Journal of Informetrics 10 (2016) 606–621

Contents lists available at ScienceDirect

Journal of Informetrics

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

Generation of topic evolution trees from heterogeneous bibliographic networks

Scott Jensen [a], Xiaozhong Liu [b,*], Yingying Yu [c], Staša Milojevic [d]

[a] San Jose State University, Lucas College and Graduate School of Business, MIS Department, One Washington Square, San Jose, CA 95192, USA
[b] Indiana University, School of Informatics and Computing, Dept. of Info. and Library Sciences, Wells Library, Room 027, 1320 E. 10th Street, Bloomington, IN 47405, USA
[c] Dalian Maritime University, College of Transportation Management, Department of Management Science and Engineering, Room 211, Guanli Building, No. 1, Linghai Road, Dalian, Liaoning Province 116026, China
[d] Indiana University, School of Informatics and Computing, Dept. of Info. and Library Sciences, Wells Library, Room 017, 1320 E. 10th Street, Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form 8 April 2016
Accepted 8 April 2016
Available online 6 May 2016

Keywords:
Topic evolution
Heterogeneous bibliographic network
Meta-path
Visualization

ABSTRACT

The volume of the existing research literature is such it can make it difficult to find highly relevant information and to develop an understanding of how a scientific topic has evolved. Prior research on topic evolution has often leveraged refinements to Latent Dirichlet Allocation (LDA) to identify emerging topics. However, such methods do not answer the question of which studies contributed to the evolution of a topic. In this paper we show that meta-paths over a heterogeneous bibliographic network (consisting of papers, authors and venues) can be used to identify the network elements that made the greatest contributions to a topic. In particular, by adding derived edges that capture the contribution of papers, authors, and venues to a topic (using PageRank algorithm), a restricted meta-path over the bibliographic network can be used to restrict the evolution of topics to the context of interest to a researcher. We use such restricted meta-paths to construct a topic evolution tree that can provide researchers with a web-based visualization of the evolution of a scientific topic in the context of interest to them. Compared to baseline networks without restrictions, we find that restricted networks provide more useful topic evolution trees.

1. Introduction

An exponential growth in the output of scientific literature (Bornmann & Mutz, 2015; Price, 1961, 1963; Van Raan, 2000) is one of the staples of contemporary science. In such an environment, scientists are becoming more specialized with narrower expertise (Jones, 2005, 2010), while the solutions of difficult problems in science and industry often require interdisciplinary approaches (Wagner et al., 2011) and team-based research (Falk-Krzesinski et al., 2010). Thus, ideally, scientists need to have deep knowledge in their own specialty and broad knowledge across a range of domains, i.e., be "T-shaped" (Barile, Franco, Nota, & Saviano, 2012; Donofrio, Spohrer, & Zadeh, 2010). While the Internet and electronic publications have made it easier to access an unprecedented volume and range of resources, this wealth of information can make it difficult to identify the best resources for learning the foundations of a specific topic or to identify the researchers, papers, and venues that have

* Corresponding author.
E-mail addresses: scott.jensen@sjsu.edu (S. Jensen), liu237@indiana.edu (X. Liu), uee870927@126.com (Y. Yu), smilojev@indiana.edu (S. Milojevic).

http://dx.doi.org/10.1016/j.joi.2016.04.002
1751-1577/

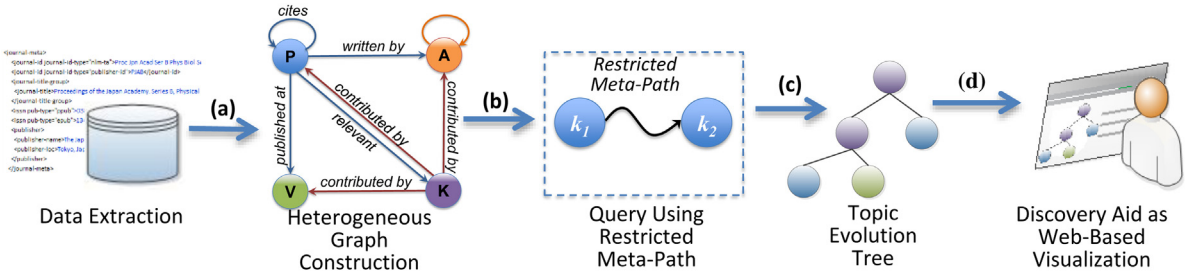

Fig. 1. Overview of the topic evolution tree (TET) methodology for identifying the evolution of scientific topics: (a) generating a heterogeneous graph from scholarly data (b) calculating the relatedness of topics in a user's context (c) generating a context-sensitive TET, and (d) visualizing the evolution of scholarly topics.

made the greatest contribution to a specific topic. As an example, each year the U.S. National Library of Medicine database of biomedical literature (MEDLINE) is growing by approximately 700,000 articles from 21,000 different journals and as of 2015 contains reference data on over 25 million resources.[1] While this abundance of resources provides an incredible opportunity, it can also overwhelm researchers, whether they are searching for information as part of learning a specialization or trying to develop a broad understanding of different fields.

While the current data deluge has exacerbated the problem of retrieving relevant documents, the problem itself is not new. The field of information retrieval has been proposing mostly topical solutions to the problem of retrieving relevant documents since the 1950s. At the same time, the field of bibliometrics/informetrics/scientometrics started harvesting "vast quantities of knowledge about knowledge, or metaknowledge" (Evans & Foster, 2011; p. 721) from journal articles. While the two subfields of information science mostly developed in parallel (Glänzel, 2015; Wolfram, 2015), more recently there have been efforts to bring these two subfields closer in ways that would benefit both (e.g., Glänzel, 2015; Mayr & Scharnhorst, 2015; Mutschke & Mayr, 2015; White, 2007a, 2007b, 2015; Wolfram, 2015). This paper aims to contribute to those efforts, not only by utilizing knowledge from both subfields, but by proposing solutions for identifying related research that would be useful in building information systems and delineating reference sets for informetrics research.

In this research, we propose a topic evolution tree (TET) that builds on prior research to present different evolutionary paths for topics to different individuals, based on the context that is relevant to their research. To achieve this we use a heterogeneous bibliographic network (Sun, Han, Yan, Yu, & Wu, 2011) constructed from four types of entities present in a scientific papers repository: papers (P), venues (V), authors (A), topics or keywords (K), and the relationships between them. In TET we utilize multiple meta-paths[2] between topic nodes in the heterogeneous graph to identify the topics that a given topic has evolved from. In constructing the TET for a topic, we calculate a score for each meta-path instance based on the edge weights of the relationships. This approach shares similarities with the calculation of path instance scores as proposed in PathSim (Sun et al., 2011). However, we propose that for topic evolution, better performance results can be obtained by adding meta-path restrictions based on a new "contribution" edge as proposed for citation recommendation (Liu, Yu, Guo, & Sun, 2014b). Namely, we use not only the edges previously used in bibliographic networks, such as "written by", "cited by", "published in", or "used (topic)" (Lee & Adorna, 2012; Shi, Kong, Yu, Xie, & Wu, 2012; Sun et al., 2011;), but also contribution edges derived from PageRank calculations for each type of node (Liu et al., 2014b). For example, the "contributed-by-author" edge is calculated for each topic over a graph of authors citing other authors. We employ contribution edges as a means to restrict the context of a meta-path so that a random walk of the heterogeneous graph generates a TET showing the evolution of a topic in a particular context. In other words, we obtain the evolution of each branch of the TET based on the context or query topic that is specifically of interest to each scholar. The workflow used to generate a topic evolution tree is shown in Fig. 1.

As an example, in the information retrieval domain, a researcher might be interested in the evolution of the topic "Cloud computing"[3] (user query) and the papers, authors, or venues that made the most significant contribution to the evolution of that topic. The topic "Cloud computing" would be the root node of the TET and each edge from that root to a child node represents the evolution of "Cloud computing" from a contributing topic. One of the topics contributing to "Cloud computing" is the Big Data topic "MapReduce", so there would be an edge in the TET from "Cloud computing" to a child node labeled "MapReduce". Since the restricted meta-path uses the contribution edge from the bibliographic network, the subtree in the TET for "MapReduce" focuses on the evolution of that topic in the context of "Cloud computing". Thus, the evolution of the topic "MapReduce" could be different in the context of cloud computing than in the context of data security, since the papers, authors, or venues covering "MapReduce" will have made at least slightly different contributions in one context versus the

[1] http://www.ncbi.nlm.nih.gov/pubmed.

[2] A meta-path P is a path defined between two nodes on the heterogeneous graph and specifies the types of nodes and relationships in the graph that are on that path (Sun et al., 2011).

[3] We select this topic as a case study because it is a relatively new and important topic in the ACM Digital Library and in a domain that we are familiar with. In the future, we plan a more comprehensive study to evaluate the quality of TETs for a larger number of randomly selected topics.