Big Data and Tourism, *The SAGE International Encyclopedia of Travel & Tourism*

Saskia Cousin
Gael Chareyron
Sébastien Jacquot

Introduction

For about 10 years now, drawing on their travel experiences, tourists have been using dedicated networks and social media sites to post comments, photographs and reviews. A decline in the frequency of slideshows and travelogs perhaps, but such photographs and comments emerge from tens of millions of users and today amount to hundreds of millions of elements of data available online. The size, the complexity and the heterogeneous nature of this body of data present a challenge for analysts and traditional statistical and business analysis tools are no longer relevant. There is a need to turn to methods that developed from the early year of the millennium which dealt with huge datasets (massive data) or "big data": cloud computing, data mining, graph analysis, knowledge discovery, new databases, opinion mining, etc. Whether the objective be scientific or commercial, such work is about engaging with and studying digital traces connected to tourism, its practices and changing nature across space and through time.

Part 1: From Web 2.0 to digital trails

The emergence of the term e-tourism is premised on the notion that the Internet plays an increasingly important role in tourism, not only as a valuable tool for promotion and distribution but also as an essential support for the journey itself. Emerging at the start of the new millennium, the Web 2.0 transformed e-tourism by allowing Internet users to become the producers of content itself (user-generated content) as in Wikipedia, Internet forums, blogs, etc. Pioneers in this movement, the earliest sites specialising in user generated evaluations appeared during the 1990s, for example Amazon in 1996, followed in 2000 by Cityvox and L'internaute (specialising in restaurants) and, above all, Trip Advisor which brought together assessments by tourists of a variety of activities. This was followed by applications for the sharing of photos and videos: Flickr (created in 2004, bought and developed by Yahoo in 2005), Panoramio (created in 2005 in Spain, bought by Google in 2007), then Instagram (2010, bought by Facebook in 2012). New applications were continually developed. These sites which allow for the sharing of both content and every day assessments have built up an exceptional body of photographs, videos and comments which form a particular electronic evidential corpus of knowledge on touristic visits to a particular territory, to hotels, restaurants or sites. In the case of digital social networks, voluntarily submitted evidence such as this forms digital trails which can be identified, provided they are associated with the appropriate metadata, in other words the information accompanying a photograph, comment or assessment: GPS coordinates of the places photographed or commented upon, the precise date, the profile of the user (gender, place of origin, language,…), tags and folksonomies, etc. These partial and heterogeneous digital trails need to be analysed and completed in order to produce knowledge about touristic practices, movements and densities. Other trails, left

in an involuntary fashion, can also be used. Neither public nor visible, they are connected to the use of a digital and electronic profile which automatically produces evidence: mobile phones, payments by bank card, transport networks, travel cards or season tickets (the Amsterdam Pass for example).

The growing number of active users of Flickr, Instagram and Tripadvisor show the huge dimension of the practices of on line sharing. To this must be added the far larger number of non-active Internet users, who read the opinions or look at photographs before choosing a destination, a hotel or restaurant. Such vast amounts of data turn the relationship of supply and demand in tourism upside down: from this perspective it is tourists themselves, a group of peers, who define the activities, while at the same time traditional experts lose their exclusivity, even their legitimacy to recommend sites, restaurants and destinations. The importance of the web 2.0 and social networks becomes quite evident when the results of search engines relating to requests for information about touristic destinations are consulted. This massive amount of online data constitutes a challenge for the tourist industry at a variety of different levels. Firstly, they are used to inform research into the e-reputation of tourist businesses, hotels in particular. For businesses as with destinations, it is no longer enough to keep an eye on and manage the negative effects of social networks (negative comments), but to use them for communication. Aware of the prescriptive and performative character of people's opinions and of the importance of the digital word of mouth (eWOM), tourist operators, notably the DMO (Destination Management Organisations) for example, encourage their tourists to submit photographs and hashtags on social networks. Thus the city of Amsterdam installed playful street furniture picking up on tourist marketing slogans in order to generate a significant flow of photographs by tourists on social networks. Montréal drew on Instagram users by encouraging them to spread the word about their experiences with a unique hashtag (MTLMOMENTS) and reproduced these images on their website.

Part 2: Why should we talk about Big Data and its applications?

If the social and marketing uses of information emerging from big data is multiplying, these enormous datasets are also contributing to restructuring the cartographic and quantitative research characteristic of tourism studies. From a scientific perspective, in order to be considered as big or massive, the data must have 5 characteristics (the 5 Vs): volume, variety, variability, velocity (the fact that new data is continually appearing) and veracity. The data which emerges from social networks relevant to tourism perfectly matches these characteristics. In terms of volume, from the beginning of 2015 there have been more than 200 million recommendations and opinions on TripAdvisor relating to 57,000,000 members. By the end of 2014 at least 213 million photographs have been geo-tagged on Flickr, relating to 2,000,000 users, while in 2013 Flickr announced that they had more than 3 1/2 billion photographs in total. On Instagram at least 700 million photographs were geo-tagged across the 53 million users. Besides this, such data is being continually augmented (new photographs posted, new comments), and they show great variety (photos, texts, notes, etc), requiring methods specific to big data in order for them to be structured and analysed.

In this context, big data constitutes a fundamental challenge for tourism studies: on one hand it renews cartographic and quantitative approaches to tourism; on the other, it provides access to new knowledge. In fact, digital data produced by tourists provides unprecedented information on densities, movements and mobilities among tourists, their timetables, most photographed sites, relationships between sites and touristic experiences. However, while several recent works do draw on social networking sites, they are largely satisfied with a comparatively superficial analysis which does not make the most of the extensive body of data, only selecting, for example, the first 100 comments on TripAdvisor, or a random sample of photographs. The scientific challenge is the development of methods to automate the treatment of data and metadata in a variety of directions.

In the first place, such research aims to overcome the technical and methodological difficulties inherent in the treatment of big data. It is necessary to develop ways of collecting such complex data using a variety of techniques of *crawling*. Once established, one of the greatest difficulties in the analysis of digital trails is the incomplete character of the data. For example, we never have access to a complete picture of the tourist's route but simply a collection of fragments which we combine to produce usable information: tourists don't post everything they do, there are gaps in their profiles, no information on their place of origin, etc. When a user's profile has very little information, a question which might appear basic, for example knowing whether we are dealing with a tourist or local inhabitant might necessitate a set of very complex protocols in order to treat the data in an automised fashion. In order to respond to such a problem, with the aim of enhancing the profiles, it is necessary to implement learning algorithms on a probabalistic basis.

What type of knowledge and what applications could thus be developed? The geo-localised metadata associated with photographs or comments allows for the identification of zones of density in relation to the number of posts, and thus the production of maps of density at whatever scale is relevant: from the distribution of photographs in the Château de Versailles Gardens to comments posted on TripAdvisor regarding San Francisco or the whole of the world. The cross referencing of geo-localised data with other metadata may also provide an automatic classification of touristic spaces. For example, the geo-localised data from Twitter differentiates urban zones according to their uses, distinguishing between business zones, leisure/weekend, nightlife and residential, and makes apparent popular/touristic zones in Manhattan, London and Madrid.

Such analyses take a more dynamic turn when they characterise the journeys and itineraries of tourists. So, the combination of certain metadata (geo-localisation and chronogram) by user makes apparent the nodes (each site photographed or commented upon) of a touristic route. The automatic aggregation of all of these many journeys builds a map of touristic movements within a single destination. Such results are stimulating for the knowledge of touristic practices, notably because of the variety of scales such studies illuminte coherent spaces of touristic practice, frequently different from the administrative limits used by the DMO. Beyond this, for the same territory, such

cartography allows for a differentiation between mainstream circuits and more alternative routes. Finally, it allows questions to be asked about the degree, the capacity and the modalities of movement away from a tourist site by locating tourists who have moved away. For example, for World Heritage sites, we could map out the routes taken by tourists around Angkor and the degree of distancing which they permitted themselves.

Based on tracking technologies, the data emerging from mobile telephones offers similar perspectives in regards to the characterisation of densities and routes. However, such data is not directly accessible to researchers, something which requires agreements with mobile telephone companies. Moreover, the use of this data raises ethical questions as the traces which have been left are involuntary.

Following the localisation of densities and the identification of routes, research is currently directed to variables concerning tourists: the examination of the metadata of the users of TripAdvisor allows the characterisation of practice profiles of a destination according to nationality, gender, and even according to age groups. Cluster analysis results in tourist typologies.

Finally, comments, blogs and forums are just as much sources for experiential accounts worthy of study as they are a body of data. Extending the studies of guest books and travel diaries, research into tourist comments online is also frequently limited to a restricted data set, using textual analysis software, for example for studying the touristic experience of 149 anglophones tourists in the silk market of Beijing who had posted on TripAdvisor. Also, based on a huge body of texts, techniques of *opinion mining*, well developed over the past few years, permit the characterisation of emotions and opinions etc. These find their use in the touristic domain through the analysis of on line comments by tourists, on Twitter, TripAdvisor, travel forums, etc.

Part 3: The challenge of the link between the data and tourism

The consideration and analysis of digital trails through big data makes apparent several challenges which concern the tourism sector. It poses new ethical questions and an examination of the purpose of these new practices which themselves depend on and generate masses of data.

The applications of big data open a completely new field of study which is drawing the attention of tourist operators. Tourism research institutes are being renewed through taking account of the information coming out of big data. Thus several DMOs are establishing partnerships with mobile telephone operators to track the coming and going of French and foreign visitors to France (Tourism Office and the Congres de Paris, the French regions, etc.). At the same time the development of mobile applications serves to generate data which is likely to improve knowledge of the practices and itineraries of visitors.

The great mass of data analysed gives the impression of comprehensiveness; the cartography permitted by their extraction is often seductive. The interpretation and the presentation of results are however fragile in a number of ways.

Firstly, the group of tourists posting photographs and comments on social networks are neither the same, nor representative of the total tourist population. Big data needs to be backed up by a sociology and geography of the usage of social networks: not all social classes, age groups, nationalities, etc, practice digital sharing of experiences in the same fashion. Not all of them do this on the same networks. For example, studies carried out amongst Hong-Kong residents show that those who share information on their touristic breaks are younger, more educated and more autonomous during their trips. The challenge is thus to achieve a consideration of the results of big data that takes into account more classical research (statistics linked to accommodation, research based on questionnaires) which, of course, has its own bias.

The second precaution concerns the very content of the body of texts and images. Tourists choose photographs which they would like to share or the sites upon which they comment. Touristic big data can thus not give access to the totality of touristic practice or experience but to a selected element. This opens up new research perspectives, notably ethnographic, in order to understand and analyse those elements of shared experience and those which rest within a private or restricted sphere, off-line.

Finally, acute questions of ethics and deontology are at the heart of this sort of research. The issue of the respect of private life is raised in a crucial fashion in the context of the extraction of involuntary traces emerging from electronic devices such as mobile phones. Assuming anonymity is sufficiently well respected should such data be used for an uniquely commercial objective or made available to the research community (the open data thematic).? Digital traces coming from social networks are voluntary and thus can be interpreted as constituting public space on the web.

However, this public space is controlled, organised by algorithms deployed by businesses which gather the data. Research into big data also implies a decrypting of logic which underpins the production and presentation of data by tourists. To avoid the effects of the imposition of "systems of calculation "and the construction of confidence, the preferred research route consists in addressing, on the one hand, the way in which Internet users appropriated the platforms, and on the other, on the content of accounts or images and not their aggregation. For example, by abandoning the score attributed to the site by TripAdvisor in order to concentrate on comments, their styles or their evaluative registers.

Cross-References

See also “Media and Tourism”, “Social networking”, “Travel Writers”, “Tourism Statistics”, “Technologies issues in Travel and Tourism”, “Quantitative Tourism Research”

Further readings

Borel, Simon. “Les liaisons numériques. Dangereuses ou vertueuses ?” *Revue du MAUSS*, 2011/2, 38, 349-368, 2011.
Buhalis, Dimitrios and Robert Law. “Progress in information technology and tourism management: 20 years on and 10 years after the Internet - The state of eTourism research.” *Tourism management*, 29, 609-623, 2008.
Cardon, Vincent. “Des chiffres et des lettres, Evaluation, expressions du jugement de qualité et hiérarchies sur le marché de l’hôtellerie”, *Réseaux*, *Evaluations profanes*, 2014/1, 183, pp. 207-245, 2014.
Cousin, Saskia, Gael Chareyron, Jérôme Da-Rugna and Sébastien Jacquot. "Studying TripAdvisor. Or how to Trip-patrol holidays maps", *EspacesTemps.net*, 2015.
Chareyron Gael, Jérome Da Rugna and Bérangère Branchet. “Mining tourist routes using Flickr traces”. *ASONAM* 2013: 1488-1489, 2013.
Donaire José, Raquel Camprubi, and Nuria Gali. “Tourist clusters from Flickr travel photography”, *Tourism Management Perspectives*, 11, 26-33, 2014.
Frias-Martinez Vanessa and Enrique Frias-Martinez. “Spectral clustering for sensing urban land use using Twitter activity”, *Engineering Applications of Artificial Intelligence*, Volume 35, October 2014, 237–245, 2014.
Jeacle Ingrid and Chris Carter. “In TripAdvisor we trust : Rankings, calculative regimes and abstract systems”, *Accounting, organizations and society*, 36 : 293-309, 2011.
Law, Robert. “Internet and Tourism, Part XXI”, *Journal of Travel & Tourism Marketing*, 20:1, 75-77, 2008.
Leung Daniel, Rob Law, Hubert van Hoof and Dimitrios Buhalis. “Social Media in Tourism and Hospitality: A Literature Review.” *Journal of Travel & Tourism Marketing*, 30:1-2, 3-22, 2013.
Lee, Hee A., Rob Law and Jamie Murphy. “Helpful Reviewers in TripAdvisor, an Online Travel Community”, *Journal of Travel & Tourism Marketing*, 28:7, 675-688, 2011.
Liu Bing, *Opinion mining and sentiment analysis*, Morgan & Claypool Publishers, 180 p., 2012.
Marrese-Taylor, Edison, Juan Velasquez and Felipe Bravo-Marquez. “A novel deterministic approach for aspect-based opinion mining in tourism products reviews”, *Expert systems with applications*, 41, 7765-7775, 2014.
Munar Ana Maria, and Jens Steen Jacobsen. “Motivations for sharing tourism through social media”, *Tourism management*, 43, 46-54, 2014.
Shoval, Noam and Michael Isaacson. *Tourist mobility and advanced tracking technologies*, Routledge, 2010.
Van Laere, Olivier, Steven Schockaert and Bart Dhoedt. “Georeferencing Flickr resources based on textual meta-data”, *Information Sciences*, 238, 52-74, 2013.
Wu, Mao-Ying, Geoffrey Wall and Philip Pearce. “Shopping experiences : International tourists in Beijing Silk’s market”, *Tourism Management*, 41, 96-106, 2014.
Xiang Zheng and Ulrike Gretzel. “Role of social media in online travel information search”, *Tourism managemen*t, 31, 179-188, 2010.