Decision Support Systems xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

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

Academic paper recommender system using multilevel simultaneous citation networks

Jieun Son, Seoung Bum Kim *

Department of Industrial Management Engineering, Korea University, 145 Anam-Ro, Seoungbuk-Gu, Seoul 136-713, South Korea

ARTICLE INFO

Article history:
Received 20 February 2017
Received in revised form 19 October 2017
Accepted 19 October 2017
Available online xxxx

Keywords:
Academic paper recommender
Citation networks
Recommender systems
Text mining

ABSTRACT

Researchers typically need to filter several academic papers to find those relevant to their research. This filtering is cumbersome and time-consuming because the number of published academic papers is growing exponentially. Some researchers have focused on developing better recommender systems for academic papers by using citation analysis and content analysis. Most traditional content analysis is implemented using a keyword matching process, and thus it cannot consider the semantic contexts of items. Further, citation analysis-based techniques rely on the number of links directly citing or being cited in a single-level network. Consequently, it may be difficult to recommend the appropriate papers when the paper of interest does not have enough citation information. To address these problems, we propose a recommendation system for academic papers that combines citation analysis and network analysis. The proposed method is based on multilevel citation networks that compare all the indirectly linked papers to the paper of interest to inspect the structural and semantic relationships among them. Thus, the proposed method tends to recommend informative and useful papers related to both the research topic and the academic theory. The comparison results based on real data showed that the proposed method outperformed the Google Scholar and SCOPUS algorithms.

1. Introduction

Recommender systems have become popular and have begun to attract increasing attention from both academia and industry [1]. However, compared with other recommender applications, such as those for books, movies, and music, fewer studies have examined recommender systems for academic papers. Researchers typically need to filter a substantial number of academic papers to find those relevant to their research. This filtering is cumbersome and time consuming because the number of published academic papers is growing exponentially. Consequently, there is an urgent need for efficient academic paper recommender systems [2].

Previous studies have focused on finding better recommender systems for academic papers relevant to specific research topics. One of the most commonly used recommendation approaches is collaborative filtering. However, extensive studies indicated that this approach has inherent problems, such as data sparsity and ratings imbalance [3]. To address these problems, related recommendation techniques, such as content-based filtering, network analysis, and information retrieval, are being studied [4,5].

Content-based filtering approaches create relationships between items by analyzing their inherent characteristics. In most content-based filtering systems for textual applications (e.g., academic papers), item descriptions are keywords extracted from text. However, keyword-based systems create numerous complications that originate from natural language ambiguity. Further, keyword-based systems are unable to capture the semantics of user interests because they are primarily driven by string matching operations [6].

Citation networks based on citation-related connections within scientific literature have been used to compute relatedness among academic papers. Link-based techniques, such as co-citation and bibliographic coupling, measure relevance by focusing on neighbors. Co-citation is the similarity measure for two papers cited together by other papers, and bibliographic coupling is the similarity measure for two papers that refer to the same paper. However, recommender systems that use link-based techniques cannot consider complex relationships among papers because these techniques count the number of links directly cited in single-level networks [7].

Selecting relevant academic papers is similar to information retrieval, an activity in which information resources relevant to the desired information are selected from amongst a collection of such resources. The mainstream tool used for information retrieval research is ranking algorithms. The PageRank algorithm, in particular, has been used to produce a better global ranking of search results [8]. However, although

* Corresponding author.
E-mail addresses: jieunson@korea.ac.kr (J. Son), sbkim1@korea.ac.kr (S.B. Kim).

https://doi.org/10.1016/j.dss.2017.10.011
0167-9236/

PageRank can estimate the authority of a paper, one of its drawbacks is that it ranks papers based on the citation count. Therefore, recent articles always score low and consequently are not recommended by the algorithm [9].

1.1. Contributions of the paper

The main contributions of our research can be summarized as follows. First, we propose a novel recommendation system for academic papers that combines citation analysis and network analysis. The proposed method is based on multilevel citation networks that compare all indirectly linked papers to the paper of interest to inspect the structural and semantic relationships among them. Our main research objective in this study is to consider the mutual relationships among the papers in a broad network beyond a single level, and to evaluate the significance of each paper through certain centrality measures. Second, the lack of a citation count notwithstanding, the proposed method can find influential papers using centrality measures that are derived from a citation network. Finally, we found that the proposed method outperformed the existing methods, Google Scholar and SCOPUS, based on user satisfaction data. We asked users to receive recommendations from these algorithms and rate the recommended item lists based on their satisfaction with the results.

1.2. Organization of the paper

The remainder of this paper is organized as follows. In the related work section, we review the academic paper recommendation approaches proposed in existing literature. In the proposed method section, we detail our approach to academic paper recommendation by using multilevel simultaneous citation networks. The experimental evaluation section presents and discusses our experimental results and evaluation. The last section concludes this paper.

2. Related work

2.1. Content-based filtering and collaborative filtering

Two types of algorithms are typically used in recommender systems: collaborative filtering and content-based filtering [10]. Collaborative filtering algorithms match users in a system based on the similarity of the past ratings provided by each user, and then recommend items that similar users have liked. Collaborative filtering requires a matrix consisting of user ratings for a particular item [11]. Therefore, collaborative filtering cannot generate accurate recommendations without sufficient initial ratings from users. This problem also occurs in the domain of academic papers because it is very difficult to gather user score information in digital libraries. To resolve this problem, researchers have focused their attention on creating a matrix of collaborative filtering ratings from the citation web between academic papers. In this ratings matrix, authors are represented as rows and papers as columns. Each entry is a rating for certain papers that the authors have cited [12]. However, in many cases the authors have published papers in various technical fields or changed their areas of interest during their careers. For example, a researcher who started his or her research activities in electrical engineering may later write many papers about statistics. Thus, it may be difficult to find a similar author group and recommend proper papers to target users when a database contains many such cases.

Content-based filtering algorithms recommend items to users based on their description [13]. Applications of content-based filtering in academic paper recommender systems rely on the ability to compare the similarities of complete text or keywords because text-based features are excellent for classifying papers [14]. The "related documents" function of SCOPUS is one example of a content-based filtering approach. The SCOPUS system defines keywords for a research paper, and the indexed keywords can be automatically imported as tags. Papers that contain one or more words in common with those in the paper of interest are returned as relevant. For the recommendation, the papers are selected in the order of the highest matching frequency of the keywords [15]. However, in some cases, text features are not as good at finding a related paper. Although a paper may be conceptually similar to the paper a user may be interested in, it may use a different vocabulary. In this case, the relevant paper may be overlooked. Conversely, the same word can be used in papers in many different fields; this can then result in the wrong papers being recommended to users. For example, "port" is an endpoint of communication in a computer operating system. However, "port" also has other meanings in unrelated contexts; it can mean "harbor" in the context of shipping, and it is also used to refer to a type of wine.

2.2. Information retrieval techniques

PageRank is one of the methods that Google uses to evaluate the importance of webpages to improve the quality of web search engines [7]. This algorithm has been widely applied not only to rank web search results, but also to recommend academic papers [16]. Google Scholar primarily uses PageRank techniques to identify papers related to the paper of interest. Quoting from https://scholar.google.com/intl/en/scholar/about.html, "Google Scholar aims to rank papers the way researchers do, weighing the full text of each paper, who it was written by, where it was published, as well as how often and how recently it has been cited in other papers" [17]. To provide recommendations, the "related articles" function of Google Scholar presents a list of closely related papers, ranked primarily by how similar these papers are to the paper of interest [18]. Although PageRank is a good method for determining the authority of a paper, it tends to rank papers based primarily on the number of citations. As a result, recent papers are always ranked low, even when the paper is known as eminent literature. This is an important limitation because recent papers may be important to researchers who wish to understand current issues and to set research directions. Bethard and Jurafsky [19] proposed integrating a keyword-based algorithm and citation information for learning literature search models. The main idea of their integrated approach is to look for similar terms and topics among the articles. Therefore, they include the classic term frequency-inverse document frequency (TF-IDF), which represents both the user query and the word counts in document and latent Dirichlet allocation (LDA). PageRank and citation count are only used to boost the article.

2.3. Citation networks

Citation analysis, used in large applications such as patent analysis and document analysis, refers references in one item to another item. While the two approaches presented above are based on similarity, the citation network is based on relational information. Therefore, it is useful for understanding the relationship between subjects, the flow of history, and publication trends [20]. The citation analysis of academic papers in particular is important because it can directly reveal papers closely related to the query paper. Several previous studies recommended papers for a manuscript containing a partial list of citations. Co-citation analysis, introduced by Small [21], is one of the first applications of co-occurrence. Small suggested that the more two papers are related to each other, the more often they are co-cited. Liang [22] presented graph networks that show how the papers are connected through citations. Connections are based on bibliographic coupling and co-citation strength [23,24]. Once a graph was built, graph metrics were used to find recommendation candidates. One or several input papers are given as the paper of interest and random walks were conducted to find the most popular items in the network graph [25,26]

Much of the literature on citation analysis considers just one level, directly linked to nodes [27]. However, in single-level analysis, the

Please cite this article as: J. Son, S.B. Kim, Academic paper recommender system using multilevel simultaneous citation networks, Decision Support Systems (2017), https://doi.org/10.1016/j.dss.2017.10.011