Transportation Research Part C 77 (2017) 49–66

Contents lists available at ScienceDirect

Transportation Research Part C

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

Discovering themes and trends in transportation research using topic modeling

Lijun Sun [a,*], Yafeng Yin [b]

[a] The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[b] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 19 June 2016
Received in revised form 15 January 2017
Accepted 17 January 2017

Keywords:
Transportation research
Topic modeling
Publication data
Research policy

ABSTRACT

Transportation research is a key area in both science and engineering. In this paper, we present an empirical analysis of 17,163 articles published in 22 leading transportation journals from 1990 to 2015. We apply a latent Dirichlet allocation (LDA) model on article abstracts to infer 50 key topics. We show that those characterized topics are both representative and meaningful, mostly corresponding to established sub-fields in transportation research. These identified fields reveal a research landscape for transportation. Based on the results of LDA, we quantify the similarity of journals and countries/regions in terms of their aggregated topic distributions. By measuring the variation of topic distributions over time, we find some general research trends, such as topics on sustainability, travel behavior and non-motorized mobility are becoming increasingly popular over time. We also carry out this temporal analysis for each journal, observing a high degree of consistency for most journals. However, some interesting anomaly, such as special issues on particular topics, are detected from temporal variation as well. By quantifying the temporal trends at the country/region level, we find that countries/regions display clearly distinguishable patterns, suggesting that research communities in different regions tend to focus on different sub-fields. Our results could benefit different parties in the academic community—including researchers, journal editors and funding agencies—in terms of identifying promising research topics/projects, seeking for candidate journals for a submission, and realigning focus for journal development.

1. Introduction

With the rapid urbanization globally, transportation has become an increasingly important ingredient in the quality of life, making a major impact on human well-being. Aiming to provide better transportation systems and services, transportation research has long been a key area in both science and engineering. This has been reflected in both the rising application of emerging technologies, the growth in interdisciplinary collaborations, and the increasing number of conferences organized, journals created and research articles published (Banister, 2014; Button, 2015).

Scientific publication is often considered a key proxy to reflect the trend of research development in both theory and practice. In terms of transportation research, the problems and challenges encountered have been constantly changing over time, and the scope of transportation research has also become more diverse, with a widening and inter-disciplinary coverage of

* Corresponding author at: 75 Amherst Street, E14-474D, Cambridge, MA 02142, USA.
E-mail addresses: sunlijun@mit.edu (L. Sun), yafeng@umich.edu (Y. Yin).

http://dx.doi.org/10.1016/j.trc.2017.01.013
0968-090X/

topics, ranging from those long lasting questions such as traffic congestion and signal control, to emerging technologies such as autonomous vehicles, connected vehicles, big data analytics and artificial intelligence, to societal problems such as sustainability and environmental justice. The field is evolving given the specific questions raised and the advances in solutions/technologies developed. As a result, transportation research has witnessed an explosion of research publications in last decades.

There exists a great body of literature studying publication data with quantitative methods, which are often referred to as scientometrics (e.g., see Heilig and Voß (2015) for a study on public transportation). Although scientometric analysis offers a good tool to quantify the importance of articles and authors from citation data, it fails to provide topic related information for us to better understand different research context in detail. In fact, the content of scientific publications is often of more importance to study a field, in the sense that it could help us to obtain solutions to targeted problems, understand the development of particular technology, and learn the motivation and creation of new ideas. The abstract of an article is the first but concise piece of content-related information we can get, since it essentially reveals the whole picture of an article from a reader's point of view. In other words, an abstract can be considered a condensed representation of an article, and it has been successfully used to identify and interpret scientific themes. For example, Griffiths and Steyvers (2004) investigated abstract data from articles published in the *Proceedings of the National Academy of Science (PNAS)* from 1991 to 2001 and compared research topics/areas obtained from topic modeling with existing categories. Blei and Lafferty (2006) applied dynamic topic models on historical literature from the journal *Science* during 1880–2000 to investigate how individual topics change over time. Gatti et al. (2015) applied topic modeling on article metadata from 20 journals in the field of operations research and management science, and quantified the generality and specificity of different journals. To the best of our knowledge, there is little work done in the field of transportation with an exception that Das et al. (2016) applied topic modeling on a sample of abstracts from papers presented at the *Transportation Research Board (TRB) Annual Meeting* and investigated topics changes from 2008 to 2014.

In this paper, we investigate research topics and their trends to understand the field of transportation research from 1990 to 2015 using publication metadata obtained from 22 scientific journals. We follow a similar framework as what Gatti et al. (2015) has applied in the field of operations research and management science. The purpose of this work is to better identify, quantify and understand themes and trends in transportation research over the last 25 years, and to provide a valuable tool to researchers, journal editors, publishers and funding agencies to make more informed decisions. We also hope this work could stimulate more discussion on the state of publishing in transportation research (e.g., see a recent discussion in Button (2015)).

The remainder of this paper is organized as follows. Section 2 summarizes the notations used throughout this paper. In Section 3, we introduce the concept of topic modeling and latent Dirichlet allocation (LDA). We also present various measures to quantify topic distribution by aggregating the result at the levels of journal, country/region, and time. Section 4 introduces the article abstract data extracted from Web of Science and the software package we used for topic inference. In Section 5, we conduct extensive analysis on the extracted topic and word distributions using those defined measures. Finally, Section 6 summarizes our study and suggests some future research directions.

2. Notations

We use the notations listed in Table 1 throughout this paper.

3. Methodology

In this section, we first introduce the concept of latent Dirichlet allocation and its application in topic modeling. We follow a similar analytical framework and use similar measures as the work of Gatti et al. (2015), which focuses on the field of operations research and management science, to quantify the variation of topics across journals, countries/regions and time. In doing so we introduce various measures based on the posterior document-topic distribution θ_d: (1) the topic composition of each journal, (2) the topic composition of each country/region, and (3) topic composition over time for each journal or country/region. These measures are used in the analyses presented in Section 4.

3.1. Latent Dirichlet allocation (LDA)

LDA is a generative probabilistic model introduced by Blei et al. (2003) for the purpose of topic modeling. It is built on the classical probabilistic latent semantic analysis (pLSA) model (Hofmann, 1999) and focuses on discovering main themes from multinomial document-word observations. However, LDA itself is a general statistical model and can be applied in various domains and settings, such as finding patterns in genetic data, images, music, and social networks (see Blei (2012) for a short review). For example, in travel behavior and activity research, LDA has been used to analyze human location and activity data to discover structural daily routines (Huynh et al., 2008; Farrahi and Gatica-Perez, 2011; Hasan and Ukkusuri, 2014). As an unsupervised model, LDA does not require any prior annotations or labeling of the documents. All the topics emerge naturally from the statistical structure of document-word data itself.