Tourism Management 68 (2018) 301–323

Contents lists available at ScienceDirect

Tourism Management

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

Big data in tourism research: A literature review

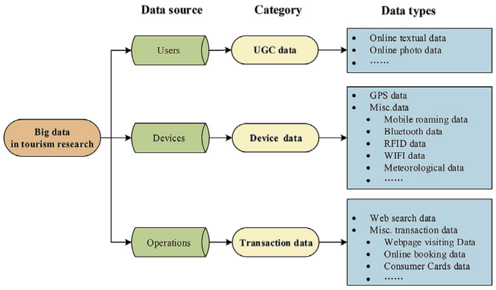
Jingjing Li [a], Lizhi Xu [b, c], Ling Tang [a, *], Shouyang Wang [d, e], Ling Li [a]

[a] *School of Economics and Management, Beihang University, Beijing 100191, China*
[b] *Collaborative Innovation Center of eTourism, Beijing Union University, Beijing 100101, China*
[c] *Tourism College, Beijing Union University, Beijing 100101, China*
[d] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*
[e] *School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China*

HIGHLIGHTS

- Various big data have been applied to tourism research, making a great improvement.
- A comprehensive review on different types of big data in tourism is presented.
- Carrying different information, different types address different tourism issues.
- Each type's research focuses, data characteristics and analytic tools are analyzed.
- The corresponding major challenges and further directions are further investigated.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 September 2017
Received in revised form
24 January 2018
Accepted 8 March 2018

Keywords:
Tourism research
Big data
Literature review
Tourism management
Tourist behavior

ABSTRACT

Even at an early stage, diverse big data have been applied to tourism research and made an amazing improvement. This paper might be the first attempt to present a comprehensive literature review on different types of big data in tourism research. By data sources, the tourism-related big data fall into three primary categories: UGC data (generated by users), including online textual data and online photo data; device data (by devices), including GPS data, mobile roaming data, Bluetooth data, etc.; transaction data (by operations), including web search data, webpage visiting data, online booking data, etc. Carrying different information, different data types address different tourism issues. For each type, a systematical analysis is conducted from the perspectives of research focuses, data characteristics, analytic techniques, major challenges and further directions. This survey facilitates a thorough understanding of this sunrise research and offers valuable insights into its future prospects.

1. Introduction

With the rapid development of computer science and Internet techniques, massive-scale data in both structured and unstructured styles are generated, recorded, stored and accumulated, forming the *big data* and opening a new age (Kambatla, Kollias, Kumar, & Grama, 2014). In such a big data era, a variety of big data,

[*] Corresponding author. School of Economics and Management, Beihang University, 37 Xueyuan Road, Haidian District, Beijing 100191, China.
E-mail address: lingtang@buaa.edu.cn (L. Tang).

https://doi.org/10.1016/j.tourman.2018.03.009
0261-5177/© 2018 Elsevier Ltd. All rights reserved.

together with the conceptual and technological innovations, have been used in extensive areas of science, engineering, healthcare, management, business, tourism, etc. (Hashem et al., 2015). However, there has not existed a uniform definition of big data yet, with different researchers giving a variety of definitions. A famous and original definition is the *3V*, which characterizes big data as *Volume*, *Variety* and *Velocity* (Laney, 2001). Gantz and Reinsel (2011) extended the 3 V's definition to a 4 V's concept, by introducing *Value* to highlight the validity and usefulness of big data. Even though controversial in definition, big data and big data technologies have made great contributions to improving scientific researches, with tourism research as an emerging, typical example.

Even at an early stage, a rich mine of tourism-related big data have been generated from three primary sources—users, devices and operations. First, the Internet has fostered a rapid rise in social media, offering a capacious platform to spread user-generated content (UGC) data (in terms of texts, photos, etc.) (Xiang, Du, Ma, & Fan, 2017). Second, due to the vigorous development of Internet of things (IoT), diverse sensor devices have been developed and employed to track tourist movements and environmental conditions, providing considerable spatial-temporal big data (such as global position system (GPS) data, mobile roaming data, Bluetooth data, etc.) (Shoval & Ahas, 2016). Third, tourism is a complex system covering a series of operations (i.e., transactions, activities or events in tourism market) such as web searching, webpage visiting, online booking & purchasing, etc., thus producing the corresponding transaction data of web search data, webpage visiting data, online booking data etc., for understanding tourist behavior and improving tourism marketing. Based on the big data from these three main sources, tourist behavior and tourism market can be better explored and understood by both academia and industries.

Using the aforementioned valuable big data, tourism research has stepped into the big data era and brought forth amazing improvements. For instance, Yang, Pan, Evans, and Lv (2015) advocated that the large scale of big data could finely make up for the limitation of sample size issues faced by survey data users, providing a new way to understand tourist behavior. Similarly, Li, Pan, Law, and Huang (2017) argued that big data analytics could provide sufficient data without sampling bias, helping both academia and industries better understand tourist behavior. Xiang, Schwartz, Gerdes, and Uysal (2015) claimed that big data analytics could develop new knowledge to reshape the understanding of hospitality industry and to support the corresponding decision making. With the aforementioned superiorities, big data allowed a better understanding of tourism demand, tourist behavior, tourist satisfaction and other tourism issues.

Given that big data have substantially changed the traditional tourism research based on traditional data, a comprehensive review on such sunrise research, i.e., using big data in tourism research, is extremely desired. On the one hand, although facing a growing volume of publications, this sunrise research field was still not very clearly known by potential researchers. For example, it was still uncertain about what particular types of big data have been used in tourism and how to take advantage of these new data. On the other hand, when comparing big data and traditional data, the former might be much more informative and structure-complex, thus appearing different data characteristics, focusing on different research issues, and requiring different analytic techniques. Therefore, a comprehensive review on full-scale types of big data in tourism research is strongly required in terms of research focuses, data characteristics and analytic techniques, in order to present a historical tour of how these particular big data have contributed to tourism research and to offer helpful insights into the future prospects.

However, a systematic literature review on the big data in tourism research was still lacking. The existing literature reviews on tourism research mainly focused on the following issues: tourism within a certain country such as China (e.g., Bao, Chen, & Ma, 2014; Huang & Chen, 2016; Sun, Wei, & Zhang, 2017; Zhang, Lan, Qi, & Wu, 2017); tourism in particular types such as event tourism and volunteer tourism (Getz & Page, 2016; Wearing & McGehee, 2013); tourism demand (Goh & Law, 2011; Song & Li, 2008); tourist behavior (Bhati & Pearce, 2016; Pomfret & Bramwell, 2016); tourism attraction (Leask, 2016); tourism risk (Yang, Khoo-Lattimore, & Arcodia, 2017). Nevertheless, regarding applying big data to tourism research, to the best of our knowledge, there existed only three relevant reviews: Rashidi, Abbasi, Maghrebi, Hasan, and Waller (2017) explored the capacity of social media data for modelling travel movements; Schuckert, Liu, and Law (2015b) reviewed the studies on online reviews in tourism and hospitality; Shoval and Ahas (2016) conducted a literature review on tracking data in tourism research. Obviously, the three studies focused on a certain type of big data (social media data, online reviews, or tracking data), without an overall analysis on full-type big data. Moreover, all the three studies were conducted mainly from the dimension of research fields, without a full consideration of data characteristics and analytic techniques. However, a different type of big data (from a different data source and with different information) is certainly different from any of other types, in terms of research focuses (on different tourism issues), data characteristics and processing techniques. Therefore, this paper attempts to fill in such a literature gap to present a comprehensive literature review on different types of big data in tourism research, and provide a systematical analysis on each type from the perspectives of research focuses (on tourism issues), data characteristics, analytic techniques, challenges and further directions.

The main goal of this paper is to present a comprehensive literature review on the application of big data to tourism research. Relative to the existing studies, the major contributions of this paper can be summarized into three aspects: (1) to the best of our knowledge, it might be the first attempt to review the full-scale types of big data used in tourism research; (2) given that a different type of big data (carrying different information) might address different tourism issues, appear different data characteristics and require different analytic techniques, a systematical analysis for each type is conducted from the perspectives of research focuses (on tourism issues), data characteristics and analytic techniques; (3) based on such a thorough survey, the major challenges and future prospects are carefully investigated.

The remaining part of this paper is organized as follows. Section 2 presents the general findings (or statistics) of the reviewed literature, together with the analytical framework of this paper. By following this framework, Sections 3–5 thoroughly investigate the big data in tourism research derived from the three primary sources, i.e., users, devices and operations, respectively. Section 6 concludes the main findings of the review and points out the further directions of applying big data to tourism research.

2. General findings

This section displays the general findings (statistics) of the review. First, Section 2.1 presents the literature collection. Section 2.2 provides a descriptive statistical analysis of the selected literature. Finally, Section 2.3 formulates the general analytical framework of this paper.

2.1. Databases

The articles on tourism research using big data are collected from the following academic databases: Web of Science, ScienceDirect, SAGE Journals Online, Emerald Insight, Springer and Wiley Online Library. Additionally, the powerful search engine, Google

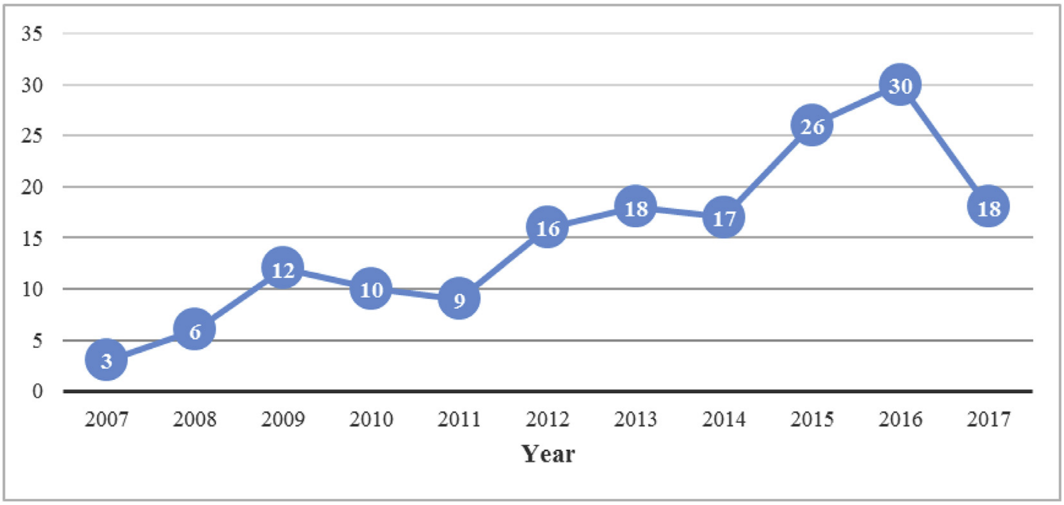


Fig. 1. Distribution of the published year.

Scholar, is also employed. To get a full-scale result, we use not only the keywords of *tourism* and *big data*, but also other related words regarding tourism research (*tourism* OR *travel* OR *tourist* OR *visit* OR *hospitality* OR *hotel* OR *destination* etc.) coupled with a term concerning big data (*big data* OR *volume data* OR *vast data* OR *data mining* OR *UGC* OR *text mining* OR *sentiment analysis* OR *consumer reviews* OR *geotagged photos* etc.). Time limitation of publications was not considered in this study.

Only full-length articles are included in our samples. Thus, book reviews, reports, viewpoints, research notes and short communications are excluded from the analysis. Articles are re-checked individually for relevancy. Finally, a total of 165 publications are selected in this study. For each article, important attributes, such as author(s), title, publication year, keywords, source title, country/territory and data type(s), are recorded.

2.2. Descriptive statistical analysis

Based on a descriptive statistical analysis, some interesting general findings regarding the overall growth, publication sources, research regions and data types are deduced for the existing tourism research using big data.

2.2.1. Overall growth

Fig. 1 illustrates the annual numbers of published articles on tourism research using big data, and two important results can be easily concluded. On the one hand, the tourism research using big data was still at an early stage, in terms of a short history (just beginning since 2007) and a small number of annual publications (at most 30). On the other hand, there appeared a generally growing trend in the annual numbers of published articles during 2007–2016, indicating an increasingly wide attention to such an emerging research. Interestingly, there existed one obvious growth point in 2015, after which the annual number of articles has stepped into a higher level with a peak of 30 in 2016.

2.2.2. Publication sources

Fig. 2 presents the numbers of articles in different publication sources. Majority of the articles on big data application to tourism were journal papers (144 publications, accounting for

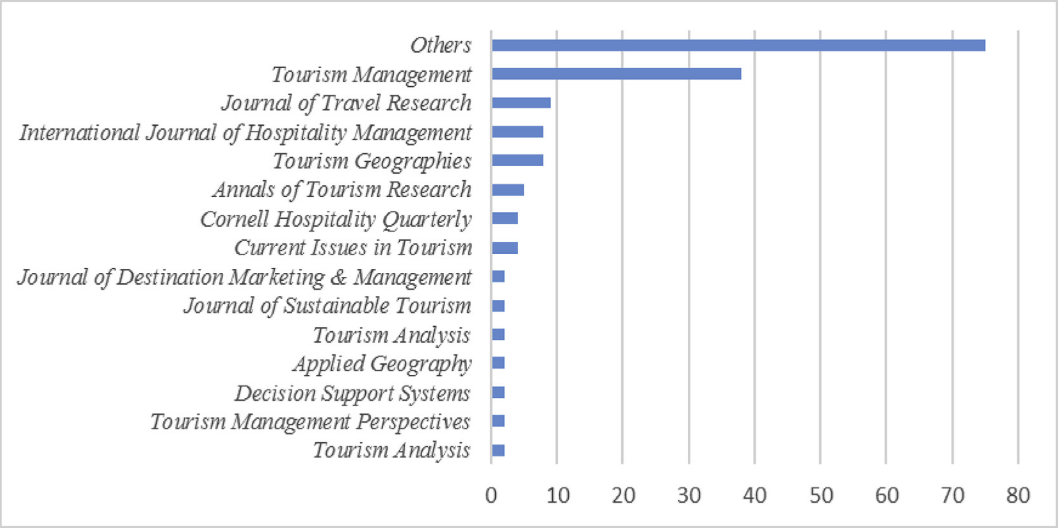


Fig. 2. Distribution of the published sources.

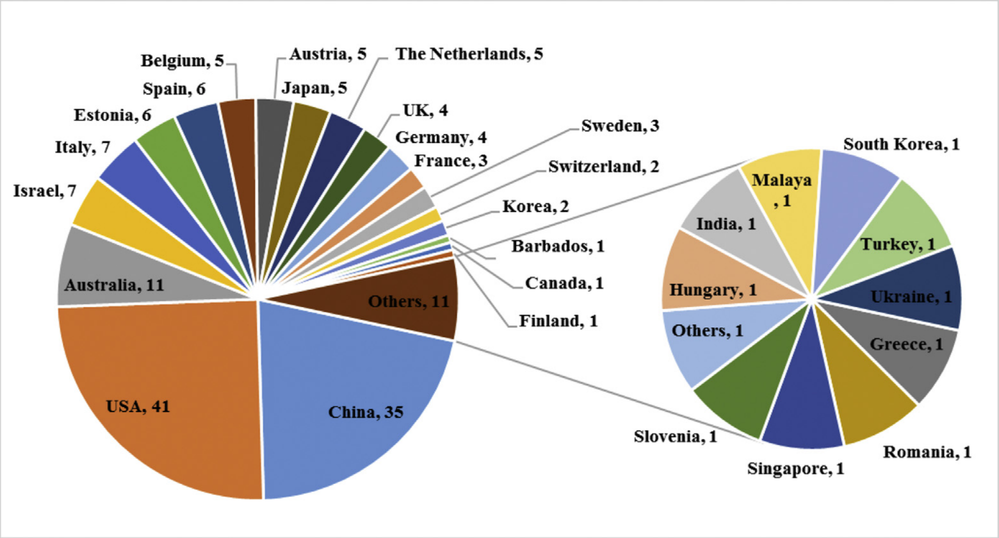


Fig. 3. Coverage of research regions.

approximately 87%), and only 21 publications were conference papers. The results indicate that various journals have started paying attention to such an emerging topic, applying big data to tourism research. Top four leading journals were *Tourism Management* (38 publications), *Journal of Travel Research* (9), *International Journal of Hospitality Management* (8) and *Tourism Geographies* (8).

2.2.3. Research regions

Fig. 3 shows the coverage of research regions (identified as the affiliation of the first author). It can be found that at least 33 countries or regions have been involved in the tourism research using big data. Asia, Europe and North America made the greatest contributions to applying big data to tourism research. Among countries, USA ranked the first (identified in 41 publications), followed by China and Australia (in 35 and 11, respectively). These results generally accord with the development levels of big data technologies (particularly social media) and the scale of users in different countries and regions.

2.2.4. Data types

The types of big data used in the existing tourism research are identified, as the results shown in Fig. 4. It can be obviously seen that the big data used in tourism research were mainly derived from three data sources: users (accounting for approximately 47%), devices (36%) and operations (17%). Hidden reason for the distribution of data sources lies in data availability. Specifically, the success of applying UGC data to tourism research might be attributed to the low cost and easy access to such online data; in contrast, the lower using level of transaction data is mainly due to that most of them are private information, which only in the possession of tourism organizations or government sectors. As for data types, online textual data made up the largest shares of UGC data (55%) as well as all the kinds of big data (26%). GPS data was ranked the top of big data in tourism research (accounting for 21%) and the first of device data (58%).

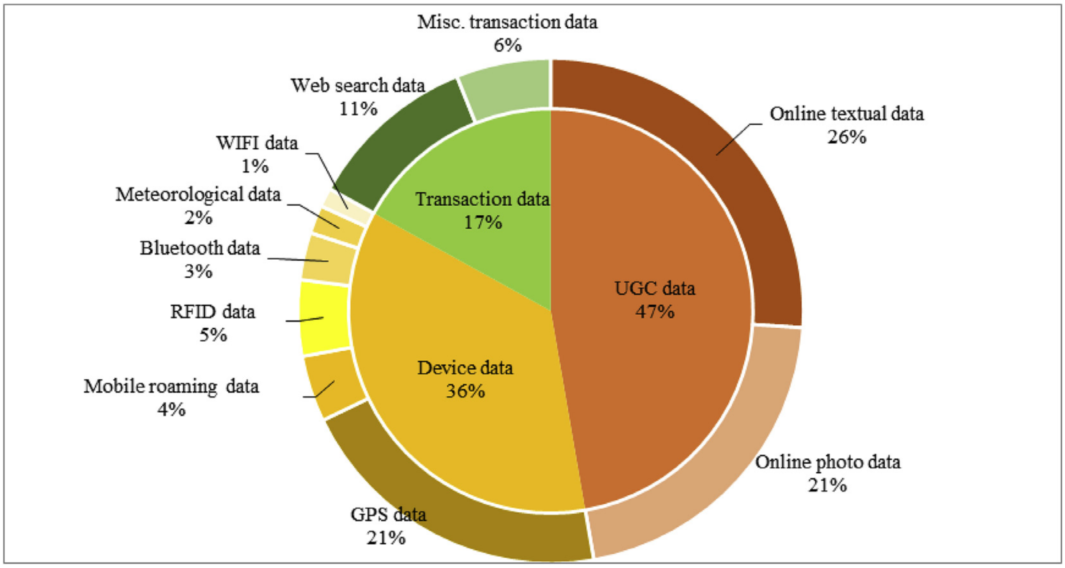


Fig. 4. Distributions of data sources and data types.

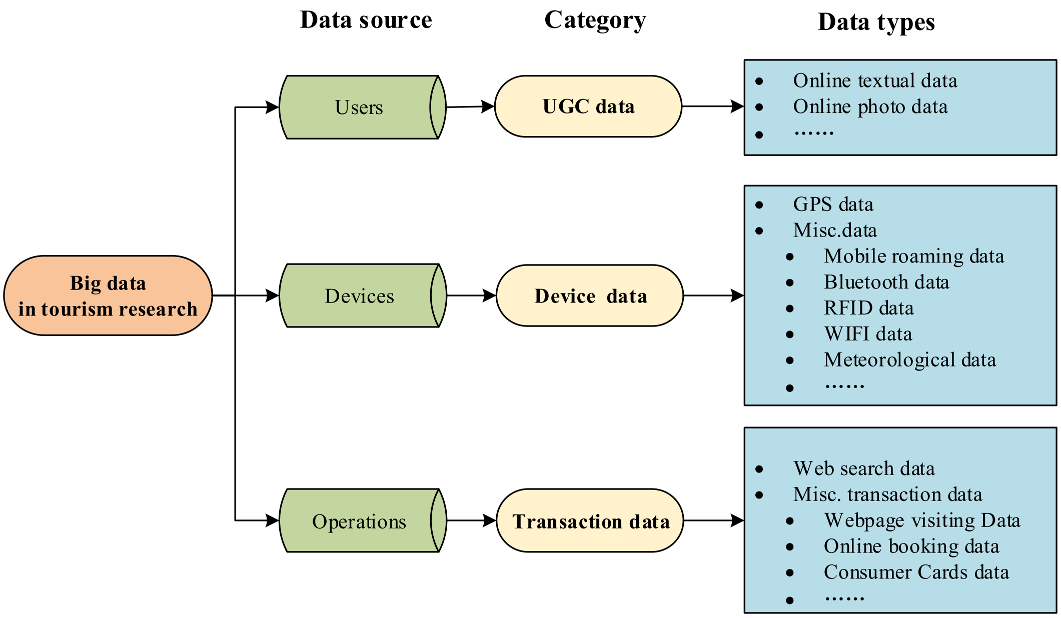


Fig. 5. Analytical framework of the literature review.

2.3. Analytical framework

According to the existing literature, a variety of big data have been applied to tourism research. According to Fig. 4, these big data can be classified into three categories by data sources. (1) UGC data (by users): online textual data and online photo data actively provided by users. (2) Device data (by devices): GPS data, mobile roaming data, Bluetooth data, RFID data, WIFI data, etc. collected passively through devices. (3) Transaction data (by operations): web search data, webpage visiting data, online booking data, etc. recording all the user related online operations such as web searching, booking & purchasing and webpage visiting. According to data sources and their corresponding data types, the analytic framework of our literature review can be formulated, as illustrated in Fig. 5.

The analytic framework is deduced based on not only the characteristics of each data type but also the existing surveys on big data. First, the development of web 2.0 and social media has offered a spacious platform for the users to share their tourism experiences, in terms of online textual data (including product reviews and blogs) and online photo data. These data posted actively by users have widely been regarded as UGC data in the existing related studies (Hu, Chen, & Chou, 2017; Lu, Wu, & Sang, 2017; Shi, Serdyukov, Hanjalic, & Larson, 2011; Xiang et al., 2017). Second, with the development of IoT, diverse devices (e.g., GPS loggers, telecommunication base stations, Bluetooth sensors, electronic readers and WIFI scanners) have been employed to track tourist movements. The related special-temporal big data recorded passively by different devices can be treated as one category namely device data (or tracking data) (Hardy et al., 2017; Shoval & Ahas, 2016). Third, tourism is a complex system covering a series of operations (i.e., transactions, activities or events in tourism market) such as web searching, online booking and purchasing, etc. Accordingly, the corresponding data such as web search data, webpage visiting data, online booking data, etc. are typical cases of transaction data (Pan, Xiang, Law, & Fesenmaier, 2011).

Based on the analytical framework in Fig. 5, Sections 3–5 further investigate the big data in tourism research, which are from the three data sources of users, devices and operations, respectively. For each data source and the corresponding big data in different types, a systematical analysis is conducted in terms of research focuses (on tourism issues), data characteristics, analytic techniques, as well as major challenges and further directions.

3. UGC data

In the digital age, the boom in web and social media has dramatically changed the way of traveling, providing a spacious platform to share UGC data. As a major category of big data, UGC data have been used to promote tourism research, including two types: (1) online textual data like product reviews and blogs released on social media; (2) online photo data posted on photo-sharing websites.

3.1. Online textual data

Social media, thriving on the vigorous development of Internet, offers an ample platform for tourists to spread a variety of tourism-related information, such as traveling reviews and experiences. For example, travelers can express their satisfaction and dissatisfaction toward tourism products, generating a rich mine of online reviews data. Tourists can also share their traveling perspectives and experiences in blogs such as Twitter and Sina Weibo, providing valuable information for potential tourists. These online reviews data, blogs data and other related data in a textual style constitute a special type of big data in tourism research—online textual data, conveying feelings, sentiments and moods of tourists.

3.1.1. Research focuses

Online textual data applied to tourism research mainly included two types, i.e., reviews data and blogs data. Carrying distinctive information, reviews data and blogs data have their own respective research focuses.

Reviews data, expressing tourists' attitudes toward tourism products, have mainly been applied to measuring tourist satisfaction. The specific hot topics included exploring the attributes of tourist satisfaction (e.g., Guo, Barnes, & Jia, 2017; Liu, Teichert, Rossi, Li, & Huet, 2017; Lu & Stepchenkova, 2012; Xu & Li, 2016),

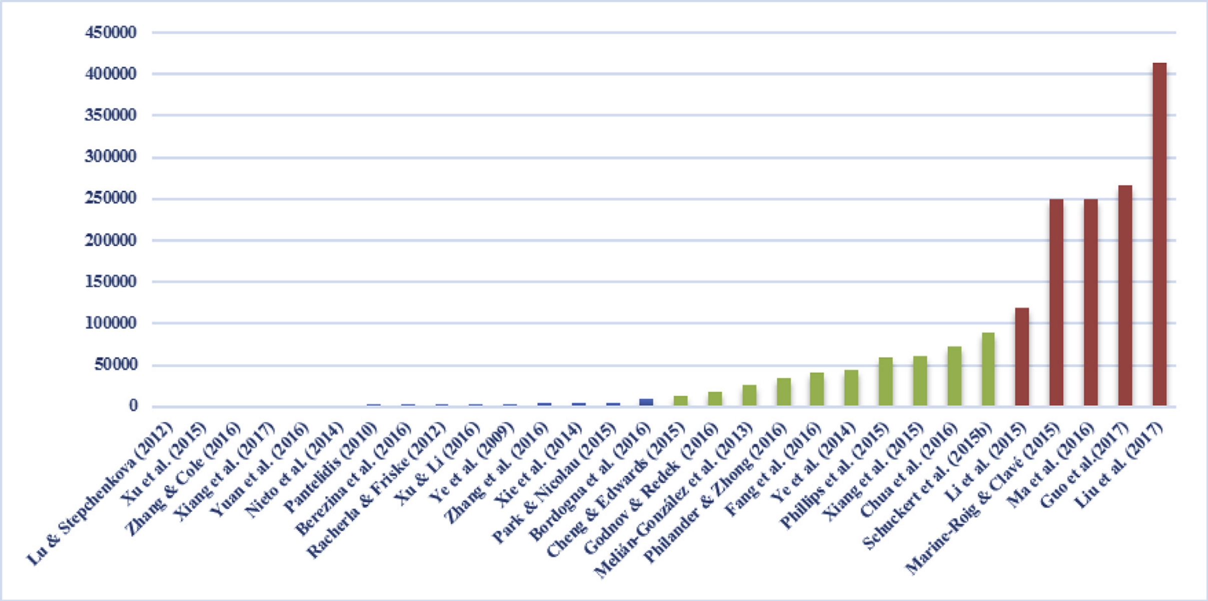


Fig. 6. The sample size of existing tourism research using online textual data (Godnov and Redek, 2016; Nieto, Hernández-Maestro, & Muñoz-Gallego, 2014; Pantelidis, 2010).

and the relationship between tourist satisfaction and other related factors (such as guest experience and competitive position) (Crotts, Mason, & Davis, 2009; Xiang et al., 2015). As for the reviewed targets, hotels (including rural lodgings), restaurants and attractions were mostly focused on. Hotel reviews were popularly adopted for evaluating and improving the e-word of mouth of a hotel (Berezina, Bilgihan, Cobanoglu, & Okumus, 2016; Guo et al., 2017; Hu et al., 2017; Ma, Luo, Yao, Cheng, & Chen, 2016; Melián-González, Bulchand-Gidumal, & González López-Valcárcel, 2013; Phillips, Zigan, Silva, & Schegg, 2015; Xiang et al., 2015, 2017; Zhang & Cole, 2016). Restaurant reviews were for tourism satisfaction judgment (Park & Nicolau, 2015; Zhang & Cole, 2016). Attraction reviews could helpfully improve attraction management (Fang, Ye, Kucukusta, & Law, 2016; Pearce & Wu, 2015). The main reason for the popularity of these three reviewed targets might be that lodging, eating and traveling might be the top three important factors that travelers are mostly concerned about in a tour. The corresponding review-based research can provide insightful implications for hospitality industry (covering hotels and restaurants) and attraction marketing, in order to attract more tourists (Guo et al., 2017).

Blogs data, recording travelling stories and tourist feelings, mainly addressed tourism recommendation and tourist sentiment analysis. For tourism recommendation, Yuan, Xu, Qian, and Li (2016) used tourist blogs data to mine tourism locations and travel sequences for an efficient travel schedule. By processing tourism blogs, Xu, Yuan, Ma, and Qian (2015) explored the valuable information about where to go and what to play. Regarding tourist sentiment analysis, Philander and Zhong (2016) and Kontopoulos, Berberidis, Dergiades, and Bassiliades (2013) demonstrated the application of tourist sentiment series from Twitter data.

3.1.2. Data characteristics

Focusing on data sources, the online reviews data in tourism research generally originated from diverse social media such as TripAdvisor (e.g., Fang et al., 2016; Hu et al., 2017; Li, Law, Vu, Rong, & Zhao, 2015; Liu et al., 2017; Ma et al., 2016; Schuckert, Liu, & Law, 2015a; Xie, Zhang, & Zhang, 2014; Ye, Li, Wang, & Law, 2014), Yelp (Park & Nicolau, 2015; Racherla & Friske, 2012), Expedia (Xiang et al., 2015), Ctrip (Ye, Law, & Gu, 2009), Qunar (Zhang, Zhang, & Yang, 2016), Booking (Xu & Li, 2016), Dianpin (Zhang, Ye, Law, & Li, 2010), etc. Among them, TripAdvisor, one of the largest tourism social media, was the most popularly used. For blogs data, Twitter and Sina Weibo were the two primary sources. For example, Twitter data were used by Chua, Servillo, Marcheggiani, and Moere (2016), Bordogna, Frigerio, Cuzzocrea, and Psaila (2016) and Philander and Zhong (2016) to mine tourist geographic information and capture tourist sentiment. Sina Weibo, the Chinese alternative to Twitter, was also employed by Cheng and Edwards (2015) to explore potential tourist generating regions, the life span of travel news, and tourists' attitudes toward travel policy changes.

The sample size varied across the existing articles, ranging from 373 (Lu & Stepchenkova, 2012) to 412,784 (Liu et al., 2017), in tourism research using online textual data. According to Fig. 6, half of the related studies (15 out of 30 articles) relied on a small quantity of samples, with the sizes below 10,000; only 5 articles utilized relatively large-scale data, with the sizes above 100,000.

3.1.3. Analytic techniques

To extract and utilize the useful information hidden in online textual data, diverse text mining techniques have extensively been adopted in tourism research, including two typical stages: data collection and data mining comprising two sub-steps of data preprocessing and pattern discovery. These processes, together with the related techniques, are displayed in Fig. 7.

The first step is to collect online textual data (including tourism-related reviews and blogs) from the related social media sites, via web crawling technology (Xiang et al., 2015, 2017; Xu et al., 2015). In particular, a web crawler (or robot and spider), in terms of a program or a suite of programs, is implemented to iteratively and automatically download web pages, extract uniform resource locators (URLs) from their hypertext markup language (HTML) and fetch them (Thelwall, 2001). For example, Xiang et al. (2017) used the web crawlers, in Python and Java programming languages, to gain hotel-related reviews. Guo et al. (2017) developed a web crawler to collect reviews data from TripAdvisor periodically. Yuan et al. (2016) employed the web crawling technology to obtain user-generated tourist blogs from tourism websites.

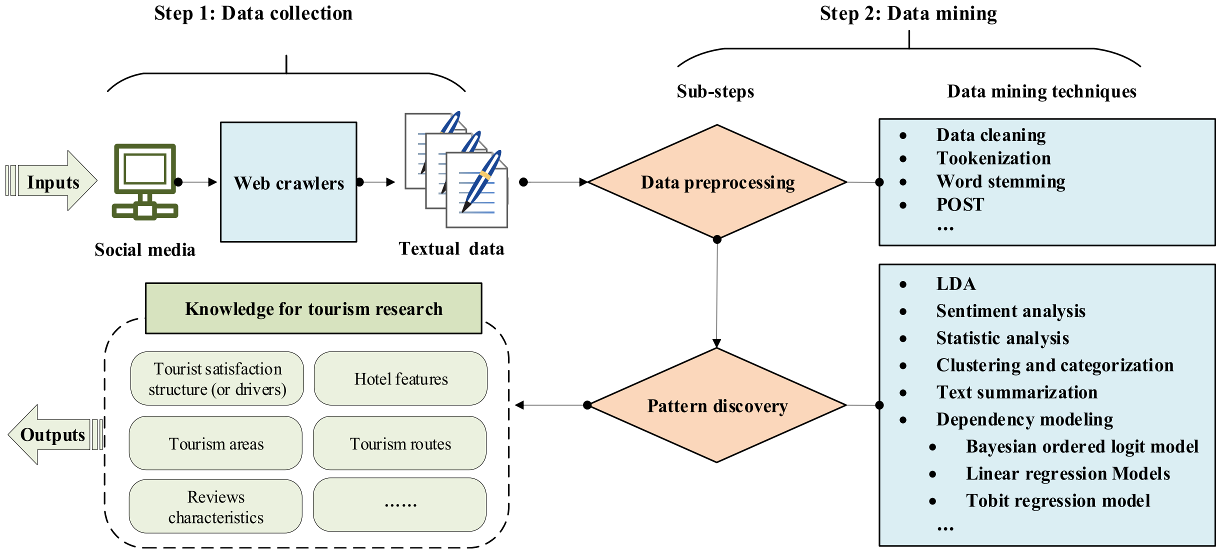


Fig. 7. Process of using online textual data in tourism research.

In the second step, i.e., data mining, the collected online textual data are analyzed to extract useful knowledge for tourism research, through two sub-stages: data preprocessing and pattern discovery. In data preprocessing, different techniques were adopted for different research purposes, and popular operations are data cleaning, tokenization, word stemming and part-of-speech tagging (POST) in the existing tourism literature using online textual data.

- *Data cleaning*, for detecting and removing inaccurate or useless records from online textual data, such as misspelling (Xiang et al., 2015), stop words (Xiang et al., 2015; Xu & Li, 2016; Xu et al., 2015), non-target language and low frequency words (Guo et al., 2017), in order to leave the valuable tourism-relevant information.
- *Tokenization*, with the aim to break up the travel-related textual data into words, phrases or other meaningful elements, namely tokens. Through this process, the tourism-related keywords regarding tourism spots, travel feelings, etc. can be filtered from massive sentences (Guo et al., 2017; Xiang et al., 2017; Xu & Li, 2016).
- *Word stemming*, to identify the word's roots and regard all words with the same root as one token, for modelling simplicity (Xu & Li, 2016).
- *POST*, to label each word in a sentence with a POS tag, i.e., noun, adjective or adverb. For example, given that hotel reviews are expressed mainly by nouns, adjectives and negative adverbs, unimportant words with other tags can be removed (Guo et al., 2017; Hu et al., 2017).

Pattern discovery, another crucial stage of textual data mining, aims at exploring interesting information in the text documents, and typical techniques in the existing tourism studies were latent Dirichlet allocation (LDA), sentiment analysis, statistical analysis, clustering and categorization, text summarization and dependency modeling.

- *LDA*, a topic model for identifying the abstract “topics” in textual data. For example, Guo et al. (2017) used LDA to rapidly discover a mixture of topics, e.g., the aspects influencing hotel customers' satisfaction, from a huge number of reviews.
- *Sentiment analysis*, to identify the tourists' attitudes toward tourism products or scenic spots, by classifying textual data into sentiment categories: positive, negative or neutral. For example, recent studies trended to use sentiment analysis as a useful tool for investigating travelers' opinions toward hotel services (e.g., Hu et al., 2017; Philander & Zhong, 2016) and hot locations (Xu et al., 2015).
- *Statistical analysis*, the most basic technique for analyzing various data, including textual data. In tourism research, descriptive statistics (e.g., mean, variance, etc.) (Racherla & Friske, 2012), *t*-test (Schuckert et al., 2015a), correlation matrix (Racherla & Friske, 2012), Kruskal-Wallis test, Mann-Whitney *U* test (Lu & Stepchenkova, 2012) and correspondence analysis (Költringer & Dickinger, 2015) were popularly used to describe diverse information contained in online textual data, such as the number of identity disclosures, reviewer reputation and hotel ratings.
- *Clustering and Categorization*, to group a set of objects in a way that the objects in the same group are more similar to each other than to those in other groups. For example, Bordogna et al. (2016) used clustering technology to group the trips whose strings of *geoslot* IDs were similar into a class, based on the geo-tagged messages in Twitter. Marine-Roig and Clavé (2015) employed categorization to group the hidden information in travel blogs and reviews on Barcelona into meaningful categories according to keywords.
- *Text summarization*, to automatically produce a succinct summary of a single or multiple document(s), for refining key information from original texts. A recent application to tourism area is the work by Hu et al. (2017), in which a multi-text summarization technique was proposed to identify the most informative sentences of hotel reviews.
- *Dependency modeling*, for capturing the relationship between textual data (particular online reviews) and tourism factors like hotel performance (Xie et al., 2014; Ye et al., 2009), restaurant performance (Zhang et al., 2010) and traveler behaviors (Zhang et al., 2016). A variety of regression models have been applied, such as Bayesian ordered logit model (Zhang et al., 2016), linear regression model (Xie et al., 2014; Ye et al., 2009; Zhang et al., 2010) and Tobit regression model (Fang et al., 2016).

Noticeably, many useful data mining tools and software packages have been developed for text processing. General multi-function tools include Waikato Environment for Knowledge Analysis (WEKA), LingPipe and TextBlob. In particular, WEKA, developed by the University of Waikato, can handle most text-processing problems, such as data cleaning, tokenization, word frequency analysis, clustering and categorization, pattern mining, etc. LingPipe, developed by Alias, is a powerful toolkit for text data cleaning, tokenization, POST, word frequency analysis, sentiment analysis, clustering and categorization, pattern mining, etc. TextBlob, a Python library, offers an application programming interface (API) for common natural language processing, such as POST, noun phrase extraction, sentiment analysis, classification, translation, etc. Moreover, there also exist some professional tools with particular functions, such as Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) and Jieba for Chinese tokenization and POST, and Stanford Loglinear POS Tagger for POST (Hu et al., 2017).

Finally, the interesting information extracted through data mining are converted into useful knowledge to further serve tourism research. According to the related studies, valuable knowledge covered tourist satisfaction structure (or drivers) (Liu et al., 2017; Zhang & Cole, 2016), hotel preferences (Li et al., 2015), tourism areas (Fuchs, Höpken, & Lexhagen, 2014), tourism routes (Chua et al., 2016), reviews characteristics (Xiang et al., 2017), etc. Such insightful knowledge was extremely helpful in improving tourism management and offering tourism recommendations (Xu et al., 2015; Yuan et al., 2016).

3.1.4. *Challenges and future directions*

Recent prosperity in social media has aroused an increasing attention to applying online textual big data, in terms of reviews and blogs, to tourism research. However, there was still a considerable amount of room to expand and develop such a sunrise research. For sample size, the datasets used in most existing studies were on a relatively small scale (see Fig. 6). However, small samples are prone to selection biases and estimation biases, leading to incorrect analysis results that might be opposite to the real situations and the generalized findings. Therefore, future research with a larger-scale sample size is desirable. For research focuses, extracting valuable knowledge from massive online textual data, e.g., tourist sentiments (or opinions) toward a tourism product or destination, has been studied sufficiently. However, how to use such insightful knowledge to tourism product design and tourism marketing in practice were somewhat lacking. Moreover, other interesting issues regarding the travel-related online textual data can be also considered, such as data reliability. For example, sometimes visitors might give a fake positive review, for avoiding unnecessary troubles or gaining kickbacks (Schuckert et al., 2015b). Such a tourist online behavior is worth a deep study in the future.

Due to the unstructured and complex features of tourism-related online reviews and blogs data, the processing and analytic techniques were confronted with some tricky challenges. For instance, the existing studies mainly addressed the texts in a single language, while non-target language reviews were otherwise removed and ignored (Guo et al., 2017). However, the travelers from a different cultural background seek a different travel benefit and have a different preference (Schuckert et al., 2015a). Therefore, effective analytic methods for multi-language texts are certainly imperative to avoid information losses. Furthermore, in sentiment analysis for tourist reviews, nouns, adjectives and negative adverbs were mainly regarded as important words. However, sentiment is also impacted by the polarity of verbs (e.g., 'like', 'hate' and 'love') and the degree (e.g., 'very much', 'more' and 'excessively') (Hu et al., 2017). Therefore, such important information should be also taken into consideration in tourist sentiment analysis.

3.2. *Online photo data*

Apart from textual data, other UGC data are also posted and spread on social media, especially photos. The photos uploaded by tourists contain a rich of useful information in relation to users, locations and time, providing a new perspective to study tourist behavior, tourism recommendation (e.g., tourism spots and traveling plans) and tourism marketing. Therefore, online photo data in an unstructured type has aroused an increasingly large attention in tourism research.

3.2.1. *Research focuses*

In the realm of tourism research, online photo data are serviceable in analyzing tourist behavior, presenting tourism recommendations and promoting tourism marketing. For tourist behavior, Vu, Li, Law, and Ye (2015) explored visitors' activities in Hong Kong by using geo-tagged photos. Lu et al. (2017) conducted a comparative study on the behaviors of overlapped tourists on different photo-sharing websites. Da Rugna, Chareyron, and Branchet (2012) discovered tourists' origins based on geo-tagged photos.

Exploring travel recommendations from photo data, regarding tourism destinations, travel routes and tourism duration, has been sufficiently studied. To discover popular tourism destinations (i.e., where to travel), Lee, Cai, and Lee (2014) extracted the associative points-of-interest patterns from the geo-tagged photos in Queensland, Australia. Zhou, Xu, and Kimmons (2015) detected the places of interest in multiple cities based on the spatial and temporal features of Flickr images. To find effective travel routes (how to travel), diverse travel route recommendation systems were proposed with the aim at extracting travel paths from a mass of photos (Kurashima, Iwata, Irie, & Fujimura, 2013; Lu, Wang, Yang, Pang, & Zhang, 2010; Okuyama & Yanai, 2013). As for tourism duration (how long to travel), Popescu and Grefenstette (2009) deduced visiting durations for the tourism attractions in four major cities, based on Flickr photo data. Similarly, Lee et al. (2014) identified the best time and the corresponding time duration for specific destinations based on the geo-tagged photos in Queensland and Australia.

Online photo data have also been used in promoting tourism marketing. Taking destination marketing for example, online photo data have increasingly become an effective vehicle to form tourism destination image (TDI) for potential visitors (Deng & Li, 2018). For example, Deng and Li (2018) proposed a machine learning based model to select photo elements from the viewers' perspective and assist destination marketing organizations (DMO) in photo selection. Hunter (2013) conducted a visual analysis of Hunan Province online destination image based on online photo data. Stepchenkova and Zhan (2013) compared the photos collected from DMO and Flickr, and found that the photos behave differently in reflecting users' perceptions of a destination.

3.2.2. *Data characteristics*

Online photo data convey a rich of useful messages in terms of metadata, i.e., the heterogeneous information embedded in photos. Valuable metadata for tourism research fall into four primary categories: user-related information (like photo ID and user ID), temporal information (taken date and uploaded date), geographical information (latitude & longitude) and textual information (title, descriptions and tags), as described in Table 1.

The photo big data for tourism research mainly originated from three photo-sharing websites or platforms: Flickr, Panoramio and Instagram. Useful APIs are provided by the three platforms, to allow an easy access to the photo data and the embedded metadata. Among the three platforms, Flickr with sufficient tourism photos

Table 1
Valuable metadata in online photos data for tourism research.

Metadata	Descriptions
Photo ID	ID of the photo downloaded from websites
User ID	ID of the tourist uploading the photo
Taken date	The date and time to take the photo
Uploaded date	The date and time to upload the photo
Geographical information	The latitude and longitude to take the photo
Textual metadata	Tourist-defined textual information such as title, descriptions and tags of the photo

was the dominant source of the photo data used in tourism research (e.g., De Choudhury et al., 2010; Kisilevich, Krstajic, Keim, Andrienko, & Andrienko, 2010; Kurashima et al., 2013; Lee et al., 2014; Majid et al., 2013; Mamei, Rosi, & Zambonelli, 2010; Quack, Leibe, & Van Gool, 2008; Shi et al., 2011; Shi, Serdyukov, Hanjalic, & Larson, 2013; Vu et al., 2015; Zhou et al., 2015; Önder, 2017). As for the other two sources, typical examples were Lu et al. (2010), Oku, Hattori, and Kawagoe (2015) and Okuyama and Yanai (2013) focusing on Panoramio datasets, and Lu et al. (2017) for Instagram datasets. An interesting question arises of why Flickr still dominant the photo-based tourism research, even when Instagram or other mobile photo-sharing platforms have more users and photos today. The possible reason might be that Flickr (released in 2005) has a relatively long history, with the corresponding data having been studied somewhat sufficiently, compared with Instagram (in 2010). Such an interesting phenomenon indicates that more attention could be paid to the rising stars in the photo-sharing family such as Instagram, in the future research.

3.2.3. *Analytic techniques*

To explore the valuable information hidden in photo data for tourism research, a variety of photo data mining techniques were implemented to construct a tourism recommendation system, including three major steps: data preprocessing, metadata clustering and trajectory discovery. Fig. 8 illustrates the typical process of using online photo data in tourism research, together with the corresponding techniques.

First, the raw data collected from photo-sharing websites are preprocessed in terms of data cleaning, formation and text mining, to leave valuable metadata for further analyses. For instance, Lee et al. (2014) removed duplicates of photos and converted the data into a formatted type for clustering analysis. Önder (2017) took a data cleaning method to identify whether the photos were posted by residents of the destination or tourists. As for the textual metadata in photos, text mining methods were utilized as preprocess techniques to explore tourists' interests and motivations to take the photos (Miah, Vu, Gammack, & McGrath, 2017). Through data preprocessing, valuable metadata can be extracted for the further studies in the following two steps.

Second, a clustering analysis was conducted on the extracted metadata from three primary perspectives: spatial dimension for discovering tourism spots (e.g., Kisilevich et al., 2010; Lee et al., 2014; Mamei et al., 2010; Okuyama & Yanai, 2013; Zhou et al., 2015); user dimension for travelers' origins (Da Rugna et al., 2012); temporal dimension for tourism durations (Lu et al., 2010; Popescu, Grefenstette, & Moëllic, 2009). Existing clustering methods for analyzing the metadata in tourism research can be classified into three main categories: centroid-based, density-based and connectivity-based methods. Regarding centroid-based methods, *k*-means is a popular algorithm in tourism research (e.g., Chen, Battestini, Gelfand, & Setlur, 2009; Kennedy, Naaman, Ahern, Nair, & Rattenbury, 2007; Kurashima et al., 2013). However, as argued by Lee et al. (2014), such a centroid-based partitioning method has one obvious drawback that it requires an exhaustive search to find the best cluster center, which is inefficient in data-rich environments; in contrast, density-based clustering might be more suitable for photo big data, which requires the minimum domain knowledge and effectively filters outliers even in the presence of noise points. Therefore, density-based spatial clustering of applications with noise (DBSCAN) and its variants have been developed and extensively applied to tourism photo clustering (e.g., Lee et al., 2014; Majid et al., 2013; Miah et al., 2017; Oku et al., 2015; Vu et al., 2015; Zhou et al., 2015). To build a hierarchy of clusters, connectivity-based clustering (also known as hierarchical clustering), a flexible and fast algorithm based on a dissimilarity matrix irrespective of metric spaces (Quack et al., 2008), has also

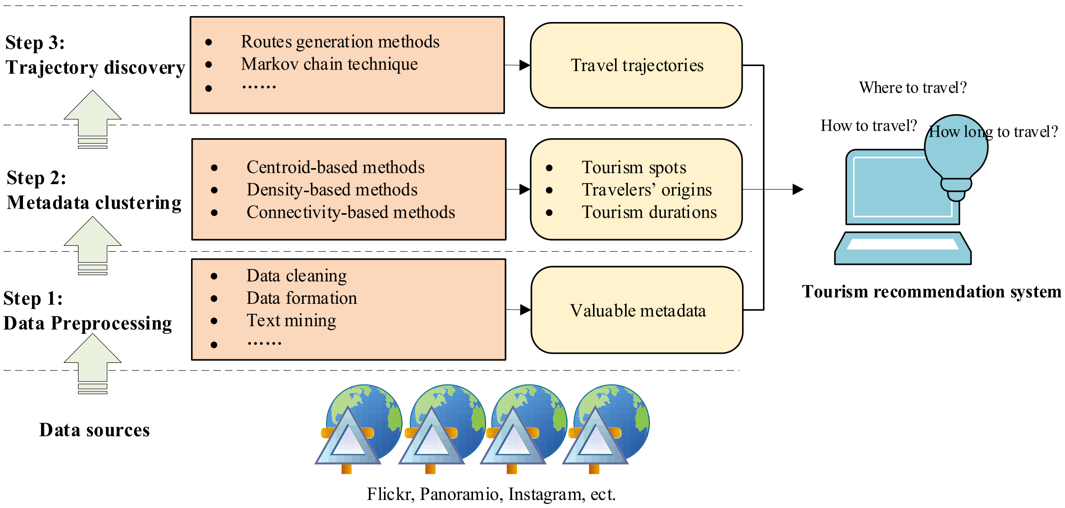


Fig. 8. Process of using online photo data in tourism research.

been introduced to detect tourist places and to group the cities that were visited by the same tourists (e.g., Okuyama & Yanai, 2013; Önder, 2017).

Finally, travel trajectories, i.e., the sequence of tourism spots and the time intervals between them, are investigated for making appropriate travel plans. As for analytic techniques, a variety of routes generation methods have been created and utilized (e.g., Lu et al., 2010; Okuyama & Yanai, 2013). Furthermore, Markov chain technique, a probabilistic model well handling sequential information, was another promising algorithm to capture more detailed knowledge of travel routes and to predict the next tourism spots based on tourists' current locations (e.g., Kurashima et al., 2013; Lu et al., 2010; Vu et al., 2015). Through this step, travel trajectories can be extracted from massive travelers' movements (Okuyama & Yanai, 2013).

Through the above three steps, a powerful tourism recommendation system can be formulated not only to extract valuable information from online photos regarding tourist behavior, tourism spots, tourism durations and travel trajectories, but also to present useful recommendations for travelling— where to travel, how to travel and how long to travel.

3.2.4. *Challenges and future directions*

According to the above analyses, the tourism research using online photo data can be improved similarly by expanding research areas and perfecting analytic techniques. As for research areas, discovering tourism spots (i.e., where to travel) from the spatial dimension of online photos has been studied much more sufficiently; in comparison, the research from the temporal dimension was otherwise insufficient, with only three related works (i.e., Lu et al., 2010; Popescu et al., 2009; Popescu & Grefenstette, 2009) to the best of our knowledge. However, temporal patterns, such as the best visiting time and duration, are also important factors in making appropriate travel plans. In addition, the sources of photo data were limited to three photo-sharing websites (i.e., Flickr, Panoramio and Instagram), while other social media, such as Sina Weibo, can also be considered.

Regarding analytic techniques, clustering and sequencing analyses have popularly been employed to investigate tourism-related photo data in the existing tourism research, and other competitive big data mining techniques for estimation, classification, prediction, affinity grouping and association rules can be also introduced into such a sunrise research. Furthermore, most existing studies conducted analyses indirectly on the metadata embedded in photos. However, photo data themselves involve a wealth of interesting information apart from metadata, and powerful photo mining techniques acting directly on images are needed.

4. Device data

With the vigorous development of IoT, diverse devices (or sensors) have been developed and employed to track tourists' movements, providing massive high-quality data for tourism management, such as GPS data, mobile roaming data, Bluetooth data, RFID data and WIFI data (Shoval & Ahas, 2016). Moreover, automatic weather station sensors have collected a rich mine of meteorological data to serve travel decision making, given that weather is an important factor in tourism. These above informative big data in both structured and unstructured types have already been applied to tourism research and appeared their respective superiorities.

4.1. *GPS data*

GPS is basically a series of satellites that orbit the earth broadcasting signals picked up by a system of receivers (Shoval & Isaacson, 2007). As a tool for tracking travelers' special movements, the distinctive advantage of GPS lies in being both global and accurate. The existing studies have fully shown feasibility and superiority of GPS data in tourism research (Bauder & Freytag, 2015).

4.1.1. *Research focuses*

Generally, the application of GPS data to tourism research has experienced three main stages, as shown in Fig. 9. In Stage I, similar to other sunrise researches at an early stage, feasibility and usefulness were the main research focus. For example, Van der Spek, Van Schaick, De Bois, and De Haan (2009) and Hallo et al. (2012) demonstrated the usefulness of GPS in tracking tourists. East, Osborne, Kemp, and Woodfine (2017) proved that combining GPS and survey data can improve understanding of visitor behavior. Furthermore, Tchetchik, Fleischer, and Shoval (2009) developed a method to collect GPS data, which can be used to segment visitors.

In Stage II, tourist behavior was mainly concentrated on, covering spatial behavior, temporal behavior and spatial-temporal behavior. To explore tourist spatial behavior, Edwards and Griffin (2013) adopted GPS tracking data to find out how tourists moved around a city. For temporal behavior, Birenboim, Anton-Clavé, Russo and Shoval (2013) explored the temporal activity patterns of theme park visitors. However, far more existing research preferred spatial-temporal behavior by combining spatial behavior with temporal behavior, with the typical works by McKercher, Shoval, Ng, and Birenboim (2012), Zakrisson and Zillinger (2012), Grinberger, Shoval, and McKercher (2014) and Shoval, McKercher, Birenboim, and Ng (2015).

In Stage III, GPS data has been extended to tourism recommendation. For instance, Yoon, Zheng, Xie, and Woo (2010)

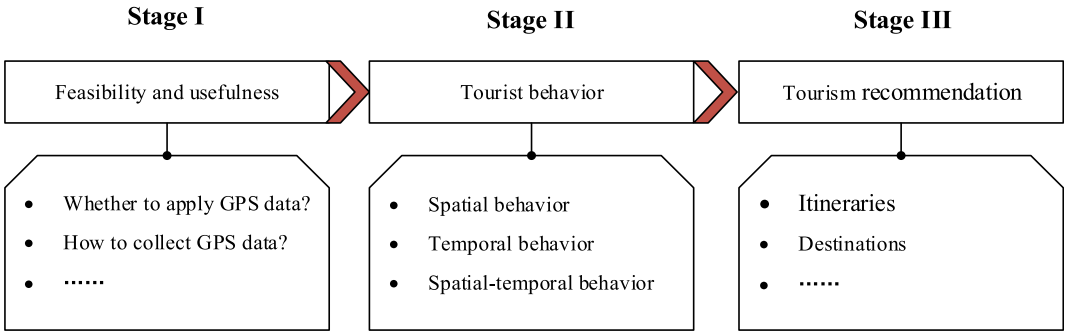


Fig. 9. Historical tour of applying GPS data to tourism research.

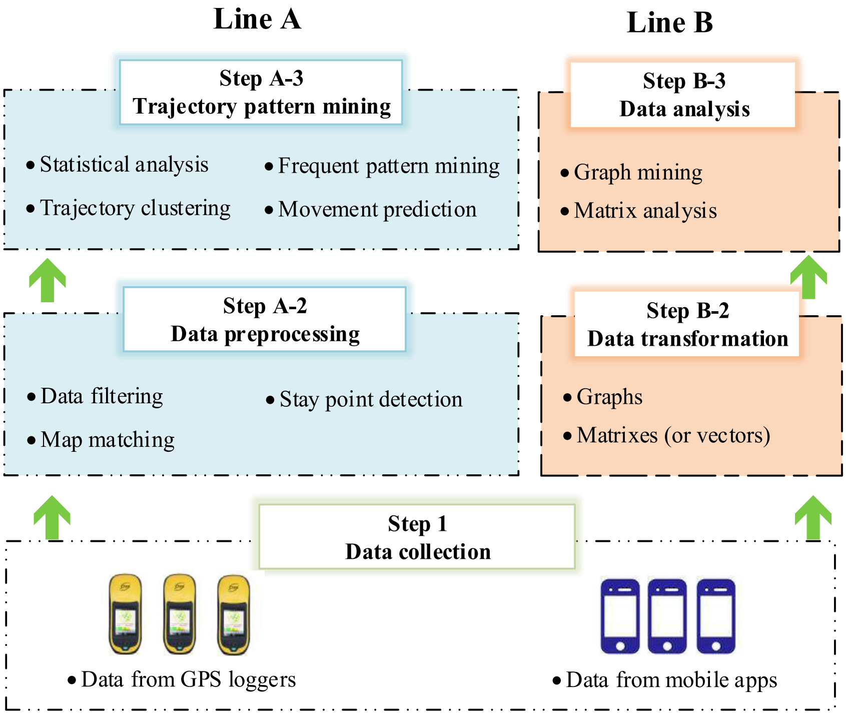


Fig. 10. Process of using GPS data in tourism research.

proposed a smart recommendation for highly efficient and balanced itineraries based on multiple GPS trajectories. Zheng, Huang and Li (2017) predicted the next destination of individual tourists using the GPS tracking data.

4.1.2. *Data characteristics*

GPS loggers carried by volunteers and GPS-enabled mobile applications installed in smartphones are two primary channels to collect GPS big data in tourism research. On the one hand, asking participants to take along GPS loggers during their visits was a dominant method to collect GPS data in the existing tourism research (e.g., Beeco et al., 2013; McKercher et al., 2012; Zakrisson & Zillinger, 2012). Generally, attributes of such GPS data include longitude, latitude, time stamp, speed, direction, etc. Absolutely, the GPS loggers can potentially avoid many problems (such as inaccuracy) of other tracking methods (Zheng et al., 2017). However, recruiting volunteers to carry the GPS loggers requires a high research cost (including both devices and labor costs), and this data collection method is somewhat sample-biased and goal-oriented.

On the other hand, a much more low-budget and flexible channel, i.e., GPS-enabled mobile applications, was also used for tourism research to obtain the tourist spatial-temporal behavior. For instance, Brovelli, Minghini, and Zamboni (2016) applied the GPS data collected from a mobile client application, Open Data Kit (ODK) Collect that is available for Android mobiles, to map tourist destinations. Ayscue, Boley, and Mertzlufft (2016) obtained GPS data via the smartphone application of GPS & Map Toolbox for iPhones, and argued that mobile GPS data can help better understand resident attitudes toward tourism.

4.1.3. *Analytic techniques*

Various data mining techniques have been developed and employed for investigating the GPS data in tourism research. Based on different data formats, these techniques fall into two technical lines, with the first one (marked as Line A) directly processing raw GPS data and the second (Line B) for the transformed formats (e.g., graphs, tensors and matrixes), as their respective processes compared in Fig. 10.

In both Lines A and B, the first step is to collect data. Generally, the GPS data used in tourism research are mainly from two sources: GPS loggers carried by volunteers and GPS-enabled mobile applications installed in smartphones. Accordingly, GPS data can be collected by either recruiting volunteers to carry GPS loggers during their tourism activities, or asking participants to share their locations via GPS-enabled mobile applications.

As for Line A, after data collection, another two steps are conducted to process raw GPS data, i.e., data preprocessing and trajectory pattern mining, as listed in the left part of Fig. 10. In data preprocessing (Step A-2), a set of tasks, such as data filtering, map matching and stay point detection, are performed. *Data filtering* aims to remove noise points caused by poor signals. For example, Birenboim, Reinau, Shoval, and Harder (2015) removed meaningless points in the closed exhibitions (where GPS reception is low) from the original data, to explore the real tourists' movements. *Map matching* projects the points of a history trajectory onto the segments of a road network where the points are generated, for analyzing tourist behavior (Li, Wu, Peng, & LV B, 2016). For instance, East et al. (2017) projected GPS points onto a park map for tourist behavior understanding. *Stay point detection* identifies the location where a moving object has stayed for a while (Zheng, 2015). In tourism research, a stay point might correspond to a specific place such as a tourist spot, a restaurant and a shopping mall that a tourist has been to, carrying more semantic meanings than other points in a trajectory. According to the existing studies, the stay point detection has successfully been used to explore main places of interest (Orellana, Bregt, Ligtenberg, & Wachowicz, 2012) and time allocation during a tour (Birenboim et al., 2013).

In Step A-3, the movement patterns of tourists are further explored based on the preprocessed GPS data, and popular technologies include statistical analysis, trajectory clustering, frequent pattern mining and movement prediction. *Statistical analysis*, the basic data analysis method, has widely been used to capture tourist trajectory patterns from GPS data. Representative examples were: descriptive analysis, to estimate average travel time (Shoval, McKercher, Ng, & Birenboim, 2011) and GPS data points as well as their precision (Hallo et al., 2012); circular statistics, to compute circular mean times and dispersions for different groups of tourists (Chhetri, Corcoran, & Arrowsmith, 2010). *Trajectory clustering* identifies representative paths or common trends shared by different tourists by grouping similar trajectories into one cluster. The popular algorithms were hierarchical cluster (Zakrisson & Zillinger, 2012) and *K*-means cluster (Huang & Wu, 2012). *Frequent pattern mining* explores frequently generalized sequences in a time-ordered set of events. For instance, Orellana et al. (2012) adopted frequent pattern mining to explore generalized sequential patterns from the tourists' movements in natural recreational areas. *Movement prediction* estimates the next tourist location based on history GPS tracking data. For example, Zheng et al. (2017) utilized a probabilistic analysis to forecast the next potential locations of individual tourists.

Instead of directly processing raw GPS data, Line B transforms them into other formats, such as graphs and matrixes (or vectors). Such new representations of trajectories largely expand and diversify the approaches to GPS data mining. For instances, when transforming GPS data into graphs, powerful graph mining methods could be introduced, for instance, to identify tourism areas and travel routes (Tchetchik et al., 2009). Orellana et al. (2012) generated a movement vector based on movement parameters such as speed and bearing, to investigate tourist behavior in space and time.

4.1.4. Challenges and future directions

Many improvements were still required to enrich the research focuses and analytical methods of using GPS data in tourism research. For research focuses, although recent research has started applying GPS data to tourism recommendation, most of the related works focused on exploring feasibility of the GPS data in tourism research and analyzing tourist behavior. Therefore, interesting studies that apply the informative GPS data to tourism recommendation and other issues (such as tourism product design) are strongly suggested. Furthermore, the GPS data in tourism research were collected mainly through two channels, i.e., GPS loggers carried by volunteers and GPS-enabled mobile applications installed in smartphones. With the development of tourism service facilities, other GPS loaders, such as GPS-enabled watches, sharing bicycles (and cars), renting cars and tourist yachts, have emerged and might provide new significant information about tourist movements.

As for analytic methods, the existing technologies were multifarious and diverse; however, some other useful processing tools for GPS data in other research fields can also be introduced. For example, we can record a time-stamped geographical coordinate for tourists every second, to enhance data precision. However, many applications do not really need such a high precision of location in practice. Therefore, data compression methods can be implemented as a preprocessing method, which can reduce the overhead for communication, computing and data storage. Moreover, tourists may have periodical patterns when they go to the same tourism destinations periodically. Such a periodic behavior can be used to forecast tourist movements and thus deserves more attention (Zheng, 2015).

4.2. Misc. Device data

In addition to GPS data (accounting for the largest proportion of device data in tourism research), other device data are also important for tourism research, including mobile roaming data, Bluetooth data, RFID data, WIFI data and meteorological data.

4.2.1. Mobile roaming data

With the rapid development of telecommunication technology, roaming service provided by mobile network operators has become a relatively new but increasingly popular tool in tracking tourist behavior. Since travel is the movement of people between relatively distant geographical locations, travelers trend to use their mobile phones in other place or even foreign countries inevitably. Roaming service allow tracking tourist locations when they use the mobile phone in non-registered places, in terms of mobile roaming data.

In view of privacy concerns, i.e., that both tourists and mobile network operators do not wish to share their private information, mobile roaming data have not been widely used in tourism research yet, as the available works listed in Table 2. In comparison to GPS data, the application of mobile roaming data to tourism research was still at a starting stage, with two focuses: data applicability and tourist behavior. On the one hand, Ahas, Aasa, Roose, Mark, and Silm (2008) investigated the usefulness of mobile roaming data in tourism research. On the other hand, roaming data have been applied to analyzing tourist behavior such as tourist flows (Raun, Ahas, & Tiru, 2016), travel distances (Nilbe, Ahas, & Silm, 2014), repeat visiting (Kuusik, Tiru, Ahas, & Varblane, 2011), destination loyalty (Tiru, Kuusik, Lamp, & Ahas, 2010) and space consumption (Ahas, Aasa, Mark, Pae, & Kull, 2007).

Mobile roaming data are collected via radio waves, which are sent and received by telecommunication base station and stored automatically in the memory or log files of mobile network operators (Raun et al., 2016). When a mobile phone registered in a place but is used in another place, its user might be recognized as a potential tourist, and the corresponding information, e.g., country of origin, location coordinates and the time making a phone, are recorded as mobile roaming data for reflecting tourist spatial-temporal behavior. It is worth noticing that even both effective in tracking tourist movements outdoors, mobile roaming data and GPS data appear different data characteristics (Ahas et al., 2008). In particular, GPS data can trace spatial movements continuously during the whole travel, whereas mobile roaming data record the locations only on active uses of mobile phones in the mobile network: outgoing and incoming calls; sending and receiving messages; using the Internet and data services (Raun et al., 2016). Moreover, the precision of mobile roaming data are somewhat lower than GPS data (Ahas et al., 2008).

Mobile roaming data has both advantages and disadvantage, especially for tourism research. As for advantages, mobile roaming data can cover a larger spatial area, even including some less visited

Table 2
Typical works on application of mobile roaming data to tourism research.

Authors	Analytic technique(s)	Tourism issue
Raun et al. (2016)	Econometric model	Tourist flows
Nilbe et al. (2014)	Econometric model	Travel distances
Kuusik et al. (2011)	Statistical analysis	Segmentation of repeat visitors
Tiru et al. (2010)	Statistical analysis	Destination loyalty of tourists
Ahas et al. (2008)	Statistical analysis	Data applicability in studying tourism
Ahas et al. (2007)	Case study & statistical analysis	Foreign tourists' space consumption

Table 3
Typical works on application of Bluetooth data to tourism research.

Authors	Event(s)/place(s)	Tourism issue	Sensors' number	Devices' number	Detection duration
Versichele et al. (2014)	Cathedral, church, indoor market, etc.	Tourist attractions	29	17,496	15 days
Yoshimura et al. (2014)	Museums	Tourist movements	7	24,452	24 days
Delafontaine et al. (2012)	Trade fair	Tourist movements	22	14,498	5 days
Versichele et al. (2012)	Ghent festivities event	Tourist movements	22	102,467	10 days
Stange et al. (2011)	Tribune and shopping center	Tourist movements	30	12,700	2 days

places. Furthermore, some valuable tourist information such as origin countries are contained in mobile roaming data. Therefore, applying mobile roaming data to tourism research can provide a new analysis perspective for tourist behavior and tourism plans. For disadvantages, obtaining mobile roaming data is extremely difficult due to privacy and surveillance concerns, which largely suppressed its applications. In addition, the location information of mobile roaming data might be at a relatively low level of accuracy, particularly when the Global System of Mobile communication (GSM) is poorly equipped in the tourism destinations (Ahas et al., 2007).

Popular processing techniques for mobile roaming data are statistical analyses, econometric models and case study, in tourism research. As for statistical analyses, Kuusik et al. (2011) applied statistical methods to detect repeat visitors. Tiru et al. (2010) conducted a statistical analysis on foreign visits and uses of roaming phones. Ahas et al. (2008) presented a statistical analysis on the tourist roaming calls in Estonian. In terms of econometric models, Raun et al. (2016) used a binary logistic regression to compare the overall visits to Estonia with the visits to two smaller studying areas in Estonia. Nilbe et al. (2014) constructed a linear regression to evaluate travel distance. By combining case study and statistical analysis, Ahas et al. (2007) analyzed the seasonality of foreign tourists' space consumption in Estonia.

The tourism research with mobile roaming data was still at a starting stage—the related studies were relatively few in comparison to those with GPS data, and the analyses were somewhat simple. Thus, there was a lot of room to develop such an emerging area. As for research areas, the existing studies still focused on data applicability (similar to Stage I for GPS data), and tourist behavior analysis (similar to Stage II). Accordingly, interesting studies at a deeper level and on a larger range (such as covering the factors in tourism recommendation) are suggested for the tourism research using mobile roaming big data. As for analytic techniques, descriptive statistical analyses, econometric models and case study were the dominant approaches to analyzing mobile roaming data for tourism research. Other even more effective analysis tools, such as data mining algorithms and artificial intelligences, are strongly recommended to investigate such informative data.

4.2.2. Bluetooth data

Bluetooth, invented by Ericsson in 1994, is an open and wireless communication technology. To collect Bluetooth data, Bluetooth sensors are preplaced in the target area, and personal carry-on devices (such as mobile, mp3-player and headsets) could be detected. Since Bluetooth technology can well monitor non-participatory and unannounced movements of massive individuals in terms of positions and trajectories, Bluetooth data thus opens a new analysis dimension of tourist behavior.

Even with the extensive development and application of Bluetooth technology, the tourism studies using Bluetooth data were somewhat insufficient in terms of a small number of works and a limited research area. Table 3 lists the available works, and two interesting findings can be obtained. First, mainly due to the finite monitoring range of Bluetooth, these studies were limited to: a planned event-tourism activity, such as a festivity event (Versichele, Neutens, Delafontaine, & Van de Weghe, 2012) and a trade fair (Delafontaine, Versichele, Neutens, & Van de Weghe, 2012); or an indoor place, such as, a cathedral, a church, an indoor market (Versichele et al., 2014), a museum (Yoshimura et al., 2014) a tribune and a shopping center (Stange, Liebig, Hecker, Andrienko, & Andrienko, 2011). Second, Bluetooth data were used to capture two tourism issues, tourist spatial-temporal movements (Delafontaine et al., 2012; Stange et al., 2011; Versichele et al., 2012; Yoshimura et al., 2014) and tourist attractions (Versichele et al., 2014).

Bluetooth data record valuable information for tracking tourist spatial-temporal behavior, including time stamp and signal strength intensity (RSSI) of detection, and media access control (MAC) address and class of device (COD) code of the detected device. As for sample sizes, the numbers of sensors and detected devices, and detection duration in each related study are presented in Table 3. Obviously, a small number of Bluetooth sensors (from 7 to 30) sufficed to detect a large number of individuals (12,700–102,467). In addition, because the related studies mainly focused on event-tourism activities, the sampling periods were relatively short, all within a month.

Bluetooth technique has both superiorities and weakness in tracking visitors' movements. Regarding superiorities, compared to GPS, Bluetooth tracking is cost-effective and convenient, neither hiring the participants to take along the loggers nor requiring previous registration (Yoshimura et al., 2014). Moreover, Bluetooth allows unannounced tracking, in which the tracked targets are ignorant of being tracked (Versichele et al., 2014). Third, Bluetooth can be used in crowed indoor scene (e.g., inside buildings) or in the proximity of tall structures where GPS connectivity cannot be guaranteed, which is superior to GPS. Considering weaknesses, a Bluetooth proximity sensor just let us know the time-stamped

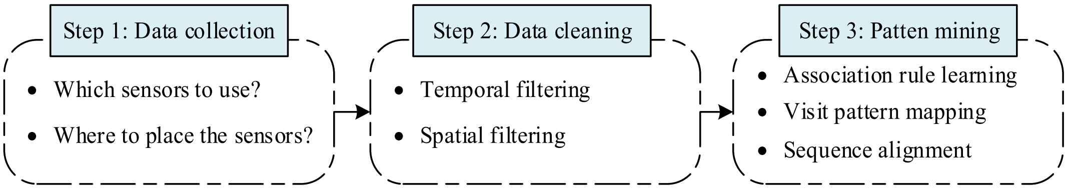


Fig. 11. Process of using Bluetooth data in tourism research.

sequence of individual transitions between nodes rather than all the movements of a device as GPS provided (Yoshimura et al., 2014). Furthermore, the coverage of the study areas is limited due to node radio ranges (Delafontaine et al., 2012). In addition, like other device data, the use of Bluetooth data might lead to some ethic problems concerning privacy losses, i.e., revealing the private information of tourists without announcement and agreement (Hardy et al., 2017).

To process the Bluetooth data, three main steps were involved in the existing tourism studies: data collection, data cleaning and pattern mining (in Fig. 11). In *data collection*, an appropriate type of Bluetooth sensors is chosen in relation to the given detection range (Versichele et al., 2012); the placing plan of the sensors is designed to not only well balance efficiency and economy, but also strategically cover the whole desired area. In *data cleaning*, the raw Bluetooth data are processed by temporal and spatial filtering to remove noise or correct mistakes, leaving the precise spatial-temporal movements of visitors (Stange et al., 2011). In *pattern mining*, the meaningful information, with respect to the tourist spatial-temporal movements and tourist attractions, are explored by using effective data mining methods, including association rule learning (Versichele et al., 2014), visit pattern mapping (Versichele et al., 2014) and sequence alignment (Delafontaine et al., 2012).

According to the existing literature, the application of Bluetooth data to tourism research was insufficient, in terms of few related studies and narrow research area. Since Bluetooth has two distinctive superiorities, i.e., availability in crowed indoor scenes and allowing unannounced tracking, Bluetooth data deserve a further investigation in terms of expanding research area (such as indoor tourism recommendation for travel routes and tourism duration) and mining techniques (such as those for GPS data).

4.2.3. RFID data

The commercial use of RFIDs started more than 20 years ago and have been proved to be useful in improving service operations such as tracking movements, automation for labor replacement, cycle time reduction, and personal security and safety (Ferrer, Dew, & Apte, 2010; Roh, Kunnathur, & Tarafdar, 2009). Recently, RFIDs have also shown benefits in perfecting tourism processes in hotels, cruise ships, resorts and theme parks (Hozak, 2012).

The application of RFID data to tourism research have two focuses, data feasibility analysis and tourism recommendation. As for data feasibility analysis, Hozak (2012) discussed the application of RFID data in tourism. Öztayşi, Baysan, and Akpinar (2009) investigated the possibility of utilizing RFID in hospitality industry to enhance service quality, customer satisfaction, market share and profitability. Zeni, Kiyavitskaya, Barbera, Oztaysi, and Mich (2009) proposed a lightweight framework to gather RFID data on the places that tourists visit during an event. In addition, RFID reader deployment strategy is also an important research issue in data feasibility analysis. For instance, Tsai, Chang, and Kuo (2017) proposed an ant colony based optimization for RFID reader deployment in theme parks. As for tourism recommendation, a variety of recommendation systems have been designed. For instance, Wan (2009) proposed a RFID and personalized recommendation based tourism information system, in which RFID data were used for user identification. Tsai and Chung (2012) provided a personalized route recommendation service for theme parks using RFID information. Deka et al. (2016) developed a system, in which RFID tags were used to guide and to inform tourists about the given location and surroundings.

The RFID system consists of two components, a reader and a tag. The tag is also known as transponder, namely a combination of transmitter and receiver for receiving a specific radio signal and automatically transmitting a reply. In tourism research, the tags can be attached to E-passports, public transport cards, Trento card, mobile, luggage, hotel items, etc. (Lucia, 2013; Wan, 2009; Öztayşi et al., 2009). The reader is a transceiver, for querying the sensor tags, collecting the information and taking a reaction. RFID data appear two primary advantages in tracking tourists. First, RFID data, especially those obtained from passive tags without a power source, are cost-competitive compared to other device data such as GPS, WIFI and Bluetooth data (Lucia, 2013). Second, the precision of RFID data is at a rather high level, similar to GPS data (Lucia, 2013).

The analytic technologies for RFID data in tourism research were mainly qualitative analysis, whereas quantitative data mining technologies were far less used. The typical qualitative analysis method, case study, has widely been applied to both the data feasibility of RFID data in tourism research and tourism recommendation (Deka et al., 2016; Öztayşi et al., 2009; Wan, 2009). Some data mining methods such as data cleaning, clustering and route generation were used to investigate tourist behavior (Tsai & Chung, 2012).

Unfortunately, the potential of RFID data in serving tourism research has not been fully exploited yet. On the one hand, most of existing studies have paid an attention to the data feasibility in tourism research; however, other interesting issues, such as tourist behavior and consumer preference based on RFID payment data in hotels (Öztayşi et al., 2009), also deserve investigation. On the other hand, the dominant analytic methods for RFID data in tourism research were qualitative analysis; in comparison, the helpful big data mining approaches that have already been applied to other device data (particularly GPS data, see Section 4.1-(2)) have not been introduced.

4.2.4. WIFI data

WIFI can be considered as a vigorous alternative to Bluetooth in tracking tourist movements. Although similar to Bluetooth in allowing unannounced tracking, WIFI otherwise appears a far more convenient and low-cost virtue (Bonné, Barzan, Quax, & Lamotte, 2013). First, unlike Bluetooth using a general application for older generation smartphones, WIFI is enabled by default on modern smartphones. Second, WIFI does not depend on a smartphone being set into a discoverable mode, nor require the smartphone to be connected to a wireless network. As for disadvantages, WIFI data are confronted with small-range coverage and privacy concerns, similar to Bluetooth data.

The tourism research using WIFI data was much more insufficient than those with Bluetooth data, and there only existed two related studies to the best of our knowledge. Bonné et al. (2013) might be the first attempt to introduce WIFI tracking data into tourism research, for capturing tourist behavior in a popular music festival and a university campus. Chilipirea, Petre, Dobre, and van Steen (2016) investigated the data cleaning techniques for the WIFI data recorded by the WIFI scanners located in the city center during a festival. In view of its distinctive merits, WIFI data have a bright prospect for tourism research, especially for tourist behavior in a tourism event (such as a concert, a museum visit and a sports meeting), and hence tourism recommendation and emergency management.

4.2.5. Meteorological data

Meteorological data are also a typical kind of big data in terms of complexity and large-scale, in a structured, unstructured or hybrid type. Meteorological data are collected by automatic weather station sensors, and generally fall into fifteen categories according to contents and attributes: upper air data, surface data, radiation data, marine data, agricultural data, cryosphere data, chemistry and physics data of atmosphere, hydro-meteorology data, solar-terrestrial physics data, analytic data, meteorological disaster data, historical data, soil and vegetation data, radar data and satellite data (Guo, 2016). These data are in different formats: office document, text, image, XML, HTML, audio and video. Therefore, processing meteorological data in different categories and formats is a challenging task.

According to the existing literature, such valuable big data, nevertheless, had not an extensive application to the field of tourism research yet, with two main focuses: effect estimation of weather in tourism and tourism recommendation. Joo, Kang, and Moon (2014) measured the effect of rain on the behavior of theme-park visitors based on a two-stage decision-making process. Using the dynamic heterogeneous panel data technique, Falk (2010) analyzed the relationship between the number of overnight stays and different measures of snow depth based on panel data covering 28 Austrian ski resorts. Recently, Guo (2016) used meteorological big data to calculate tourism indexes of tourism industry and provide tourism recommendations to travelers, via the Hadoop architecture and MapReduce algorithm.

However, the value of meteorological data might be underestimated, given that a satisfactory travel is high in relation to a fine weather. For example, mining meteorological data will offer useful insights into various tourism decisions, such as traveling plan making and tourism product design. Moreover, investigating how meteorological data influence short-term tourist decisions and behaviors may be another interesting topic for further study.

5. Transaction data

Transaction data, another valuable type of big data for tourism research, record tourism-related operations (or transaction, activities and event in tourism market), such as web searching, webpage visiting, online booking & purchasing, etc. The corresponding transaction data have already been employed to promote tourism prediction, search engine optimization (SEO), tourism behavior understanding and tourism marketing.

5.1. Web search data

Search engines are an emerging source of big data for tourism research, allowing and recording the web searching operations for tourism-related contents. In particular, tourists can seek travel information through a search engine, leaving searching traces on the websites; in return, such traces are recorded and processed to form a valuable type of big data—web search data, directly reflecting public attention toward a tourism item thereby helpfully understanding tourism market.

5.1.1. Research focuses

According to the existing studies, web search data have appeared an excellent performance in tourism research, especially for capturing tourist online behavior and making the related decisions (Li et al., 2016). In particular, tourism prediction was the top hot issue by using web search data, as typical works listed in Table 4.

From Table 4, three important findings can be easily obtained. First, as for prediction targets, the existing tourism researches using web search data focused on the most essential factor in tourism market—tourism demand, covering tourist volume, tourist flows and hotel demand. The result indicates that because web search data can effectively capture public attention toward tourism, thus, they can be used as powerful predictors for tourism demand. Second, given that the first work was published in 2011 (i.e., Gawlik et al., 2011), using web search data in tourism prediction had a relatively short history and remained at an early stage. The result implies that there might be a considerable amount of room to improve such a sunrise research and the analytic techniques. Third, among search engines, Google and Baidu were the dominant ones, with the corresponding web search data of Google trends and Baidu index, respectively. Interestingly, the latter was all applied to China's tourism market, and the hidden reasons will be discussed in the subsection of data characteristics.

In addition, web search data have also been proved useful in SEO or search engine marketing (SEM)—a promising measure for travel and tourism (online) marketing. In particular, SEO and SEM aim to help businesses and organizations to gain visibility on the search engine result pages through paid or non-paid means, and have become one of important strategic tools for promoting and advertising tourism products (Pan et al., 2011). Web search data, reflecting travelers' searching behavior on the Internet, obviously have a great relationship with SEO and SEM. Some existing studies have discussed it in depth and proved that it is promising. For example, Xiang & Pan (2011) identified the patterns in online travel queries across tourist destinations, and offered insightful implications to SEM for tourist destinations. Pan et al. (2011) synthesized the research on SEM in tourism and related fields, and presented a model to describe the evolving dynamics in SEM.

5.1.2. Data characteristics

From Table 4, the Google trends and Baidu index were the most popular web search data used in tourism research. In particular, Google trends and Baidu index get the statistical big data by sending website traffic data to analytics servers via a snippet (tracking code) which is included on the website and activated when a visitor views a page on somebody's website (Boswell, 2011, p. 135). The Google trends and Baidu index have both similarities

Table 4
Typical works on applications of web search data to tourism prediction.

Author(s)	Search engine(s)	Data frequency	Prediction target	Region(s)
Peng, Liu, Wang, and Gu (2017)	Baidu	Daily	Tourist volume	Jiuzhaigou, China
Li et al. (2017)	Baidu	Monthly	Tourist volume	Beijing, China
Huang, Zhang, and Ding (2017a)	Baidu	Daily	Tourist volume	The Forbidden City, China
Li et al. (2016)	Baidu	Daily	Tourist volume	Jiuzhaigou, China
Rivera (2016)	Google	Weekly	Hotel demand	Puerto Rico
Gunter and Önder (2016)	Google	Monthly	Tourist volume	Vienna, Austria
Yang et al. (2015)	Google & Baidu	Monthly	Tourist volume	Hainan Province, China
Park, Lee, and Song (2017)	Google	Monthly	Tourist inflows	Japan to South Korea
Bangwayo-Skeete and Skeete (2015)	Google	Weekly	Tourist volume	US, Canada & UK to five recipient Caribbean countries
Pan, Chenguang Wu, and Song (2012)	Google	Weekly	Hotel demand	Charleston, US
Choi and Varian (2012)	Google	Monthly	Tourist volume	Hong Kong
Gawlik, Kabaria, and Kaur (2011)	Google	Monthly	Tourist volume	Hong Kong
Artola, Pinto, and de Pedraza García (2015)	Google	Monthly	Tourist inflows	Spain

Table 5
Characteristic comparison of Google trends and Baidu index.

Characteristic	Google trends	Baidu index
Using area	The whole world	China only
Period	2004 to date	PC index: 2006 to date; Mobile index: 2011 to date
Frequency	Monthly & weekly	Weekly & Daily
Maximal number of terms searched at once	5	5
Search volume reported	Relative volume	Absolute volume
Method of term matching	Partial match	Partial match

and differences in data characteristics (Vaughan & Chen, 2015), as listed in Table 5.

Based on data characteristics (see Table 5), some interesting findings for research focuses (Table 4) can be finely explained. As for data frequency, because the Google trends and Baidu index are generated at different frequencies, the corresponding tourism predictions are different frequency dominant. In particular, all existing tourism predictions with Google trends used monthly data or weekly data, while those with Baidu index mainly focused on daily data in 3 out of 5 articles. These results are in line with the respective frequency features of the two data types. However, as high-frequency data (especially daily data) are extremely sensitive to various uncertain factors (such as weather and extreme events), failing in reflecting tourists' real behaviors. Thus, lower-frequency data (such as quarterly data) are recommended for future research.

As for using areas, it is noticeable that Baidu index was limited to China, while the Google trends was used worldwide. Therefore, in the existing research for China's tourism prediction, Baidu index was used broadly, whereas those for other regions preferred Google trends. In particular, Google engine exited from the mainland China in 2010. Thus, although Google is the dominant search engine in the world (taking up approximately 66.7% of global market share in 2013), it is not the case in China (only 2.1% of Chinese market share) (Yang et al., 2015). In comparison, Baidu engine makes up the largest proportion in China (69% in 2013), thus Baidu index might be more suitable in the case of China's tourism prediction than Google trends (Yang et al., 2015).

5.1.3. *Analytic techniques*

As for how to apply the valuable web search big data to tourism prediction, two main steps are taken, i.e., keywords selection and predictors introduction, as shown in Fig. 12.

In the first step, keywords are carefully selected to obtain the appropriate web search data $x_{1,t}$, $x_{2,t}$,…, $x_{k,t}$ from a search engine, where $x_{i,t}$ denotes the web search data of the ith keyword at time t and k is the total number of selected keywords. Keywords selection is the core process in the tourism research using web search data, and the results are highly dependent on the selection methods (Peng et al., 2017). According to the related studies, three kinds of keywords selection methods were popularly used in tourism research: empirical (or experiential), territorial and technological approaches.

Empirical approach (i.e., direct selection) determines keywords simply based on the knowledge and experiences of researchers. Accordingly, empirical method becomes the simplest way of keywords selection, and has popularly been employed to retrieve the web search data in tourism research. For instance, based on empirical method, Bangwayo-Skeete and Skeete (2015) directly implemented the two keywords, i.e., 'hotels' and 'flights', to obtain the corresponding Google trends as the predictors for the tourist arrivals in Caribbean. Similarly, Pan et al. (2012) used empirical method and choose five relevant Google trends to forecast the demand for hotel rooms. Though simple and easy to operate, such a subjective method is prone to ignoring important keywords or even selecting incorrect ones.

Territorial keywords selection is an extension of the empirical one, first using empirical method to determine base keywords and then adding other keywords pertinent to both the base and the recommended ones (via the powerful recommendation function of search engines), with the main aim of getting full-scale keywords. In tourism research, Huang et al. (2017a) utilized territorial method to seek Baidu index data for forecasting the tourism flows of the Forbidden City in China, with the base keywords of 'the Forbidden City', 'the Palace Museum', 'the Forbidden City in Beijing', 'tickets of the Forbidden City', 'the Palace Museum ticket price', 'the Palace Museum picture', etc. Obviously, territorial method can capture far

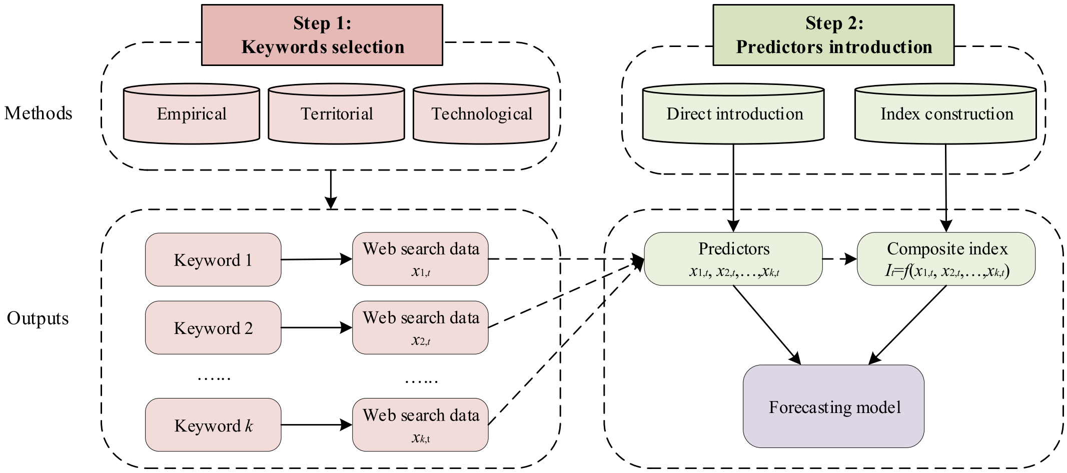


Fig. 12. Process of using web search data in tourism prediction.

more keywords than empirical method; however, it might otherwise suffer a large level of noises due to covering excessive keywords and even unrelated ones.

Technological method sifts keywords from a large selection scope based on the predictive ability, in terms of the correlations with the predictive variables. For instance, Yang et al. (2015) selected the keywords of Google trends and Baidu index, based on Pearson correlation coefficient between each candidate web search data and the predictive variable (i.e., China's tourist volume) at different lags. Peng et al. (2017) combined Hurst exponent and time difference correlation analysis to select keywords of Baidu index for predicting the volume of tourism visitors in the Jiuzhai Valley scenic area. Interestingly, finely coupling territorial approach (for generating sufficient candidates) with technological method (for sifting the most predictive keywords) might be a relatively systematic method for selecting keywords of web search data.

As for how to introduce the web search data of the selected keywords into forecasting models, most of tourism prediction researches directly used the raw data, $x_{1,t}$, $x_{2,t}$,…, $x_{k,t}$, as predictors (e.g., Choi & Varian, 2012; Huang et al., 2017a; Pan et al., 2012). However, some recent studies preferred index construction, i.e., combining them into one or a few composite index(s), $i_t = f(x_{1,t}$, $x_{2,t}$,…, $x_{k,t})$. For example, Li et al. (2017) employed generalized dynamic factor model and principal component analysis to construct two indexes, respectively, in the prediction for China's tourism demand. Focusing on China's tourist volume, Yang et al. (2015) aggregated web search data into one index by using shift and sum method. Park et al. (2017) further extended this method by introducing mean square error and mean absolute error as criteria to determine baseline regression models instead of pre-determined thresholds. When the keywords are excessive, index construction method appears superiority in effectively avoiding collinearity in regression over direction introduction method.

5.1.4. Challenges and future directions

Generally, using web search data in tourism prediction was still at an early stage (just since 2011), and there existed a lot of room to improve such a sunrise research, particularly from the perspectives of research area expansion and analytic technique innovation. As for research areas, such a valuable type of big data, web search data, has focused on tourism prediction and SEO (or SEM). However, as web search data can provide a better understanding for tourist attention toward a tourism product or destination, they could also helpfully serve other interesting tourism issues, e.g., tourism product design, tourism precaution, etc. In the existing studies on tourism prediction, tourist demand (covering tourist volume, tourist flows and hotel demand) was concentrated on, while other essential factors in tourism market, e.g., tourism revenue and environmental carrying capacity, were neglected. However, an exact prediction for these crucial factors is also extremely important for managers to make various decisions in tourism industry.

Regarding analytic techniques, keywords selection is the most crucial process to seek appropriate web search data, and most of the existing researches relied on empirical and territorial methods. However, empirical approach, selecting keywords subjectively and directly, might ignore certain important ones or even select incorrect ones; territorial approach might otherwise suffer a high level of noises due to covering excessive keywords and uninfluential ones. Therefore, finely coupling territorial approach (for generating sufficient candidates) with technological methods (for sifting the most powerful keywords) is strongly recommended as a relatively systematic method in tourism research for selecting web search data. Furthermore, in technological method, besides correlation analyses, other relationship investigations can be also introduced to find the most predictive web search data. Finally, when the keywords (and thus the corresponding series of web search data) are too many, index construction, i.e., combing them into one or a few composite index(s), is strongly suggested to avoid collinearity in regression, rather than introducing all of them as predictors directly. The existing methods of index construction require innovations, particularly by carefully considering the linkage mechanism between web search data and prediction targets.

5.2. Misc. transaction data

In addition to web search data, other transaction data, related to the operations of tourist webpage visiting, booking, purchasing, etc. are also attractive for tourism researchers. However, these data have been used far less in tourism research, because most of them are in the control of tourism organizations (such as hotels, travel agencies and attraction managers) and government sectors. The available related researches are as follows.

- *Webpage visiting data*: webpage visiting (or browsing) data can help understand the online browsing behavior of visitors, e.g., how they find the website and how they interact with it, thereby improving online marketing in terms of adjusting the site's content and design. For example, based on regression models, Plaza (2011) investigated how new visits and return visits to tourism websites affect pages per visit, and how direct visits, reference sites visits and search engines visits affect returns visits.
- *Online booking data*: Important information about online booking operations are recorded by hotel websites, and has proved to be useful for both hotel managers and investors. For example, based on a nested logit model, Saito, Takahashi, and Tsuda (2016) analyzed the choice behavior of visitors by using the online booking data of major four hotels near Kyoto station, which were collected from a Japanese booking website by National Institute of Informatics. Ghose, Ipeirotis, and Li (2012) used a dataset of US hotel reservations (sales price and volume) together with social media data to infer the economic impact of various location and service characteristics of hotels, via a random coefficient hybrid structural model.
- *Consumer Cards data*: Consumer cards data are captured when tourists make purchases, and can help corporations to study tourist purchasing behavior and design customized products. For instance, Weaver (2008) explored consumer cards data (including credit cards data, reward cards data and payment cards data) to understand consumption related behaviors and experiences within contained environments (such as casinos and cruise ships). Sobolevsky et al. (2014) demonstrated the applicability of bank card transactions in analyzing tourist mobility patterns and identifying tourist origins, based on a standard community detection approach.
- *Attractions sales data*: The ticket sales data of attractions can improve destination management. For instance, Shih, Nicholls, and Holecek (2009) estimated the influence of daily weather variations on daily ski lift ticket sales at two Michigan ski resorts, based on regression models.
- *Highway traffic data*: The highway traffic data can accurately capture spatial features of the visitors in a self-driving tour. For instance, based on the express highway data obtained from Jiangsu Provincial Communication Department, Huang, Cao, Jin, Yu, and Huang (2017b) measured the carbon emissions of self-driving tourism and the spatial relationship with scenic spots, via the network analysis and automatic classification modules of ArcGIS.

- *Hotel stays data*: As an important transaction data, hotel stays data are valuable for hospitality management. For instance, Falk (2010) investigated the relationship between the number of overnight stays and the snow depth at 28 Austrian ski resorts, via a panel data analysis.
- *Hotel consumption data*: Hotel chains provide an access to the big data of individual hotels' monthly electricity and water consumption. By analyzing the hotel electricity consumption data, Kahn and Liu (2016) revealed inefficiency in hotel energy use.

According to the above existing studies, transaction data have been introduced to tourism research, and appeared their respective advantages. However, such informative big data failed to make a great contribution in terms of few related articles. The hidden reason might be that most transaction data are mainly in the control of tourism organizations and government sectors, and are difficult to obtain for privacy concerns. Under such a background, a reciprocal cooperation between academia and industries becomes a promising way to not only largely promote this newborn research (i.e., using transaction data in tourism research) but also effectively address practical problems in tourism industry in return.

6. Discussions

In the new age of big data, missive-scale big data in different types have been employed to enrich and promote tourism research. This paper might be the first attempt to present a comprehensive literature review on full-type big data in tourism research. Given that different big data types carry different information, address different tourism issues and require different analytic techniques, a systematical analysis is conducted for each data type from four main perspectives: research focus, data characteristics, analytic techniques, and major challenges and future directions. This survey allows a thorough understanding of such a sunrise research, i.e., applying big data to tourism research, and provides insightful perspectives into the future prospects.

6.1. Main findings

Generally speaking, the application of big data to tourism research was still at an early stage, in terms of a short history (just since 2007) and a small number of annual publications (30 at most). However, such a sunrise research is undergoing a rapid growth with a general upward trend in annual number of articles. Leading journals with the most related articles are *Tourism Management*, *Journal of Travel Research*, *International Journal of Hospitality Management* and *Tourism Geographies*.

A variety of big data have been applied to tourism research, making amazing employments and innovations. These big data originated from three primary sources: (1) users, producing UGC data (such as online textual data and online photo data); (2) devices, for device data (such as GPS data, mobile roaming data, Bluetooth data, RFID data, WIFI data and meteorological data); (3) operations, for transaction data (such as web search data, online booking data, webpage visiting data etc.).

Carrying different information, a different data type addresses different tourism issues (i.e., research focuses). Among them, UGC data were the dominant type in tourism research (accounting for approximately 47%), which have greatly served tourist sentiment analysis, tourist behavior analysis, tourism marketing and tourism recommendation. The device data (36%) were still at a starting application stage to tourism research, but has appeared a significant superiority in investigating tourist spatial-temporal behavior. In contrast, the tourism research using transaction data was relatively few (17%). The main reason lies in the difficulty in accessing such private data which are mainly in the control of tourism organizations or government sectors.

For each big data type, an in-depth systematical review is conducted (see Sections 3–5), from the perspectives of research focuses, data characteristics and analytic techniques, along with major challenges and further directions. Through such a thorough analysis, we found that each type of data has its own advantages and disadvantages, and it is, therefore, applicable to specific tourism research fields. Accordingly, a comparative analysis across different types of big data in tourism research is conducted, as the results listed in Table 6.

On the one hand, the research focuses of different types of big data were highly dependent on the data characteristics (especially advantages). For example, online textual data (conveying tourist sentiment) was helpful in analyzing tourist satisfaction about tourism products or destinations. Web search data (directly reflecting public attention toward tourism markets, corresponding to potential demands) have improved tourism demand prediction and online marketing. As weather is an important factor in tourism, meteorological data (recording weather factors) have been employed to estimate the effects of weather on tourism, and to give the related tourism recommendations. Transaction data, especially those recording tourists' online operations, have mainly been applied to online tourism marketing.

It is worth noticing that although some types of big data have been used in the same issue, each of them has its own distinctive analysis perspectives. For example, online photo data (covering geo-information) and various device data were all specialized in modelling tourist spatial-temporal behavior (see Table 6). However, they behaved differently in tourism research due to their different data characteristics. For example, online photo data and mobile roaming data were only applicable to the macro level due to the accuracy limitation; in comparison, other tracking data with higher precision (such as GPS data, RFID data, Bluetooth data and WIFI data) can be used to model the tourist movement at the micro level.

On the other hand, regarding disadvantages, the use of big data in tourism research was still confronted with some tricky challenges concerning data quality, data cost and privacy concerns, as follows.

- *Data quality*: Although sufficient in data volume, the big data in tourism research are usually suppressed by quality problems. For example, reliability concerns of online textual data cannot be avoided as some visitors might somehow give fake reviews (Schuckert et al., 2015b). The location information of photo data and mobile roaming data is at a lower level of accuracy than that of GPS data. Web search data of Google trends might be biased due to data sampling and approximation methods (Pan et al., 2012).
- *Data cost*: The cost of data collection became another block to the tourism research with big data, especially for device data. To obtain the device data, a high cost has to be spent not only for purchasing sensor devices (e.g., GPS loggers and Bluetooth sensors) but also for recruiting volunteers. Moreover, the tracking range are often restricted due to fund scarcity. In contrast, at a respectively low cost, UGC data and web search data have accounted for approximately 58% of the application of big data in tourism research.
- *Privacy concerns*: Privacy concerns severely hamper tourism stakeholders (e.g., tourists, online travel agencies, hotels, government sectors and attraction organizations) to share their private data. Actually, privacy concerns are a common challenge in the age of big data, in which the sensitive information concerning customer movements and trade secret

Table 6
Comparison among different types of big data in tourism research.

Data source	Data type	Research focuses	Advantages	Disadvantages
Users	Online textual data	Tourist sentiment analysis; Tourism recommendation	Low cost; Conveying tourist sentiment	Reliability concerns

	Online photo data	Tourist behavior analysis;
		Tourism marketing;
		Tourism recommendation
			Low cost;
			Containing multi-metadata
				Low precision of location;

Devices	GPS data	Data feasibility in tourism; Tourist spatial-temporal behavior analysis; Tourism recommendation	Global; High precision	High cost; Privacy concerns
	Mobile roaming data		Availability in less visited places; Allowing tracking tourist origins	Privacy concerns; Low precision of location
	Bluetooth data		Availability in crowed indoor scenes; Allowing unannounced tracking	Small-range coverage; Privacy concerns

	RFID data		Low cost;
			High precision;
			Availability in crowed indoor scenes
				Small coverage range;
				Privacy concerns

	WIFI data		Availability in crowed indoor scenes; Allowing unannounced tracking	Small coverage range; Privacy concerns
	Meteorological data	Effect estimation of weather on tourism; Tourism recommendation	Containing weather factors	Difficulty in processing multi-format data
Operations	Web search data	Tourism demand prediction; Search engine optimization	Low cost; Reflecting public attention	Possible estimation biases
	Other transaction data	Tourist behavior analysis; Tourism marketing	Recording tourists' operations in tourism markets	Privacy concerns

are well kept secret. For instance, the use of most device data (including mobile roaming data, Bluetooth data, RFID data, WIFI data) is prevented due to privacy concerns from revealing tourist movements. Transaction data (such as consumer cards data), which reveal trading secret of tourists, are also difficult to obtain.

To effectively overcome the above challenges regarding data availability, a reciprocal cooperation between academia and industries becomes an extremely promising way. This method will not only largely guarantee data availability and reduce data cost of the tourism research using big data, but also effectively address practical problems in the tourism industry in return. Furthermore, signing confidentiality agreements and/or excluding sensitive information might be another practical way to solve the tough problem of privacy concerns.

6.2. Future directions

Even with amazing improvements and innovations, there still existed ample room to develop the tourism research using big data, especially from the perspectives of expanding research area and developing analytic techniques.

First, some other valuable big data and interesting tourism issues can be also considered to enrich the tourism research using big data. Among big data, device data (except for GPS data) and transaction data had made a relatively small contribution to tourism research, mainly due to the tricky challenges of high cost and privacy concerns (as discussed previously). Therefore, the application of these big data to tourism research can be largely improved by strengthening the corporation between researchers and tourism industries. Other even more valuable types of big data, such as audio data, video data, cross-domain data, multi-type data, etc., have not been introduced to tourism research yet. However, these big data contain a rich mine of distinctive information, and can also be utilized to provide new perspectives of understanding tourist behavior, tourism management and tourist market. For example, with the boom in smart tourism, most scenic spots have been equipped with monitoring systems, which can generate a large volume of video data. These interesting data directly record tourist behavior thereby helpfully serving the corresponding tourism management. In addition to the big data in tourism, peripheral data or cross-domain data such as health, insurance and education data are also valuable for tourism research, because these external factors may influence tourist preference significantly. The majority of existing related studies focused on a single type of big data; however, it might be somewhat insufficient to capture characteristics of the complex tourism system. Therefore, multi-type data is strongly recommended to reveal other potential interesting results for tourism research.

Hot issues using big data in tourism research were tourism demand prediction, tourist sentiment analysis, tourist behavior analysis and tourism recommendation. However, some other important issues, such as tourism precaution, tourism online marketing, scene spots programing, tourism product design and tourism carrying capacity estimation, can be also well addressed by using the valuable big data. In particular, tourism precaution and tourism emergency deserve a deep and extensive research. For example, various extreme events, including natural disasters (e.g., the earthquake in Jiuzhaigou Valley of China in August, 2017), public health emergencies (e.g., SARS in China in 2003) and political events (e.g., the deployment of terminal high altitude area defense (THAAD) system in Korea in 2016), have not only largely impacted the tourism markets, but also highlighted the importance of response policies. Therefore, the related big data, i.e., online news and online opinions regarding these extreme events, are quite useful for tourism precaution and tourism emergency. Moreover, other than these traditional aspects of tourism research, online marketing and data driven marketing using more intelligent approaches are really interesting and with practical implications.

Regarding analytic techniques, given that big data are commonly in a massive-scale, unconstructed (or semi-structured) and complex form, an appropriate analytic technique is essential for guaranteeing the effectivity and validity of tourism research with big data. However, in such a newborn research, traditional methods (e.g., statistical analysis and econometric models) still made up a considerable proportions of big data analysis techniques. Therefore, powerful techniques particularly for big data are strongly recommended in the future research to diversify the existing methods. Promising examples are trajectory indexing, retrieval and outlier detection for device data (Zheng, 2015), speech analysis for audio data, video content analysis (including server-based and edge-based architectures) for video data (Gandomi & Haider, 2015), and hybrid techniques for multi-type and/or multi-characteristic data (Tang, Yu, Liu, & Xu, 2013). Moreover, advanced analytical method such as machine learning and deep learning integrated with large-scale data are really promising for tourism research.

Acknowledgments

The authors would like to thank Bob McKercher, Chris Ryan and the three anonymous reviewers for helpful comments to earlier versions of the article. This work is supported by grants from the National Science Fund for Outstanding Young Scholars (NSFC No. 71622011), the National Natural Science Foundation of China (NSFC No. 71301006), and the construction project of Collaborative Innovation Center of e-Tourism of Beijing (822139917160102).

References

Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management, 28*(3), 898–910.

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management, 29*(3), 469–486.

Artola, C., Pinto, F., & de Pedraza García, P. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower, 36*(1), 103–116.

Ayscue, E. P., Boley, B. B., & Mertzlufft, C. E. (2016). Mobile technology & resident attitude research. *Tourism Management, 52*, 559–562.

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.

Bao, J., Chen, G., & Ma, L. (2014). Tourism research in China: Insights from insiders. *Annals of Tourism Research, 45*, 167–181.

Bauder, M., & Freytag, T. (2015). Visitor mobility in the city and the effects of travel preparation. *Tourism Geographies, 17*(5), 682–700.

Beeco, J. A., Huang, W. J., Hallo, J. C., Norman, W. C., McGehee, N. G., McGee, J., et al. (2013). GPS tracking of travel routes of wanderers and planners. *Tourism Geographies, 15*(3), 551–573.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. *Journal of Hospitality Marketing & Management, 25*(1), 1–24.

Bhati, A., & Pearce, P. (2016). Vandalism and tourism settings: An integrative review. *Tourism Management, 57*, 91–105.

Birenboim, A., Anton-Clavé, S., Russo, A. P., & Shoval, N. (2013). Temporal activity patterns of theme park visitors. *Tourism Geographies, 15*(4), 601–619.

Birenboim, A., Reinau, K. H., Shoval, N., & Harder, H. (2015). High-resolution measurement and analysis of visitor experiences in time and space: The case of Aalborg Zoo in Denmark. *The Professional Geographer, 67*(4), 620–629.

Bonné, B., Barzan, A., Quax, P., & Lamotte, W. (2013). WiFiPi: Involuntary tracking of visitors at mass events. In *2013 IEEE 14th international symposium on a world of wireless*. Madrid, Spain: Mobile and Multimedia Networks (WoWMoM).

Bordogna, G., Frigerio, L., Cuzzocrea, A., & Psaila, G. (2016). Clustering geo-tagged tweets for advanced big data analytics. In *2016 IEEE international congress on big data, San Francisco, USA*.

Boswell, P. (2011). *Google analytics: Measuring content use and engagement*. Society for Technical Communication.

Brovelli, M. A., Minghini, M., & Zamboni, G. (2016). Public participation in GIS via mobile applications. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 306–315.

Chen, W. C., Battestini, A., Gelfand, N., & Setlur, V. (2009). Visual summaries of popular landmarks from community photo collections. In *2009 conference record of the 43th asilomar conference on signals*. Pacific Grove, USA: Systems and Computers.

Cheng, M., & Edwards, D. (2015). Social media in tourism: A visual analytic approach. *Current Issues in Tourism, 18*(11), 1080–1087.

Chhetri, P., Corcoran, J., & Arrowsmith, C. (2010). Investigating the temporal dynamics of tourist movement: An application of circular statistics. *Tourism Analysis, 15*(1), 71–88.

Chilipirea, C., Petre, A. C., Dobre, C., & van Steen, M. (2016). Presumably simple: Monitoring crowds using WiFi. In *2016 IEEE 17th international conference on Mobile Data Management (MDM), Porto, Portugal*.

Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *The Economic Record, 88*(1), 2–9.

Chua, A., Servillo, L., Marcheggiani, E., & Moere, A. V. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management, 57*, 295–310.

Crotts, J. C., Mason, P. R., & Davis, B. (2009). Measuring guest satisfaction and competitive position in the hospitality and tourism industry: An application of stance-shift analysis to travel blog narratives. *Journal of Travel Research, 48*(2), 139–151.

Da Rugna, J., Chareyron, G., & Branchet, B. (2012). Tourist behavior analysis through geotagged photographies: A method to identify the country of origin. In *2012 IEEE 13th international symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary*.

De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., & Yu, C. (2010). Automatic construction of travel itineraries using social breadcrumbs. In *Proceedings of the 21st ACM conference on hypertext and hypermedia, Toronto, Canada*.

Deka, M. J., Joshi, J., Sinha, N., Tyagi, A., Kushal, A., & Jain, A. (2016). Indoor and outdoor position identification using RFID. In *International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India*.

Delafontaine, M., Versichele, M., Neutens, T., & Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography, 34*, 659–668.

Deng, N., & Li, X. R. (2018). Feeling a destination through the "right" photos: A machine learning model for DMOs' photo selection. *Tourism Management, 65*, 267–278.

East, D., Osborne, P., Kemp, S., & Woodfine, T. (2017). Combining GPS & survey data improves understanding of visitor behaviour. *Tourism Management, 61*, 307–320.

Edwards, D., & Griffin, T. (2013). Understanding tourists' spatial behaviour: GPS tracking as an aid to sustainable destination management. *Journal of Sustainable Tourism, 21*(4), 580–595.

Falk, M. (2010). A dynamic panel data analysis of snow depth and winter tourism. *Tourism Management, 31*(6), 912–924.

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management, 52*, 498–506.

Ferrer, G., Dew, N., & Apte, U. (2010). When is RFID right for your service? *International Journal of Production Economics, 124*(2), 414–425.

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.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iView*, 1–12.

Gawlik, E., Kabaria, H., & Kaur, S. (2011). *Predicting tourism trends with google insights*. Retrieved from http://cs229.stanford.edu/proj2011/GawlikKaurKabariaPredictingTourismTrendsWithGoogleInsights.pdf.

Getz, D., & Page, S. J. (2016). Progress and prospects for event tourism research. *Tourism Management, 52*, 593–631.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science, 31*(3), 493–520.

Godnov, U., & Redek, T. (2016). Application of text mining in tourism: Case of Croatia. *Annals of Tourism Research, 58*, 162–166.

Goh, C., & Law, R. (2011). The methodological progress of tourism demand forecasting: A review of related literature. *Journal of Travel & Tourism Marketing, 28*(3), 296–317.

Grinberger, A. Y., Shoval, N., & McKercher, B. (2014). Typologies of tourists' time–space consumption: A new approach using GPS data and GIS tools. *Tourism Geographies, 16*(1), 105–123.

Gunter, U., & Önder, I. (2016). Forecasting city arrivals with Google analytics. *Annals of Tourism Research, 61*, 199–212.

Guo, X. (2016). Application of meteorological big data. In *In 2016 16th International Symposium on Communications and Information Technologies (ISCIT), Qingdao, China*.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichl et al location. *Tourism Management, 59*, 467–483.

Hallo, J. C., Beeco, J. A., Goetcheus, C., McGee, J., McGehee, N. G., & Norman, W. C. (2012). GPS as a method for assessing spatial and temporal use distributions of nature-based tourists. *Journal of Travel Research, 51*(5), 591–606.

Hardy, A., Hyslop, S., Booth, K., Robards, B., Aryal, J., Gretzel, U., et al. (2017). Tracking tourists' travel with smartphone-based GPS technology: A methodological discussion. *Information Technology & Tourism, 17*(3), 255–274.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115.

Hozak, K. (2012). RFID applications in tourism. *International Journal of Leisure and Tourism Marketing, 3*(1), 92–108.

Huang, Z., Cao, F., Jin, C., Yu, Z., & Huang, R. (2017b). Carbon emission flow from self-driving tours and its spatial relationship with scenic spots–A traffic-related big data method. *Journal of Cleaner Production, 142*, 946–955.

Huang, S. S., & Chen, G. (2016). Current state of tourism research in China. *Tourism Management Perspectives, 20*, 10–18.

Huang, X., & Wu, B. (2012). Intra-attraction tourist spatial-temporal behavior patterns. *Tourism Geographies, 14*(4), 625–645.

Huang, X., Zhang, L., & Ding, Y. (2017a). The Baidu Index: Uses in predicting tourism flows–A case study of the Forbidden City. *Tourism Management, 58*, 301–306.

Hu, Y. H., Chen, Y. L., & Chou, H. L. (2017). Opinion mining from online hotel reviews–A text summarization approach. *Information Processing & Management, 53*(2), 436–449.

Hunter, W. C. (2013). China's chairman Mao: A visual analysis of Hunan province online destination image. *Tourism Management, 34*, 101–111.

Joo, H. H., Kang, H. G., & Moon, J. J. (2014). The effect of rain on the decision to visit a theme park. *Asia Pacific Journal of Tourism Research, 19*(1), 61–85.

Kahn, M. E., & Liu, P. (2016). Utilizing "Big Data" to improve the hotel sector's energy efficiency: Lessons from recent economics research. *Cornell Hospitality Quarterly, 57*(2), 202–210.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing, 74*(7), 2561–2573.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., & Rattenbury, T. (2007). How flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of the 15th ACM international conference on multimedia, Augsburg, Germany*.

Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., & Andrienko, G. (2010). Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections. In *2010 14th international conference on Information Visualisation (IV), London, UK*.

Költringer, C., & Dickinger, A. (2015). Analyzing destination branding and image from online sources: A web content mining approach. *Journal of Business Research, 68*(9), 1836–1843.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications, 40*(10), 4065–4074.

Kurashima, T., Iwata, T., Irie, G., & Fujimura, K. (2013). Travel route recommendation using geotagged photos. *Knowledge and Information Systems, 37*(1), 37–60.

Kuusik, A., Tiru, M., Ahas, R., & Varblane, U. (2011). Innovation in destination marketing: The use of passive mobile positioning for the segmentation of repeat visitors in Estonia. *Baltic Journal of Management, 6*(3), 378–399.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70.

Leask, A. (2016). Visitor attraction management: A critical review of research 2009–2014. *Tourism Management, 57*, 334–361.

Lee, I., Cai, G., & Lee, K. (2014). Exploration of geo-tagged photos through data mining approaches. *Expert Systems with Applications, 41*(2), 397–405.

Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management, 46*, 311–321.

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

Liu, Y., Teichert, T., Rossi, M., Li, H., & Hu, F. (2017). Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. *Tourism Management, 59*, 554–563.

Li, X., Wu, Q., Peng, G., & LV B. (2016). Tourism forecasting by search engine data with noise-processing. *African Journal of Business Management, 10*(6), 114.

Lucia, M. D. (2013). Economic performance measurement systems for event planning and investment decision making. *Tourism Management, 34*, 91–100.

Lu, W., & Stepchenkova, S. (2012). Ecotourism experiences reported online: Classification of satisfaction attributes. *Tourism Management, 33*(3), 702–712.

Lu, X., Wang, C., Yang, J. M., Pang, Y., & Zhang, L. (2010). Photo2trip: Generating travel routes from geo-tagged photos for trip planning. In *Proceedings of the 18th ACM international conference on multimedia, Firenze, Italy*.

Lu, D., Wu, R., & Sang, J. (2017). Overlapped user-based comparative study on photo-sharing websites. *Information Sciences, 376*, 54–70.

Majid, A., Chen, L., Chen, G., Mirza, H. T., Hussain, I., & Woodward, J. (2013). A context-aware personalized travel recommendation system based on geotagged social media data mining. *International Journal of Geographical Information Science, 27*(4), 662–684.

Ma, J., Luo, S., Yao, J., Cheng, S., & Chen, X. (2016). Efficient Opinion summarization on comments with Online-LDA. *International Journal of Computers, Communications & Control, 11*(3), 414–427.

Mamei, M., Rosi, A., & Zambonelli, F. (2010). Automatic analysis of geotagged photos for intelligent tourist services. In *2010 6th international conference on Intelligent Environments (IE), Kuala Lumpur, Malaysia*.

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.

McKercher, B., Shoval, N., Ng, E., & Birenboim, A. (2012). First and repeat visitor behaviour: GPS tracking and GIS analysis in Hong Kong. *Tourism Geographies, 14*(1), 147–161.

Melián-González, S., Bulchand-Gidumal, J., & González López-Valcárcel, B. (2013). Online customer reviews of hotels: As participation increases, better evaluation is obtained. *Cornell Hospitality Quarterly, 54*(3), 274–283.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information & Management, 54*(6), 771–785.

Nieto, J., Hernández-Maestro, R. M., & Muñoz-Gallego, P. A. (2014). Marketing decisions, customer reviews, and business performance: The use of the Toprural website by Spanish rural lodging establishments. *Tourism Management, 45*, 115–123.

Nilbe, K., Ahas, R., & Silm, S. (2014). Evaluating the travel distances of events visitors and regular visitors using mobile positioning data: The case of Estonia. *Journal of Urban Technology, 21*(2), 91–107.

Oku, K., Hattori, F., & Kawagoe, K. (2015). Tweet-mapping method for tourist spots based on now-tweets and spot-photos. *Procedia Computer Science, 60*, 1318–1327.

Okuyama, K., & Yanai, K. (2013). A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web. In *The Era of interactive media* (pp. 657–670). New York: Springer.

Önder, I. (2017). Classifying multi-destination trips in Austria with big data. *Tourism Management Perspectives, 21*, 54–58.

Orellana, D., Bregt, A. K., Ligtenberg, A., & Wachowicz, M. (2012). Exploring visitor movement patterns in natural recreational areas. *Tourism Management, 33*(3), 672–682.

Öztayşi, B., Baysan, S., & Akpinar, F. (2009). Radio frequency identification (RFID) in hospitality. *Technovation, 29*(9), 618–624.

Pan, B., Chenguang Wu, D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology, 3*(3), 196–210.

Pantelidis, I. S. (2010). Electronic meal experience: A content analysis of online restaurant comments. *Cornell Hospitality Quarterly, 51*(4), 483–491.

Pan, B., Xiang, Z., Law, R., & Fesenmaier, D. R. (2011). The dynamics of search engine marketing for tourist destinations. *Journal of Travel Research, 50*(4), 365–377.

Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using google trends data. *Journal of Travel & Tourism Marketing, 34*(3), 357–368.

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research, 50*, 67–83.

Pearce, P. L., & Wu, M. Y. (2015). Entertaining international tourists: An empirical study of an iconic site in China. *Journal of Hospitality & Tourism Research*. https://doi.org/10.1177/1096348015598202.

Peng, G., Liu, Y., Wang, J., & Gu, J. (2017). Analysis of the prediction capability of web search data based on the HE-TDC method–prediction of the volume of daily tourism visitors. *Journal of Systems Science and Systems Engineering, 26*(2), 163–182.

Philander, K., & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management, 55*(2016), 16–24.

Phillips, P., Zigan, K., Silva, M. M. S., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management, 50*, 130–141.

Plaza, B. (2011). Google Analytics for measuring website performance. *Tourism Management, 32*(3), 477–481.

Pomfret, G., & Bramwell, B. (2016). The characteristics and motivational decisions of outdoor adventure tourists: A review and analysis. *Current Issues in Tourism, 19*(14), 1447–1478.

Popescu, A., & Grefenstette, G. (2009). Deducing trip related information from flickr. In *Proceedings of the 18th international conference on world wide web, Madrid, Spain*.

Popescu, A., Grefenstette, G., & Moëllic, P. A. (2009). Mining tourist information from user-supplied collections. In *Proceedings of the 18th ACM conference on information and knowledge management, Paris, France*.

Quack, T., Leibe, B., & Van Gool, L. (2008). World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 international conference on content-based image and video retrieval, Niagara falls, Canada*.

Racherla, P., & Friske, W. (2012). Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications, 11*(6), 548–559.

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies, 75*, 197–211.

Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management, 57*, 202–212.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management, 57*, 12–20.

Roh, J. J., Kunnathur, A., & Tarafdar, M. (2009). Classification of RFID adoption: An expected benefits approach. *Information & Management, 46*(6), 357–363.

Saito, T., Takahashi, A., & Tsuda, H. (2016). Optimal room charge and expected sales under discrete choice models with limited capacity. *International Journal of Hospitality Management, 57*, 116–131.

Schuckert, M., Liu, X., & Law, R. (2015a). A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently. *International Journal of Hospitality Management, 48*, 143–149.

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

Shih, C., Nicholls, S., & Holecek, D. F. (2009). Impact of weather on downhill ski lift ticket sales. *Journal of Travel Research, 47*(3), 359–372.

Shi, Y., Serdyukov, P., Hanjalic, A., & Larson, M. (2011). Personalized landmark recommendation based on geotags from photo sharing sites. In *Proceedings of the 5th international AAAI conference on weblogs and social media, Barcelona, Spain*.

Shi, Y., Serdyukov, P., Hanjalic, A., & Larson, M. (2013). Nontrivial landmark recommendation using geotagged photos. *ACM Transactions on Intelligent Systems and Technology (TIST), 4*(3), 47.

Shoval, N., & Ahas, R. (2016). The use of tracking technologies in tourism research: The first decade. *Tourism Geographies, 18*(5), 587–606.

Shoval, N., & Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research, 34*(1), 141–159.

Shoval, N., McKercher, B., Birenboim, A., & Ng, E. (2015). The application of a sequence alignment method to the creation of typologies of tourist activity in time and space. *Environment and Planning B: Planning and Design, 42*(1), 76–94.

Shoval, N., McKercher, B., Ng, E., & Birenboim, A. (2011). Hotel location and tourist activity in cities. *Annals of Tourism Research, 38*(4), 1594–1612.

Sobolevsky, S., Sitko, I., Des Combes, R. T., Hawelka, B., Arias, J. M., et al. (2014). Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in Spain. In *2014 IEEE international congress on big data, Anchorage, USA*.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—a review of recent research. *Tourism Management, 29*(2), 203–220.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., & Andrienko, N. (2011). Analytical workflow of monitoring human mobility in big event settings using Bluetooth. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on indoor spatial awareness, Chicago, Illinois*.

Stepchenkova, S., & Zhan, F. (2013). Visual destination images of Peru: Comparative content analysis of DMO and user-generated photography. *Tourism Management, 36*, 590–601.

Sun, Y., Wei, Y., & Zhang, L. (2017). International academic impact of Chinese tourism research: A review based on the analysis of SSCI tourism articles from 2001 to 2012. *Tourism Management, 58*, 245–252.

Tang, L., Yu, L., Liu, F., & Xu, W. (2013). An integrated data characteristic testing scheme for complex time series data exploration. *International Journal of Information Technology and Decision Making, 12*(3), 491–521.

Tchetchik, A., Fleischer, A., & Shoval, N. (2009). Segmentation of visitors to a heritage site using high-resolution time-space data. *Journal of Travel Research, 48*(2), 216–229.

Thelwall, M. (2001). A web crawler design for data mining. *Journal of Information Science, 27*(5), 319–325.

Tiru, M., Kuusik, A., Lamp, M. L., & Ahas, R. (2010). LBS in marketing and tourism management: Measuring destination loyalty with mobile positioning data. *Journal of Location Based Services, 4*(2), 120–140.

Tsai, C. Y., Chang, H. T., & Kuo, R. J. (2017). An ant colony based optimization for RFID reader deployment in theme parks under service level consideration. *Tourism Management, 58*, 1–14.

Tsai, C. Y., & Chung, S. H. (2012). A personalized route recommendation service for theme parks using RFID information and tourist behavior. *Decision Support Systems, 52*(2), 514–527.

Van der Spek, S., Van Schaick, J., De Bois, P., & De Haan, R. (2009). Sensing human activity: GPS tracking. *Sensors, 9*(4), 3033–3055.

Vaughan, L., & Chen, Y. (2015). Data mining from web search queries: A comparison of google trends and Baidu index. *Journal of the Association for Information Science and Technology, 66*(1), 13–22.

Versichele, M., De Groote, L., Bouuaert, M. C., Neutens, T., Moerman, I., & Van de Weghe, N. (2014). Pattern mining in tourist attraction visits through association rule learning on bluetooth tracking data: A case study of Ghent, Belgium. *Tourism Management, 44*, 67–81.

Versichele, M., Neutens, T., Delafontaine, M., & Van de Weghe, N. (2012). The use of bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography, 32*(2), 208–220.

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.

Wan, Z. (2009). Personalized tourism information system in mobile commerce. In *Proceedings of the international conference on management of e-commerce and e-government, Nanchang, China*.

Wearing, S., & McGehee, N. G. (2013). Volunteer tourism: A review. *Tourism Management, 38*, 120–130.

Weaver, A. (2008). When tourists become data: Consumption, surveillance and commerce. *Current Issues in Tourism, 11*(1), 1–23.

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., & Pan, B. (2011). Travel queries on cities in the United States: Implications for search engine marketing for tourist destinations. *Tourism Management, 32*(1), 88–97.

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.

Xu, X., & Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International Journal of Hospitality Management, 55*, 57–69.

Xu, H., Yuan, H., Ma, B., & Qian, Y. (2015). Where to go and what to play: Towards summarizing popular information from massive tourism blogs. *Journal of Information Science, 41*(6), 830–854.

Yang, E. C. L., Khoo-Lattimore, C., & Arcodia, C. (2017). A systematic literature review of risk and gender research in tourism. *Tourism Management, 58*, 89–100.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397.

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., Li, H., Wang, Z., & Law, R. (2014). The influence of hotel price on perceived service quality and value in e-tourism: An empirical investigation based on online traveler reviews. *Journal of Hospitality & Tourism Research, 38*(1), 23–39.

Yoon, H., Zheng, Y., Xie, X., & Woo, W. (2010). Smart itinerary recommendation based on user-generated GPS trajectories. In *Proceedings of the 7th international conference on ubiquitous intelligence and computing, Xi'an, China*.

Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardin, F., Carrascal, J. P., Blat, J., et al. (2014). An analysis of visitors' behavior in the Louvre museum: A study using bluetooth data. *Environment and Planning B: Planning and Design, 41*(6), 1113–1131.

Yuan, H., Xu, H., Qian, Y., & Li, Y. (2016). Make your travel smarter: Summarizing urban tourism information from massive blog data. *International Journal of Information Management, 36*(6), 1306–1319.

Zakrisson, I., & Zillinger, M. (2012). Emotions in motion: Tourist experiences in time and space. *Current Issues in Tourism, 15*(6), 505–523.

Zeni, N., Kiyavitskaya, N., Barbera, S., Oztaysi, B., & Mich, L. (2009). RFID-based action tracking for measuring the impact of cultural events on tourism. In *Proceedings of the international conference on information and communication Technologies in Tourism, Amsterdam, The Netherlands*.

Zhang, Y., & Cole, S. T. (2016). Dimensions of lodging guest satisfaction among guests with mobility challenges: A mixed-method analysis of web-based texts. *Tourism Management, 53*, 13–27.

Zhang, L., Lan, C., Qi, F., & Wu, P. (2017). Development pattern, classification and evaluation of the tourism academic community in China in the last ten years: From the perspective of big data of articles of tourism academic journals. *Tourism Management, 58*, 235–244.

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management, 29*(4), 694–700.

Zhang, Z., Zhang, Z., & Yang, Y. (2016). The power of expert identity: How website-recognized expert reviews influence travelers' online rating behavior. *Tourism Management, 55*, 15–24.

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology (TIST), 6*(3), 29.

Zheng, W., Huang, X., & Li, Y. (2017). Understanding the tourist mobility using GPS: Where is the next place? *Tourism Management, 59*, 267–280.

Zhou, X., Xu, C., & Kimmons, B. (2015). Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Computers, Environment and Urban Systems, 54*, 144–153.

Jingjing Li received the M.S. degree from School of Economics and Management, Beijing University of Chemical Technology (BUCT), Beijing, China, in 2016. Currently, she is a Ph.D student in management science and engineering at School of Economics and Management, Beihang University, Beijing, China. Her research interests include big data mining and tourism management. She has published two journal papers on big data mining in *Energy Economics* and *International Journal of Global Energy Issues*.

Lizhi Xu received the Ph.D. degree in management science and engineering from School of Economics and Management, Beihang University (BUAA), Beijing, China, in 2012. Currently, she is a researcher at Collaborative Innovation Center of eTourism, Beijing Union University, Beijing, China. Her research interests are tourism economics, tourism management, and tourism forecasting. She has published more than 10 papers in journals including *Journal of Systems Science and Complexity*, *East Asian Economic Perspectives*, *Journal of Systems Science and Information*, etc. In addition, she has published more than 10 books.

Ling Tang received the Ph.D. degree in management science and engineering from Institute of Policy and Management, Chinese Academy of Sciences (CAS), Beijing, China, in 2012. Currently, she is a Professor at School of Economics and Management, Beihang University (BUAA), Beijing, China. Her research interests include big data mining, modeling and forecasting, artificial intelligence, and tourism management. She has published more than 40 papers in journals including *International Journal of Forecasting*, *Journal of Forecasting*, *IEEE Transactions on Knowledge and Data Engineering*, *Neural Computing & Applications*, *Chaos Solitons & Fractals*, *Annals of Operations Research*, *Computers & Operations Research*, *Computers & Industrial Engineering*, etc.

Shouyang Wang, Ph.D., is a Professor at Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS), Beijing, China. His research interests are big data mining, modeling and forecasting, artificial intelligence, and tourism management. He has published more than 340 papers in leading journals including *International Economic Review*, *Journal of Banking & Finance, IEEE Transactions on Engineering Management*, *International Journal of Forecasting*, etc. He is also editor-in-chief, associate editor, or editorial board member of 15 international journals.

Ling Li received the M.S. degree from School of Economics and Management, Beijing University of Chemical Technology (BUCT), Beijing, China, in 2015. She is currently pursuing the Ph.D. degree in management science and engineering at School of Economics and Management, Beihang University (BUAA), Beijing, China. Her research interests include big data mining and tourism management. She has published two journal papers in *Journal of Cleaner Production* and *Online Information Review*.