Information Processing and Management 54 (2018) 203–218

Contents lists available at ScienceDirect

Information Processing and Management

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

Ranking themes on co-word networks: Exploring the relationships among different metrics

Zhao Wanying[a], Mao Jin[a], Lu Kun[b,*]

[a] School of Information Management, Wuhan University, No.299 Bayi Street, 430072 Wuhan, China
[b] School of Library and Information Studies, University of Oklahoma, 401 W. Brooks St., Norman, OK 73019-6032, USA

ARTICLE INFO

Keywords:
Co-word analysis
Network structure
Frequency-based metrics
Network-based metrics
Co-word network simulation

ABSTRACT

As network analysis methods prevail, more metrics are applied to co-word networks to reveal hot topics in a field. However, few studies have examined the relationships among these metrics. To bridge this gap, this study explores the relationships among different ranking metrics, including one frequency-based and six network-based metrics, in order to understand the impact of network structural features on ranking themes on co-word networks. We collected bibliographic data from three disciplines from Web of Science (WoS), and generated 40 simulation networks following the preferential attachment assumption. Correlation analysis on the empirical and simulated networks shows strong relationships among the metrics. Their relationships are consistent across disciplines. The metrics can be categorized into three groups according to the strength of their correlations, where Degree Centrality, H-index, and Coreness are in one group, Betweenness Centrality, Clustering Coefficient, and frequency in another, and Weighted PageRank by itself. Regression analysis on the simulation networks reveals that network topology properties, such as connectivity, sparsity, and aggregation, influence the relationships among selected metrics. In addition, when comparing the top keywords ranked by the metrics in the three disciplines, we found the metrics exhibit different discriminative capacity. Coreness and H-index may be better suited for categorizing keywords rather than ranking keywords. Findings from this study contribute to a better understanding of the relationships among different metrics and provide guidance for using them effectively in different contexts.

1. Introduction

Keywords of scientific articles, either manually assigned (author keywords and subject descriptors) or automatically generated, are widely used to reveal themes, structures, and development of a field, for example, through co-word analysis (Callon, Courtial, Turner, & Bauin, 1983). Unlike other bibliometric methods, such as co-citation analysis or co-author analysis, co-word analysis is a content-based method from which the results can be directly interpreted according to their semantics. Term frequency, defined as the number of occurrences of a term in a collection, is often used to identify important themes of a field (Khasseh, Soheili, Moghaddam, & Chelak, 2017). The assumption is that a frequently investigated topic could be an important theme in the field. Identifying themes by frequency is simple and convenient. However, this metric ignores the co-occurrence relationships among keywords, which can be captured by co-word networks. The structures of co-word networks carry information beyond term frequency, which can be used to measure the importance of keywords.

As social network analysis becomes popular, co-word analysis shifts to network-based metrics for measuring important themes

* Corresponding author.
E-mail address: kunlu@ou.edu (K. Lu).

https://doi.org/10.1016/j.ipm.2017.11.005
Received 29 June 2017; Received in revised form 10 October 2017; Accepted 14 November 2017

(Song & Cai, 2012). Network-based metrics rank the nodes (i.e. keywords) in co-word networks by considering the network topology. For example, Degree Centrality measures the importance of nodes through the number of incident edges. Nodes with higher values of degree centrality are regarded as more important. Some researchers have applied network-based metrics to co-word networks for identifying important themes in a research field (Ronda-Pupo & Guerras-Martin, 2012; Song & Cai, 2012). However, few studies have examined the relationships among traditional metrics (e.g., frequency) and network-based metrics on co-word networks. In this study, we explore the relationships among frequency and a number of popular network-based metrics, including Degree Centrality, Betweenness Centrality, Clustering Coefficient, Coreness, H-index, and Weighted PageRank. Our first research question is:

- What are the differences and correlations of these seven metrics in ranking themes revealed by keywords in a research field? In particular, are there any differences between frequency-based and network-based methods?

A further goal of this study is to examine the impact of structural features of co-word networks on the relationships among frequency-based and network-based metrics. The research question is:

- Do the relationships among different metrics differ depending on different disciplines or the different properties of networks, such as density, the number of vertices, the number of edges, etc.? If so, what are the factors that influence their relationships?

To address these questions, the seven metrics were compared using two groups of data. One group was the empirical data from three fields, and the other group was a collection of simulated co-word networks with different scales and other structural features. The simulated co-word networks were generated by following the generation process of real co-word networks, and were used to investigate the impact of structural features on the relationships between the seven metrics. The findings of this study help to understand the relationships among different methods in identifying important themes on a co-word network and provide guidance on how to use them effectively in different contexts.

This study extends previous work that has attempted the random walk method on co-word networks (Chiu & Lu, 2015), and aims to provide a more comprehensive understanding of the metrics. In particular, this study includes more datasets, a more comprehensive list of network-based metrics, and a simulation method for co-word networks.

2. Literature review

Work related to this study can be found in the following areas: co-word analysis method, metrics for co-word networks, and network simulation.

2.1. Co-word analysis method

As a commonly used method in information science over the past decades, co-word analysis is well known for its ability to reveal themes, structures, and development of a field by examining co-occurrences of term pairs from different parts of papers. Terms from titles (Besselaar & Heimeriks, 2006), abstracts (Ravikumar, Agrahari, & Singh, 2015), and full-texts (Janssens, Leta, Glänzel, & Moor, 2006) are frequently used. Also, author keywords (Cho, 2014) and subject terms (Ocholla, Onyancha, & Britz, 2010) are recently used in co-word analysis, the results of which have shed light on the structure and development of a research field.

Co-word analysis has been through two stages: the first stage is characterized by frequency-based analysis methods, while the second stage places more emphasis on term co-occurrences and the resulting network structure. In the first stage, co-word analysis is often combined with multidimensional scaling (MDS) or other clustering methods. It has been used to reveal the development of concepts (Ronda-Pupo & Guerras-Martin, 2012) or domains (Gan & Wang, 2015; Viedma-Del-Jesus, Perakakis, Muñoz, López-Herrera, & Vila, 2011), and to find hot topics (Liu, Chen, Liu, & Xie, 2016) or hidden topics (Milojević, Sugimoto, Yan, & Ding, 2011; Muñoz-Leiva, Sánchez-Fernández, Liébana-Cabanillas, & López-Herrera, 2012). However, research in this stage does not consider network structures. In the second stage, network analysis is applied to co-word networks (Hong et al., 2016; Liu, Hu, & Wang, 2012), which provides different metrics to measure network properties. These metrics can be categorized into global indicators that describe the overall properties of a network (e.g. density, diameter, and average degree), and local indicators that delineate the importance of individual vertices in a network (e.g. centrality). Many network-based metrics can be used to identify important themes on a co-word network. However, previous studies have not comprehensively investigated the relationships among different metrics for this purpose.

2.2. Metrics for co-word networks

Global metrics describe the entire network, especially the topology of a network. For example, the numbers of vertices and edges are the two simplest global metrics that define the size of a network. According to Wang, Li, and Chen (2012), connectivity, sparsity, aggregation, uniformity, and assortativity are five perspectives that reflect network topology. Connectivity concerns how strongly vertices connect with each, and thus it mainly focuses on the largest component of a network, which is a connected component of a network that has the most vertices. Common metrics of connectivity include metrics related to edges, such as the number of edges whose weights are 1. Unlike connectivity, sparsity focuses on network degree, which is often revealed by average degree and density. Aggregation reflects how closely vertices are connected with each other, which is usually measured by average distance and