Journal of Informetrics 8 (2014) 776–790

Contents lists available at ScienceDirect

Journal of Informetrics

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

Time gap analysis by the topic model-based temporal technique

Do-Heon Jeong[a,1], Min Song[b,*]

[a] Korea Institute of Science and Technology Information (KISTI), 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, South Korea
[b] Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, South Korea

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form 10 July 2014
Accepted 14 July 2014

Keywords:
Text mining
Topic modeling
Latent Dirichlet Allocation (LDA)
Content analysis
Temporal analysis
Multiple resources

ABSTRACT

This study proposes a temporal analysis method to utilize heterogeneous resources such as papers, patents, and web news articles in an integrated manner. We analyzed the time gap phenomena between three resources and two academic areas by conducting text mining-based content analysis. To this end, a topic modeling technique, Latent Dirichlet Allocation (LDA) was used to estimate the optimal time gaps among three resources (papers, patents, and web news articles) in two research domains. The contributions of this study are summarized as follows: firstly, we propose a new temporal analysis method to understand the content characteristics and trends of heterogeneous multiple resources in an integrated manner. We applied it to measure the exact time intervals between academic areas by understanding the time gap phenomena. The results of temporal analysis showed that the resources of the medical field had more up-to-date property than those of the computer field, and thus prompter disclosure to the public. Secondly, we adopted a power-law exponent measurement and content analysis to evaluate the proposed method. With the proposed method, we demonstrate how to analyze heterogeneous resources more precisely and comprehensively.

1. Introduction

Research on citation analysis in the bibliometrics area started in the mid-1950s, when co-citation relationships between academic journals were studied (Garfield, 1955). Traditionally, main information resource utilized in bibliometrics was a scholarly paper (Ball & Tunger, 2006), whereas patent information was another main information resource used for the technology-competitive analysis and promising technology finding (Daim, Rueda, Martin, & Gerdsri, 2006). As the importance of World Wide Web has been highlighted since the late 1990s, the concept of webometrics which involved the intellectual structure of the Web has also emerged as a mainstream research (Björneborn & Ingwerson, 2004). Meanwhile, through scientometrics studies, the quantitative analysis area expanded into scientific technologies, research policies, and the social studies of science (Schoepflin & Glänzel, 2001). Currently, informetrics which covers various information resources and methodologies is widely used from the quantitative perspective (Egghe, 2005).

Several limitations have existed in the various fields of trend analysis. Firstly, most analysis-related studies only analyzed subjects using a single information resource. Xu, Zhu, Qiao, Shi, and Gui (2012) and Sajjad et al. (2013) argued that such

* Corresponding author. Tel.: +82 2 2123 2416.
E-mail addresses: heon@kisti.re.kr (D.-H. Jeong), min.song@yonsei.ac.kr (M. Song).
[1] Tel.: +82 42 869 1792.

http://dx.doi.org/10.1016/j.joi.2014.07.005
1751-1577/

an approach narrowly led to one-sided results. Secondly, although a few attempts have been made to identify meaningful connections between heterogeneous resources, most studies were still limited to connecting resources through citation information between patents and papers (Finardi, 2011; Shibata, Kajikawa, & Sakata, 2010). Thirdly, no major studies have been made on content-focused interlinking Web news articles with other resources such as papers or patents. A large number of studies have attempted to analyze recent trends based on Web news articles (Amitay, Carmel, Herscovici, Lempel, & Soffer, 2004; Kim & Oh, 2011; Vaughan & You, 2008), but integrative studies considering various resources were still insufficient.

This paper aims to achieve the following objectives by providing an integrative method for analyzing multiple resources such as papers, patents, and Web news articles. Firstly, the proposed method of the time gap analysis aims to shed light on how a specific resource precedes others. If a consistent time gap phenomenon is found between resources, then these resources can be arranged according to the time taken from creation to publication. Secondly, this study aims to identify whether the coherent time gap phenomenon occurs between academic areas. If a time gap between resources in a specific academic area is consistently shorter than other areas, topics of the academic area change faster than ones of other academic areas.

This paper is organized as follows. In the related work section, previous studies of recent trend and temporal analysis are discussed. In addition, quantitative research and text mining-based studies are reviewed from the perspective of the trend analysis and multiple resources. In the Section 3, we propose the topic model-based method for analyzing the time gap phenomenon, which is the core part of this study. In the Section 4, multiple resources are plotted on one time axis; the time gap phenomenon between Web news articles, papers, and patents is revealed, and trends between academic fields are identified through the experiments based on the optimization model. In the Section 5, we propose two measurements to interpret experimental results with the statistical method and interpretation of significant topic changes. Finally, in the conclusion section, the proposed method and various experimental results are summarized. Study limitations and future research directions are also discussed.

2. Related work

2.1. Trend analysis with various resources

To examine trends or find competition strategies, quantitative analysis method of informetrics was mainly utilized to determine co-authorship and citation relationships (Chua & Yang, 2008) and correlations of Web-link information (Vaughan & You, 2008). Other researchers made use of a combined research method using both quantitative analysis and content analysis based on text mining (Song, Kim, Zhang, Ding, & Chambers, 2014). Whereas, computer science field has been mainly interested in terminological trend analysis based on natural language processing (NLP) technique. Topic Detection and Tracking (TDT) is conducted on the basis of corpora built mainly through Web news articles and papers (Mei & Zhai, 2005).

Analysis studies by the type of resource can be classified as follows; firstly, research based on paper analysis primarily looks at citation relationships between papers. Therefore, many researchers analyzed the intellectual structures of scholarly fields using citation information (Chua & Yang, 2008), or revealed co-authorship in specific academic disciplines (Levitt & Thelwall, 2008); and still others found core articles on promising technologies (Shibata et al., 2010). Secondly, studies that focus on patents tend to analyze technological innovations or R&D trends. Lee and Lee (2013) attempted to detect emerging technologies in the energy-engineering field with patent data collected from the U.S. Patent and Trademark Office (USPTO). Guan and Zhao (2013) tried to select university-industry collaboration partner based on patent data in the field of nanobiopharmaceuticals. Thirdly, Web data-based research can be classified into Web structure analysis using Web links and Web content analysis based on text mining. With regard to the Web-link-associated traditional informetrics, which also means webometrics, Amitay et al. (2004) conducted trend detection by analyzing temporal link for Web site search and event detection. Furthermore, Vaughan and You (2008) conducted a study to find business competition using the co-link information using the content analysis of Web resource.

The aforementioned studies have the limitation of only addressing a single information resource. In the mid-1980s, Lancaster and Lee (1985) conducted topic analysis by examining topic diffusion of various journals. They also analyzed the time gap phenomenon in a domain, but still dealt with a single resource. Experimental research has recently attempted to investigate these problems and enable the use of heterogeneous resources. The mainstream of studies on heterogeneous data attempts to coordinate research papers and patents as the fundamental indicators for scientific research and development. Narin, Hamilton, and Olivastro (1997) investigated correlations between terms co-occurring in both papers and patents. Kim, Hwang, Jeong, and Jung (2012) categorized all technical terms into five development stages using academic papers and patents based on decision tree algorithm. Xu et al. (2012) constructed semantic linkages between patents and papers based on LDA topic modeling. Sajjad et al. (2013) determined the time differences in technological trends in the fields of bio and medical sciences based on papers, patents, and Web news articles. The results yielded the findings that the terms appearing in Web news articles were the slowest in start point but the fastest in growth and extinction, whereas papers marked the fastest start points and the most sustainable graph curves.

From the review of the related studies, it can be concluded that research efforts using multiple resources have not yet begun in full force. This study aims to propose a new temporal analysis method for performing integrative analysis of