Journal of Informetrics 10 (2016) 1092–1107

Contents lists available at ScienceDirect

Journal of Informetrics

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

Searching bibliographic data using graphs: A visual graph query interface

Yongjun Zhu*, Erjia Yan

College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 12 May 2016
Received in revised form 19 August 2016
Accepted 24 September 2016

Keywords:
Information retrieval
Bibliographic queries
Graph query interface
Information visualization
Graph databases

ABSTRACT

With the ever-increasing scientific literature, improving the efficiency of searching bibliographic data has become an important issue. With a lack of support of current bibliographic information retrieval systems in expressing complicated information needs, getting relevant bibliographic data is a demanding task. In this paper, we propose a visual graph query interface for bibliographic information retrieval. Through this interface, users can formulate bibliographic queries by interacting with a graph. Visual graph queries use a set of nodes with constraints and links among nodes to represent explicit and precise bibliographic information needs. The proposed visual graph query interface allows users to formulate several complex bibliographic queries (e.g., bibliographic coupling) that are not attainable in current major bibliographic information retrieval systems. In addition, the proposed interface requires less number of queries in completing everyday bibliographic search tasks.

1. Introduction

Graph data are prevalent in the real world as data from a variety of domains (e.g., physics, chemistry, biology, sociology, and computer science) can be represented by graph data models (Aggarwal & Wang, 2010). Graph data models can represent relational information and enable a number of applications by supporting efficient searching and mining (Cook & Holder, 2006). Because of this, a few studies have investigated ways of generating graphs from arbitrary data (e.g., Baeza-Yates, Brisaboa, & Larriba-Pey, 2010). Bibliographic data are graph data in nature because they can be represented in the form of interconnected papers, authors, terms, sources, and organizations. Recent bibliometric studies, including searching bibliographic data (Zhu, Yan, & Song, 2016), measuring scholarly impact (Yan & Ding, 2009), and mining bibliographic networks (Sun, Barber, Gupta, Aggarwal, & Han, 2011) have taken the advantage of the graphical representation of bibliographic data. Regardless of the physical representations (e.g., relational databases) of graph data, efficient searching of graph data is one of primary tasks for the information retrieval community (e.g., Kacholia et al., 2005; Jiang, Wang, Yu, & Zhou, 2007; Yuan, Wang, Chen, & Wang, 2013). User interface is an integral part of searching (Hearst, 2009), and a variety of user interfaces have been proposed to support searching graph data, including regular expression- (Giugno & Shasha, 2002), keyword- (Tran, Wang, Rudolph, & Cimiano, 2009), and natural language-based (Pradel, 2012) interfaces. A recent movement towards efficient graph data searching is to adopt graph queries (e.g., Zhang, Zhang, Tang, Rao, & Tang, 2010; Han, Finin, & Joshi, 2012). Graph queries are a way of searching graph data by taking a graph pattern with a few constraints over nodes and edges as input,

* Corresponding author.
E-mail addresses: zhu@drexel.edu (Y. Zhu), ey86@drexel.edu (E. Yan).

http://dx.doi.org/10.1016/j.joi.2016.09.005
1751-1577/

which is a natural fit to graph data (He and Singh, 2008). Graph queries are known to convey richer information than other forms of queries and thus improve search performance (e.g., Zhou, Wang, Xiong, Wang, & Yu, 2008). Traditionally, systems with a graph query interface relied on textually represented graph queries (i.e., graph query languages). For example, He and Singh (2008) proposed a graph algebra-based query language to explore graph data, and, likewise, Leser (2005) proposed a pathway query language for biological networks. Even though textually represented graph queries are an effective way of searching graph data, writing these queries requires substantial efforts and causes a hindrance to users (Jayaram, Khan, Li, Yan, & Elmasri, 2015). As an alternative, a few studies proposed searching graph data using visually represented graph queries (e.g. Ceri, Comai, Damiani, & Fraternali, 1999). These visual graph queries are seen as more user-friendly because users do not need to remember the syntax of textual graph queries (Ykhlef & Alqahtani, 2011).

Despite the improved performance of graph queries, searching bibliographic data still faces a critical challenge-the proliferation of data has made it increasingly burdensome to retrieve relevant literature. Major bibliographic search systems provide forms, keywords, and Boolean queries as the main interfaces for searching bibliographic data. A typical search scenario is that users need to go through multiple refining processes after sending the first query; even so, they usually end up in getting too many search results than can be absorbed. Therefore, it is imperative to enable queries to covey more explicit information and represent more complicated information needs to only return the most pertinent search results. Aforementioned search user interfaces have limitations in representing such precise information needs. For example, they cannot directly represent queries such as "*papers on information retrieval, which were cited by John's papers that had been presented in SIGIR*". This type of queries, on the other hand, can be easily represented by visual graph queries with a set of nodes with constraints and links. Visual graph query interfaces are thus seen as a reliable solution for users to express precise and explicit information needs and receive more relevant search results. With this motivation in mind, this study aims to propose a visual graph query interface for bibliographic information retrieval. Specifically, this study aims to address the following research questions:

- How to design and implement a visual graph query interface to search bibliographic data?
- What features does a visual graph query interface need to have in order to improve bibliographic data retrieval? And
- How to integrate a visual graph query interface with back-end databases to build a streamlined system?

The work is built on our previous work (Zhu et al., 2016), in which, we proposed a framework for graph-based bibliographic information retrieval. In the present work, we focus on visual graph query formulation and processing while using the same graph schema proposed in our previous work.

2. Literature review

Earlier studies on visual graph queries were carried out by taking a specific data structure—XML-in mind (e.g., Ceri et al., 1999; Erwig, 2003; Ni & Ling, 2003; Ykhlef & Alqahtani, 2009). These studies proposed visual graph queries for querying and restructuring XML data. As XML data are quite complex with multiple nested structures, visual graph queries are seen as an efficient solution. Because the main goal of these studies was to build efficient languages of visual graph queries by investigating the structural aspects of XML documents, they are intended to be used by other systems but not the end users.

Recent studies (Hogenboom et al., 2010Hogenboom, Milea, Frasincar, & Kaymak, 2010; Schweiger, Trajanoski, & Pabinger, 2014) proposed visual graph query interfaces for users to query graph data. However, these visual graph queries were designed only to search for data that are stored as Resource Description Framework (RDF) triples, which is a standard data format of Semantic Web. Because SPARQL is the de facto standard RDF query language, those visual graph queries were designed to be translated into SPARQL, which limits their applicability. The proposed visual graph query interface in this study, however, does not stick to a specific database, and is more flexible (e.g., relational databases, graph databases).

Besides the abovementioned differences, a clear distinction between our study and the earlier studies is that we focus on a specific domain—bibliographic information retrieval. By focusing on this domain, we are able to deliver more customized solutions that can be used in real-world cases. To the best of our knowledge, the only one that is the most relevant to our study is the study by Gómez-Villamor et al. (2008). They proposed a bibliographic exploration tool based on a graph query engine. The tool employed visual graphs, while the actual queries are formulated by clicking one of three predefined queries other than a graph. But the work is largely different from our study in that we propose visual graph queries for bibliographic information retrieval, in which users visually draw graph queries to search bibliographic data.

3. Bibliographic graph queries

Bibliographic data are by nature a directed graph with nodes and links. For example, a link named "*WRITES*" is a directed link, in which the source is an "*Author*" and the target is a "*Paper*". There are a variety of ways to model bibliographic data using graphs. The one shown in Fig. 1 shows a typical schema of bibliographic data with five bibliographic entities. In the schema, "*Source*" denotes a journal or a conference in which authors publish or present papers. "*Term*" denotes a keyword, a topic; or a concept that describes a paper. In the rest of this paper; we will use this typical schema to explain the visual graph query interface.