Tourism Management Perspectives 33 (2020) 100608

Contents lists available at ScienceDirect

Tourism Management Perspectives

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

Network analysis of big data research in tourism

Xin Li[a,*], Rob Law[b]

[a] Donlinks School of Economics and Management, University of Science and Technology Beijing, 30 Xueyuan Road, Haidian District, Beijing, 100083, China
[b] School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:
Big data
Tourism studies
Co-citation analysis
Network analysis
Research trends

ABSTRACT

This study aims to provide a comprehensive network analysis to understand the current state of big data research in tourism by investigating multi-disciplinary contributions relevant to big data. A comprehensive network analytical method, which includes co-citation, clustering and trend analysis, is applied to systematically analyse publications from 2008 to 2017. Two unique data sets from Web of Science are collected. The first data set focuses on big data research in tourism and hospitality. The second data set involves other disciplines, such as computer science, for a comparison with tourism. Results suggest that applications of social media and user-generated content are gaining momentum, whereas theory-based studies on big data in tourism remain limited. Tourism and other relevant domains have similar concerns with the challenges involved in big data, such as privacy, data quality and appropriate data use. This comparative network analysis has implications for future big data research in tourism.

1. Introduction

The advancement of Internet technology and adoption of mobile devices have created massive user-generated big data, in which this popular buzzword has resulted in changes in the tourism and hospitality fields (Mariani, Baggio, Fuchs, & Höepken, 2018). Given the increasing popularity of big data and analytics, an accumulating number of studies on big data, such as social media data, are being conducted in various research fields, particularly tourism and hospitality (Boyd & Crawford, 2012; Li, Xu, Tang, Wang, & Li, 2018). Researchers and industries are exploring the theories and applications of big data by using high-performance algorithms to maximise the fullest potential of big data. Previous studies have generally considered big data a favourable supplement than the traditional small data; hence, they utilised the big data to detect patterns, understand consumer satisfaction and predict different outcomes, such as tourist arrivals and hotel occupancy (Li & Law, 2019; Li, Pan, Law, & Huang, 2017; Xiang, Schwartz, Gerdes, & Uysal, 2015).

Topics related to big data and social media in the tourism and hospitality fields have received increased attention from researchers. Numerous studies in tourism- and hospitality-related journals can be identified by inputting keywords, such as 'big data', 'social media', 'social network', 'user-generated content (UGC)' and 'online reviews', into professional databases (Xiang & Gretzel, 2010; Ye, Law, & Gu, 2009). General databases include Science Direct, Scopus by Elsevier, Google Scholar and Web of Science (Law, Qi, & Buhalis, 2010). Researchers can collect data from these databases and analyse the retrieved information by adopting bibliometric techniques to identify important issues and trends in big data-related research (Lu & Stepchenkova, 2015; Schuckert, Liu, & Law, 2015).

The current line of study contributes to the literature that involves the analysis of research foci and trends. However, existing studies on tourism and hospitality are limited to two streams. Firstly, the existing analyses emphasise on the systematic investigation and evaluation of big data or social media in the tourism and hospitality fields (Leung, Law, van Hoof, & Buhalis, 2013). That is, minimal effort has been devoted to compare the current state of big data research in tourism and hospitality with that in other relevant fields, such as management and marketing, from a comprehensive global perspective. The selected articles focused on the tourism and hospitality field and seldom considered the publications from other fields. This approach restricts the quantity of publications and completeness of evaluation in a broad social science domain. Secondly, the majority of the review articles conducted content analysis on the development of specific topics within a selected period. However, a comprehensive visualisation analysis of the overall development of big data research, particularly co-citation, clustering and evolving trends, has yet to be conducted. Such a visualised analytical tool can be used to obtain an improved understanding of the research nature and emerging and evolving trends in big data research in tourism and hospitality. Dynamic changes in research trends

* Corresponding author.
E-mail addresses: drxinli@ustb.edu.cn (X. Li), rob.law@polyu.edu.hk (R. Law).

https://doi.org/10.1016/j.tmp.2019.100608
Received 13 August 2018; Received in revised form 28 October 2019; Accepted 6 November 2019

over the years can be identified to further follow the entire research progress.

The concept of big data has gained increasing popularity in the social sciences, where it has been used to create values and enhance the performance of industries (Mariani et al., 2018). The methodological development of big data and big data analytics requires knowledge from various disciplines, including computer science and mathematics (Bryson, Kenwright, Cox, Ellsworth, & Haimes, 1999). Tourism studies have been described as inter-disciplinary perspectives (Coles, Hall, & Duval, 2006). Ritchie, Sheehan, and Timur (2008) reported that tourism studies are the intersections of multiple disciplines, such as sociology, business, economics, culture and management. Therefore, advancements in other disciplines, such as computer science, mathematics, engineering and business and management, can promote the development of tourism studies. By comparing the developments of big data research in tourism with those in other domains, we can formulate a comprehensive understanding of the state of big data research in the tourism and hospitality fields.

The primary objectives of this study are to provide a comprehensive network analysis of big data research in the tourism field by assessing the overall research trends and foci. Furthermore, the current research aims to compare the current big data applications in tourism with those in other relevant fields from a multidisciplinary perspective. Existing review studies, such as that of Mariani et al. (2018), have systematically reviewed business intelligence and big data in the tourism and hospitality field until 2016. The current study differs from the existing ones and contributes to the tourism and hospitality literature by performing network analysis to provide a comprehensive evaluation of big data research in other research fields, including computer science, management and marketing. We conducted a systematic network analysis of the relevant publications and established a complete database through Web of Science. A co-citation network connects articles with authors and their cited references, which has been demonstrated effective to identify the research state and reveal the intellectual structure of the tourism field (Benckendorff & Zehrer, 2013). Additionally, the different research foci and evolving trends during the studied years can be revealed on the bases of algorithms in a time-zone view of a network. Particularly, network analysis is proposed to answer the following questions: What are the most cited articles related to big data research in tourism? What are the most prominent topics in big data-related research? How did big data research evolve from 2008 to 2017? What can big data research in tourism draw from other disciplines from a social science perspective? Hence, this study contributes to the existing literature by providing answers to the abovementioned research questions, and offering potential directions for future research on big data in tourism and hospitality.

Two data sets were collected from various research domains, including tourism and hospitality, business, management and marketing. For the first data set, we focused on publications in the tourism and hospitality fields. We did not limit the research domains for the second data set, which covered relevant publications in comprehensive domains, including management and marketing, computer science, mathematics and geography. Thereafter, we analysed the two data sets and compared their results to provide a comprehensive understanding of big data research in tourism. Network and comparative analyses, including co-citation, clustering and trend analysis, were performed to investigate the overall research state of big data in tourism from 2008 to 2017.

The remainder of this paper is organised as follows. Section 2 presents a review of the relevant literature and identifies research gaps. Section 3 introduces the network analytical method. Section 4 describes the data and network analysis. Lastly, Section 5 provides the conclusions, study limitations and future research directions.

2. Literature review

Firstly, we analysed the existing systematic reviews related to big data and social media in tourism and hospitality, with focus on the data and methods used. Secondly, we introduced an analytical tool that visualises the network of scientific publications to systematically quantify the reviews.

2.1. Evaluation of publications: data and methods

Numerous empirical studies and review articles on big data, information technique, Web 2.0 and social media have increasingly become popular topics in tourism and have attracted the attention of researchers (Buhalis & Law, 2008). Systematic reviews are particularly useful in revealing research progress, identifying existing research gaps and offering agenda for future studies by using systematic methods and establishing a data set that represents the studied tourism domain (Leung et al., 2013). Various selections of databases, adoption of review methods and study periods may yield various outcomes in evaluating research progress.

The database used to evaluate the publications was generated from search engines and popular databases, such as Google Scholar, Scopus, EBSCOHost and Web of Science (Leung et al., 2013). Each data provider has its advantages. McKercher (2012) argued that Google Scholar was useful for citation analysis owing to its vast database and ease of use. McKercher (2012) acquired data from Google Scholar, which included 54 journals in hospitality and tourism, and used Publish or Perish software to propose an influence ratio measure to assess the impact of the journals. Lee, Law, and Ladkin (2014) examined authorship, length, collaboration and citation counts in selected publications by using data from Google Scholar. Published journals or conference proceedings can also serve as useful benchmarks when evaluating the quality of articles. However, Google Scholar was criticised for its inaccuracy and duplication issues (Lu & Stepchenkova, 2015). The database retrieved from Web of Science was commonly used by existing studies because of its wide coverage and authority (Li, Ma, & Qu, 2017; Li, Qiao, & Wang, 2017).

Bibliometric methods are typically employed to investigate the influences of publications in a systematic review (Lee et al., 2014; McKercher, 2012). The impact of publications generally depends on several characteristics. For example, researchers used citation count as an important indicator to highlight influence. Several quantifiable elements, such as word count, authorship and cooperation network, can be used to evaluate publications (Lee et al., 2014). Schuckert et al. (2015) reviewed 50 published online reviews in tourism and hospitality academic journals from 2004 to 2013 and conducted content analysis to classify the selected articles and examine the methodological trends. Lu and Stepchenkova (2015) reviewed 122 articles that utilised UGC as a research mode and presented topics and challenges in UGC research. Law et al. (2010) noted that analysis related to references is important to ascertain the relationship among published articles. However, a single attribute, such as citation count, only describes articles from one dimension and is insufficient to assess the entire research progress. A visualised network that combines the co-citation relationship among articles and evolving trends remains inadequate.

Although existing review articles have investigated big data related progress, a comprehensive review in tourism with a complete database and visualisation method has yet to be conducted. For example, Wamba, Akter, Edwards, Chopin, and Gnanzou (2015) analysed 64 selected articles from 2008 to 2012 and synthesised big data applications through an in-depth case study. Sagiroglu and Sinanc (2013) reviewed big data from several important issues until 2012, such as big data content, scope, samples, methods, advantages and challenges. The aforementioned researchers argued that companies and organisations should attach considerable importance to big data because big data analysis results in accurate prediction and marketing strategies. Chen,

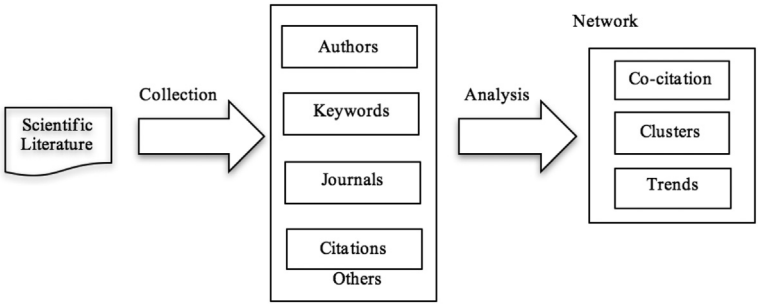


Fig. 1. Analytical framework to evaluate publications.

Chiang, and Storey (2012) completed a bibliometric study of critical business intelligence and analytic publications, researchers and topics using academic and industry publications from 2000 to 2011. Their study targeted six articles to introduce the research framework in business intelligence and big data analytics. The preceding reviews contributed to the development of big data applications, although they did not particularly reveal the recent research state of big data in tourism and hospitality.

2.2. Modelling and visualisation tool

The modelling and visualisation of review findings are useful in obtaining comprehensive results to reveal the research state and evolving research trends (Van den Besselaar & Heimeriks, 2006). Particularly, the connections among the reviewed articles are difficult to depict because of the complicated relationship between the citing and cited articles. Small (1973) indicated that if two references are cited together in a publication, then these references are related. For example, one article is cited by two articles, which are also cited by other articles. Co-citation is an important indicator that reflects the relationship among references during publication modelling. Co-citation network determines the intellectual structure of the reviewed field (Benckendorff & Zehrer, 2013).

Meanwhile, considering one single attribute only describes articles from one dimension and is insufficient in assessing the entire research progress. Analytical tools have been developed to solve the problem related to the connections among references. Particularly, these tools generate networks that use the attributes of publications and their references. CiteSpace, which is a Java application, is an ideal example. CiteSpace was developed on the bases of bibliometric constructs, such as co-citation analysis and evolving networks (Morris, Yen, Wu, & Asnake, 2003). CiteSpace was adopted in previous studies to detect and visualise emerging trends and patterns in the scientific literature. Certain articles are actively cited in a specific field. Thus, these articles are considered vital in determining research trends (Chen & Morris, 2003). Theories from bibliometric analysis indicate that the citing and cited articles represent the research front and intellectual base, respectively (Chen, 2006; Chen, Ibekwe-SanJuan, & Hou, 2010). Therefore, a visualised co-citation network can be constructed after modelling the reviewed articles. That is, network analysis can reveal and visualise the most cited articles of a specific topic during a given period and the major research trends and emerging patterns in a time-zone view. This approach is particularly useful for discerning the overall research trends and patterns in various academic fields (Li, Qiao, & Wang, 2017).

For example, the aforementioned approach was applied in various research domains. Li, Ma, and Qu (2017) analysed publications in three hospitality journals and identified patterns, research topics and influential researchers. Benckendorff and Zehrer (2013) conducted network analysis to identify the pioneering scholars and seminal research in three leading tourism journals. This approach has also been utilised to detect trends and patterns in different scientific fields, such as business models, geographic information systems and health care (Brailsford, Harper, Patel, & Pitt, 2009; Li, Ma, & Qu, 2017; Li, Qiao, & Wang, 2017).

In summary, although various studies, including Leung et al. (2013) and Mariani et al. (2018), have contributed to the existing research through their systematic reviews of big data, business intelligence and social media in tourism and hospitality, minimal attention has been devoted to the investigation on big data research via network analysis and evaluation of the trends of research focus from a broad social science perspective. Two research gaps are identified. Firstly, Leung et al. (2013) and Mariani et al. (2018) analysed numerous academic articles (44 from 2007 to 2011 and 96 from 2000 to 2016, respectively) in the tourism and hospitality fields. By contrast, the current study constructed a more comprehensive selection of publications in tourism and hospitality and in other relevant social science domains from 2008 to 2017. Secondly, although the aforementioned studies conducted content analysis and systematic quantitative methods, a visualised investigation on the reviewed articles remains limited. An in-depth visualised analysis that reveals recent developments in big data-related research, such as research front, research focus and evolving trends, has yet to be carried out.

3. Analytical framework

To investigate the state and progress of big data-related research, we proposed an analytical framework to assess publications by using the co-citation networks generated in the publications (see Fig. 1).

The major steps in this framework include data collection, preliminary analysis and network analysis.

3.1. Data collection

Data were acquired from full-length articles in academic journals, conference proceedings, editorials and reviews in Web of Science until 2017. Each record contains several relevant attributes of a publication, such as authorship, citation counts, published journals and citing references. Relevant keywords and search engines for retrieval should be selected to collect a complete data set. The following rules in data collection should be observed. Firstly, the keywords should be substantially relevant to the specific topic. Hence, we used ‘big data’ as a keyword because it expresses the relevant topic for the selected publications. Secondly, search engines should return authoritative and accurate results. Accordingly, we selected Web of Science as our primary search source because it retrieves publications from Science Citation Index Expanded, Social Science Citation Index, Arts & Humanities Citation Index and Emerging Source Citation Index. For the convenience of comparative design, we obtained two individual data sets related to big data. Our main data set was collected by restricting our search to tourism and hospitality industries, whereas the other comparative data set covered a general domain.

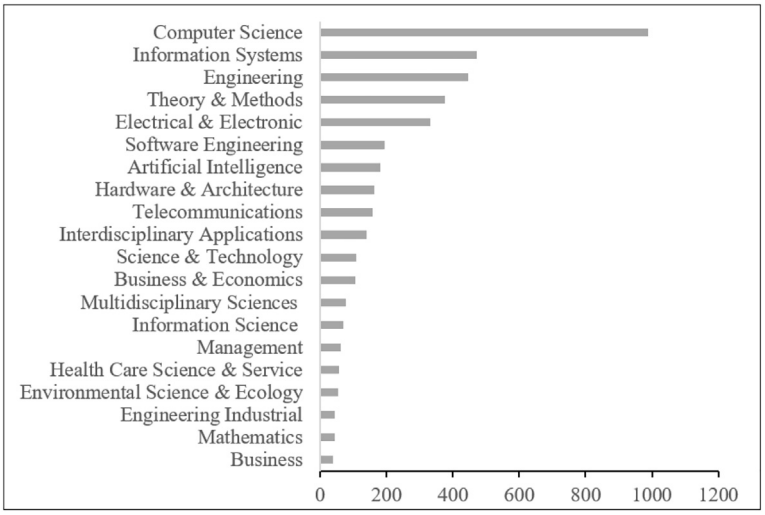


Fig. 2. Distribution of the top 20 categories in a general big data domain.

Table 1
Summary of tourism-related publications and citations (Data set #1).

	Number of publications	Total citations	Percentage of publications	Percentage of citations
2008	4	29	0.65	0.95
2009	11	60	1.79	1.97
2010	6	404	0.98	13.26
2011	14	194	2.28	6.37
2012	27	301	4.40	9.88
2013	53	749	8.63	24.58
2014	69	490	11.24	16.08
2015	153	612	24.92	20.09
2016	203	190	33.06	6.24
2017	74	18	12.05	0.59
Total	614	3047	100.00	100.00

Table 2
Summary of big data publications and citations (Data set #2).

	Number of publications	Total citations	Percentage of publications	Percentage of citations
2008	10	629	0.50	3.45
2009	2	179	0.10	0.98
2010	1	38	0.05	0.21
2011	10	249	0.50	1.37
2012	67	1903	3.35	10.44
2013	329	3672	16.46	20.14
2014	598	6070	29.91	33.30
2015	687	4022	34.37	22.06
2016	274	1389	13.71	7.62
2017	21	78	1.05	0.43
Total	1999	18,229	100.00	100.00

3.2. Preliminary data analysis

Basic descriptive analysis was conducted to pre-process the collected data sets. The specific data attributes were examined because they are useful indicators in determining the research impact of a publication. For example, being cited is strongly relevant to the impact of a publication in a specific domain. High citations indicate the strong impact of a research because it was cited by many subsequent studies. References are also beneficial in evaluating a publication because they reveal how the current study is linked to existing ones. Additionally, the attributes of authors and institutions can represent individuals and organisations that undertake specific topics. The attributes of published journals can also facilitate the identification of periodicals related to the topics. Therefore, a network that combines complex relationships should be built to substantially analyse data.

3.3. Network analysis

A network that can determine the dynamic connections between the publications and references was constructed. Thereafter, a scientific network was designed to focus primarily on the connection of the publications. This network aims to assess the following metrics: co-citation, clusters and trends. The rationale in observing these metrics is threefold: (1) the current research state and research front can be identified using co-citation analysis, (2) the different research foci can be revealed by clustering analysis based on the log-likelihood ratio algorithm and (3) the evolving trends of publications during the studied years can be shown in a time-zone view by analysing the keywords. Moreover, a comparative analysis of the two data sets was performed under the same framework.

In terms of research design, we constructed two data sets to evaluate the progress of big data in tourism and hospitality research compared with other domains. The first data set was obtained by restricting the research field to tourism and hospitality. The second data set was collected using the keyword ‘big data’ and the collected publications covered various research domains, such as computer science, social sciences and statistics.

The results of the comparative analysis of the two data sets primarily emphasised network analysis in tourism research and the connection to other relevant domains. The design of the research experiments considered the following factors. Firstly, by performing network analysis in tourism, the progress and future development of the use of big data in tourism and hospitality can be clearly presented in a quantified and visual manner. Secondly, comparing the findings against a general big data domain will enable us to acquire a global understanding and assessment of a research topic, such as the major research interests and high-impact researchers in big data, and determine how big data research in tourism is related to other relevant domains.

3.4. Data and analysis

We used the analytical software CiteSpace, which is a Java application, for the scientific literature to identify emerging trends in big data-related research (Chen, 2006). Firstly, we collected two data sets from Web of Science and presented their basic descriptions. Secondly, we performed co-citation, clustering and trend detection analyses on the data sets. Thirdly, we compared the results obtained from the two data sets to evaluate the connections and differences between tourism

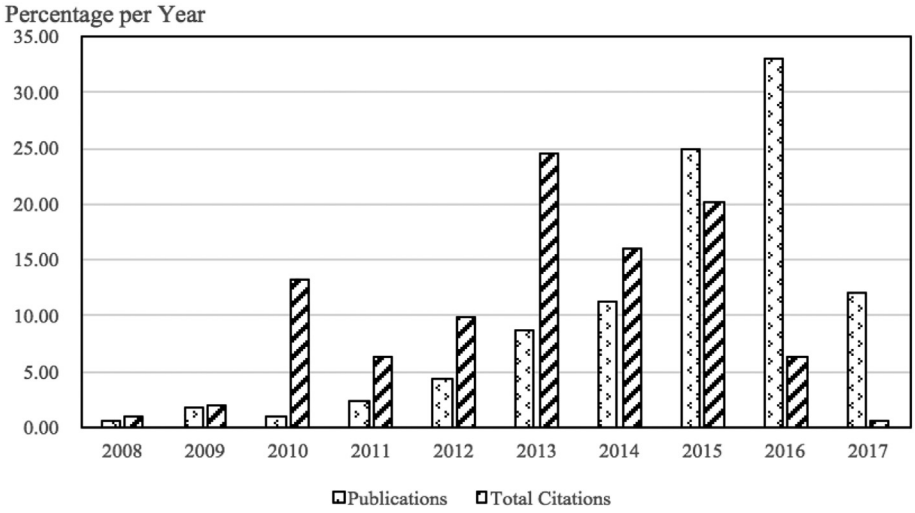


Fig. 3. Percentage of publications and citations by year in data set #1.

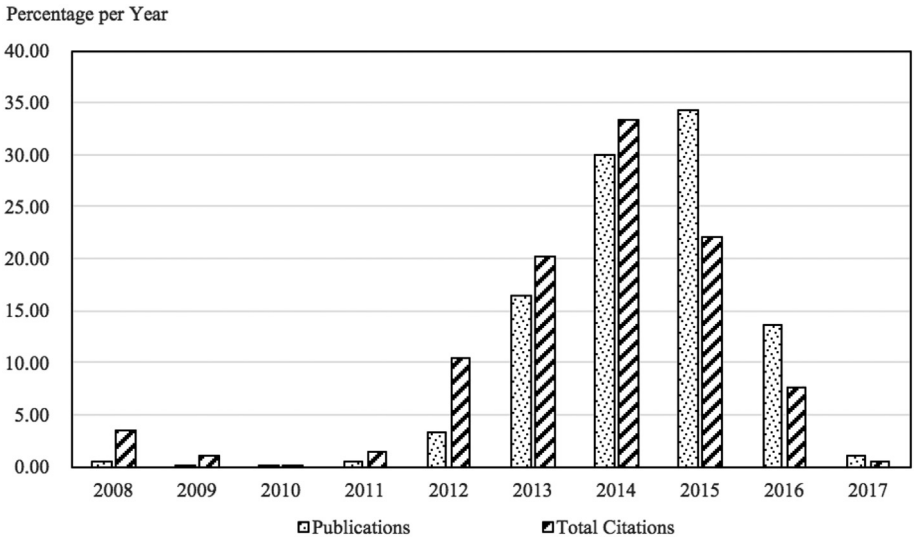


Fig. 4. Percentage of publications and citations by year in data set #2.

and other domains.

3.5. Data collection

In March 2017, all relevant publications as of 2017 were identified and collected from Web of Science. The data used in this study pertained to scientific literature, such as full-length articles, reviews and conference articles. Each data record included the following publication attributes: authors, published journals/proceedings/books, publication years, keywords, abstract, citation counts and cited references. These attributes are valuable in identifying the impact of a current publication.

Two data sets were collected to construct the emerging networks in big data research on the basis of the analytical framework. The emphasis was on the use of big data in tourism and hospitality. Thus, data set #1 was collected using *big data* combined with the terms *tourism*, *travel*, *hotel* and *hospitality* as keywords to retrieve articles from Web of Science because keywords are related to research contexts. After carefully examining the retrieved articles, we found that publications related to social media data were seldom included. Given the state of the tourism and hospitality fields, social media data are considered important big data sources generated from the Internet because such information affects tourist behaviours and tourism management (Xiang et al., 2015). Therefore, the data set was further expanded by adding 'social media' as another keyword. To obtain a comprehensive review of previous studies, the current study analysed articles published in academic journals, conference proceedings, book reviews and editorial materials. The present research differs from other review studies, such as Leung et al. (2013), which mainly analysed full-length articles in academic journals. Lastly, a data set with approximately 600 publications and 21,108 references was obtained.

The other data set for comparative analysis was constructed by focusing on a broad big data research field. Data set #2 covered computer science, information systems, engineering, economics, finance, healthcare, social science and other domains. Fig. 2 shows the top 20 categories in this data set.

We filtered the data set using the following steps to ensure that the studied articles are cited to construct a network. Firstly, we sorted the data set by using citation counts. Secondly, data with no citation counts before March 2017 were excluded to satisfy the requirement that the used data should be cited. Lastly, we acquired a data set that contained 1999 records and 58,155 references.

3.6. Data description

The number of publications and publication impact have become

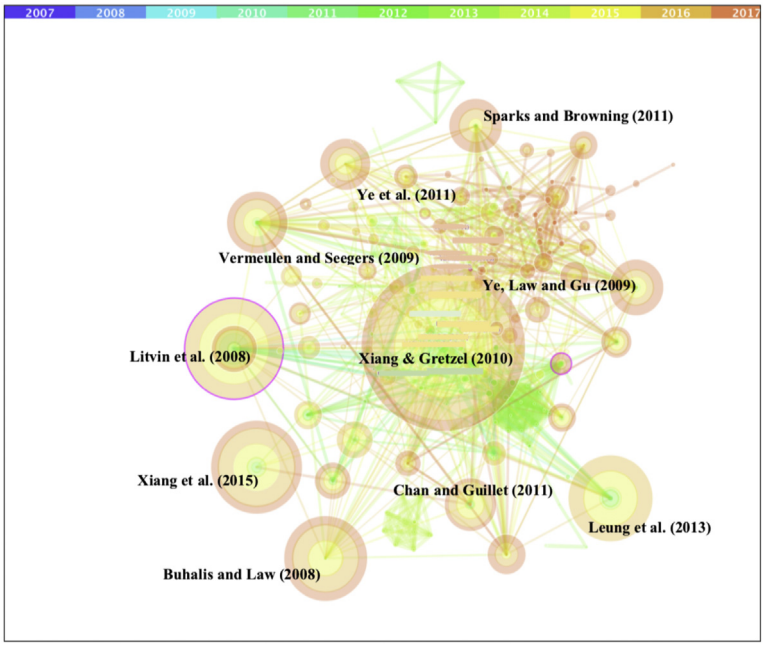


Fig. 5. Co-citation network for big data research in tourism (2008 to 2017).

Table 3
Most cited studies in dataset #1.

ID	References	Keywords
1	Xiang and Gretzel (2010)	Social media, travel information search
2	Litvin et al. (2008)	e-WOM, online marketing
3	Xiang et al. (2015)	Big data, text analytics, hotel satisfaction
4	Buhalis and Law (2008)	e-Tourism, Internet
5	Leung et al. (2013)	Social media, review, tourism research
6	Vermeulen and Seegers (2009)	Online review, hotel, e-WOM
7	Ye et al. (2009)	Online reviews, hotel, sales
8	Sparks and Browning (2011)	Online reviews, travel choice, hotel bookings
9	Chan and Guillet (2011)	Social media sites, hotels
10	Ye et al. (2011)	UGC, travel behaviour, bookings

increasingly significant because of the increasing popularity of big data research. Tables 1 and 2 summarise the number of publications and total citations that reflect the impact of publications in the two data sets in previous years.

Table 1 shows that the total number of publications and citations are 614 and 3047, respectively. The numbers in Tables 1 and 2 for 2017 only reflect publications until March because the data collection was conducted in March 2017. The results suggest that the tourism and hospitality fields exhibited significant growth in terms of the number of publications from 2013 to 2016. The total citations from 2013 to 2015 reached 749 (2013), 490 (2014) and 612 (2015), with growth percentages of approximately 25%, 16% and 20%, respectively. The number of citations in 2016 declined. Table 2 presents the substantial increase in the number of publications and citations in the general big data domains. From 2013 to 2015, the publications and citations of big data research increased by approximately 16%, 30% and 34% and 20%, 33% and 22%, respectively. The summaries in Tables 1 and 2 suggest that the popularity of big data-related research initially increased significantly and declined slightly thereafter from 2008 to 2017.

Figs. 3 and 4 depict the detailed percentage increase (i.e. trends) of publications and citations from 2008 to 2017 in the two data sets.

3.7. Network analysis in tourism research

3.7.1. Co-citation analysis

A co-citation network was developed using data set #1. Fig. 5 depicts the relationship between the literature and others that often appear together on reference lists. The size and color of the nodes denote the frequencies of co-citation and the first cited year, respectively. We identified several of the most cited publications in tourism field on the basis of the frequency reflected by the node sizes.

Table 3 lists the 10 most co-cited studies based on the co-citation network. The majority of the studies (i.e. 9 out of 10) were published in tourism and hospitality-related journals, except for Ye, Law, Gu, and Chen (2011) (No. 10) which appeared in *Computers in Human Behavior*, a non-tourism journal. The results indicate that although these top-cited papers have several relationships with big data or social media related to tourism and hospitality, several of these papers did not directly use the extensive data collected from big data sources. For example, Xiang and Gretez (2010); Litvin, Goldsmith, and Pan (2008); Buhalis and Law (2008) as well as Leung et al. (2013) (Nos. 1, 2, 4 and 5, respectively) examined several important issues on electronic word-of-mouth (e-WOM), blogs, virtual communities, e-tourism after the Internet, travel

Table 4
Most cited studies in data set #2.

ID	References	Keywords
1	Fuchs, Höpken, and Lexhagen (2014)	Big data analytics, destination
2	Mayer-Schönberger and Cukier (2013)	Big data, revolution
3	Marine-Roig and Clavé (2015)	Smart city, big data, business intelligence
4	Bangwayo-Skeete and Skeete (2015)	Google data, forecasting
5	Vu, Li, Law, and Ye (2015)	Geotagged photo, data mining
6	Yang, Pan, and Song (2014)	Web traffic data, big data, hotel demand

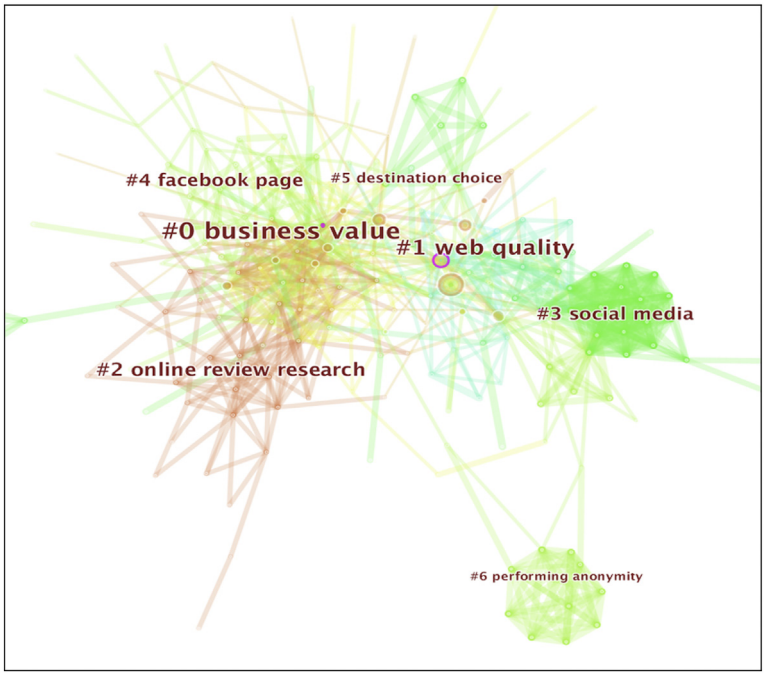


Fig. 6. Clusters in the co-citation network in data set #1.

Table 5
Description of the five largest clusters in data set #1.

Cluster ID	Cluster labels	Cluster sizes
0	Business value	46
1	Web quality	37
2	Online review research	31
3	Social media	28
4	Social network site	27
5	Destination choice	18
6	Performing anonymity	12

Table 6
Major keywords in data set #1.

Year	Keywords	Frequency	Year	Keywords	Frequency
2010	Social media	280	2014	Online review	29
	Internet	60		Destination	21
	Management	48		Industry/product	20
	Model	47		Service quality	19
2011	Word of mouth	104	2015	Online review	31
	Facebook	34		Quality	15
	Satisfaction	30		Loyalty/trust	14/13
	Twitter	16		Smart tourism	12
2012	UGC	39	2016	Network	12
	Information technology	32		Hotel industry	10
	Experience	29		Engagement	7
	Impact	68	2017	Consumer review	5
2013	Big data	59		Customer satisfaction	5
	Hospitality	47		Analytics	4
	Behaviour	42		Online hotel review	2
	Review	29		Mobile technology	2
	Destination image	26		Cultural tourism	2
	TripAdvisor	17		Challenge	2

information search and online marketing. However, these studies did not adopt any big data or social media data. Nevertheless, the preceding studies are highly cited in big data or social media research because they represent several fundamental issues in tourism and hospitality. These pioneering publications have inspired numerous researchers to investigate the contribution of social media and big data to the online tourism domain.

Table 3 shows that other primary research topics have concentrated on the adoption of online reviews generated from social media sites and big data analytics. Xiang et al. (2015) (No. 3) explored the use of big data analytics to comprehend the relationship between hotel guest experience and stratification. Various studies (Nos. 6–10) have investigated the utility of online reviews as important exogenous variables to predict and understand important issues, such as hotel sales, online bookings and social media marketing.

The findings show that the current tourism and hospitality literature emphasises social media-related research. In practice, big data-related publications that collect and analyse this type of data are relatively limited. Therefore, the citations of big data-related studies are fewer than those of social media-related ones. Accordingly, only a few big data-related studies have been identified because of the strong impact of social media research (see Table 3). Table 4 presents the most cited big data studies.

These studies used big data, such as geotagged photos, web traffic data and Google data, to investigate forecasting, smart tourism design and consumer behaviour. Analytics, including data mining and statistics, were also used in the big data-related research.

3.7.2. Clustering analysis

We ascertained 23 clusters by using the generated co-citation network. The log-likelihood ration algorithm was utilised to determine the labels of clusters. We obtained seven large clusters, namely, business value, web quality, online review research, social media, social network site, destination choice and evaluation (see Fig. 6). The clusters are named as such because of their most cited literature. For example, Xie, Zhang, and Zhang (2014) is the most cited study in Cluster #0. Therefore, this study was labeled accordingly with the extracted distinctive keyword 'business value'. Table 5 provides detailed information on the generated clusters.

The outcomes of the clustering analysis revealed that the majority of the clusters are closely connected to one another. The studies in these clusters are related to big data and social media and have similarities with general research topics despite the likelihood of having dissimilar

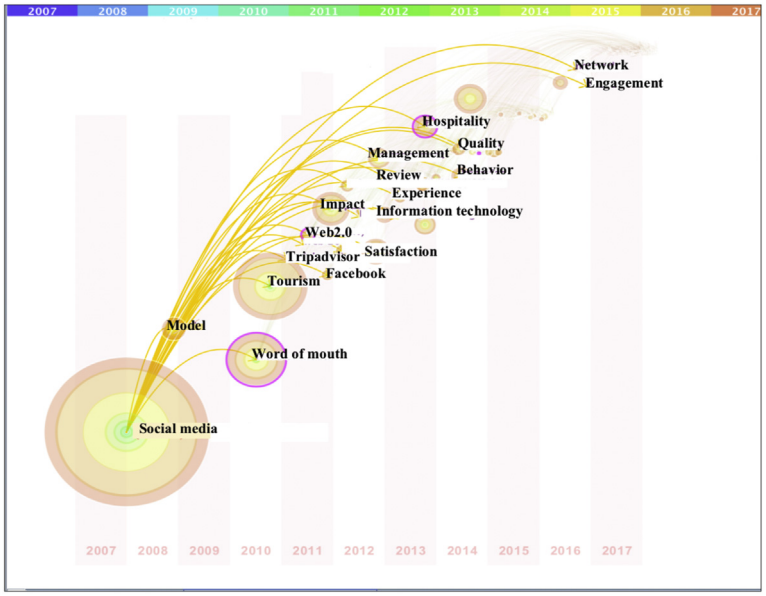


Fig. 7. Trend of keywords in data set #1.

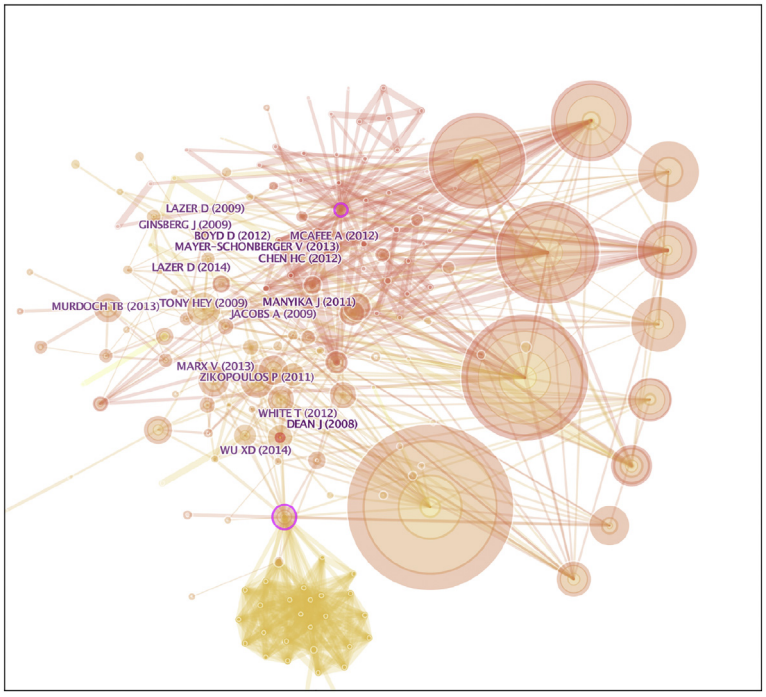


Fig. 8. Co-citation network in data set #2.

focal points. Particularly, Cluster #0 includes research that investigated social media contents from the perspective of consumer-generated online reviews. For example, Xie et al. (2014) verified the relationship between online reviews and offline hotel performance. In Cluster #1, such studies as Pulvirenti and Jung (2011) examined the effects of social networks on web quality and stratification in tourism destination marketing. Online review research is a large cluster in the data set and covers the literature that used online reviews to understand consumer behaviour and predict hotel sales. Clusters #3 and #4 are identified as social media and social media sites (Facebook), respectively, which have similar meanings but different emphases. Cluster #3 involves reviews or research agendas. For example, Munar, Gyimóthy, and Cai (2013) proposed a research agenda in tourism social media. By contrast, Cluster #4 comprises research on the applications of social media sites, such as Facebook and Twitter, in tourism and hospitality.

Cluster #6 is distinct from the other major clusters because certain studies in this cluster are not found in tourism journals. For example, Orlikowski and Scott (2013) explored the evaluation of products and services through social media from the organization viewpoint and was published in *Organization Science*.

3.7.3. Trend analysis

Table 6 shows the major keywords and their frequencies in data set #1 and Fig. 7 depicts the evolving trends of these keywords from 2008 to 2017 with different nodes and links. The node sizes in Fig. 7 are consistent with the term frequencies in Table 6. For example, the term 'social media' is marked by a green color with a frequency of 280 in 2010 in Fig. 7.

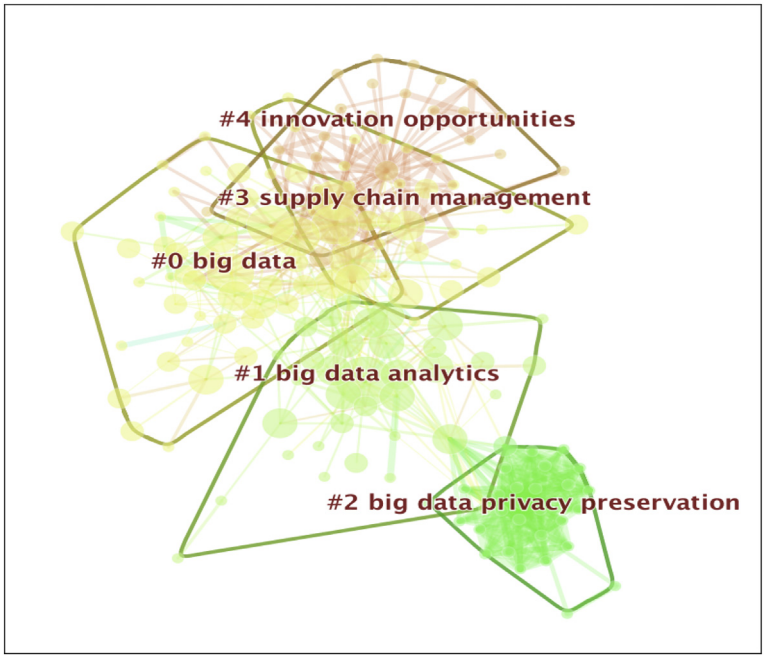


Fig. 9. Network of the major clusters.

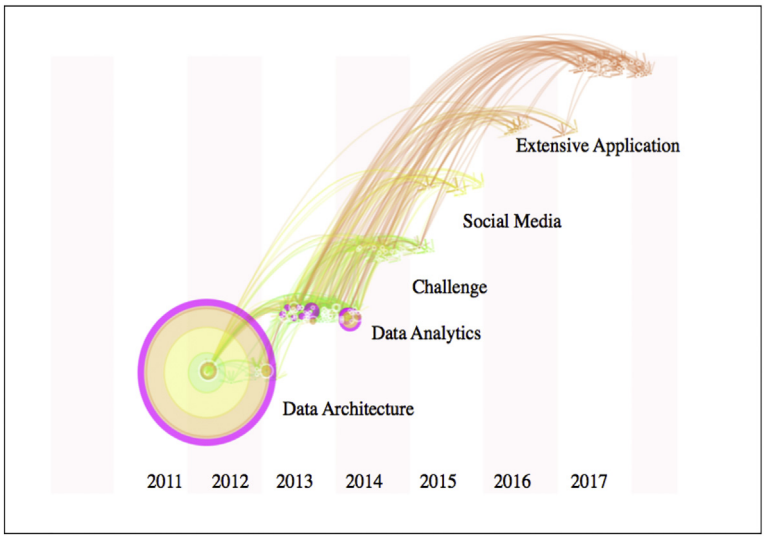


Fig. 10. Trend of keywords in data set #2.

The results showed that the keywords with high frequencies included social media, Internet, word of mouth and big data. Thus, the current popular research topics in this study are related to Internet applications and UGC. Particularly, researchers have investigated the effects of big data and social media techniques on traditional tourism and hospitality industries.

Fig. 7 shows that early research concentrated on different social media sources (i.e. Facebook, Twitter and TripAdvisor) to investigate consumer behaviour, such as satisfaction or expectation, on the basis of the changes in keywords in previous years. Researchers have also discussed the impact of big data and social media on tourism and hospitality from 2010 to 2013. After 2013, diverse research topics, including service quality, loyalty, trust and smart tourism, have been explored. Research on big data in tourism experienced expansive growth. Researchers discussed the application and influence of big data and social media in tourism and hospitality.

Such keywords as 'mobile technology' and 'challenge' were often used in 2017. Mobile technology is an emerging popular topic that can significantly affect tourism and hospitality. Several researchers have also focused on the potential challenges in big data. To reduce errors, data should be carefully analysed and used appropriately (Xiang, Du, Ma, & Fan, 2017).

3.8. Comparative analysis

A comparative analysis was conducted to explore the relationship of big data research in tourism and relevant domains. We provided the results generated using the second data set, which was constructed from a general domain (see Figs. 8 to 10 and Table 7). We emphasised the similarities and differences in the research topics and trends between tourism and other relevant domains.

Firstly, general big data research can be classified into theoretical formulation and extensive applications. The results of the network analysis of data set #2 suggest that general big data research presents

Table 7
Major keywords in data set #2.

Year	Keywords	Frequency	Year	Keywords	Frequency
2012	Big data	722	2015	Internet of things	23
	MapReduce	104		Perspective	18
	Cloud computing	97		Pattern	14
	Visualisation	20		Epidemiology	11
2013	System	125		Feature selection	7
	Model	92		Intelligence	7
	Network	86	2016	Internet of things	10
	Algorithm	69		Opportunity	10
	Data mining	64		Health care	9
	Management	62		Recognition	8
	Challenge	58		Support vector machine	7
	Big data analytics	54		Genetic algorithm	7
	Classification	50	2017	Integration	3
	Privacy	43		Industry	3
	Machine learning	43		Smart city	3
	Impact	42		Supply chain	2
	Prediction	31		Gene expression	2
2014	Information	46		Greenhouse gas emission	2
	Internet technology	37		Life cycle	2
	Social media	25		Policy	2
	Behaviour	13		Decision making	2
	Regression	11		Business analytics	2

significant interests in algorithm design, analytical tools and diverse applications. The most influential publications reflect the three perspectives of big data research, namely, data algorithms, applications and potential concerns (see Fig. 8). Big data algorithms and techniques, such as MapReduce and Hadoop, are discussed in general big data research. Table 7 shows that researchers have also discussed the various applications of big data, such as supply chain management, smart city, healthcare and gene expression (Waller & Fawcett, 2013). Fig. 10 depicts the five terms identified from the network analysis: data architecture, data analytics, challenge, social media and extensive applications, which also demonstrates that general big data research has formed theories, algorithms, methods and diverse applications.

By contrast, big data research in tourism and hospitality presents several application-based investigations and limited theory-based studies. Tables 5 and 6 show that the results from data set #1 indicate that the major concentrations include social media and big data applications. Particularly, researchers have applied analytical methods from econometrics, data mining and business intelligence to solve problems in tourism and hospitality. These applications include accurate forecasting of hotel or tourism demands, consumer experiences and satisfaction and decision-making in travel planning. However, only a few common theories are available related to the use of big data in tourism and hospitality for researchers and industries. Ruths and Pfeffer (2014) and Xiang et al. (2017) demonstrated that an improved understanding of big data or social media data is considerably required for their analysis. For example, numerous studies may apply big data, such as online reviews to assess the impact of this type of data on hotel performance. Nevertheless, only a few studies have illustrated why big data have an impact on hotel performance from the theoretical perspective (Xiang et al., 2017). Therefore, the theoretical foundations of big data in tourism and hospitality remain to be improved in the future.

Secondly, the findings from the two data sets indicate similar concerns in big data research. These findings confirm that big data can generate traps and misunderstandings, which have been demonstrated by previous studies, such as Lazer, Kennedy, King, and Vespignani (2014) and Xiang et al. (2017). The results from the keywords in the two data sets demonstrate these findings. The keywords in Table 7 with relatively high frequencies include challenge, traps and failures. The clusters in Fig. 9 show that big data privacy is commonly discussed in research related to big data. For example, Lazer et al. (2014) analysed the traps in Google Flu Trend (GFT), which is a popular tool that uses search engine data, and proposed traps in big data analysis. By tracking the forecasting performance of GFT, the researchers verified that the approach overlooks considerable information that can be extracted by traditional statistical methods. Researchers have also focused on the potential negative effects of big data on tourism. Xiang et al. (2017) discussed the quality of big data and argued for the appropriate use of data.

Therefore, we are convinced that further effort should be exerted for the correct use of big data and toward establishing the theoretical foundation that regulates the practical applications of big data. In this manner, we promote the development of tourism and hospitality. Advancements in multiple disciplines should also be integrated into big data research.

4. Conclusions

In recent years, big data-related research has become increasingly popular in various fields, including tourism and hospitality. To improve the use of big data, researchers and industries have continuously sought to develop the appropriate algorithms and solutions that draw from multi-disciplinary knowledge, such as mathematics, computer science and geography. By analysing big data-related publications in tourism and other relevant fields, this study contributes to the literature by answering the following research questions: (1) How did big data-related research evolve in recent years? (2) What can big data research in tourism draw from other disciplines in a social science perspective? The comprehensive investigation on big data research from 2008 to 2017 is conducted through a visualised network analysis. The findings indicate the influential studies on big data, major clusters of research topics and trends of research keywords in tourism and hospitality.

The present study offers several implications for practice. Big data should be carefully identified and used in tourism and hospitality. The results related to the two data sets demonstrated that concerns with data traps or failures have elicited increasing attention. Big data cannot replace all data sources and industries should not disregard traditional observations or domain knowledge when making decisions. Further effort should be directed toward the appropriate comprehension of the meaning of big data. New theories and improved algorithms should be incorporated into big data research in the tourism and hospitality fields.

Although this study is not the first to review the current progress of big data, it emphasised the network analysis of recent publications and provided insightful results within and potentially beyond tourism and hospitality. This research contributes to the existing literature and provides empirical results. Firstly, the current study collected two data sets that cover several important citation indexes. A distinct advantage is that the cited references were used to depict the connections among publications and reveal research trends. Secondly, this study used a quantitative approach for modelling the publications to present the relationship between citations and cited articles in previous years. Accordingly, this research provides the results of co-citation networks, clusters and trends in keywords in big data-related publications. Furthermore, the current study performs an alternative evaluation of big data research in tourism with various multidisciplinary domains. Findings suggest that applications of social media and user-generated content are gaining momentum, whereas theory-based studies on big data in tourism remain limited. Therefore, the study also contributes to the academia by offering future research directions for big data research in tourism and hospitality.

This study also has several limitations. Although we collected 1999 publications and 58,155 references, we could not cover all big data-related research in our data sets owing to the selected keywords. We used *big data* and *social media* with terms related to tourism and hospitality as the major keywords to retrieve data from Web of Science, thereby ensuring that the collected publications are related to big data.

Future studies can incorporate other relevant keywords and generate a comprehensive database of big data research from various providers, such as Google Scholar. Additionally, future studies can apply machine-learning methods, instead of merely performing manual examination, to filter irrelevant publications and improve the efficiency and accuracy of database construction.

Acknowledgements

This work was supported by the funding from the National Natural Science Foundation of China (Grant No. 71601021 and No. 71974010). The authors would like to thank the Editor and four anonymous referees for all the constructive, insightful, and helpful suggestions and comments. The first author would like to thank Dr. Daniel Fesenmaier and Dr. Zheng Xiang for their helpful suggestions for improving the early version of this paper.

References

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46*, 454–464.

Benckendorff, P., & Zehrer, A. (2013). A network analysis of tourism research. *Annals of Tourism Research, 43*, 121–149.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Brailsford, S. C., Harper, P. R., Patel, B., & Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation, 3*(3), 130–140.

Bryson, S., Kenwright, D., Cox, M., Ellsworth, D., & Haimes, R. (1999). Visually exploring gigabyte data sets in real time. *Communications of the ACM, 42*(8), 82–90.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the internet—The state of eTourism research. *Tourism Management, 29*(4), 609–623.

Chan, N. L., & Guillet, B. D. (2011). Investigation of social media marketing: How does the hotel industry in Hong Kong perform in marketing on social media websites? *Journal of Travel & Tourism Marketing, 28*(4), 345–368.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the Association for Information Science and Technology, 57*(3), 359–377.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the Association for Information Science and Technology, 61*(7), 1386–1409.

Chen, C., & Morris, S. (2003). Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. *Proceedings of the IEEE Symposium on Information Visualization, October 19–21, Seattle, Washington, USA* (pp. 67–74). .

Coles, T., Hall, C. M., & Duval, D. T. (2006). Tourism and post-disciplinary enquiry. *Current Issues in Tourism, 9*(4–5), 293–319.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations–a case from Sweden. *Journal of Destination Marketing & Management, 3*(4), 198–209.

Law, R., Qi, S., & Buhalis, D. (2010). Progress in tourism management: A review of website evaluation in tourism research. *Tourism Management, 31*(3), 297–313.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science, 343*(6176), 1203–1205.

Lee, H. A., Law, R., & Ladkin, A. (2014). What makes an article citable? *Current Issues in Tourism, 17*(5), 455–462.

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing, 30*(1–2), 3–22.

Li, X., & Law, R. (2019). Forecasting tourism demand with decomposed search cycles. *Journal of Travel Research*. https://doi.org/10.1177/0047287518824158.

Li, X., Ma, E., & Qu, H. (2017). Knowledge mapping of hospitality research − a visual analysis using CiteSpace. *International Journal of Hospitality Management, 60*, 77–93.

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66.

Li, X., Qiao, H., & Wang, S. (2017). Exploring evolution and emerging trends in business model study: A co-citation analysis. *Scientometrics, 111*(2), 869–887.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management, 68*, 301–323.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29*(3), 458–468.

Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing & Management, 24*(2), 119–154.

Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: A systematic literature review. *International Journal of Contemporary Hospitality Management, 30*(12), 3514–3554.

Marine-Roig, E., & Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management, 4*(3), 162–172.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin: Harcourt.

McKercher, B. (2012). Influence ratio: An alternate means to assess the relative influence of hospitality and tourism journals on research. *International Journal of Hospitality Management, 31*(3), 962–971.

Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Timeline visualization of research fronts. *Journal of the American Society for Information Science and Technology, 55*(5), 413–422.

Munar, A. M., Gyimóthy, S., Cai, L., & Marı'a, M. A. (2013). Tourism social media: a new research agenda. In G. Szilvia, & L. Cai (Vol. Eds.), *Tourism Social Media (Tourism Social Science Series). Vol. 18. Tourism Social Media (Tourism Social Science Series)* (pp. 1–15). Bingley: Emerald Group Publishing Limited.

Orlikowski, W. J., & Scott, S. V. (2013). What happens when evaluation goes online? Exploring apparatuses of valuation in the travel sector. *Organization Science, 25*(3), 868–891.

Pulvirenti, M., & Jung, T. (2011). Impact of perceived benefits of social media networks on web quality and E-satisfaction. *Information and Communication Technologies in Tourism* (pp. 513–524). Vienna: Springer.

Ritchie, J. R., Sheehan, L. R., & Timur, S. (2008). Tourism sciences or tourism studies? Implications for the design and content of tourism programming. *Téoros. Revue de recherche en tourisme, 27*(27–1), 33–41.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13). San Diego, California, USA* (pp. 42–47). The Institute of Electrical and Electronics Engineers (IEEE).

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing, 32*(5), 608–621.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology, 24*(4), 265–269.

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*(6), 1310–1323.

Van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics, 68*(3), 377–393.

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management, 30*(1), 123–127.

Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2015). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tourism Management, 46*, 222–232.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77–84.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How "big data" can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234–246.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51–65.

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management, 31*(2), 179–188.

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International Journal of Hospitality Management, 44*, 120–130.

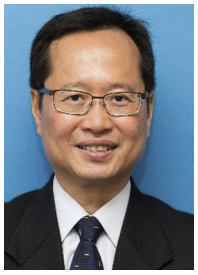
Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. *International Journal of Hospitality Management, 43*, 1–12.

Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research, 53*(4), 433–447.

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management, 28*(1), 180–182.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior, 27*(2), 634–639.

Xin Li, Ph.D., is an Assistant Professor at Donlinks School of Economics and Management, University of Science and Technology Beijing. Dr. Li's research interests are big data analytics, econometric modeling, data mining and forecasting. Dr. Li focuses on understanding tourism activities by combing user-generated contents with econometric and machine learning techniques. She also participated in many research projects on monitoring, forecasting, and early-warning of economy and industries in China.



Rob Law, Ph.D., is a Professor at the School of Hotel and Tourism Management, the Hong Kong Polytechnic University. His research interests are information management, modelling and forecasting, artificial intelligence and technology applications. He has received many research related awards and honors, as well as millions of USD external and internal research grants. Prof. Law serves in different roles for 140+ research journals, and is a chair/committee member of more than 130 international conferences.