Journal of Informetrics 8 (2014) 594–605

Contents lists available at ScienceDirect

Journal of Informetrics

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

Knowledge diffusion path analysis of data quality literature: A main path analysis

Yu Xiao[a], Louis Y.Y. Lu[b,c,*], John S. Liu[d], Zhili Zhou[a]

[a] School of Management, Xi'an Jiaotong University, Xi'an 710049, China
[b] College of Management, Yuan Ze University, Taoyuan 32003, Taiwan
[c] Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan 32003, Taiwan
[d] Graduate Institute of Technology Management, National Taiwan University of Science and Technology, Taipei 10607, Taiwan

ARTICLE INFO

Article history:
Received 20 January 2014
Received in revised form 1 May 2014
Accepted 5 May 2014

Keywords:
Data quality
Main path analysis
Knowledge diffusion
Citation analysis
Social network analysis
Big data

ABSTRACT

This study presents a unique approach in investigating the knowledge diffusion structure for the field of data quality through an analysis of the main paths. We study a dataset of 1880 papers to explore the knowledge diffusion path, using citation data to build the citation network. The main paths are then investigated and visualized via social network analysis. This paper takes three different main path analyses, namely local, global, and key-route, to depict the knowledge diffusion path and additionally implements the *g*-index and *h*-index to evaluate the most important journals and researchers in the data quality domain.

1. Introduction

The term 'Big Data' has become more popular recently with the development of the Internet and cloud technology. In this era of massive data explosion, data as the most valuable asset is especially important for enterprises and may even become the biggest trading commodity in the future. Good data quality can help organizations reduce costs, operate more efficiently, and decrease risk. Data quality is a multidisciplinary theme involving statistics, knowledge representation, data mining, management information systems, and data integration (Batini & Scannapieca, 2006). As a result, academia and companies are very interested in data quality development in the last several decades.

Researchers have put forth many interesting studies in the field of data quality, related to its dimensions, models, frameworks, measurement techniques, methodologies, and applications (Lee, Strong, Kahn, & Wang, 2002; Pipino, Lee, & Wang, 2002; Wang & Madnick, 1990). In the beginning, researchers focused on how to manage datasets to find and solve data quality problems. After the rapid development of computer technology, scientists began to consider the problem of defining, measuring, and improving the quality of electronic data. Currently, even more researchers are discussing the applications of data quality management from different perspectives. Thousands of papers have been published in journals, meetings, and books on data quality, among which are several review papers such as Wang, Reddy, and Kon (1995), Batini and Scannapieca

* Corresponding author at: College of Management, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li, Taoyuan 32003, Taiwan. Tel.: +886 3 4638800x2624; fax: +886 3 4633824.

E-mail addresses: xiaoyu1029@stu.xjtu.edu.cn (Y. Xiao), louislu@saturn.yzu.edu.tw (L.Y.Y. Lu), johnliu@mail.ntust.edu.tw (J.S. Liu), zlzhou@mail.xjtu.edu.cn (Z. Zhou).

http://dx.doi.org/10.1016/j.joi.2014.05.001
1751-1577/© 2014 Elsevier Ltd. All rights reserved.

(2006), and Madnick, Wang, Lee, and Zhu (2009). These authors have provided comprehensive introductions and interpretations covering the development of this knowledge domain and have presented clear and concise frameworks for data quality management. However, these review papers only examined less than three hundred articles using traditional survey techniques. When the volume of the dataset that surveyed by a review paper is small, the review task is manageable, but once the volume is large, researchers are unable to collate and summarize the papers in a clear and complete manner.

In this study, we adopt an effective method to handle thousands of papers in data quality. We perform a citation-based main path analysis to trace the knowledge diffusion paths in the field of data quality, from which those papers that play important roles in this field are identified. At the same time, we discover the development trends of data quality management. In addition, this paper uses the *g*-index and the *h*-index to evaluate the main journals and researchers in the data quality domain.

The rest of the article is organized as follows. In Section 2 we briefly explain the methodology used in this study. Section 3 discusses how the data are acquired and presents the basic statistics, followed by a presentation and discussion of the analysis results. The last section concludes.

2. Methodology

This study applies an integrated approach of main path analysis to explore the development and knowledge diffusion trajectories of the data quality literature. The characteristics of the approach come from the fact that it views the diffusion process from a variety of perspectives.

Citations are important resource for exploring how knowledge diffuses and for evaluating the level of contribution a scientist makes to the practice of science. Garfield (1970) claimed that a citation index is an effective and efficient tool to perform citation analyses.

From the perspective of time, a citation network is a time sequence diagram that reflects the historical evolution and knowledge diffusion of a scientific or technology field. Citation network analysis portrays the historical development and evolutionary path in a specific research field or even across research fields. It provides a visual aid of what role the literature and authors play and what theories and methods represent the mainstream research in this field.

Hummon and Doreian (1989) first introduced the theory of the main path analysis (MPA). They developed new methods based on search algorithms to analyze a citation network describing the development of DNA theory. They called the sequences of links and nodes in the network as search paths, and calculated a traversal count for each link to quantify the connectivity. Finally, they proposed that the main goal of the main path analysis is to find the development trend in research fields through identifying the maximum connectivity from a series of studies in the literature. Subsequently, the MPA method was applied to social network analysis (Hummon & Carley, 1993; Hummon & Doreian, 1990). These research works proved that the MPA method is correct and effective, but these studies used only a small sample size with low complexity. It is nearly impossible to manually generate large and complicated networks, but a computer program can easily build the network. Batagelj and Mrvar (1998) presented some approaches to analyze and visualize a large network and implemented them in a program named Pajek. Batagelj (2003) proposed an approach to analyze a very large citation network and presented a better algorithm, named Search Path Count (SPC), for the main path search. De Nooy, Mrvar, and Batagelj (2005) published the book “Exploratory social network analysis with Pajek”, in which they proposed a clear concept of the main path in an acyclic network. They defined that a main path is a path from a source vertex to a sink vertex with the highest traversal weights on its arcs.

Many researchers have used main path analysis to explore the path of technological development by using bibliographical citation data and/or patent citation data. Verspagen (2007) adopted patent citation data to explore the history of fuel cell research. Mina, Ramlogan, Tampubolon, and Metcalfe (2007) used both bibliographic and patent citation data to map the emergence, growth, and transformation of medical knowledge. Lucio-Arias and Leydesdorff (2008) combined path dependent transitions and main path analysis with HistCite^{TM} to highlight the development of a structural backbone in the field of fullerenes and fullerene-like structures of nanotubes. Harris, Luke, Zuckerman, and Shelton (2009) used main path analysis to identify the gap between the discovery of risk factors and the delivery of interventions by reviewing articles on secondhand smoke published between 1965 and 2005. Liu, Lu, Lu, and Lin (2013) applied main path analysis to find the development trajectories of the data envelopment analysis (DEA) literature. Lu and Liu (2013) utilized the MPA method to identify the knowledge diffusion path of the resource-based theory.

How to measure the weights of each citation link from a set of starting nodes to the set of ending nodes is an important step in the main path analysis. Several indices have been proposed, including the NPPC (Node Pair Projection Count), the SPLC (Search Path Link Count), and the SPNP (Search Path Nodes Pair) introduced by Hummon and Doreian (1989), and the SPC (Search Path Count) suggested by Batagelj and Mrvar (1998) and Batagelj (2003). These indices are similar, but subtle differences exist among them. Batagelj (2003) observed that the SPC, SPLC and NPPC methods produced almost the same results, but the SPC has additional properties and is the first choice. Hence, we choose the SPC to count the weight of each citation link in this study.

Fig. 1 shows how SPCs for each link are calculated. In a citation network, the original nodes of the knowledge diffusion paths are the ‘source’ nodes and the end nodes are the ‘sink’ nodes. Obviously, a ‘source’ is a node that is cited, but cites no other nodes; a ‘sink’ is a node that cites other nodes, but is not cited.