Contents lists available at ScienceDirect

Journal of Destination Marketing & Management

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

Research paper

Tourism analytics with massive user-generated content: A case study of Barcelona

Estela Marine-Roig [a],*, Salvador Anton Clavé [b]

[a] Department of Business Administration and Economic Management of Natural Resources (AEGERN), University of Lleida (UDL), C/ Jaume II, 73, Campus de Cappont, 25001 Lleida, Catalonia, Spain
[b] Rovira i Virgili University, Catalonia, Spain

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form
17 June 2015
Accepted 18 June 2015
Available online 29 July 2015

Keywords:
User-generated content
Smart city
Smart tourism destination
Big data
Business intelligence
Catalonia

ABSTRACT

The aim of this paper is to highlight the usefulness of big data analytics to support smart destinations by studying the online image of Barcelona (a leading smart city and tourist destination) as transmitted via social media through the analysis of more than 100,000 relevant travel blogs and online travel reviews (OTRs) written in English by tourists who have visited the city in the last 10 years. The proposed methodology used in this paper facilitates the massive gathering, cleaning up, and analysis of tourism-related user-generated content (UGC) from the most suitable sources, and helps to define the transmitted image of the city through collecting and processing large volumes of digital data. It is also used to extract business intelligence (BI) from OTRs concerning visits to Barcelona's main landmark/attraction, La Sagrada Familia. The findings of this massive content analysis of information from a trustworthy source, UGC data, is very useful in appling BI to destination management, both in order to develop and assess marketing strategies and to improve branding and positioning policies among tourism and marketing organizations. It reinforces the ability of cities such as Barcelona to develop a smart city and destination concept, as well as a strategy for themselves.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

The concept of 'big data' refers to the massive accumulation of information and to the systems that manipulate these large datasets. Gandomi and Haider (2015) highlight the three Vs (volume, variety, and velocity) that characterize big data, and claim that traditional data-management systems are insufficient to manage it, giving rise to big data technologies capable of creating real-time intelligence from high volumes of various data. In this respect, Sanz (2013) asserts that cities with an appropriate operating system can store, analyse, and generate near-real-time business intelligence (BI) with big data collected from social media feeds, among other sources.

The spectacular growth of social media and user-generated content (UGC) on the Internet provides a huge quantity of information that allows for the firsthand ascertaining of the experiences, opinions, and feelings of tourism 'users' or customers (Marine-Roig & Anton Clavé, 2015; Xiang, Schwartz, & Uysal, 2015). The volume of data generated in social media has grown from terabytes to petabytes (Gandomi & Haider, 2015; He & Chen, 2014), and data stored and analysed by big companies are set to move from the petabyte to exabyte magnitude soon (Hu, Wen, Chua, & Li, 2014). Social media can be classified into blogs, review sites, media sharing, question-and-answer sites, social bookmarking, social networking, social news, and wikis (Gandomi & Haider, 2015; Marine-Roig, 2014). Lu and Stepchenkova (2015) found that the main sources for studies on UGC data are, in order of frequency: consumer review websites, blogs, media-sharing websites, social networks, and virtual communities; the main topic areas being service quality, destination image and reputation, UGC as electronic word-of-mouth (eWOM), experiences and behaviour, and mobility patterns. In recent research, Koltringer and Dickinger (2015) have found that UGC is the richest and most diverse source of online information.

In the field of tourism, most authors agree on the importance of UGC (Koltringer & Dickinger, 2015; Lu & Stepchenkova, 2015; Marine-Roig, 2015) in the construction of destination image through the eWOM effect (Hidalgo, Sicilia, & Ruiz, 2014; Jalilvand, Samiei, Dini, & Manzari, 2012), and consider travel blogs, online travel reviews (OTRs), or online consumer reviews as rich sources of UGC data (Marine-Roig, 2014; Xiang et al., 2015). In the field of tourism and hospitality, a relative decrease in travel blogs can be observed, along side a tremendous growth in OTRs, especially in

* Corresponding author.
E-mail addresses: estela.marine@aegern.udl.cat (E. Marine-Roig), salvador.anton@urv.cat (S. Anton Clavé).

http://dx.doi.org/10.1016/j.jdmm.2015.06.004
2212-571X/© 2015 Elsevier Ltd. All rights reserved.

the hospitality sector (Marine-Roig & Anton Clavé, 2015). For instance, in January 2015 TripAdvisor asserted that it had reached more than 200 million reviews and opinions, Trivago had reached 140 million integrated user hotel reviews, Booking had collected 43 million verified reviews, and Expedia had gathered 11 million customer reviews. Given such figures, UGC should be identified as a valuable source of big data that is useful for the management of smart cities and smart tourism destinations.

This paper aims to highlight the usefulness of big data analytics to support smart tourism destinations by studying the online social media-transmitted image of Barcelona (a leading smart city and tourist destination) through the analysis of more than 100,000 relevant travel blogs and OTRs written in English by tourists who have visited the city in the last 10 years. To do so, it proposes a method of gathering and analysing big UGC data composed of five stages: destination choice, Web hosting selection, data collection, pre-processing, and content analysis. More specifically, a quantitative content analysis was conducted of 117,487 travel blogs and OTRs. This method is also used to extract BI from the 7481 OTRs on visits in 2014 to Barcelona's main landmark, La Sagrada Familia. The findings of this massive content analysis of information from a trustworthy source, UGC data, is of paramount usefulness in terms of applying BI to destination management, not only to develop marketing strategies but also improve branding and positioning policies among tourism and marketing organizations. It reinforces the ability of cities such as Barcelona to develop a smart city and destination concept, as well as a strategy for themselves.

2. State of the art

According to Del Chiappa and Baggio (2015), the concept of a smart tourism destination is still emerging and is arising from that of the smart city. A smart city is a city that performs in a forward-looking way in regard to six characteristics (economy, people, governance, mobility, environment, and living), and is built on the smart combination of endowments and activities of self-decisive, independent, and aware citizens (Giffinger et al., 2007). Boes, Buhalis, and Inversini (2015) indicate that a smart city focuses on its citizens, while a smart destination intends to improve tourist experiences through information and communication technologies (ICTs). They build a framework for the dimensions of the smart tourism destination that requires fundamental constructs (leadership, human capital, entrepreneurs, innovation, and social capital) supported by technology applications and a strong ICT infrastructure. This in turn, provides the basis to support the components of tourism (tourism experience, tourism competitiveness, and the six As of tourism: attractions, accessibility, amenities, available packages, activities, and ancillary services) and smart cities.

More specifically, Del Chiappa and Baggio (2015) define a smart tourism destination as a networked system of stakeholders delivering services to tourists, complemented by a technological infrastructure aimed at creating a digital environment that supports cooperation, knowledge sharing, and open innovation. In this vein, Buhalis and Amaranggana (2014) consider that 'smartness', when referring to a tourism destination, requires the dynamic interconnection of stakeholders through a platform capable of exchanging real-time information related to tourism activities, with the objective of maximizing user or customer satisfaction and resource management efficiency. These activities produce a large multidimensional set of digital information, which is understood within the concept of big data, and allows national tourism organizations (NTOs) and destination marketing organizations (DMOs) to extract valuable insights.

According to Wang, Li, and Li (2013), the use of big data by smart tourism destinations can support business decision-making and optimal resource allocation, and can assist in the discovery of new insights in ways that affect markets and organizations. In their chapter on strengths for the tourism industry of using big data, Oliver et al. (2014) describe a set of advantages over traditional methodologies offered by analysing large amounts of data reliability, representativeness, information detail and segmentation capacity, the ability to 'hybridize' data with other current or future sources, new information flows, and the possibility of new business opportunities.

Hu et al. (2014) propose a definition for big data analytics based on the software, hardware, and aim of analysis: 'Big data analytics is the process of using analysis algorithms running on powerful supporting platforms to uncover potentials concealed in big data, such as hidden patterns or unknown correlations' (p. 656). They categorize the analysis process into two alternative paradigms: streaming and batch. Streaming processing is characterized by data being analysed as it arrives because near-real-time results are needed, such as in the case of online applications, and only a small part of the stream is temporarily stored in the memory. Conversely, batch processing is characterized by data first being stored, and then divided into chunks that are processed in parallel in a distributed system; finally, the intermediate results are aggregated.

Gandomi and Haider (2015) focus on big data analytical processes in two phases: data management (acquisition and recording; extraction, cleaning, and annotation; and integration, aggregation, and representation) and analytics (modelling and analysis, and interpretation). In the field of tourism destinations, Fuchs, Hopken, and Lexhagen (2014) propose a knowledge destination framework architecture that distinguishes between a knowledge creation layer (data sources, data extraction, data warehousing, and knowledge generation through data mining) and a knowledge application layer, where there is a destination management information system that grants stakeholders instant access to BI-based analysis results.

Lu and Stepchenkova (2015) observe that a growing number of UGC analytical works use specialized analytical and computational procedures to manage big data. However, in most studies, UGC data are manually collected; the manual handling of data is time-consuming, limits sample size, and facilitates researcher bias. These authors also found that, in general, methods to gather information are unclear, and that the technical details of data analyses are often incomplete. As an example of manually handling UGC data and limited sample size, in addition to the 122 cases gathered by Lu and Stepchenkova (2015) – He and Chen (2014), located 333 blog posts with Google Blog Search, filtered them manually, which left them with 317 relevant blog posts, using the so-called Blog Mining method, they then merge the posts in text files. Analyse these posts using the *Leximancer* programme allowed them to identify seven relevant themes.

Koltringer and Dickinger (2015) extracted destination brand identity and image from 5719 online documents through Web content mining and natural language processing. Their method was divided into the following stages: data gathering, keyword analysis, sentiment detection, category building, and correspondence analysis. As an example of the large-scale analysis of consumer-generated content, Xiang et al. (2015), analysed 60,648 online customer reviews from Expedia, corresponding to 10,537 hotels from the 100 largest cities in the United States. Using a web crawler, they gathered all available textual content for each city and each hotel, and then created a relational database with unique identifiers assigned to each hotel property, review, and unique word. Then, after a pre-processing phase, they analysed guest experiences.

In this respect, the massive analysis of UGC data is of great interest when identifying the characteristics of the transmitted image of the destination and its assets by online users. UGC has proved to be a great valuable data source for the study of destination reputation, branding, and image, as well as for analysing tourists' perceptions, experiences, and behaviour (Koltringer & Dickinger, 2015; Lu & Stepchenkova, 2015). Furthermore, this massive UGC analysis is of special relevance in studying the image that will be transmitted to other tourists through the eWOM effect (Hidalgo et al., 2014; Jalilvand et al., 2012; Koltringer & Dickinger, 2015). This image, strongly related to destination attractiveness, is considered to be highly influential for tourists' and stakeholders' decision making (Schmallegger & Carson, 2010; Yoo & Gretzel, 2010) as well as highly trustworthy (Leung, Law, Van Hoof, & Buhalis, 2013), because it is conveyed by tourists' peers and not driven by an economic interest.

There is plenty more, however, that can be done to maximize the usefulness of travel blogs and reviews as sources of information for businesses, DMOs, and academics (Pan, MacLaurin, & Crotts, 2007) by analysing big UGC data through data analysis systematization and computerization (Marine-Roig & Anton Clavé, 2015). Here, we argue that massive analysis of UGC about a destination and its attractions is of great importance for DMOs in order to operationalize and make use of the huge amounts of information provided by online users It should become a key tool for BI in a smart destination, as an added value or specificity of a smart city, in order to provide a high-quality tourist experience.

3. Background of the case study

Barcelona is the capital city of Catalonia, Which has its own tourist brand (Catalan Tourist Board, 2015) (See Fig. 1). As can be observed on the map, the sub-regional areas around Barcelona also adopt Barcelona as their main identity brand: 'Barcelona Coast' and 'Barcelona Landscapes'. According to Datzira-Masip and Poluzzi (2014), these areas surrounding Barcelona seek to benefit from the great national and international dissemination of the Barcelona brand, and share the strategic approach of the 'Greater Barcelona' brand. These brands concerning the surrounding areas of Barcelona are of interest, as they provide a complementary offer to the Barcelona central brand and generate tourism flows in relation to the city itself. In fact, in contrast to the findings of Liu (2014), in the case of Catalonia the Barcelona brand outshines the Catalonia (nation) brand and all other surrounding cities and brands. When in Catalonia, most foreign tourists only perceive the image of Barcelona.

As a tourist city, Barcelona has become one of the most-recognized international tourist destinations during the last 25 years (Datzira-Masip & Poluzzi, 2014). Tourist inflows have grown

Table 1
Foreign tourists to the city of Barcelona by country of origin.
Source: Authors, from http://professional.barcelonaturisme.com.

Year	🇩🇪	🇫🇷	🇮🇹	🇯🇵	🇬🇧	🇺🇸	World
	de	fr	it	jp	uk	us	total
2005	294,156	362,038	475,175	113,137	712,763	429,920	5,656,848
2006	355,586	449,515	610,535	134,184	764,846	483,061	6,709,175
2007	377,033	456,508	623,058	140,679	787,057	529,609	7,108,393
2008	345,596	454,381	547,609	135,506	675,040	468,395	6,659,075
2009	334,335	501,284	541,521	138,534	523,281	478,775	6,476,033
2010	361,358	567,287	563,666	151,236	531,952	549,137	7,133,524
2011	397,285	593,842	559,621	156,989	529,356	606,781	7,390,777
2012	414,539	572,259	491,103	162,887	592,713	635,386	7,440,113
2013	453,102	636,903	447,721	170,092	629,969	627,412	7,571,766

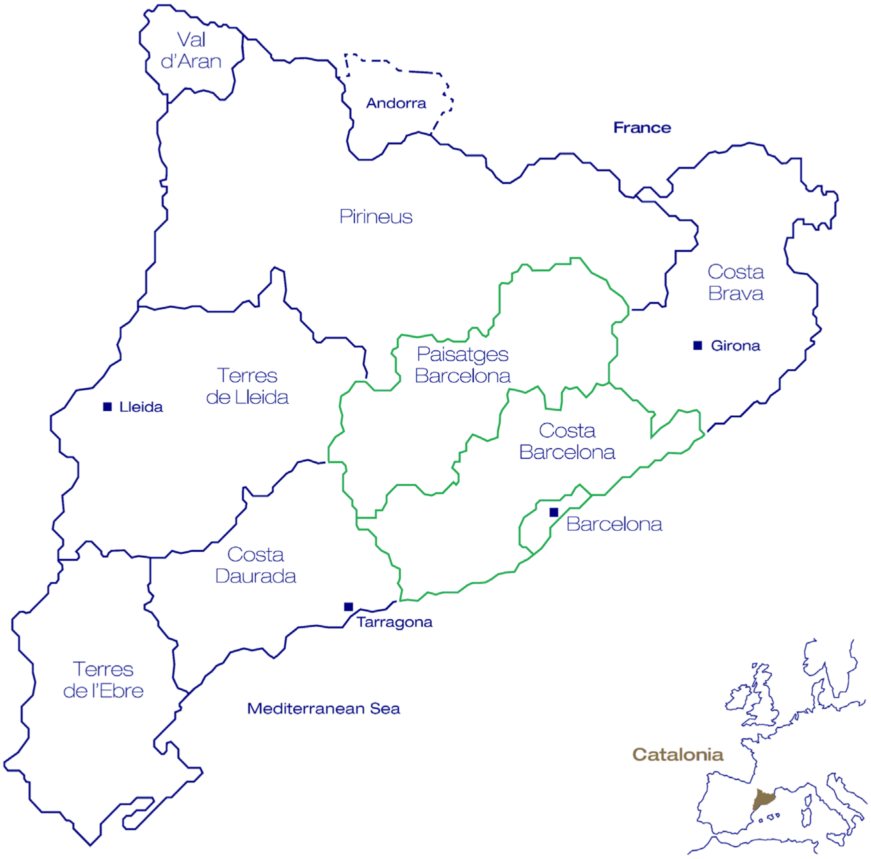


Fig. 1. Tourist brands of Catalonia (CTB, 2015).

continuously since the 1992 Olympic Games, except for the years following the 2008 global financial crisis (Table 1). Barcelona ranks fourth in the 2014 Top 10 European Cities for total number of bednights by international tourists after London, Paris, and Rome, with a growth of 4.6% in 2013 (Baudot, 2015). Moreover, Barcelona is the sixth most powerful city brand in the world (Michael, 2014). It is Europe's leading cruise port and the fourth most important cruise port in the world (Barcelona Turisme, 2014). Almost half the inhabitants and tourists in Catalonia are concentrated in the Barcelona area. Barcelona is the only city in the world with nine UNESCO World Heritage Site (WHS) buildings (Barcelona Turisme, 2014). One of them, the Basilica de La Sagrada Familia, receives over 3 million visitors per year.

Barcelona is also a leading smart city, and was chosen in 2011 as the GSMA Mobile World Capital from 2012 to 2018, and was granted European Capital of Innovation ('iCapital') status by the European Commission (2014) for introducing the use of new technologies to bring the city closer to its citizens. Barcelona also received the Bloomberg Philanthropies 2014 Mayors Challenge Grand Prize for Innovation and €5 million towards its plan to create a digital and community 'trust network' for each of its at-risk elderly residents. Barcelona won the City Climate Leadership Award 2014 in the category of Intelligent City Infrastructure for a new ICT architecture that provides a single platform interconnecting the entire city. Wakefield (2013) places Barcelona among the top 10 smart cities around the world. The European Parliament (2014) classifies Barcelona in the First Group of Smart Cities (cities with a large number of initiatives, each covering a variety of characteristics), together with Amsterdam and Helsinki. Cohen (2014) ranks Barcelona first in the annual Smartest Cities in the World ranking.

In fact, using the case of the Barcelona smart city model, Bakici, Almirall, and Wareham (2013) state that the main assets of a smart city should be grouped under four main topics: smart governance, smart economy, smart living, and smart people. This model's foundations lie on three pillars: ubiquitous infrastructures, information, and human capital. Open data and the analysis of open data are, in this vein, a principal component of the smart city strategy. Thus, the analysis of the transmitted image through social media should be understood as a transversal project for a given city in the sense in which Arup (2013) classifies transversal city projects in the case of Barcelona (telecommunications networks, urban platforms, and intelligent data); he also mentions the OpenData BCN portal. In fact, smart cities tend to open up data to the public and businesses in order to stimulate smart decisions between citizens, visitors, and stakeholders. Morabito (2015) highlights the case of OpenData BCN repository concerning big data and analytics issues, where the public and firms can access information such as population, public facilities, or the economy. Morabito (2015) concludes

> 'Barcelona's open-source, smart city platform, engages local talent in smart city development, ensures technology and provider independence and data stewardship and remains with the public, under its stewardship and safeguards civil liberties' (p. 38).

To use the case of Barcelona to test the feasibility of a UGC data analytics method for the perceived (and transmitted) image is not only convenient, but also coherent with the 'smart' nature and commitment of the city. This provides an interesting case on the potential usefulness of proposing the integration of UGC tourism data into the general big data repositories of the city in terms of tourism, business, and city intelligence. For all of the reasons above, Barcelona presents an excellent case to demonstrate how big data analytics of UGC can be useful to support smart tourism destinations.

4. Methodology

This paper introduces a complete UGC data analytics method suitable for adoption and adaptation by smart cities and destinations, such as the city of Barcelona, whose transmitted image through UGC is the content analysed. Once the destination is chosen, the proposed method (See Fig. 2) consists of four stages whose steps and constituent features are set out in detail: Web hosting selection (webpage structure and website webometrics), data collection (web structure mining, filter settings, and downloading), pre-processing (web content mining, language detection, finding the user's hometown, arranging, cleaning, and debugging), and quantitative content analysis (parser settings, and categorization).

This method is based on the batch-processing paradigm (data are first stored and then analysed). However, the relatively small quantity of data in this case study, about 250,000 pages (50 GB) before pre-processing, allows the downloaded data to be stored and processed in a few days by a single computer (i.e., there is no need to work in a distributed system).

4.1. Web hosting selection

The vast quantity of online information generated by users that this study intended to analyse requires a computerized process, and hence, there is a need for semi-structured sources. Nevertheless, given the impossibility of locating and manipulating all the existing UGC data on the Internet, a selection of the most suitable sources was required.

A. *Website structure*. Firstly, UGC sources that have a minimal

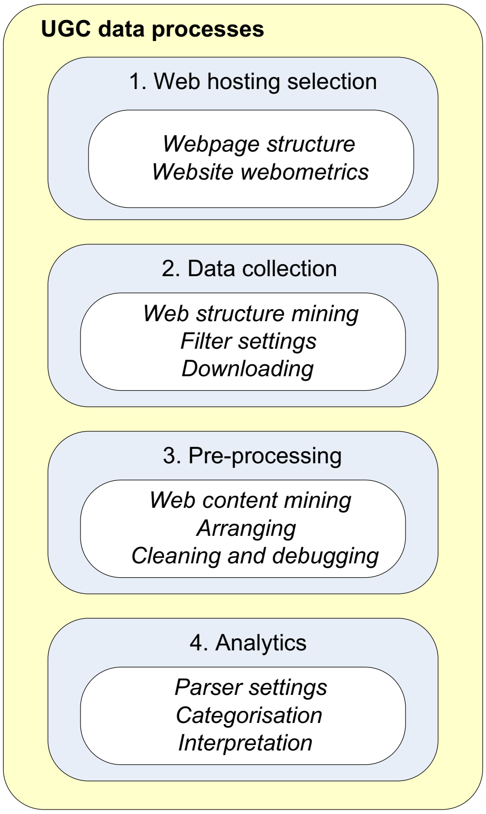


Fig. 2. Main stages of the proposed method.

structure allowing for the computerization of data collection and the systematic arrangement of information were selected. In the field of tourism, the main attributes are post-title, destination, and trip date. These pre conditions are fulfilled by most websites that host travel blogs and OTRs. It is also interesting to know the country of origin of the trip diary author and, in the case of OTRs, the topic to which they refer (things to do, attraction factor, hotel, restaurant, etc.)

B. *Website webometrics*. Secondly, website webometrics was conducted. Drawing on previous works and by using two search engines, with the query '*travel blog*' *OR* '*travel review*' on March 1, 2015, 15,800,000 results on BlogSearch.Google.com, and 1,880,000 results were obtained on Bing.com. The websites located in first position, providing they fulfilled the previously mentioned conditions and contained more than 100 entries related to the case study, were then selected. Websites and OTRs dedicated to the subject of accommodation and dining were excluded because of their great specialization, which makes them more suitable for analysing the quality of such services rather than studying destination image, as this research intends. Eleven websites were included in this first selection: GetJealous.com, MyTripJournal.com, StaTravel.com, TravBuddy.com, TravelBlog.org (TB), TravelJournals.net, TravellersPoint.com, TravelPod.com (TP), TripAdvisor.com (TA), Venere.com, and VirtualTourist.com (VT).

Numerous studies suggest that UGC, through the eWOM effect, considerably influences the formation of destination image (Hidalgo et al., 2014; Jalilvand et al., 2012; Koltringer & Dickinger, 2015; Marine-Roig, 2015). However, as a source of online information, it needs to be disseminated via the Internet to be able to influence potential tourists. It is evident that webpages containing UGC that are neither visible nor popular, nor contain information about a specific destination, cannot influence the construction of that destination's image. It is also interesting to ascertain the country of origin of the website audience in order to learn whether these coincide with the markets from which tourists come to the destination.

As a result, a ranking (Table 2) with the pre-selected websites is applied using a weighted formula: 'TBRH = 1*B(V) + 1*B(P) + 2*B(S)' (Marine-Roig, 2014), where 'B' is Borda's ordering method, 'V' is website visibility (quantity and quality of inbound links), 'P' their popularity (visitors, visits, and traffic in general), and 'S' the size (number of entries related to the case study).

Some authors base the selection of the most suitable websites for their case study only on Google PageRank (PR), but this is not a very suitable way to construct a ranking because of its low granularity (Marine-Roig, 2014). As can be observed in Table 2, TripAdvisor and VirtualTourist have the same PageRank (7), which is meaningless when all of the other rankings give a much higher position to TripAdvisor. This includes Google itself, who is responsible for PageRank and has far more indexed pages about TripAdvisor than VirtualTourist.

Table 2
Webometrics of the top four websites hosting travel diaries (as of March 1, 2015).

Webometric	Source	TA	TB	TP	VT
Indexed pages	Google.com	239,000,000	487,000	389,000	553,000
	Bing.com	3,030,000	49,900	50,800	72,400
Link-based rank	Google PR	7	6	6	7
	Yandex CY	1700	90	325	350
Visit-based rank	Compete.com	52	36,203	13,595	2488
	Quantcast.com	168	25,894	8474	4276
	Alexa.com	212	46,122	27,390	5644
Size	Entries	104,430	2589	1264	9204
TBRH	Rank	**1**	**3**	**4**	**2**

4.2. Data collection

Once a destination and the websites hosting travel blogs and OTRs had been selected, the selected webpages were located and downloaded.

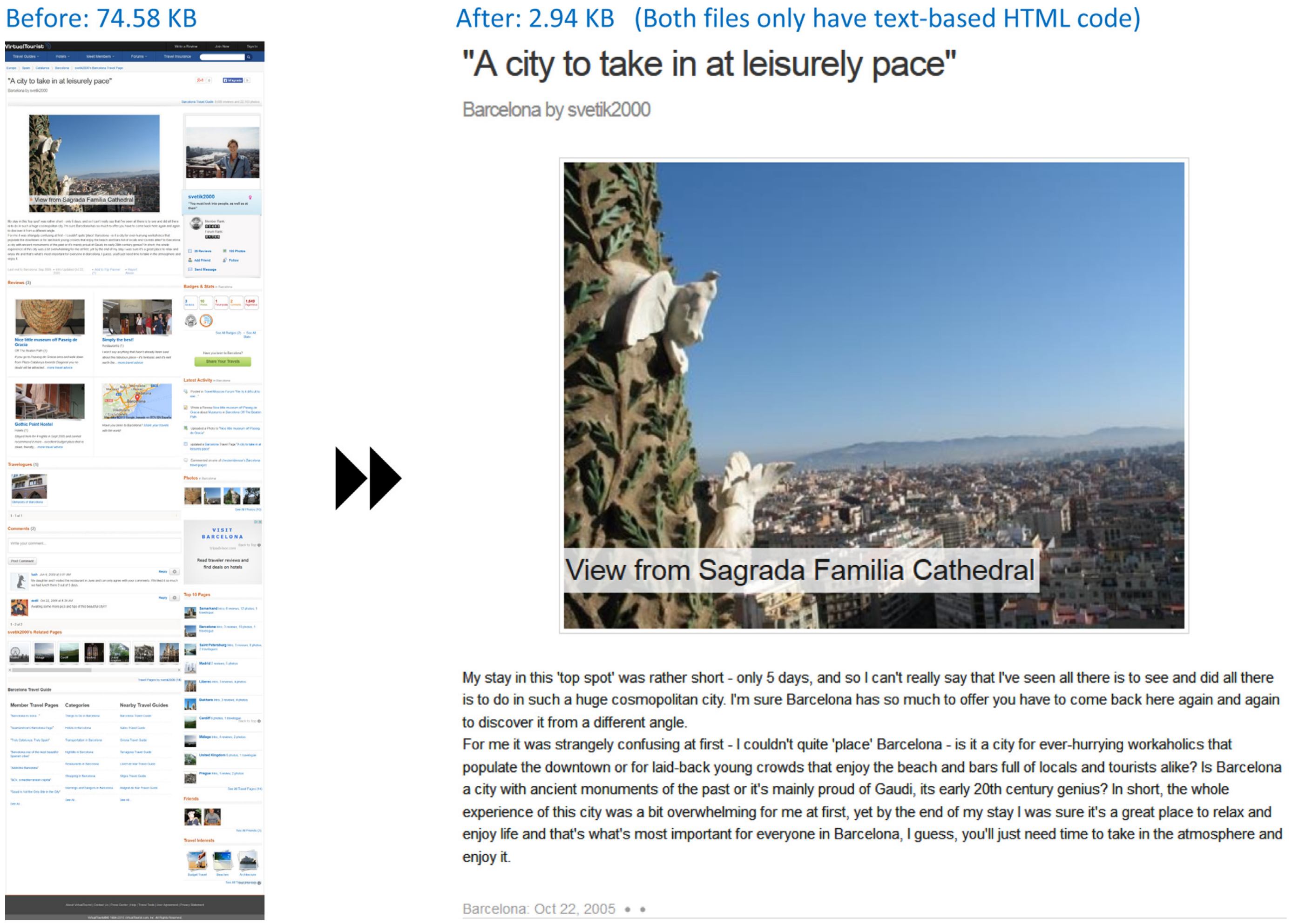
A. *Web structure mining*. To collect data, web structure mining was conducted. According to Liu (2011), web structure mining discovers practical knowledge from hyperlinks, which represent the structure of the Web, and from anchor text associated with the hyperlinks. In this research, the hyperlinks of the relevant webpages related to the case study (travel blogs, travel pages, travelogues, and travel reviews about Barcelona and other destinations in the surrounding area) were identified.

B. *Filter settings*. Once the information derived from the hyperlinks was obtained, the minimum possible quantity of files were downloaded, so as not to overload the webhost traffic and to save space on the local hard disk. For this purpose, filters were applied (level, file type, URL, and content filters). Such filters are indispensable when automating the downloading of UGC data:
   1. *Level filter*. Level value defines the depth of the search for HTML documents. In other words, level is the number of mouse clicks on hyperlinks that are necessary to get from the start page to the last page needed.
   2. *File type filter*. A filename generally consists of a name and an extension. A filename extension is a suffix (separated from the base filename by a dot) of the name applied to indicate the file format. This filter allows the download by filename extension to be restricted.
   3. *URL filter*. The uniform resource locator (URL) filter allows performing in any part of it: protocol (http, https, ftp, etc.), server, domain, subdirectories (folders), and filenames. It also allows acting in any segment of the URL using the included and excluded keywords.
   4. *Content filter*. Content filters allow the user to specify the keywords to search in all of the text of the downloaded pages; you can even search within HTML tags. Keywords can be separate words with space symbols and exact word sequences. It is possible to look for some or all of the keywords. Whereas with the filters previously seen you have to access the files on the server, the content filter has the disadvantage of having to check if the file contains the keywords you searched for before processing it.

C. *Data download*. To download the webpages related to the case study, we used a web copier was used, specifically the Offline Explorer Enterprise (OEE), because it has the capacity to download up to 100 million URLs per project. It also offers the fastest possible multi-threaded processing of downloaded files by using all CPU cores (MetaProducts, 2015).

4.3. Data pre-processing

Before the webpages in a collection can be used for retrieval, some pre-processing tasks need to be performed, such as identifying different text fields, anchor text, and main content blocks, as well as removing HTML tags (Liu, 2011). The purpose of this phase is to extract information from the downloaded webpages, which allows the local path to HTML files to be structured so they can be ordered and segmented into different criteria, such as destination, date, language, nationality, topic, etc. It also intends to reduce the webpages to the content block written by the user without losing the HTML format, and overcome encoding problems as well as



"A city to take in at leisurely pace"

Barcelona by svetik2000

View from Sagrada Familia Cathedral

My stay in this 'top spot' was rather short - only 5 days, and so I can't really say that I've seen all there is to see and did all there is to do in such a huge cosmopolitan city. I'm sure Barcelona has so much to offer you have to come back here again and again to discover it from a different angle.

For me it was strangely confusing at first - I couldn't quite 'place' Barcelona - is it a city for ever-hurrying workaholics that populate the downtown or for laid-back young crowds that enjoy the beach and bars full of locals and tourists alike? Is Barcelona a city with ancient monuments of the past or it's mainly proud of Gaudi, its early 20th century genius? In short, the whole experience of this city was a bit overwhelming for me at first, yet by the end of my stay I was sure it's a great place to relax and enjoy life and that's what's most important for everyone in Barcelona, I guess, you'll just need time to take in the atmosphere and enjoy it.

Barcelona: Oct 22, 2005 • •

Fig. 3. VT travel page before and after the cleaning stage.

most common mistakes.

A. *Web content mining*. First, data mining was conducted. Web content mining extracts useful information from webpage contents (Liu, 2011).
    1. *Finding user's hometown*. To be able to segment by nationalities, it is necessary to know the traveller's country of origin; but some bloggers' profiles, such as in TravelBlog, do not contain the hometown field. In other cases, the profile contains the town, city, region, or country of origin, but the interest here is to classify them by states. Given that all destinations have a travel guide with a geographically structured hyperlink, country extraction can be automatized.
    2. *Detecting language*. Specialized software is needed. In this research a Java programme based on the Nakatani Shuyo (shuyo.wordpress.com) *Language Detection Library* (LDL) which was used. This detects 53 languages, extendable through an included process. By means of this system, based on the Naive Bayes LDL classifier, each language is detected with a high degree of precision (probability higher than 99%). It is convenient to perform this analysis on textual content without HTML tags after the cleaning stage, because what the user has actually written represents just a minimal part of the webpage content (Fig. 3).

B. *Data arranging*. Next was data arranging – a key step in the process. The download process stored files on the local hard disk, following a structure parallel to that in the webserver. To identify and classify travel blogs and OTRs, file structure was arranged using the following pattern of folders and files: \host\brand\town\date_lang_isFrom_pageId_topic.htm
The backslashes of the pattern represent subdirectories or folders, the hyphens unite the composite words, and the underscores separate the different words or numbers from the filename. Each item also follows a pattern to facilitate the manipulation of files:
    1. *host*. Two-letter acronym that represents the hosting website (e.g. TB: TravelBlog.org, VT: VirtualTourist.com).
    2. *brand*. Five-letter abbreviation (e.g. cBarc: Costa Barcelona [Barcelona Coast]; pBarc: Paisatges Barcelona [Barcelona landscapes]) of the territorial brand (Section 3).
    3. *town*. Destination town or city (e.g. L'Hospitalet de Llobregat: L-Hospitalet-de-Llobregat)
    4. *date*. Date of travel blog or OTR in YYYYMMDD format, based on the ISO 8061 standard, which allows the numerical or alphabetical ordering of files by dates.
    5. *lang*. Two-letter language code according to the ISO 639-1 standard (e.g. ca: Catalan, en: English). If language could not be detected (e.g. multilingual entry or insufficient text) the code 'xx' was used.
    6. *isFrom*. Two-letter country code according to the ISO-3166-2 standard (e.g, GB: United Kingdom, US: United States of America). The states are unique, however. Some users write their city or region only, and cannot be classified because of the problem of homonyms, especially in English-speaking countries. Both in these cases and in those where the

hometown is not available, we can write the code ‘ZZ’ was used.

7. *pageId*. The website identified must be unique. In the case of the example shown: destination code, member code, and travelogue or OTR code separated by underscores.
8. *topic*. Four-letter abbreviation of the subject of the OTR. The most common classifications of OTRs are things to do, hotels, and restaurants.

This organization of data allows any target subset to be obtained by means of the utilities in the Operating System.

C. *Data cleaning*. A webpage usually contains a large amount of noise, such as navigation menus, advertisements, copyright notices, etc (Liu, 2011), that have no relationship with the perceived image of the destination that this study aims to analyse. In this pre-processing phase, we aim to preserve the part of the webpage content generated by the user without modifying its HTML format, and eliminating everything else (SeeFig. 3).

The text-based webpage code is delimited by HTML tags. Then, in each webhost, the HTML directives must be identified that do not affect the user’s post. This study used the *Microsoft Expression Web 4* (free version). Both the opening and closing tags of the negligible divisions, as well as the text delimited by them, were recursively removed (i.e., tags and nested tags are removed) with an *ad hoc* programme.

This stage, along with debugging, is essential in obtaining quality information in the analysis phase. Without sacrificing anything from the UGC data, or from its HTML format, needless content has been eliminated, and the content of the example webpage (Fig. 3) has been reduced by about 25 times, which is very helpful when working with more than 100,000 pages.

D. *Data debugging*. Concerning data debugging, the Catalan language uses special characters such as cedillas and vowels with accents, which are both problematic for codification in websites and puzzling for English-speaking users. The special codifications and misspellings distort the analysis and had to be corrected, especially those that affect destination and attraction names, and proper nouns in general, such as character encoding problems, misspellings, and translations.

4.4. Content analysis

Quantitative content analysis was performed through a website content parser, *Site Content Analyzer* (SCA). which parses online and offline for keywords, suggests the most relevant and weighty phrases, and analyses link structure (CleverStat, 2009). SCA discovers the frequency, density, and weight of keywords for each webpage. Afterwards, keywords are grouped into categories in order to analyse users’ perceived image.

A. *Parser settings*. SCA enables the configuring of many options as delimiters and parse zones, but the most interesting configurations are the black list, composite words, and tag and position weight.
   1. *Black list*. In the documents, there are many insignificant words that help construct sentences but do not represent any content; such stop words should be identified and removed before documents are indexed (Liu, 2011; Xiang et al., 2015). The black list is used to include these stop words, such as some adverbs, articles, prepositions, pronouns, and conjunctions.
   2. *Composite words list*. This consists of groups of words that have a joint meaning together. The composite words have priority over the black list and over simple keywords. For example, the 'Basilica of the Sagrada Familia' preserves the ‘of’ and the ‘the’ despite being included in the black list, and takes preference over the simple keywords ‘Basilica’, ‘Sagrada’, and ‘Familia’.
   3. *Keyword weight*. The weight indicates the prominence or visibility of the keyword or key phrase. For example, in general, it has been proven in studies about click stream and navigation patterns that the most disseminated content is that which is situated at the top of a webpage, rather than in the bottom area. SCA assigns a weight to keywords according to their position and the HTML tag that defines their format and features.

B. *Categorization*. For the quantitative content analysis, the frequency of all of the unique keywords was obtained. These keywords, must be grouped into categories to enable knowledge to be extracted from the data. In quantitative content analysis, it is assumed that the words most often mentioned are the words that reflect the greatest concerns (Stemler, 2001). The development of efficacious categories is based on the following principles:

Categories should be exhaustive (i.e., there should be a category for every relevant item in the text), mutually exclusive (i.e., no recording unit should be placed in more than one single category), and independent (i.e., assignment of any recording unit into a single category does not affect the classification of other data units) (Holsti, 1969, p. 95, as cited in Stepchenkova (2012)).

Quantitative content analysis research can be based on two categorization models: *a priori* based on pre-established categories or based on categories discerned from the text itself (Stepchenkova, 2012).

Based on former works (Marine-Roig, 2013) and on preliminary frequency analysis, eight broad categories related to attraction factors were developed to analyse the cognitive component of destination image: Food and Wine; Intangible Heritage; Leisure and Recreation; Nature and Active Tourism; Sports; Sun, Sea, and Sand; Tangible Heritage; and Urban Environment. Additionally, for this paper, a specific category was built regarding the 'smartness' of the city of Barcelona. This category was constructed by identifying keywords and composite words related to smart cities and smart destinations according to previous academic and professional works (Arup, 2013; Bakici et al., 2013; Buhalis & Amaranggana, 2014; Cisco, 2014; Cohen, 2014; Giffinger et al., 2007; Laursen, 2014; Manville et al., 2014; Morabito, 2015). These keywords and composite words were: big data, citizen network, cloud computing, contactless payments, citywide sensors, digital trust network, electric vehicle, electro-mobility, end-user devices, e-governance, e-services, intelligent data, intelligent sensors, internet of things, lighting plan, living labs, microgrid, o-government, open challenge, open data, orthogonal bus network, self-driving car, self-sufficient islands, smart allotment, smart building, smart city, smart destination, smart district, smart economy, smart energy, smart environment, smart governance, smart grid, smart lighting, smart living, smart mobility, smart parking, smart people, smart street, smart streetlights, smart tickets, smart tourism, smart tourism destination, smart traffic, smart water, streamlined rubbish collection, telecommunications network, telemanagement of irrigation, urban platform, and wi-fi networking.

Some of the keywords in this category are not exclusive to the smart city/destination topic, but are mutually exclusive from the other categories based on tourist attraction factors. Therefore they

Table 3
Twenty-five most frequent keywords in the entire dataset.

Rank	Keyword	Count	Site wide density (%)	Average weight
01	barcelona	292,161	3.18	60.63
02	tour	115,377	1.25	33.73
03	great	68,041	0.74	24.00
04	sagrada familia	58,478	0.64	66.04
05	gaudi	47,317	0.51	19.83
06	visit	38,418	0.42	15.48
07	city	37,376	0.41	12.04
08	amazing	36,410	0.40	24.97
09	place	35,958	0.39	16.45
10	beautiful	33,957	0.37	24.13
11	parc guell	30,757	0.33	64.42
12	good	30,225	0.33	14.68
13	way	30,217	0.33	16.53
14	nice	25,597	0.28	20.44
15	park	23,984	0.26	15.37
16	best	23,778	0.26	25.02
17	guide	23,371	0.25	10.74
18	experience	23,258	0.25	22.95
19	walking	22,641	0.25	37.39
20	casa mila	22,262	0.24	71.71
21	museum	21,286	0.23	27.52
22	people	20,802	0.23	5.01
23	walk	19,778	0.22	11.16
24	montserrat	18,976	0.21	52.02
25	building	18,835	0.21	12.97

can provide an idea of the relevance of this attribute for the perceived image of a destination.

5. Results and discussion

To demonstrate the usefulness of big data analytics to give insight into the image of destinations as transmitted by tourists, frequency analysis of all the texts in OTRs written in English about Barcelona and its surrounding areas (117,487 entries from the last 10 years) was conducted. Moreover, a closer look at the landmark Basilica de La Sagrada Familia (7481 OTRs written in 2014) was proposed in order to examine the usefulness of the analysis of massive UGC data for the management of particular attractions. La Sagrada Familia was chosen for analysis because it is considered the main landmark or emblem of Barcelona (Sobrer, 2002), it is the most visited attraction of the city (Barcelona Turisme, 2014), and in the initial frequency word count of all OTRs about Barcelona, it was by far the city's most mentioned asset (see Table 3). Each analysis produced a list of unique keywords by frequency. In order to estimate the importance of these keywords, site-wide density (i.e., taking all of the text into account, stop words included) was calculated and the average weight is presented here. This massive analysis was used to provide insight into both the transmitted image of the destination and any specific managerial issues in a specific attraction.

5.1. The transmitted image of Barcelona and its surrounding areas

Destination image, transmitted online in UGC, is a rich source to unveil destination image at the post-trip stage, when the tourist has acquired an elaborate image after evaluating their experience. At this stage, Opoku (2006) argues that counting word frequencies is a suitable method to analyse destination image in UGC, as words in text reflect what their creators think are important attributes or characteristics in an explicit way, thus indicating the extent to which the message sender is focusing on a particular image dimension or aspect. Table 3 shows the top 25 of 84,945 unique keywords. Barcelona stands out for its high frequency and weight, eclipsing the surrounding brands and destinations. In relation to its attractions, remarkably, some of the top positions correspond to the Catalan architect Antoni Gaudi and three of his masterpieces (Basilica of the Sagrada Familia, Guell Park, and Mila House/La Pedrera). These Gaudi masterworks were the leaders in average weight, and the results are congruent with the position of a leading destination enjoyed by Barcelona city and the brand, and with the fact that Gaudi's works are registered as a UNESCO WHS.

Furthermore, among the top 25 keywords, there were six good feelings, which appear almost 200,000 times in the dataset (i.e., with an average of 1.65 per post), showing a generally positive perception of the destination. In this respect, and from this frequency count analysis, distinguishing elements can be observed related to both the cognitive and affective components of the destination image: the former reflected in attraction factors and tangible elements, and the latter in feelings and emotions (Kim & Richardson, 2003; Krizman & Belullo, 2007), enabling destinations to assess and direct branding policies.

Entries were then segmented according to territorial brands, including the Barcelona city brand and the two surrounding brands, which also use the name Barcelona under the direct influence of Barcelona ('Barcelona Coast' and 'Barcelona Landscapes' brands). The previous frequency analysis was repeated for each of the brands. Next, keywords were grouped into the previously mentioned categories (Table 4). Grouping keywords into categories is what makes the technique of content analysis reliable and meaningful (Stemler, 2001), especially in the case of quantitative analyses of massive datasets. In this case, categorization into attraction factors can give insight into tourists' behaviour (Krizman & Belullo, 2007; O'Connor, 2010) by allowing the main assets or factors of the destination's attractiveness as presented in tourists' accounts upon which they make decisions. For example, the attraction factors present in images can be compared to official images or marketing campaigns, or compared to different destinations or territories.

Barcelona stands out in the *Tangible Heritage* and *Urban*

Table 4
Frequencies per category and tourist brand.

Category	Barcelona brand count #	Barcelona coast count #	Barcelona landscapes count #	Barcelona brand count %	Barcelona coast count %	Barcelona landscapes count %
Food and wine	69,436	3682	869	7.79	19.13	4.11
Intangible heritage	10,759	406	66	1.21	2.11	0.31
Leisure and recreation	65,879	2312	939	7.39	12.01	4.44
Nature and active tourism	24,907	767	4196	2.80	3.99	19.86
Sports	38,382	604	138	4.31	3.14	0.65
Sun, sea, sand	35,606	5005	307	4.00	26.01	1.45
Tangible heritage	467,389	4467	11,801	52.45	23.21	55.85
Urban environment	178,672	2003	2812	20.05	10.41	13.31
Smartness	2	0	0	0.00002	0.00	0.00

Environment categories; as previously stated, Barcelona city has nine buildings classified as a WHS. The Barcelona Coast stands out for its *Sun, Sea, and Sand*; *Tangible Heritage*; and *Food and Wine*. It has a vast area of vineyards dedicated to the production of cava (méthode champenoise). Barcelona Landscapes stand out for its *Tangible Heritage* and *Nature*; it has a monastery (Montserrat) that is highly popular and visited by both tourists and residents. These results show both a complementarity and a relative specialization of the three brands under the Barcelona name. Thus, the analysis of big UGC data can be of great utility for management organizations of complex destinations encompassing several functional spaces, directing tourism flows, complementing or diversifying the tourism offer in different areas, and directing and assessing 'brand architecture' strategies, which focus on developing and managing interrelated brands (Datzira-Masip & Poluzzi, 2014).

Finally, as a noteworthy point, results concerning 'smartness' should be considered trivial. The word 'big data' only appears twice in separate reviews in 2014 on the Centre of Contemporary Culture of Barcelona (CCCB) in the context: '*The exhibition on big data was brilliant*', and, '*We saw the big data exhibit. It was fascinating*'. There is not a single mention of 'smart city' or 'smart destination' or of any of the related concepts presented above. The adjective 'smart' appears 425 times associated with the word 'phone' in most cases, but in none of these cases was associated with the 38,952 mentions of 'city/ies', nor with the 2197 mentions of 'destination/s'.

5.2. Basilica de La Sagrada Familia in Barcelona

In this section, in order to look more closely at a specific landmark of the destination, analysis of the specific reviews in 2014 about La Sagrada Familia is conducted. These reviews were selected through the words Sagrada Familia that were found in the review link in the case of TripAdvisor, and in the anchor text of the navigation bar in the case of VirtualTourist. Table 5 shows the top 25 of 10,872 unique keywords in these specific reviews about La Sagrada Familia. As expected, Sagrada Familia and Barcelona were the most frequent and weighty keywords. With high density and weight also appeared four positive adjectives (amazing, beautiful, stunning, and breathtaking). These good feelings were associated with the artistic and architectural value of the Basilica, which was declared a UNESCO WHS.

Table 5
Twenty-five most frequent keywords in the Sagrada Familia subset.

Rank	Keyword	Count	Sitewide density (%)	Average weight
01	sagrada familia	17,051	4.20	81.18
02	barcelona	11,326	2.79	71.91
03	tickets	5022	1.24	6.98
04	amazing	4461	1.10	29.17
05	visit	3468	0.86	13.38
06	gaudi	3086	0.76	11.84
07	online	2965	0.73	10.96
08	beautiful	2628	0.65	24.99
09	line	2423	0.60	5.22
10	tower	2399	0.59	2.81
11	building	2358	0.58	10.93
12	church	2352	0.58	13.28
13	tour	2189	0.54	5.48
14	queue	2169	0.53	6.94
15	book	2012	0.50	15.43
16	architecture	1959	0.48	19.52
17	place	1894	0.47	14.37
18	basilica	1730	0.43	9.53
19	wait	1546	0.38	6.72
20	stunning	1361	0.34	33.40
21	cathedral	1301	0.32	11.93
22	guide	1236	0.31	2.98
23	breathtaking	1204	0.30	35.37
24	audio	1185	0.29	2.70
25	in advance	1149	0.28	11.98

Source: 7481 TA and VT OTRs written in 2014.

In addition, the keyword ticket/s appears in third position, related with book, on line/online, and in advance. Moreover, queue/s/d/ing, line/s/d/up/ups, and wait/s/ed/ing are also very common. These keywords were related to the management and planning of visits. In fact, in almost all of the OTRs related to La Sagrada Familia, the problems of long lines or queues and waits were mentioned, and advanced online booking of tickets was strongly recommended. Although the Basilica receives an average of 9000 daily visitors, these results are unexpected, mostly because the image perceived and transmitted by tourists visiting a WHS should be the uniqueness and beauty of its architectural and artistic elements, not problems regarding the purchase of tickets and having to wait to enter. In fact, it has a sophisticated booking, purchasing, and visiting process that could lead managers to believe that the problem is solved (even though, as the comments reflect, it is not).

In this respect, these results show the usefulness of big data analytics of UGC, not only in terms of the analysis of the transmitted image of a destination by online users, but also in terms of the management of a destination and of its specific assets, as this enables the identification of recurring issues and problems at specific attractions. The nature of review comments, which are usually associated with specific attractions or places, makes it possible to tackle the specific UGC information attributed to them by tourists. These results are in the line of Sanz (2013), who asserts that, nowadays, the information resulting from the analysis of UGC data online is now crucial for a city’s BI and management.

6. Conclusions and implications

Today, the information resulting from the analysis of UGC data online is fundamental for a city’s BI (Sanz, 2013). As Bakici et al. (2013) state, one of the main pillars of a smart city model is information. In this regard, UGC analytics should be seen as an important asset in destination smartness, as it is useful to make 'smarter' decisions in several areas such as destination planning, strategy, destination branding and imaging, and multiple territory brand architecture. This firsthand massive and diverse user-generated information – which is freely available online – should be made operational for destinations and taken into account when making decisions regarding future developments, and looked at more deeply in this specificity or added value to a smart city: being a smart destination. Massive UGC analytics enable the gathering of valuable quantitative information about territorial brands and destinations, post dates, user languages and hometowns, and post topics, prior to content analysis itself. Such analytics also enable studies focusing on about quantitative or qualitative content analysis in a place, period, language, or nationality, and also according to specific topics, singularly or combined, selecting a corresponding subset or a random sample thereof, with the advantage that the quantitative content analysis is limited to what has been written by the user. In this regard, massive UGC data analytics is not only useful in revealing the image of a destination in general, but also in obtaining insights concerning management issues at specific attractions.

Concerning the results of the case study, Barcelona and its surroundings, the main attraction factors to the city (Tangible Heritage, with the outstanding element of Gaudi, and Urban Environment) have been detected in tourists’ images, as well as remarkably different attraction factors to its surrounding areas, (Nature, in the case of Barcelona Landscapes; and Sun, Sea, and Sand, in the case of Barcelona Coast), providing interesting information for brand architecture in a complex destination. Moreover, a remarkable marketing issue has been detected. In spite of the international recognition of Barcelona as a leading smart city in academic, professional and institutional fields, and the efforts by the town council to publicize the concept of 'Barcelona, smart city' (Barcelona City Council, 2014) through multiple media channels, the idea of Barcelona as a smart city or smart destination is not mentioned or recognized at all by tourists in online UGC. This could be a sign of a significant marketing or communicational issue, since tourists did not mention, even once, in more than 100,000 posts about Barcelona, keywords associated with the idea of smart city or smart destination; this may also be an indication that this aspect is not valued by tourists. These issues should be further explored. Moreover, a management issue regarding Barcelona's main tourist attraction, Basilica de La Sagrada Familia, has been identified because visitors transmit an image that does not correspond to its status of being a WHS. Instead, the focus is, to a great extent, on issues related to getting tickets and accessing the attraction.

These findings provide valuable information for NTOs and DMOs because they offer firsthand knowledge of tourists' perceived image transmitted online to other tourists. Through massive UGC data analytics, tourist trends, perceived experiences, and attraction factors can be unravelled both at the level of destinations as a whole, and at the level of specific attractions or landmarks. For instance, the previous analysis could be repeated following a marketing campaign among foreign visitors to promote Barcelona as a smart city or smart tourism destination certain improvements in the management and planning of visits to La Sagrada Familia. That is not all – information about the images that will be transmitted to other tourists online during the post-trip phase can be obtained, and should be cross-referenced with UGC data from other stages of the trip. Furthermore, differential perceptions of several destinations and of central versus surrounding areas can be unveiled, and should also be cross-referenced with other information concerning tourist flows, transportation, and destination planning, among others. The information found in travel blogs and OTRs can assess smart cities and destinations with long-term marketing campaigns and planning policies, providing a long-term vision of the destination and the activities conducted in it from the point of view of tourists, which may be difficult to obtain otherwise. Therefore, DMOs should try to maximize the usefulness of big UGC data that are present in travel blogs and reviews, and use them as sources of information, especially through data analysis systematization and computerization (Marine-Roig & Anton Clavé, 2015). These will be key tools for BI in a smart destination, as an added value or specificity to a smart city, aiming to provide a high-quality tourist experience. In this respect, big data analytics can help in 'smart' marketing and policy decision-making, along with other sources of information; thus, making information present in online UGC is useful to a (smart) destination, and will improve residents' lives and tourists' experiences.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness [Grant id.: MOVETUR CSO2014-51785-R].

References

Arup (2013). *Global innovators: International case studies on smart cities*. London, United Kingdom: GOV.UK, Department for Business Innovation & Skills.

Bakici, T., Almirall, E., & Wareham, J. (2013). A smart city initiative: the case of Barcelona. *Journal of the Knowledge Economy*, *4*(2), 135–148. http://dx.doi.org/10.1007/s13132-012-0084-9.

Barcelona City Council (2014). *BCN Smart City*. Retrieved June 15, 2015, from http://smartcity.bcn.cat/en.

Barcelona Turisme (2014). *Barcelona tourism press file 2014*. ⟨http://professional.barcelonaturisme.com/⟩; retrieved 15.06.15.

Baudot, F. (2015, March 4). *European city tourism to resume positive growth in 2014. European Cities Marketing*. ⟨http://www.europeancitiesmarketing.com/category/press-releases/⟩;retrieved 15.06.15.

Boes, K., Buhalis, D., & Inversini, A. (2015). Conceptualising smart tourism destination dimensions In: I. Tussyadiah, & A. Inversini (Eds.), *Information and communication technologies in tourism 2015* (pp. 391–403). Cham, Switzerland: Springer. http://dx.doi.org/10.1007/978-3-319-14343-9_29.

Buhalis, D., & Amaranggana, A. (2014). Smart tourism destinations In: Z. Xiang, & I. Tussyadiah (Eds.), *Information and communication technologies in tourism 2014*. Cham, Switzerland, 553–564: Springerhttp://dx.doi.org/10.1007/978-3-319-03973-2_40.

Catalan Tourist Board (2015). *2015 Press Pack*. ⟨http://www.act.cat/press-pack⟩; retrieved 15.06.15.

Cisco (2014). *IoE-driven smart city Barcelona initiative*. ⟨http://internetofeverything.cisco.com⟩; retrieved 15.06.15.

CleverStat (2009). *Site Content Analyzer overview*. Retrieved June 15, 2015, from http://www.cleverstat.com/en/sca-website-analysis-software-index.htm.

Cohen, B. (2014, November 20). *The smartest cities in the world. CoExist*. ⟨http://www.fastcoexist.com/3038765/fast-cities/the-smartest-cities-in-the-world⟩; retrieved 15.06.15.

Datzira-Masip, J., & Poluzzi, A. (2014). Brand architecture management: The case of four tourist destinations in Catalonia. *Journal of Destination Marketing & Management*, *3*(1), 48–58. http://dx.doi.org/10.1016/j.jdmm.2013.12.006.

Del Chiappa, G., & Baggio, R. (2015). Knowledge transfer in smart tourism destinations: Analyzing the effects of a network structure. *Journal of Destination Marketing & Management*, *4*(3), 145–150. http://dx.doi.org/10.1016/j.jdmm.2015.02.001.

European Commission (2014). *Barcelona is 'iCapital' of Europe. Brussels*. Belgium: Press release March 11.

European Parliament (2014). *Mapping smart cities in the EU*. Strasbourg, France: Publications Office.

Fuchs, M., Hopken, 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. http://dx.doi.org/10.1016/j.jdmm.2014.08.002.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*, 137–144. http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., & Meijers, E. (2007). *Smart cities: Ranking of European medium-sized cities*. Vienna, Austria: Centre of Regional Science, Vienna UT.

He, W., & Chen, Y. (2014). Using blog mining as an analytical method to study the use of social media by small businesses. *Journal of Information Technology Case and Application Research*, *16*(2), 91–104. http://dx.doi.org/10.1080/15228053.2014.943092.

Hidalgo, M. C., Sicilia, M., & Ruiz, S. (2014). The effect of user-generated content on tourist behavior: The mediating role of destination image [Special issue]. *Tourism & Management Studies*, *10*, 158–164.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, *2*, 652–687. http://dx.doi.org/10.1109/ACCESS.2014.2332453.

Jalilvand, M. R., Samiei, N., Dini, B., & Manzari, P. Y. (2012). Examining the structural relationships of electronic word of mouth, destination image, tourist attitude toward destination and travel intention: An integrated approach. *Journal of Destination Marketing & Management*, *1*(1–2), 134–143. http://dx.doi.org/10.1016/j.jdmm.2012.10.001.

Kim, H., & Richardson, S. L. (2003). Motion picture impacts on destination images. *Annals of Tourism Research*, *30*(1), 216–237. http://dx.doi.org/10.1016/S0160-7383(02)00062-2.

Koltringer, 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. http://dx.doi.org/10.1016/j.jbusres.2015.01.011.

Krizman, D. & Belullo, A., (2007). Internet – An agent of tourism destination image formation: Content and correspondence analysis of Istria travel related websites. In *4th International Conference: Global Challenges for Competitiveness: Business and Government Perspective* (pp. 541–556). Pula: Juraj Dobrila University of Pula, Department of Economics and Tourism.

Laursen, L. (2014, November 18). Barcelona's smart city ecosystem. *MIT Technology Review*. ⟨http://www.technologyreview.com/news/532511/barcelonas-smart-city-ecosystem⟩; retrieved 15.06.15.

Leung, D., Law, R., Van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing*, *30*(1–2), 3–22. http://dx.doi.org/10.1080/10548408.2013.750919.

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data*. Berlin, DE: Springer.

Liu, S. (2014). Selecting a destination image for a capital city rather than for a nation: A segmentation study. *Journal of Destination Marketing & Management*, *3*(1), 11–17. http://dx.doi.org/10.1016/j.jdmm.2013.12.002.

Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing & Management*, *24*(2), 119–154. http://dx.doi.org/10.1080/19368623.2014.907758.

Manville, C., Cochrane, G., Cave, J., Millard, J., Pederson, J. K., & Thaarup, R. K. (2014). *Mapping smart cities in the EU*. Brussels, Belgium: European Parliament.

Marine-Roig, E. (2013). *From the projected to the transmitted image: The 2.0 construction of tourist destination image and identity in Catalonia* (Ph.D. dissertation). 〈http://hdl.handle.net/10803/135006〉; retrieved 15.06.15.

Marine-Roig, E. (2014). A webometric analysis of travel blogs and reviews hosting: The case of Catalonia. *Journal of Travel & Tourism Marketing*, *31*(3), 381–396. http://dx.doi.org/10.1080/10548408.2013.877413.

Marine-Roig, E. (2015). Identity and authenticity in destination image construction. *Anatolia – An International Journal of Tourism and Hospitality Research*. , http://dx.doi.org/10.1080/13032917.2015.1040814.

Marine-Roig, E., & Anton Clavé, S. (2015). A method for analysing large-scale UGC data for tourism: Application to the case of Catalonia In: I. Tussyadiah, & A. Inversini (Eds.), *Information and communication technologies in tourism 2015* (pp. 3–17). Cham, Switzerland: Springer. http://dx.doi.org/10.1007/978-3-319-14343-9_1.

MetaProducts (2015). *Offline Explorer Enterprise features*. Retrieved June 15, 2015, from http://metaproducts.com/mp/offline_explorer_enterprise.htm.

Michael, C. (2014, May 6). From Milan to Mecca: the world's most powerful city brands revealed. *The Guardian News, Cities, City brand.* 〈http://www.theguardian.com/cities/gallery/2014/may/06/from-milan-to-mecca-the-worlds-most-powerful-city-brands-revealed〉; retrieved 15.06.15.

Morabito, V. (2015). *Big data and analytics: Strategic and organizational impacts.* Cham, Switzerland: Springer.

O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management*, *19*(7), 754–772. http://dx.doi.org/10.1080/19368623.2010.508007.

Oliver, V., Garcia, E., Solana, A., Gonzalez, R., Pelaez, M. V., & Tome, M. J. (2014). *Big data and tourism: New indicators for tourism management*. Barcelona, Catalonia, Spain: Roca Salvatella and Telefonica.

Opoku, R. A. (2006). *Towards a methodological design for evaluating online brand positioning* (Ph.D. dissertation). 〈http://pure.ltu.se/portal/en/〉; retrieved 15.06.15.

Pan, B., MacLaurin, T., & Crotts, J. C. (2007). Travel blogs and the implications for destination marketing. *Journal of Travel Research*, *46*, 35–45. http://dx.doi.org/10.1177/0047287507302378.

Sanz, L. (2013). *City deploys big data BI solution to improve lives and create a smart-city template*. Barcelona, Catalonia, Spain: Microsoft.

Schmallegger, D., & Carson, D. (2010). Destination image projection on consumer generated content websites (CGC): A case study of the Flinders Ranges. *Journal of Information Technology & Tourism*, *11*(2), 111–127. http://dx.doi.org/10.3727/109830509789994838.

Sobrer, J. M. (2002). Against Barcelona? Gaudi, the city, and nature. *Arizona Journal of Hispanic Cultural Studies*, *6*, 205–219. http://dx.doi.org/10.1353/hcs.2011.0362.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, *7* (17). 〈http://PAREonline.net/getvn.asp?v=7&n=17〉; retrieved 15.06.15.

Stepchenkova, S. (2012). Content analysis In: L. Dwyer, A. Gill, & N. Seetaram (Eds.), *Handbook of research methods in tourism: Quantitative and qualitative approaches*. Cheltenham, UK: Edward Elgar Publishing.

Wakefield, J. (2013, August 18). Tomorrow's cities: Do you want to live in a smart city? *BBC News.* 〈http://www.bbc.com/news/technology-22538561〉; retrieved 15.06.15.

Wang, D., Li, X., & Li, Y. (2013). China's 'smart tourism destination' initiative: A taste of the service-dominant logic. *Journal of Destination Marketing & Management*, *2*(1), 59–61. http://dx.doi.org/10.1016/j.jdmm.2013.05.004.

Xiang, Z., Schwartz, Z., & Uysal, M. (2015). What types of hotels make their guests (un)happy? Text analytics of customer experiences in online reviews In: I. Tussyadiah, & A. Inversini (Eds.), *Information and communication technologies in tourism 2015* (pp. 33–45). Cham, Switzerland: Springer. http://dx.doi.org/10.1007/978-3-319-14343-9_3.

Yoo, K. H., & Gretzel, U. (2010). Antecedents and impacts of trust in travel-related consumer-generated media. *Journal of Information Technologies & Tourism*, *12* (2), 139–152. http://dx.doi.org/10.3727/109830510 × 12887971002701.