Information Processing and Management 53 (2017) 923–944

Contents lists available at ScienceDirect

Information Processing and Management

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

A survey on scholarly data: From big data perspective

Samiya Khan[a], Xiufeng Liu[b], Kashish A. Shakil[a], Mansaf Alam[a,*]

[a] Jamia Millia Islamia, New Delhi, India
[b] Technical University of Denmark, Denmark

ARTICLE INFO

Article history:
Received 9 November 2016
Revised 8 March 2017
Accepted 10 March 2017
Available online 14 April 2017

Keywords:
Scholarly data
Big data
Cloud-based big data analytics
Big scholarly data
Big data platform
Scholarly applications

ABSTRACT

Recently, there has been a shifting focus of organizations and governments towards digitization of academic and technical documents, adding a new facet to the concept of digital libraries. The volume, variety and velocity of this generated data, satisfies the big data definition, as a result of which, this scholarly reserve is popularly referred to as big scholarly data. In order to facilitate data analytics for big scholarly data, architectures and services for the same need to be developed. The evolving nature of research problems has made them essentially interdisciplinary. As a result, there is a growing demand for scholarly applications like collaborator discovery, expert finding and research recommendation systems, in addition to several others. This research paper investigates the current trends and identifies the existing challenges in development of a big scholarly data platform, with specific focus on directions for future research and maps them to the different phases of the big data lifecycle.

1. Introduction

The digital world is facing the aftermath of data explosion. In view of this, several terms like data deluge, which is a phrase used to describe the excessively huge volume of data generated at a regularly increasing basis in the world, have come into existence. A significant implication of data deluge is that it has made the scientific method completely obsolete (Anderson, 2008), as a result of which right questions need to be framed that this data can answer. This paradigm shift has given birth to the concept of big data analytics.

Big data analytics suffers from two fundamental challenges. Firstly, owing to the huge volume, variety and velocity of data involved, the storage and processing requirements of the system are rather overwhelming. Secondly, the analytics techniques and algorithms are complex, which makes big data analytics a computing-intensive task. In order to support the storage and processing requirements of big data analytics applications, cloud has been found as the most appropriate infrastructural solution (Chen & Zhang, 2014). Cloud computing offers a cost-effective solution for storing, processing and managing big data for analytical purposes, enabling the implementation of distributed and parallel paradigms for meeting the efficiency requirements.

Big data analytics is a vast field that has found applications in diverse domains and studies. Some of the most impactful researches that have merged big data analytics with other fields of study include business analytics (Duan & Xiong, 2015), multi-scale climate data analytics (Lu et al., 2011), banking customer analytics (Sun, Morris, Xu, Zhu, & Xie, 2014), smart

[*] Corresponding author.
E-mail addresses: samiyashaukat@yahoo.com (S. Khan), malam2@jmi.ac.in (M. Alam).

http://dx.doi.org/10.1016/j.ipm.2017.03.006
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

cities (Khan et al., 2015), e-commerce recommender systems (Hammond & Varde, 2013), social media analytics (Burnap et al., 2014), healthcare data analytics (Raghupathi & Raghupathi, 2014), intelligent transport management systems (Chandio et al., 2016) and railway assets management system (Thaduri, Galar, & Kumar, 2015). One of the lesser-explored applications of big data analytics lies in scholarly data. Moreover, the use of this synergistic approach to develop a big scholarly data platform for implementation of diverse scholarly applications needs to be explored.

The need for research in 'big scholarly data' and its analytics can be summarized as the lack of scholarly platforms and tools that can use this huge reservoir of data for creating applications that can benefit the research community, at large. Effective and efficient management of big scholarly data using the cloud infrastructure can facilitate the processes involved in big data analytics like data acquisition, storage, processing, analytics and visualization to support research data management and its analytical uses.

Scholarly documents are generated on a daily basis in the form of research papers, project proposals, technical reports and academic documents, by researchers and students from all over the world. Moreover, there have been many initiatives by Governments and organizations to digitize existing literary and academic resources (Meity, 2016; IFLA, 2016; Christenson, 2016). However, it is important to note that this is a generalized description and the definition may vary from one scholarly community to another. For instance, Google Scholar does not count patents as a scholarly resource. It is the huge reservoir of data that is popularly referred to as 'scholarly data'. Owing to the massive volume of these digital resources, the data needs to be looked upon from the big data perspective.

The use of big data analytics in the scholarly ecosystem for, what can be called 'research analytics' has far-reaching implications on the ease with which scholarly documents are managed and research is performed. Primarily, analytics for big scholarly data can be divided into five categories namely, research management, collaborator discovery, expert finder systems, recommender systems and visualization tools. Such analytics have gained immense importance and relevance lately particularly with the advent of multi-disciplinary research projects.

Such projects have increased the scale and complexity of research problems manifold and emphasize on the pressing need for collaboration among researchers as well as institutes or organizations. Research collaboration is not a neoconcept. However, there has been a recent shift in the manner in which collaborations are initiated. Traditionally, researchers and scholars used to meet periodically in conferences and symposiums to explore new research domains and possibility for collaborations.

With the increasing popularity of Internet, these platforms have been complemented with academic search web engines like Google Scholar and academic social networking portals like ResearchGate[1] and Academia.[2] While these platforms allow researchers to follow each other's research activities and interests, they have also created a sense of realization in the research community that the final published article is merely a milestone in research. Other aspects of research like dataset used and supporting material considered for the research are equally important.

In view of the overwhelming volume, variety and velocity characteristics of this data, scholarly data has been popularly named 'big scholarly data'. In order to develop advanced analytical applications for big scholarly data, several cloud-based tools and technologies can be used. Hadoop exists as the most popular framework for big data storage and processing, apart from a plethora of other tools like Zeppelin that are popularly used for data acquisition and visualization.

There are research challenges and limitations specific to big scholarly data at every stage of the big data lifecycle. However, some specific services that a big data platform needs to support include user data analytics and information extraction. Reliability and accuracy of information extraction methods remains a major area of concern for the reason that the accuracy of analytics results is directly dependent on the accuracy of the method employed. Moreover, there is a dearth of innovative applications that can make use of the big scholarly data reserve, with applications like research management, recommendation systems and time-evolution of research needing attention.

Another important aspect of big scholarly data management and analytics is the subject-specificity of data and applications. Generalised solutions that are cross-domain and generic need to be developed to create comprehensive, commercially viable analytical solutions for this domain. Other areas of research that have gained attention recently are academic social networks analysis and research evaluation. The motivation behind this survey is a lack of a comprehensive survey in the field of scholarly data that views this data reserve from the big data perspective, keeping the different stages of the big data lifecycle in consideration.

The results of the survey shall play a crucial role in putting the pieces together for integrating a big scholarly data platform for development of effective and efficient applications in this domain. The contributions of this research paper are as follows: (1) study big scholarly data with respect to the different phases of the big data lifecycle namely data management, analytics and visualization (2) identify the challenges that exist specific to every phase and their sub-phases (3) investigate the research issues for development of big scholarly data analytics applications (4) explore the future domains of research in this field with specific focus on creation of innovative applications that can find commercial ground and real-world adoption.

This paper surveys the existing literature on the challenges faced by the implementation of analytics techniques on big scholarly data using cloud computing. This paper is structured as follows – Section 2 covers the background and methodology followed for this survey, elaborating on the concepts, platforms and frameworks that rule the big data scenario, in gen-

[1] https://www.researchgate.net/.

[2] https://www.academia.edu/.