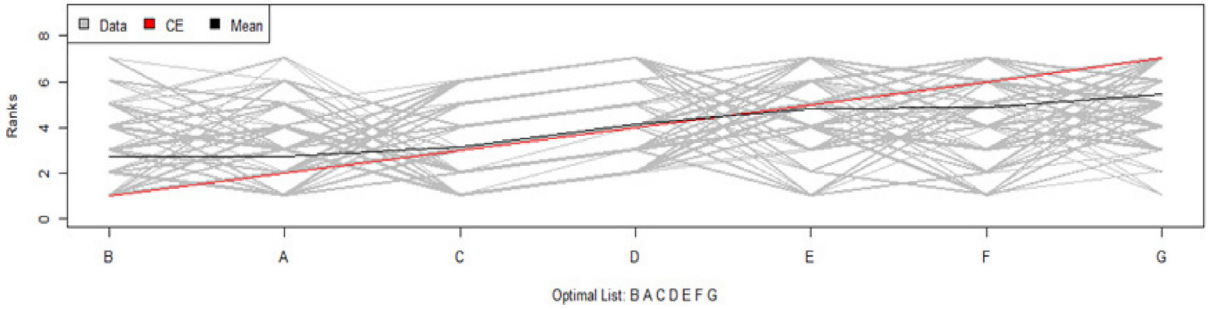



**Fig 1.** Recent papers rank aggregation comparison.

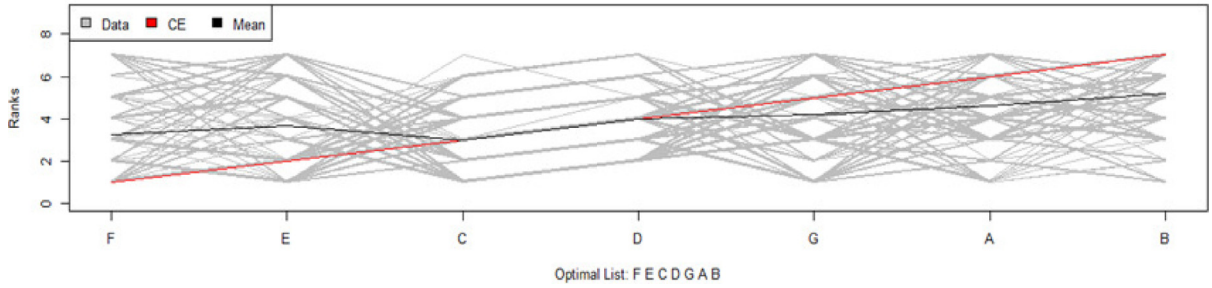



**Fig 2.** Popular papers rank aggregation comparison.

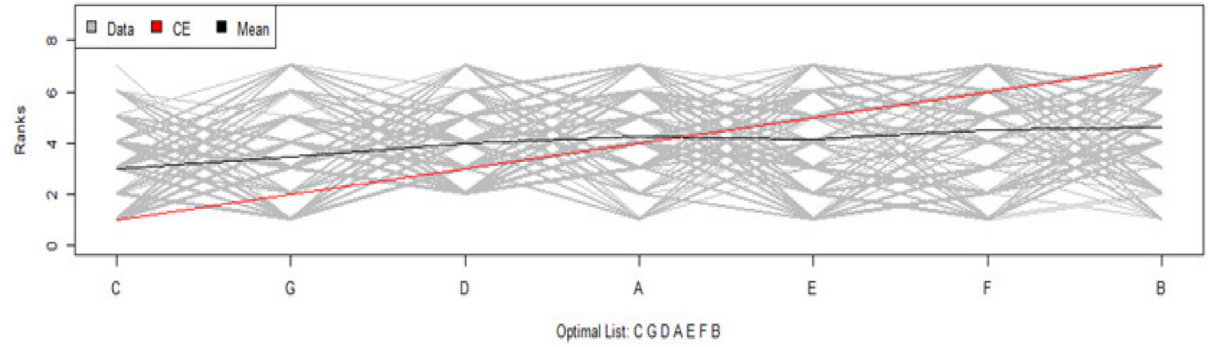



**Fig 3.** Literature survey papers rank aggregation comparison.

Both the distance measures (Spearman and Kendall) had similar results in the context of top three ranks. Table 3 is used for the interpretation of the final results. For R1, the basic variants of the AKR technique (A & B) provided the best results with most number of recent papers. The HITS enhanced AKR variants (C & D) were the next best entries in the ranked list. The basic AKR techniques consider both references and citations count by design. Therefore, recent papers with high number of references have an increased probability for being shortlisted when compared to other techniques. The benchmarked approaches (E, F & G) aren't designed to retrieve recent papers and hence, the low ranks.

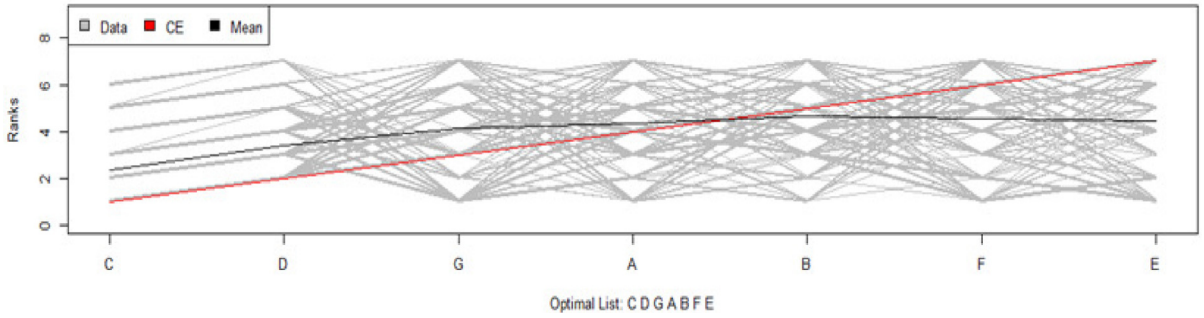



**Fig 4.** Diverse papers rank aggregation comparison.

For R2, the benchmarked approaches, particularly CFPR (F) provided the best results, thereby validating their usefulness in finding seminal/classical papers, similar to previous study results (Ekstrand et al., 2010). Among the AKR variants, the HAKRv1 technique (C) was within the top 3 ranks mainly because of the influence of paper's HITS score on the TPC value, thereby validating the main purpose behind boosting TPC with HITS score. There was a perceivable gap between HAKRv1 (C) and HAKRv2 (D) in this case. The former variant where 50% weightage is given to TPC as against 80% in the latter variant, clearly benefits papers with higher citation counts.

For R3, HAKRv1 technique (C) yields the best result because of the ability of TPC in giving precedence to papers with bigger bibliographies. Interestingly, the basic AKR techniques rank lower in the list, which may be due to lack of recent literature survey papers in the collection. For R4, HAKRv1 technique (C) gets the top rank again due to lesser degree of connections between the papers in the final list. All the variants of AKR are supposed to have lesser connections as the aim is to produce a diverse list of papers. However, the PR technique (G) seems to perform better than basic AKR variants (A & B).

From the findings, it is evident that HAKRv1 (C) is the top performing technique among the seven approaches, since it best satisfies the four requirements. This technique acquired the top most rank for R3 and R4 while it was within the top 3 ranks for R1 and R2. It is to be noted that HAKR variants (C&D) closely follow each other in three requirements with R2 as the only exception. The weight sets were important in distinguishing the performance of the variants. For all the four AKR variants, the first weight set ($W_{CC} = 0.25$, $W_{RC} = 0.25$, $W_{CO} = 0.5$) provided better results than the second weight set ($W_{CC} = 0.1$, $W_{RC} = 0.1$, $W_{CO} = 0.8$). This observation validates the importance of two absolute values of a research paper – references and citations count. TPC is a relative value since it is based on the author-specified keywords. The intuition to include all these three values in the composite rank was to consider both absolute and relative values of a research paper in the final ranking step. If TPC had been directly used for ranking the papers in the final step, the requirements R2 and R3 would have been affected as lesser number of corresponding papers would have been shortlisted in the top 20 papers. The extensibility of the TPC in AKR techniques is shown in this experiment. HITS score has been used to enhance the ability of the AKR technique in obtaining top ranks for all the four requirements. As a result, the HAKRv1 technique is the sole technique selected for the user evaluation study (described in Section 6).

## 5. Rec4LRW system

The HAKRv1 technique was implemented as a user task in a system called Rec4LRW. This implementation was performed so that the task could be evaluated by researchers. In this section, a brief introduction of the Rec4LRW system is provided. The task screens and the User interface (UI) features are also outlined.

### *5.1. Brief overview*

The Rec4LRW system has been developed to help researchers in two main search tasks of literature review and one manuscript preparatory task. The three tasks are (1) Building an initial reading list of research papers, (2) Finding similar papers based on a set of papers, and (3) Shortlisting papers from the final reading list for inclusion in manuscript based on article-type choice. The usage flow of the system is as follows: A researcher would typically run the first task at the start of the literature review, followed by selection of few relevant seed papers. The second task makes use of these seed papers to in turn find topically similar papers. In a real world scenario, this task is run multiple times until the researcher is satisfied. The third task is meant to be run when the researcher is at the stage of writing research manuscripts. A researcher would have read numerous papers while performing research. The third task helps the researcher in identifying both important and unique papers in the final list of read papers. The shortlisted papers count varies according to the article-type preference of the researcher. The tasks of Rec4LRW system are meant to be highly beneficial for novice researchers such as PhD students and also for researchers who are venturing into new research topics.

### 5.2. Task screen

A screenshot of the reading list task screen is displayed in Fig. 5. A minimalist design principle was adapted so that the user's focus is retained on evaluating the recommendations. Apart from the regular display features such as article year, author name(s), abstract, publication year and citation count, the system displays some new features which are author-specified keywords, references count and short summary of the paper (if the abstract of the paper is missing). Most importantly, information cue labels are placed beside the article title for each article. There are four labels (1) Popular, (2) Recent, (3) High Reach and (4) Survey/Review. These labels directly represent three of the reading list requirements, described in Section 3.1.

The display logic for the cue labels is described as follows. The recent label is displayed for papers published between the years 2009 and 2011 since the most recent papers in the ACM dataset is published of 2011. The survey/review label is displayed for papers which are of the type - literature survey. For the popular label, the unique citation counts of all papers for the selected research topic are first retrieved from the database. The label is displayed for a paper if the citation count is in the top 5% percentile of the citation counts for that topic. Similar logic is used for the high reach label with references count data. A word cloud generated with the author-specified keywords of the recommended papers is displayed at the end of the results section. This feature is to highlight the diversity requirement of reading list. A sample word cloud is displayed in Fig. 5.

## 6. User evaluation study

### 6.1. Study objective

Goodhue (1995) define user evaluation as an "assessment made by a user, along some continuum from positive to negative, about certain qualities of information systems". Since IR and RS techniques are implemented as part of an information system, the effectiveness of these techniques directly impact the effectiveness of the system. User evaluation studies are required for measuring the effectiveness of the system and the quality of the output so that a strong claim can be made on the success of the system and the constituent IR/RS techniques. In the current research, the overall purpose of the user evaluation study was to determine whether researchers using the reading list task can be efficient and effective in their work. In this context, researchers' perceptions of the individual characteristics of the recommended papers and overall quality of the recommendation list were measured. The specific evaluation goals put forth for the study were (i) ascertain the agreement percentages of the evaluation measures, (ii) test the hypothesis that students are more benefitted from the recommendation task in comparison to staff, and (iii) measure the correlation between the measures and build a regression model with user satisfaction as the dependent variable (DV). We believe that these evaluations goals help in adequately addressing the overall evaluation purpose.

The nuances in handling the features provided by academic search systems and task execution skills (Karlsson et al., 2012; Niu & Hemminger, 2012; Yoo & Mosa, 2015) are two factors that differentiate novice and expert researchers. These two factors support the rationale for including the hypothesis testing as one of the evaluation goals in this study. The claim is that students (novices) are expected to benefit most from suitable interventions which helps them accomplish their literature search tasks.

### 6.2. 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 also sent to mailing lists related to LIS and Information Systems. The main selection criteria was that participant should have authored at least one conference or journal paper. 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. Participants were permitted to perform the experiment from any location.

### 6.3. Study procedure

The participants were required to select a research topic from a list of 43 research topics in the task screen. Out of the provided topics, participants used 29 topics (list of used topics provided in Table 4). On selection of topic, the system provided 20 recommendations. The evaluation screen was embedded at the bottom of the screen. The participants had to answer 14 mandatory survey questions and two optional subjective feedback questions as a part of the evaluation. A five-point Likert scale was provided for measuring participant response for each question. Before the start of the evaluation study, it was hypothesized that students would rate the quality of recommendations more favourably than staff.


**Fig 5.** Reading list task screen in the Rec4LRW system.

**Table 4**
Research topics used in the evaluation study.

| Research Topics | Participant Count |
|---|---|
| Social media | 16 |
| Machine learning | 12 |
| Digital libraries | 11 |
| Data mining | 9 |
| Social networks | 8 |
| Distributed systems | 8 |
| Information retrieval | 7 |
| Human computer interaction | 6 |
| Sensor networks | 6 |
| Wireless networks | 5 |
| Embedded systems | 5 |
| Genetic algorithm | 5 |
| Computer-mediated communication | 5 |
| Recommender systems | 4 |
| Approximation algorithms | 3 |
| User-centered design | 3 |
| Cloud computing | 3 |
| Natural language processing | 3 |
| Access control | 3 |
| User experience | 2 |
| Static analysis | 2 |
| Game theory | 2 |
| Information visualization | 2 |
| Human-robot interaction | 2 |
| Collaborative filtering | 1 |
| Interaction design | 1 |
| Software architecture | 1 |
| Mobile computing | 1 |
| Eye tracking | 1 |

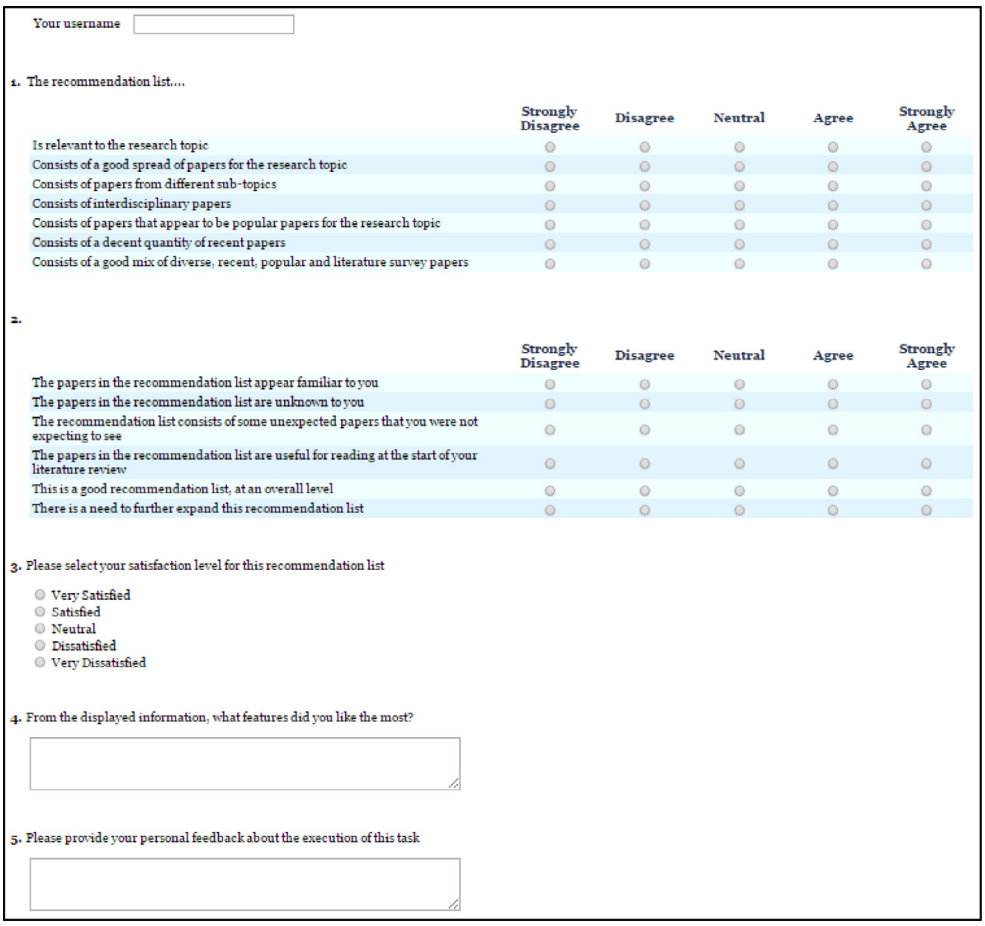
**Table 5**
Survey questions and corresponding measures from user evaluation study.

| Sl.no | Question | Measure |
|---|---|---|
| 1 | The recommendation list is relevant to the research topic | Topical_Relevance |
| 2 | The recommendation list consists of a good spread of papers for the research topic | Good_Spread |
| 3 | The recommendation list consists of papers from different sub-topics | Diversity |
| 4 | The recommendation list consists of interdisciplinary papers | Interdisciplinarity |
| 5 | The recommendation list consists of papers that appear to be popular papers for the research topic | Popularity |
| 6 | The recommendation list consists of a decent quantity of recent papers | Recency |
| 7 | The recommendation list consists of a good mix of diverse, recent, popular and literature survey papers | Good_Mix |
| 8 | The papers in the recommendation list appear familiar to you | Familiarity |
| 9 | The papers in the recommendation list are unknown to you | Novelty |
| 10 | The recommendation list consists of some unexpected papers that you were not expecting to see | Serendipity |
| 11 | The papers in the recommendation list are useful for reading at the start of your literature review | Usefulness |
| 12 | This is a good recommendation list, at an overall level | Good_List |
| 13 | There is a need to further expand this recommendation list | Expansion_Required |
| 14 | Please select your satisfaction level for this recommendation list | User_Satisfaction |

*Note*. The questions in this table were presented with the exact sentencing, at the bottom of the recommendations task screen. A five-point Likert scale was provided with each question with lowest and highest values being Strongly Disagree' and 'Strongly Agree' respectively.

### 6.4. Evaluation measures

The survey questions and the corresponding measures are provided in Table 5. A screenshot of the survey questionnaire is provided in Fig. 6. The measures popularity, recency and diversity correspond to the requirements of the reading list. The interdisciplinarity measure corresponds to the ability of the TPC technique in identifying interdisciplinary papers through the author-specified keywords of research papers. The measures good_spread, familiarity, serendipity good_list and user_satisfaction were adopted from an earlier study (Mcnee, 2006) in which recommendations for multiple information seeking tasks were provided. The measures topical_relevance and usefulness were adopted from a relevant study (Ekstrand et al., 2010). The measures good_mix, expansion_required are novel measures added for this study's requirements.


Your username

1. The recommendation list....

| | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| Is relevant to the research topic | | | | | |
| Consists of a good spread of papers for the research topic | | | | | |
| Consists of papers from different sub-topics | | | | | |
| Consists of interdisciplinary papers | | | | | |
| Consists of papers that appear to be popular papers for the research topic | | | | | |
| Consists of a decent quantity of recent papers | | | | | |
| Consists of a good mix of diverse, recent, popular and literature survey papers | | | | | |

2.

| | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| The papers in the recommendation list appear familiar to you | | | | | |
| The papers in the recommendation list are unknown to you | | | | | |
| The recommendation list consists of some unexpected papers that you were not expecting to see | | | | | |
| The papers in the recommendation list are useful for reading at the start of your literature review | | | | | |
| This is a good recommendation list, at an overall level | | | | | |
| There is a need to further expand this recommendation list | | | | | |

3. Please select your satisfaction level for this recommendation list

- Very Satisfied
- Satisfied
- Neutral
- Dissatisfied
- Very Dissatisfied

4. From the displayed information, what features did you like the most?

5. Please provide your personal feedback about the execution of this task

**Fig 6.** Task evaluation questionnaire.

### 6.5. Analysis

The response values 'Agree' and 'Strongly Agree' were the two values considered for the calculation of agreement percentages for the measures. 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, for the testing the hypothesis. The dependent variables (DVs) in the *t*-test and linear regression were normally distributed with homogeneity in variances. However the DVs were not continuous variables as per one of the assumptions of a *t*-test. We use this test as a means of approximation. Spearman correlation coefficient was used to measure the correlation between the measures since the measures use ordinal scale. Multiple linear regression was used in the analysis to understand the relationship between the evaluation measures and the user_satisfaction measure. Specifically, the method was used to identify the strong predictors. Statistical significance was set at $p < 0.05$. Statistical analyses were done using SPSS 21.0 and R.

### 6.6. Participant demographics

Out of the eligible 230 participants, around 138 participants signed the consent form. 119 of them completed the whole experiment inclusive of the three tasks in the system. The reading list task was completed by 132 participants. 62 participants were PhD/MSc students while 70 were research staff, academic staff and librarians. The average research experience for PhD students was 2 years while for staff, it was 5.6 years. 51% of participants were from the computer science, electrical and electronics disciplines, 35% from information and communication studies discipline while 14% from other disciplines.

### 6.7. Results and discussion

#### 6.7.1. Agreement on the task results

The agreement percentages for the measures by the participant group are illustrated in Fig. 7. We consider an agreement percentage above 75% as an indication of higher agreement from the participants. The students group had high agreement

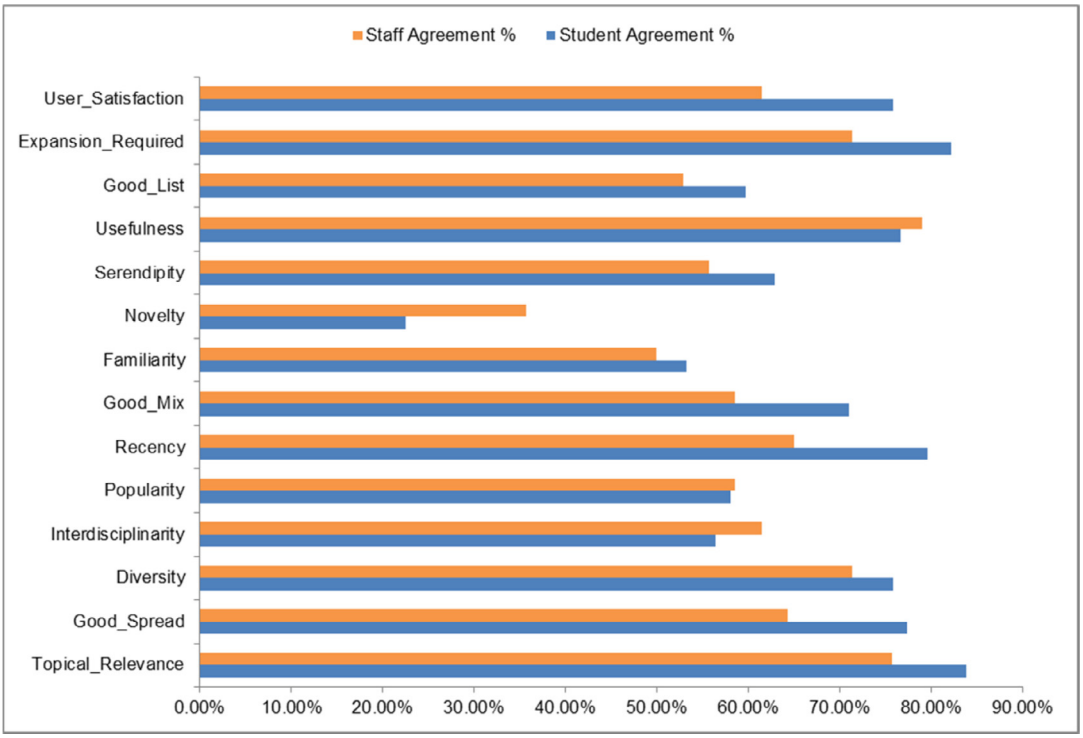



**Fig 7.** Agreement percentages for the evaluation measures.

for the measures topical_relevance (83.87%), good_spread (77.42%), diversity (75.81%), recency (79.65%), usefulness (76.72%), expansion_required (82.26%) and user_satisfaction (75.81%). Graduate research students stand to be benefitted by this automated approach of generating reading list since quality measures indicate that the list meets their expectations. On the other hand, staff group had high agreement for only two measures topical_relevance (75.71%), usefulness (78.97%). Even though the recommended papers were limited to the papers in the ACMDL dataset, many participants were expecting to see some other familiar papers from information sources such as IEEE Xplore. One such critical comment from a participant was "…the database is missing IEEE Xplore articles and from the best robotics journals and conferences which maintain high standards…" This expectation might have hampered the evaluation for the staff group. Secondly, some of the staff participants felt that the broad topics available for selection, was not convenient as they wanted to search for specific sub-topics in their area of expertise. Interestingly, the agreement level for the interdisciplinarity measure was high for the staff participants than students. This observation could be attributed to the experience of researchers in identifying certain journals and conferences as being specific to their discipline than others.

From the agreement results, it is evident that the researchers found the task to be useful in retrieving relevant, recent and diverse set of papers, thereby corroborating the results of the offline evaluation experiment. Both the groups felt that the paper count of 20 was not sufficient for a reading list (82.26% for students, 71.43% for staff). They wanted the system to allow the user to set the number of recommendations.

### *6.7.2. Hypothesis testing*

The results of the independent samples *t*-test are provided in Table 6. The difference between the groups was statistically significant for five measures topical_relevance, good_mix, novelty, good_list and user_satisfaction. Hence, the hypothesis was met for these five measures. Out of these five measures, topical-relevance, good_list and user_satisfaction are output quality measures at the recommendations list level. These measures do not ascertain user opinion on specific characteristics of the recommendations. Therefore, we can claim that students clearly preferred the recommendations more than staff based on the statistical significant differences of these three measures. The difference in topical_relevance measure could be attributed to the expectations of the participant. As mentioned earlier, broad input topics were an issue for staff and they expected the system to retrieve papers very much relevant to their usage context which they had in mind. In contrast, students evaluated the recommended papers for closeness to the input topic, hence they evaluated more favourably.

Among the 14 evaluation measures, good_spread, good_mix, good_list and user_satisfaction are the overall quality measures. Since the students rated highly for the measures good_mix and good_list, there appears to be a natural propensity to be more satisfied with the results. At an overall level, the measures with the highest mean values were topical_relevance ($M = 4.13$ for students, $M = 3.84$ for staff), good_spread ($M = 3.90$ for students, $M = 3.71$ for staff), diversity ($M = 3.94$ for stu-

**Table 6**
Independent samples *t*-test results.

| Measure | *t* value | Students *M* (SD) | Staff *M* (SD) |
|---|---|---|---|
| Topical_Relevance | 2.201* | 4.13 (0.713) | 3.84 (0.773) |
| Good_Spread | 1.401 | 3.90 (0.783) | 3.71 (0.764) |
| Diversity | 1.067 | 3.94 (0.744) | 3.80 (0.714) |
| Interdisciplinarity | 0.254 | 3.58 (0.860) | 3.54 (0.846) |
| Popularity | 0.638 | 3.61 (0.837) | 3.51 (0.928) |
| Recency | 1.296 | 3.65 (0.960) | 3.43 (0.957) |
| Good_Mix | 1.923* | 3.82 (0.800) | 3.54 (0.863) |
| Familiarity | 0.828 | 3.42 (0.860) | 3.29 (0.980) |
| Novelty | −1.676* | 2.74 (0.904) | 3.01 (0.955) |
| Serendipity | −0.455 | 3.48 (0.882) | 3.56 (0.958) |
| Usefulness | 0.985 | 3.74 (0.700) | 3.60 (0.923) |
| Good_List | 1.912* | 3.66 (0.745) | 3.37 (0.966) |
| Expansion_Required | 1.149 | 3.97 (0.677) | 3.80 (0.957) |
| User_Satisfaction | 1.818* | 3.76 (0.619) | 3.51 (0.880) |

*Note*. The differences between the two groups are statistically significant for the measures - Topical_Relevance, Good_Mix, Novelty, Good_List and User_Satisfaction.
* Indicates $p < 0.05$.

**Table 7**
Spearman correlation between measures.

| Measure | Measure | R (95% CI) |
|---|---|---|
| Popularity | Usefulness | 0.62 |
| | Good_List | 0.57 |
| | User_Satisfaction | 0.58 |
| Usefulness | Good_List | 0.67 |
| | User_Satisfaction | 0.65 |
| Good_List | User_Satisfaction | 0.72 |

*Note*. Only the measure combinations with R > 0.50 are displayed.

dents, $M = 3.80$ for staff), good mix ($M = 3.82$ for students, $M = 3.54$ for staff) and user_satisfaction ($M = 3.76$ for students, $M = 3.51$ for staff).

*6.7.3. Correlation and regression analysis*

The measure combinations with high degree of correlation (greater than 0.5) are displayed in Table 7. The popularity measure was found to be highly correlated with three output quality measures usefulness (R = 0.62), good_list (R = 0.57) and user_satisfaction (R = 0.58). This finding indicates participants' expectation of finding popular papers in the list. If more popular/seminal papers are found in the list, the participants' satisfaction with the reading list increases. The other correlation results are on expected lines with usefulness having positive correlations with good_list (R = 0.67) and user_satisfaction (R = 0.65), indicating that quality measures tend to be rated similarly.

Results from multiple linear regression testing are displayed in Table 8. The model was built with user_satisfaction as the dependent variable and the other 13 measures as the independent variables. The multiple correlation coefficient R value of 0.84 and the adjusted $R^2$ value of 0.67 indicate decent level of prediction at a statistically significant level. The model fit could potentially improve with more participants. Three independent variables familiarity, usefulness and good_list were found to be statistically significant predictors in the model. Popularity had a relatively higher estimate albeit not at a statistically significant level. Even though, the presence of familiar papers had impact on the user satisfaction, the scenario might turn out to be different when researchers are collecting papers for a new research topic. In such a situation, most of the papers would be novel; therefore the reliance on survey papers and popular papers would be high. Interestingly, this finding is different from an earlier study (Al-Maskari & Sanderson, 2010) where familiarity and user satisfaction measures were lacking correlation in a direct comparison of means. This finding was attributed due to the lack of impact of familiarity on user effectiveness. In the current study, we claim that the predictive ability of familiarity measure is restricted to evaluation environments, especially for scientific paper IR/RS tasks. In a real world scenario, when researchers search for papers, their expectation is to collect novel papers for improving their awareness on a given research topic (Athukorala, Hoggan, Lehtio, Ruotsalo, & Jacucci, 2013; Ellis & Haugan, 1997). Hence, discoverability of familiar papers in scientific paper IR/RS evaluation studies can be considered as a positive acknowledgement for the technique's performance. However, this finding is to be validated in future studies with a similar or a bigger sample size.

**Table 8**
Multiple linear regression results.

| | Estimate | SE | t value | *p* |
|---|---|---|---|---|
| Intercept | −0.051 | 0.411 | −0.125 | 0.901 |
| Topical_Relevance | 0.056 | 0.063 | 0.891 | 0.375 |
| Good_Spread | 0.055 | 0.061 | 0.893 | 0.373 |
| Diversity | −0.036 | 0.064 | −0.569 | 0.570 |
| Interdisciplinarity | 0.036 | 0.057 | 0.637 | 0.525 |
| Popularity | 0.101 | 0.062 | 1.622 | 0.107 |
| Recency | 0.017 | 0.05 | 0.342 | 0.733 |
| Good_Mix | 0.087 | 0.06 | 1.461 | 0.147 |
| **Familiarity** | **0.164** | **0.055** | **2.974** | **0.004** |
| Novelty | 0.035 | 0.053 | 0.658 | 0.512 |
| Serendipity | 0.025 | 0.044 | 0.573 | 0.568 |
| **Usefulness** | **0.175** | **0.073** | **2.407** | **0.018** |
| **Good_List** | **0.375** | **0.076** | **4.968** | **0.000** |
| Expansion_Required | −0.045 | 0.049 | −0.916 | 0.362 |

Residual standard error: 0.439 on 118 *df*
R: 0.843 Multiple $R^2$: 0.711, Adjusted $R^2$: 0.679
*F*-statistic: 22.327 on 13 and 118 *df*, *p* value: 1.391e−07

*Note*. The measures highlighted in boldface, are the statistically significant predictors with $p < 0.05$.

## 7. Limitations

There are certain limitations with the input mode of the retrieval technique and the user evaluation study. The retrieval technique 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 research topics. This scenario could have caused certain relevant papers to be missed during retrieval. The latest papers in the dataset were published in 2011. Few participants indicated that they expected to see recently published papers. 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.

## 8. Conclusions

Literature review (LR) is a crucial part of research projects as the subsequent stages in the scientific publication lifecycle are dependent on its outcomes. Each activity in LR requires a certain amount of dexterity, right from searching for papers in the relevant sources to summarizing and synthesizing the state-of-the-art literature. At the start of LR, the intent is to gain a holistic understanding of the research area so that focussed and directed searching could be performed subsequently. Researchers attempt to accumulate a representative set of papers for the aforementioned purpose. This set of papers can be called as the reading list. Prior studies have taken the approach of mainly focussing on seminal papers for automatically building the reading list. In this paper, we have raised the case for other types of papers that need to be included as part of a reading list. Accordingly, four requirements of a reading list are put forth. A novel paper metric called as Topical and Peripheral Coverage (TPC) is conceptualized for meeting the four requirements. This metric is calculated based on the author-specified keywords provided in research papers. In order to exhibit the utility of the metric, a corresponding retrieval technique called the AKR technique, has been proposed towards building a reading list. This technique belongs to the paradigm of IR and RS studies where the final retrieval of resources is seen as a ranking problem.

An offline evaluation experiment was conducted with four different variants of the AKR technique along with three other baseline approaches from earlier studies. HITS enhanced variant of the AKR technique was found to the best technique that satisfied most of the requirements. This technique was implemented as one of the three tasks in the recently developed Rec4LRW system, a prototype meant for assisting researchers in LR and manuscript preparation tasks. A user evaluation study was conducted with participation from 132 researchers, for evaluating the task performance. Study results show that students group found the task to be more useful than staff group. Participants felt that the system was able to put together a good mix of different types of papers which would be useful at the start of LR. Among the evaluated measures, popularity was found to be highly correlated with the output quality measures. Regression test results indicated that familiarity, usefulness and 'agreeability on a good list' were strong predictors for increased user satisfaction. The limited count of recommended papers and restrictive nature of the ACMDL dataset to certain disciplines, were raised as the key concerns during the study. Novice researchers (especially students) are expected to appreciate this type of automated tasks since they are in the process of learning the required search skills. These skills are not only restricted to knowing about the different information sources and methods of conducting advanced search but also in evaluating the information objects (research

papers). For the task of building reading list at start of LR, the proposed technique and the prototype implementation are to be considered as a mitigatory method for bridging the skills gap between novices and experts.

There are various implications of the current study. As part of theoretical implications, the four requirements identified for reading lists, can be used as a theoretical basis for future studies concentrating on the task of building reading lists for LR. The TPC is a novel metric for measuring the relative position of a paper in citation network, based on author-specified keywords metadata field. This metric is meant to address the extended set of requirements of a reading list. It also helps in giving higher weightage to interdisciplinary research papers. Through the current study, we have also shown its extensibility by combining with HITS algorithm. The composite ranking in AKR technique is a novel ranking approach where we have shown the process to use TPC value for ranking papers in the final list. The AKR technique can be practically implemented in scientific paper digital libraries where the author-specified keywords are separately indexed with other metadata fields. The AKR technique is a contribution to the IR studies that focus on the ranking aspect of the IR paradigm, since we perceive the task of building a reading list as a ranking problem in our research. As a part of research implications, the user evaluation study results show that (i) the presence of popular or seminal papers is largely correlated to user satisfaction and (ii) students prefer this type of automated task more than staff. These two findings are consistent with earlier studies (Du & Evans, 2011; Ekstrand et al., 2010). It is to 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.

As a part of future work, we are planning to evaluate the next release of Rec4LRW system with more UI features so that researchers could sieve through the results for better understanding the recommended papers. Additionally, we plan to set user roles in the system so that personalization and customization features are made available for users. The corpus of the Rec4LRW system will be enriched with research papers from other prominent sources in the next release.

## Acknowledgments

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

## References

Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology, 61*(5), 859–868. *Published* http://doi.org/10.1002/asi.21300.

Amolochitis, E., Christou, I. T., Tan, Z.-H., & Prasad, R. (2013). A heuristic hierarchical scheme for academic search and retrieval. *Information Processing and Management, 49*(6), 1326–1343. http://doi.org/10.1016/j.ipm.2013.07.002.

AnyStyle (2015). *AnyStyle.io* Retrieved July 22, 2015, from http://anystyle.io/.

Athukorala, K., Hoggan, E., Lehtio, A., Ruotsalo, T., & Jacucci, G. (2013). Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. *Proceedings of the American Society for Information Science and Technology, 50*(1), 1–11. http://doi.org/10.1002/meet.14505001041.

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.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407–424.

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.

Beel, J., & Gipp, B. (2009). Google scholar's ranking algorithm: The impact of citation counts (An empirical study). In *Third international conference on research challenges in information science, 2009. RCIS 2009* (pp. 439–446).

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in Human Behavior, 21*(3), 487–508.

Bullock, S. M. (2013). Using digital technologies to support Self-Directed Learning for preservice teacher education. *Curriculum Journal, 24*(1), 103–120. http://doi.org/10.1080/09585176.2012.744695.

Castells, P., Vargas, S., & Wang, J. (2009). Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *Proceedings of international workshop on diversity in document retrieval (DDR)* (pp. 29–37). Retrieved from http://hdl.handle.net/10486/666094.

Chen, C.-H., Mayanglambam, S. D., Hsu, F.-Y., Lu, C.-Y., Lee, H.-M., & Ho, J.-M. (2011). Novelty paper recommendation using citation authority diffusion. In *Technologies and applications of artificial intelligence (TAAI), 2011 international conference on* (pp. 126–131). IEEE.

Collins (2016). Definition of "reading list". *Collins english dictionary* Retrieved February 18, 2016, from http://www.collinsdictionary.com/dictionary/english/reading-list.

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.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web* (pp. 613–622).

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.

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation, 49*(4), 356–369. http://doi.org/10.1108/eb026919.

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.

Feng, M.-H., Chan, K.-H., Chen, H.-Y., Tsai, M.-F., Yeh, M.-Y., & Lin, S.-D. (2016). An efficient solution to reinforce paper ranking using author/venue/citation information - The winner's solution for WSDM Cup 2016. In *Proceedings of the WSDM Cup 2016 - Entity ranking challenge workshop* Retrieved from https://doc.co/LLcDRM.

Fidel, R. (2012). *Human information interaction: An ecological approach to information behavior*. MIT Press.

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on digital libraries* (pp. 89–98).

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM, 35*(12), 61–70.

Goodhue, D. L. (1995). Understanding user evaluations of information systems. *Management Science, 41*(12), 1827–1844. http://doi.org/10.1287/mnsc.41.12.1827.
Hjørland, B. (2013). Citation analysis: A social and dynamic approach to knowledge organization. *Information Processing & Management, 49*(6), 1313–1325. http://dx.doi.org/10.1016/j.ipm.2013.07.001.
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).
Jardine, J. G. (2014). *Automatically generating reading lists* Retrieved from https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-848.pdf.
Jardine, J., & Teufel, S. (2014). Topical PageRank : A model of scientific expertise for bibliographic search. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 501–510). Retrieved from.
Jesson, J., Matheson, L., & Lacey, F. M. (2011). *Doing your literature review: Traditional and systematic techniques*. SAGE Publications Retrieved from https://books.google.com/books?hl=en&lr=&id=LUhdBAAAQBAJ&pgis=1.
Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments: Part 2. *Information Processing & Management, 36*(6), 809–840. http://doi.org/10.1016/S0306-4573(00)00016-9.
Karlsson, L., Koivula, L., Ruokonen, I., Kajaani, P., Antikainen, L., & Ruismäki, H. (2012). From novice to expert: Information seeking processes of university students and researchers. *Procedia - Social and Behavioral Sciences, 45*, 577–587. http://doi.org/10.1016/j.sbspro.2012.06.595.
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.
Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM Computing Surveys (CSUR), 31*(4). 10.1145/345966.345982. Article No. 5.
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.
Lawrence, S., Lee Giles, C., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer, 32*(6), 67–71. http://doi.org/.1109/2.769447.
Leedy, P. D., & Ormrod, J. E. (2005). *Practical research: Planning and design*. Upper Saddle River, N.J: Prentice Hall Retrieved from https://books.google.com.sg/books/about/Practical_Research.html?id=MipiQgAACAAJ&pgis=1.
Lehmann, S., Lautrup, B., & Jackson, A. D. (2003). Citation networks in high energy Physics. *Physical Review E, 68*(2). http://doi.org/10.1103/PhysRevE.68.026113.
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.
Liu, Y., & Lin, Y. (2007). Supervised HITS algorithm for MEDLINE citation ranking. In *Proceedings of the 7th IEEE international conference on bioinformatics and bioengineering, BIBE* (pp. 1323–1327).
Markauskaite, L. (2007). Exploring the structure of trainee teachers' ICT literacy: The main components of, and relationships between, general cognitive and technical capabilities. *Educational Technology Research and Development, 55*(6), 547–572. http://doi.org/10.1007/s11423-007-9043-8.
Mcnee, S. M. (2006). *Meeting user information needs in recommender systems*. Proquest.
Merton, R. K. (1968). The Matthew effect in science. *Science*. http://doi.org/10.1126/science.159.3810.56.
Niu, X., & Hemminger, B. M. (2012). A study of factors that affect the information-seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology, 63*(2), 336–353.
Page, L., Brin, S., Motwami, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web* Retrieved from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.
Pihur, V., & Datta, S. (2009). RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* Retrieved from http://www.biomedcentral.com/1471-2105/10/62.
Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to recommender systems handbook*. US: Springer Retrieved from http://link.springer.com/.
Ridley, D. (2012). *The literature review: A step-by-step guide for students*. Sage.
Runelöv, M. (2015). *Finding seminal scientific publications with graph mining finding seminal scientific publications with graph mining*. KTH Royal Institute of Technology Retrieved from http://www.diva-portal.org/smash/get/diva2:847503/FULLTEXT01.pdf.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. http://doi.org/10.1016/0306-4573(88)90021-0.
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).
Singh, A. P., Shubhankar, K., & Pudi, V. (2011). An efficient algorithm for ranking research papers based on citation network. In *2011 3rd Conference on data mining and optimization (DMO)* (pp. 88–95).
Spezi, V. (2016). Is information-seeking behavior of doctoral students changing?: A review of the literature (2010–2015). *New Review of Academic Librarianship*, 1–29. http://doi.org/10.1080/13614533.2015.1127831.
Strohman, T., Croft, W. B., & Jensen, D. (2007). Recommending citations for academic papers. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07* (p. 705). New York, New York, USA: ACM Press.
Tabatabai, D., & Shore, B. M. (2005). How experts and novices search the Web. *Library & Information Science Research, 27*(2), 222–248. http://doi.org/10.1016/j.lisr.2005.01.005.
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 Deursen, A. J. A. M., & Van Dijk, J. A. G. M. (2009). Using the Internet: Skill related problems in users' online behavior. *Interacting with Computers, 21*(5), 393–402. http://doi.org/10.1016/j.intcom.2009.06.005.
Waltman, L., & Yan, E. (2014). Measuring scholarly impact. In *Measuring scholarly impact: Methods and practice* (pp. 285–320). Springer.
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.
White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163–171.
Yang, S., & Han, R. (2015). Breadth and depth of citation distribution. *Information Processing and Management, 51*(2), 130–140. http://doi.org/10.1016/j.ipm.2014.12.003.
Yoo, I., & Mosa, A. S. M. (2015). Analysis of PubMed user sessions using a full-day PubMed query log: A comparison of experienced and nonexperienced PubMed users. *JMIR Medical Informatics, 3*(3). http://doi.org/10.2196/medinform.3740.
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.