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.