Information Sciences 334–335 (2016) 273–292

Contents lists available at ScienceDirect

Information Sciences

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

SimCC: A novel method to consider both content and citations for computing similarity of scientific papers

Masoud Reyhani Hamedani[a], Sang-Wook Kim[a,*], Dong-Jin Kim[b]

[a] Department of Computer and Software, Hanyang University, Seoul, Republic of Korea
[b] NHN Institute of The Next Network, Republic of Korea

ARTICLE INFO

Article history:
Received 7 November 2014
Revised 20 November 2015
Accepted 5 December 2015
Available online 11 December 2015

Keywords:
Citations
Content
Contribution score
Scientific papers
Similarity

ABSTRACT

To compute the similarity of scientific papers, text-based similarity measures, link-based similarity measures, and hybrid methods can be applied. The text-based and link-based similarity measures take into account *only* a single aspect of scientific papers, content or citations, respectively. The hybrid methods consider both content and citations; however, they do not carefully consider the *relation* between the content of a pair of papers involved in a citation relationship. In this paper, we propose a novel method, *SimCC* (similarity based on content and citations), that considers *both* aspects, content and citations, to compute the similarity of scientific papers. Unlike previous methods, SimCC effectively reflects both *content* and *authority* of scientific papers *simultaneously* in similarity computation by applying a new RA (relevance and authority) weighting scheme. Also, we propose an RA+R weighting scheme to consider the *recency* of papers and an RA+E weighting scheme to take into account the *author expertise* of papers in similarity computation. The effectiveness of our proposed method is demonstrated by extensive experiments on a real-world dataset of scientific papers. The results show that our method achieves more than 100% improvement in accuracy in comparison with previous methods.

1. Introduction

Scientific papers are one of the primary sources to share information and knowledge among researchers. As the number of scientific papers being published is increasing dramatically, researchers face a serious problem to find relevant papers to the areas of their interest and to make sure whether their research problems are truly novel. Scientific literature search engines such as Google Scholar[1], CiteSeerX[2], and Microsoft Academic Search[3] are useful to alleviate this problem by providing search and/or recommendation services. The similarity measure is one of the most challenging issues in scientific literature search engines to find those papers relevant to the user requirement.

A scientific paper contains two related aspects: content and citations. The content is the main aspect of a paper that demonstrates its context. The citations, selected *carefully* by the authors according to the content, are a set of references to the *related* and *authoritative* papers called *cited papers*. If paper q cites paper p, as shown in Fig. 1, we say that there is a *citation relationship*

[*] Corresponding author. Tel.: +82 10 6749 6392.
E-mail addresses: masoud@agape.hanyang.ac.kr (M. Reyhani Hamedani), wook@hanyang.ac.kr (S.-W. Kim), djkim@nhnnext.org (D.-J. Kim).

[1] http://scholar.google.co.kr

[2] http://citeseerx.ist.psu.edu

[3] http://academic.research.microsoft.com

http://dx.doi.org/10.1016/j.ins.2015.12.001

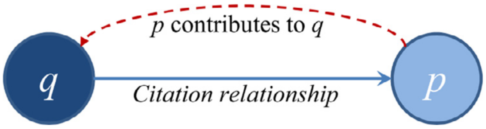

Fig. 1. Citation relationship between two papers.

from paper q to paper p. The contents of two papers in a citation relationship are likely to be *related* and not completely independent. By inserting citations in a paper, we can improve the content of the paper and make it more understandable for readers [8]. In the case of Fig. 1, we say that the cited paper p *contributes* to the citing paper q because p improves the content of q. In order to design accurate similarity measures for scientific papers, *both of these related aspects*, content and citations, should be taken into account.

To compute the similarity of scientific papers, various methods have been proposed. The text-based similarity measures such as Cosine [23], Dice Coefficient [23], Kullback–Leibler Distance (KLD) [1,4,33], and BM25 [3,21,34] focus only on the *content* of papers to compute the similarity but neglect the citations completely. The link-based similarity measures such as Bibliographic coupling [20], Co-citation [30], Amsler [2], SimRank [17], P-Rank [36], and rvs-SimRank [36] consider only the *citation relationship* among scientific papers to compute the similarity and totally ignore their content. The hybrid methods [6,26,31,35] consider *both* content and citations to compute the similarity. In references [6,31,35], the content of papers is enriched by adding additional terms from those papers that have direct citation relationship with them. Then, similarities between papers are computed based on enriched contents. These methods *simply combine* the content of a paper with the content of those papers that cite it and are cited by it. In reference [26], the similarities of scientific papers are computed based on the content and citation relationships *separately* and their resulting similarities are combined into a single value. This method does not exploit the relation between the content of two papers involved in a citation relationship.

In the literature, the utilization of content and citations is not only restricted to the topic of similarity computation. In reference [16], content and citations both are utilized to reveal the linkage between research topics in the domain of technology and social sciences. Reference [13] proposes a method for detecting emerging topics by using core papers, which are obtained by utilizing both content and citations. In reference [10], the effectiveness of utilizing weighted citation graphs assigned with different weights depending on the content and citations is investigated in detecting research fronts. In reference [5], the effectiveness of different link-based similarity measures and a hybrid method is evaluated when employed in clustering of scientific papers. In reference [29], link-based and content-based measures are utilized to calculate the relatedness between distinct communities in a citation graph. In reference [18], the similarity between a pair of co-cited authors is computed by considering the similarities between citation sentences in their citing papers. Reference [8] provides a comprehensive study on theory, approaches, and applications of content-based citation analysis (CCA). NetPLSA [24] considers both content and a graph structure in topic modeling.

In this paper, we propose a novel method called SimCC[4] (similarity based on content and citations) that considers both *content* and *citations* to compute the similarity of scientific papers. SimCC performs both *feature extraction* and *similarity computation*. For feature extraction, a paper is represented as an n-dimensional feature vector, called *F-vector*. To compute the weight of a term in the F-vector, called *RA score*, SimCC applies a new RA (relevance and authority) weighting scheme, where the weight not only represents how *relevant* the term is to the paper but also how much the paper is *authoritative* on that term. To measure the relevance score of a term to a paper, any weighting schemes such as TF, TF-IDF, and BM25 can be applied. To measure the authority of a paper on a term, SimCC analyzes the relation between the content of the paper and the content of those papers that cite it directly or indirectly as a contribution from the cited paper to the citing ones. As an example, when paper q cites paper p as shown in Fig. 1, it means p is an authoritative paper on some topics discussed in q and contributes to q to improve the content of q. The number of citations to a paper is not enough to measure its authority. *Rather, it is more important to measure how much a paper contributes to other papers that cite it directly or indirectly to improve their content*. To this end, we define the notion of a *contribution score* as the degree of authority for a paper over another *single paper* on a *specific term* via a *specific citation path*. The contribution score is the key factor in our method and is computed for a term in a paper *individually*.

The summation of all contribution scores of a paper on a specific term over other papers is considered as the authority score of the paper on that term. The RA scheme combines the relevance score and the authority score of a term in a paper together as the RA score of that term. More specifically, SimCC *supplements* the content of a paper (relevance scores) with its authority (authority scores). Therefore, SimCC reflects *both* content and authority *simultaneously* in F-vectors and consequently in the similarity computation as well. For similarity computation, SimCC can employ *any* text-based similarity measures to compute the similarity of a pair of papers. We also propose the RA+R and RA+E weighting schemes as the extensions to the basic RA scheme that consider other factors, in addition to the content and citations, in similarity computation. The RA+R scheme considers the *recency* of papers to alleviate the lack of citations to the recently-published papers. The RA+E scheme considers the *author expertise* to retrieve those papers that are written by expert authors.

[4] The initial idea has been presented with some preliminary experimental results in ACM CIKM 2013 as a short paper [25]. The current paper is an extended version that contains some new ideas, formulations, and more extensive experimental results.