

A detailed method for destination image analysing using user-generated content

Estela **Marine-Roig**^{1,*} Author's pre-print. Please see: Marine-Roig E. & Anton Clavé, S. (2016). A detailed method for destination image analysis using user-generated content. *Information Technology & Tourism*, 15(4), 341-364 . doi: 10.1007/s40558-015-0040-1

Email estela.marine@aegern.udl.cat

Salvador **Anton Clavé**²

Email: salvador.anton@urv.cat

¹ Department of Business Administration and Economic Management of Natural Resources (AEGERN), University of Lleida, Catalonia, Spain

² Research Group on Territorial Analysis and Tourism Studies (GRATET), Rovira i Virgili University, Catalonia, Spain

Abstract

Social media and user-generated content (UGC) have revolutionized tourism and hospitality communication and are seen as rich sources of information for destination image analysis. Many articles have been published about travel-related UGC, in particular, quantitative and qualitative content analysis of travel blogs and online travel reviews (OTR). Researchers have typically analysed small samples of population-representative travel diaries (tens or hundreds of files), which allow for manual processing. However, the enormous growth of OTRs requires operationalization through computerized methods, and the aim of this article is to propose a detailed method for semi-automatic downloading, arrangement, cleaning, debugging, and analysis of large-scale travel blogs and OTR data. This enables the classification of collected webpages by dates and destinations and offline content analysis of the text as written by the tourist. More than 130,000 useful trip diaries of tourists who visited Catalonia between 2004 and 2014 have been gathered, and significant results have been obtained in terms of content analysis in relation to destination image.

Keywords

Travel blog

Online travel review

Web harvesting

Web data mining

Massive content analysis

Catalonia

This article is an extended and updated version of a conference paper previously published in ENTER Proceedings of the International Conference in Lugano, Switzerland, February 3–6, 2015.

1. Introduction

Destination image is a complex construct, and is usually considered to be the sum of beliefs, ideas and impressions that people have of a place (Kotler et al. 1993). It is a complex construct resulting from both the projected and the perceived images of a destination (Marine-Roig 2015a) and is said to be composed of two primary components, the cognitive component (mental responses to the stimuli in the environment, related to physical attributes) and the affective components (how one feels about this knowledge). A third component, the conative (related to behaviour), was later introduced by several researchers (Marchiori and Onder 2015; Serna et al. 2015).

In recent years the image of tourism destinations has depended on multiple sources of information and content generated by travellers, suppliers and residents, especially online social media (Anton Clavé and Gonzalez 2008; Llodra-Riera et al. 2015; Munar and Jacobsen 2013). The rise of social media has facilitated the exponential growth of user-generated content (UGC). Today online UGC strongly and rapidly influences the formation of destination images, since on the one hand, tourists increasingly and actively post the recounts of their personal experiences online, in the form of both textual and visual contributions, and, on the other hand, they increasingly rely on UGC to reduce their uncertainty when making decisions concerning travel and tourism (Amaral et al. 2014; Llodra-Riera et al. 2015; Marchiori and Onder 2015).

Due to the influence of UGC sources in destination image formation, it is

important for tourist destinations and businesses to assess the extent to which this content contributes to the creation of and changes in destination image (Llodra-Riera et al. 2015; Marchiori and Onder 2015) and in promoting single operators and the destination as a whole (Albastroiu and Felea 2014). Moreover, UGC data in tourism is considered a good source of information for national tourism organizations (NTOs) and other policymakers, destination marketing organizations (DMOs) and other stakeholders, as well as for future travellers, because it consists of opinions freely expressed by tourists who have visited the destination. UGC also offers huge possibilities for e-commerce, business intelligence, marketing, and social studies; a growing number of commercial firms contribute to hosting, developing, distributing, rating, and mining UGC (Moens et al. 2014); and increasingly tourism firms exploit UGC for improving services that analyse travellers' post-trip experiences (Sigala et al. 2012).

According to Standing et al. (2014), Travel 2.0 approaches have spawned large amounts of UGC in the form of travel blogs and online travel reviews (OTR). Travel blogs are a cost-effective means to gather rich, authentic, and unsolicited visitor feedback (Pan et al. 2007). Travel blogs and OTRs allow for experiencing firsthand events that travellers narrate during their stay at the destination as well as viewing their photos and video uploads. Moreover, review sites such as TripAdvisor were found to have the highest level of trust by online users among several online social media sources (Munar and Jacobsen 2013); however, because of the ease of publishing fake reviews, concerns have arisen (Minazzi 2015; Schuckert et al. 2015b).

Regarding UGC as a research source in tourism, Lu and Stepchenkova (2015) state that, in general, the methods used to collect data are unclear and mostly rely on manually selecting and gathering information. In addition to the articles surveyed by these authors, other recent examples, which have not fully developed their methodologies include Wang et al. (2013), Schmunk et al. (2014), and Guo et al. (2015) who mention an *ad hoc* web crawler to collect data, without further details; Koltringer and Dickinger (2015) indicate a commercial web crawler (WebLyzard.com); whereas Fang et al. (2016) do not show how they collected their data.

To fill this gap in UGC-related research, this article aims to propose a detailed method for semi-automatic downloading, arrangement, cleaning,

debugging, and analysis of large-scale travel blog and OTR data. The methodology is applied in the case of Catalonia to analyse more than 130,000 useful travel diaries written by foreign tourists who visited Catalonia between 2004 and 2014.

2. State of the art

Social media is characterized by using ever-changing tools; new platforms are continuously created, and many destinations have difficulties achieving a long-term strategy, requiring innovative strategies and methodologies of analysis (Munar 2012). Dwyer et al. (2014) classified research methods in tourism data analysis into quantitative, qualitative and mixed approaches, but it must be considered that in the last two decades the application of quantitative techniques in both academic and non-academic (NTOs, DMOs, and other stakeholders) research has intensified. The most popular research methods for the analysis of travel blogs have been content analysis, both qualitative and quantitative, and narrative analysis (Banyai and Glover 2012); and with respect to OTRs, Schuckert et al. (2015a) found that most studies are based on quantitative content analysis.

The topic areas of the studies on tourism and hospitality are very diverse. Zeng and Gerritsen (2014) analysed 165 articles, published between 2007 and 2013, and found that the most frequently occurring keywords were “marketing,” followed by “customer/consumer behaviour,” “UGC,” “information search,” “destination management,” and “(electronic)word-of-mouth” (WOM and eWOM). Gursoy et al. (2015) presented over 200 measurement scales gleaned from studies published in selected top journals in the field, between 1992 and 2013, and grouped them into seven main categories: motivation; residents’ perceptions and attitudes; destination image; performance, evaluation, quality assessment, loyalty, and satisfaction; tourist behaviour; human resources; and hospitality and tourism operations. Lu and Stepchenkova (2015) surveyed 122 articles, published between 2001 and 2013, and classified them, in order of frequency, into service quality, destination image and reputation, UGC such as eWOM, experiences and behaviour, and mobility patterns. Schuckert et al. (2015a) analysed 50 articles related to online reviews, published in academic journals between 2004 and 2013, and grouped them into five topical clusters: online buying, satisfaction and management, opinion mining/sentiment analysis, motivation, and the role of reviewers.

Such user (consumer)-generated content data have grown exponentially in recent years, especially in the case of hospitality OTRs. For instance, in May 2015, TripAdvisor asserted that it had reached more than 225 million reviews and opinions, Booking almost 49 million verified reviews from real guests, and Trivago indicated that it had received 140 million integrated user hotel reviews. It is now considered that their processing requires the use of Big Data technologies (Krawczyk and Xiang 2015). However, Lu and Stepchenkova (2015), in an exhaustive work about UGC as a research mode, have proven that, in most studies, UGC data were collected manually, limiting the sample size. Small samples can hardly represent the population, and Banyai and Glover (2012) contend that usually samples are not selected by chance, which also questions their representativeness. Moreover, the collection of UGC data via manual copying is extremely time-consuming (Johnson et al. 2012; Lu and Stepchenkova 2015). As examples of relatively small sample size (tens or hundreds of entries), Wang and Morais (2014) examined the tourists' identity in 69 travel blogs from 16 websites utilising critical discourse analysis; these weblogs were selected after a preliminary reading of all blogs on the surveyed destinations, located through the Google search engine; Amaral et al. (2014) unveiled tourists' profiles and preferences in 813 consumer-generated reviews on 20 restaurants from TripAdvisor.com; Kladou and Mavragani (2015) assessed the cognitive, affective and conative components of the image of Istanbul, through 203 reviews posted on TripAdvisor; and Lai and To (2015) using Leximancer analysed 68 keywords from 440 entries collected in the same websites studied in this paper to show the destination image of Macao.

Many studies focus on user (consumer)-generated content on specialized websites. For instance, commissioned by the Catalan NTO, Gonzalez (2010) analysed a total of 28 sources of online information, in different formats, selected on the basis of the criteria of importance of the webpage inside the parameters of Web 2.0, presence of UGC and of a specific section for Catalonia, to unveil the image of the Catalan tourism brands on the Internet. Schmunk et al. (2014) collected 1441 reviews about hotels from TripAdvisor.com and Booking.com to extract decision-relevant knowledge from UGC. To do so, these authors proposed a sentiment analysis process. Johnson et al. (2012), using a web harvesting programme, claimed to automatize the collection of 5730 OTRs from TripAdvisor.com about Nova Scotia, but recognized that they had to

manually eliminate reviews of many other destinations because they first identified the initial pages by searching “Nova Scotia” in TripAdvisor.com and all the destinations that residents in this Canadian province had visited also appeared. These authors performed a basic classification of OTRs (attractions, hotels, and restaurants) and grouped them according to destinations. Koltringer and Dickinger (2015) collected 5719 relevant documents from online sources such as TravelBlogs.com, by the WebLyzard crawling agent, and extracted destination brand identity and image through web content mining and natural language processing, including keyword analysis and automatic sentiment detection. Li et al. (2015) gathered 1033 verified travel blog articles from CTrip.com to analyse the destination image of Taiwan. As examples of massive downloads using *ad hoc* crawling or scraping software, Serna et al. (2015) examined 4123 OTRs from MiNube.com to analyse the cognitive-affective and conative destination image of a Spanish region applying categorization techniques; Wang et al. (2013) with 834,304 OTRs from TripAdvisor.com computed two impact indexes to measure reviewer credibility based on the number of reviews and destinations on which a reviewer had posted reviews, and the number of helpful votes received by the reviews; and Xiang et al. (2015) investigated 60,648 online customer reviews from Expedia.com corresponding to 10,537 hotels from the 100 largest USA cities to analyse guest experiences.

Researchers have suggested various methods to extract knowledge from online sources. Abburu and Babu (2013) proposed a framework for web data extraction and analysis based on three basic steps: finding URLs (uniform resource locator) of webpages, extracting information from webpages, and data analysis. This suggested system architecture is divided into three modules: web crawling, information extraction, and mining. Chau and Xu (2012) proposed a framework for collecting and analysing business intelligence in blogs on a topic of interest divided into four steps: identify the explicit communities (blogging URLs), collect information about bloggers (blogging pages and blogger pages), analyse the content posted (blogging information, blogger information, and blog entries: term and opinion analysis), analyse the interaction networks and implicit communities formed by the bloggers. To mine valuable information from the structured tourism blogs, Guo et al. (2015) recommended a framework divided into blog extraction and data pre-processing, mining of frequent departure cities, mining of frequent travel spots, and spot associated

service detection. Schmunk et al. (2014) proposed a five-stage process, which consists of selecting and collecting review pages, document processing (information extraction, removing reviews with no text, filtering English texts, and generating sentences), mining, evaluation, and usage. Lai and To (2015) suggested a four-phased methodology: definition of goal and scope, data collection (identifying online information sources, determining the sample size, and downloading text files), data transformation (creating a master file, using WordSmith computer-aided lexical software to identify keywords and their frequency counts, and using Leximancer computer-aided lexical concept mapping software to project a holistic view of the study) and the interpretation of results. Koltringer and Dickinger (2015) apply a method, which includes data gathering, keyword analysis, sentiment detection, category building, and correspondence analysis. However, they converted data into a machine-readable format and only plain text remains for analyses. Li et al. (2015) utilised two approaches for analysing traveller-generated content: text mining (coding and classifying blog entries, string matching processing, and translation from simplified Chinese to English) and content analysis (keyword cleaning, keywords analysis, correspondence analysis, and affective analysis).

According to Lu and Stepchenkova (2015), in general, the studies on UGC as a research source in tourism and hospitality are vague with respect to how they collect data. They further indicate that the technical details of data analyses are often incomplete and that descriptions of the software functionalities are limited (p 142). That is why the method proposed in the next section presents a series of contributions to research on the collection and analysis of massive UGC data that are not recorded or detailed in previous works. Some of these contributions are:

- *Detailed methods* Examples of significant problems that can arise and how to solve them; detailed explained are given.
- *Suitable websites* The selection of hosting websites most suitable for the case study is warranted.
- *Web mining* Web structure mining and the process of downloading webpages are illustrated.
- *Dataset arrangement* The organization of downloaded trip diaries

allowing for multiple classifications of tens of thousands of entries via operating system utilities.

- *Data quality* Analysed data is reduced to what has been written and posted by the user (blogger or reviewer) saving anomalies (character encoding problems, most common mistakes, and translations) of proper names (especially, destination and attraction names).
- *Weight of keywords* Besides the frequency analysis, the proposed method calculates the weight of keywords and key phrases based on their emphasis (impact within the webpage).

3. Methodology

This section presents the analysis methodology that is proposed, developed and described in the following subsections, consisting of the steps for destination and hosting websites selection, data downloading, data arrangement, data cleaning, data debugging and content analysis.

The main feature of the method we propose is in the web data extraction phase, because instead of simply extracting the information, we add the cleaning and debugging phases to eliminate the noise present in the webpage to be able to reach the content analysis phase with quality information in the original HTML (HyperText Markup Language) format (see Fig. 5). That is to say, the resulting webpages only contain what the users wrote and preserve their semi-structured format to assess the HTML emphasis of keywords and key phrases and thus their potential impact on the Internet.

To demonstrate the effectiveness of the method to manage large-scale data placed in time and space, first a destination with a large tourist inflow and territorial division was selected. Consecutively, the most suitable websites for hosting travel blogs and OTRs for the case study were selected. To illustrate this method in detail, we used an example, which does not contain any personal information, of a foreign tourist who visited the selected destination and wrote one travel blog, one travelogue and three travel reviews on one of the selected webhosts. Finally, the different phases (data downloading, data arrangement, data cleaning and data debugging) are described until we reach content analysis. The *results*

section sets out the outcome of the application of the methodology to the case study.

3.1. Destination selection

Although this method can be applied to any destination, more accurate results are obtained when working with large-scale data. To classify entries, the destination needs to be divided territorially, as for example a continent into countries, or a state into regions or provinces. But it could also be applied directly to a destination city like Barcelona. Catalonia has been selected because it fulfils the following terms.

Catalonia is a Mediterranean destination with a millenary history, its own culture and language, and a rich historical and natural heritage. Catalonia offers many attractions for all sorts of visitors: culture, relaxation, nature, family-friendly facilities, sports, business, etc. Its great capacities and excellent facilities place it among Europe's prime tourist areas (CTB 2015). Catalonia is the third European region in number of overnight stays (Eurostat 2014⁵). In 2014, it welcomed more than 20 million tourists, more than three-quarters of whom came from abroad (Table 1).









Year	 be	 ca	 de	 fr	 it	 jp	 nl	 ru
2009	441	88	1052	3773	1110	118	637	216
2010	423	81	1018	3920	1175	114	652	344
2011	491	103	1010	3571	1255	111	868	496
2012	515	150	1162	3816	1168	162	691	741
2013	562	130	1280	4172	1117	207	695	979
2014	593	117	1430	4604	1346	273	815	834

Table 1

Foreign tourists (thousands) with Catalonia as principal destination

Source Trends in the main tourism magnitudes (Catalan Tourism Observatory)



The Catalan territory is made up of nine tourist brands gathered under and promoted by tourist boards, which facilitates the study of delimited spaces with a relatively homogeneous tourist offering (Fig. 1). Catalonia is not an Anglophone region, and therefore, the problems related to character codification beyond ASCII 127 (American Standard Code for Information Interchange) should be considered, specifically, those related to existing accent marks in destination and tourist attraction factor names.

Fig. 1

Tourist brands of Catalonia (CTB 2015)

Tourist brand	Abbr.
Barcelona	<i>Barna</i>
Costa Barcelona	<i>cBarc</i>
Costa Brava	<i>cBrav</i>
Costa Daurada	<i>cDaur</i>
Paisatges Barcelona	<i>pBarc</i>
Pirineus	<i>Pyren</i>
Terres de l'Ebre	<i>tEbre</i>
Terres de Lleida	<i>tLlei</i>
Val d'Aran	<i>vAran</i>
(unclassified)	<i>unCla</i>



3.2. Hosting websites selection

By means of a popular specialized search engine (BlogSearch.Google.com: “travel blog” OR “travel review”), on 2015-06-01, 14,700,000 indexed pages were initially obtained. The problem is that blogs come from diverse sources and websites, and do not have homogeneous structures, which makes it impossible to automatize the process of downloading, classification and refinement, as intended in this study.

Therefore, based on previous works, a group of websites hosting travel blogs and reviews (OTRs) with at least 100 entries about Catalonia during the studied period were selected. We also verify that the entries have a creation or modification date and that we can deduce the destination to which they refer. Entries focusing on hotels (Booking.com, Expedia.com, etc.) and restaurants were discarded because of their great volume and

specialization. Finally, eleven websites remained: GetJealous.com (GJ), MyTripJournal.com (MT), StaTravel.com (ST), TravBuddy.com (TY), TravelBlog.org (TB), TravelJournals.net (TJ), TravellersPoint.com (TS), TravelPod.com (TP), TripAdvisor.com (TA), Venere.com (VN), and VirtualTourist.com (VT).

Most authors consider that the hyperlink-based Google's PageRank (PR) algorithm is adequate to classify websites according to their levels of "prestige" or "authority" (Baggio and Corigliano 2009; Liu 2011; Ying et al. 2014) and researchers often use it to select the webpages more suitable for a case study (Law and Cheung 2010; Pan et al. 2007). However the PR is insufficient to arrange the webs for its low granularity and because it does not take into account the existing traffic (visitors, visits, etc.) on the website. As seen in Table 2, TA and VT have the same PR, while the other metrics point to a large difference between the two webhosts. Moreover, it is crucial to consider the volume of data that contains the website in relation with the period, region and/or topic studied.

Table 2

Webometrics of the top four websites hosting travel diaries (2015-05-31)

Q1

		TA	TB	TP	VT
Indexed pages	Google.com	144,000,000	453,000	334,000	554,000
	Bing.com	37,900,000	157,000	157,000	1050,000
Link-based rank	Google PR	7	6	6	7
	Yandex CY	1800	80	325	350
Visit-based rank	Compete.com	53	39,033	21,669	2162
	Quantcast.com	161	25,044	11,663	2798
	Alexa.com	192	44,789	30,595	5050
Size	Entries	119,016	3148	2375	7963
TBRH	Rank	1	3	4	2

Therefore, from the above websites, a ranking was built by applying the weighted formula "TBRH = 1 × B(V) + 1 × B(P) + 2 × B(S)" (Marine-Roig 2014), where 'B' corresponds to Borda's ordering method; 'V' to the visibility of the website (quantity and quality of inbound links); 'P', its popularity (received visits and traffic in general); and 'S', the size (number of entries related to the case study). Next the first four in the ranking were

selected (Table 2).

Previously, partial rankings are calculated with the same method:

- *Visibility (V)* = $1 \times B(\text{Google}) + 1 \times B(\text{Bing}) + 1 \times B(\text{PR}) + 1 \times (\text{CY})$. This is composed of four metrics without weighting: The indexed pages of the two search engines (Google and Bing) of major traffic in the West (Alexa.com: top sites on the web) and two well-known rankings based on the quantity and quality of incoming links, Google PageRank and Yandex citation index rank.
- *Popularity (P)* = $1 \times B(\text{Compete}) + 1 \times B(\text{Quantcast}) + 2 \times B(\text{Alexa})$. Alexa because its rank is based on world traffic, while the other two are limited to (Schmunk et al. 2014), but the sum of the two first ranks is equivalent to Alexa considering audience comes from the USA (Table 3).





Country	TA		TB		TP	
	Percent	Rank	Percent	Rank	Percent	Rank
 us	58.6	59	14.9	50,165	23.8	20,70
 in	1.9	1267	32.5	12,335	21.2	11,78
 uk	?	?	4.7	33,522	4.0	24,04
 ca	?	?	?	?	4.0	16,06

Table 3

Visitors by country (Alexa.com, 2015-05-31)

- *Size (S)* As stated previously, reviews concerning hotels and restaurants were not counted, nor were empty reviews (Schmunk et al. 2014) about Catalonia with title, date and author, but without content. For example, in the case of TA, more than 100,000 attraction reviews were empty and were removed. Also, more than 250,000 hotel reviews and more than 300,000 restaurant reviews were discarded. Finally, 132,502 relevant entries about Catalonia remained for subsequent processing. In Table 4, strong growth by TA, but a decline in TB and VT, can be observed. Notably, TP and VT were acquired a few years ago by TripAdvisor, Inc.

Table 4

Trends in web hosting

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
TA	38	81	117	204	608	1421	5933	28,387	36,045	46,142
TB	139	254	427	662	415	328	362	231	148	160
TP	100	236	276	258	226	238	218	189	346	259
VT	1498	1023	1031	762	413	398	635	306	251	172

The four selected websites are best suited to the case study according to Marine-Roig (2014). These hosting websites (TA, TB, TP, and VT) coincide with those selected by Lai and To (2015) to identify the destination image of Macao (China). Moreover, they represent a variety of trip diaries: TB and TP host travel blogs; TA hosts OTRs about hotels, restaurants and attractions; and VT, which is a virtual community, hosts travel pages, travelogues, and OTRs about hotels, restaurants, things to do, favourites, nightlife, off the beaten path activities, tourist traps, warnings or dangers, transportation, local customs, what to pack, shopping, and sports and outdoors.

3.3. Data downloading

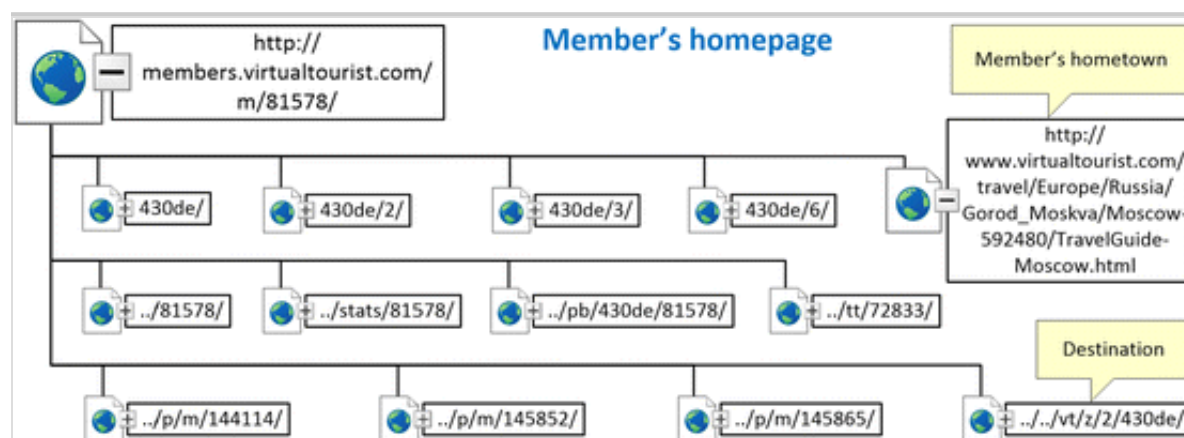
Liu (2011) considers that web mining, using data mining techniques intends to find useful information or to extract knowledge from the hyperlink structure and content of webpages. To automatize the process of extraction, first a Web crawler programme is needed, capable of roaming the hyperlink structure and downloading the linked webpages.

- *Web structure mining* The first step to download data is to navigate the selected websites manually to identify the initial pages, that is to say, those containing hyperlinks which lead to the individual blogs and OTR pages, and save their complete URLs. 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 hyperlinks. Figure 2 is constructed using a sitemap generator and only the hyperlinks related to the example of VT member 81,578 and the Barcelona destination remain. However, the only links that are used are member's profile (../81578/), travel blog (430de/),

travelogue (../tt/72833), review topic (430de/6/), travel review (../p/m/145865), and travel guide of member's hometown (see Fig. 4), because we have discarded the reviews on restaurants (../p/m/145852/& 430de/2/) and hotels (../p/m/144114/& 430de/3/), photos-backup (../pb/430de/81578/), and statistics (../stats/81578/).

Fig. 2

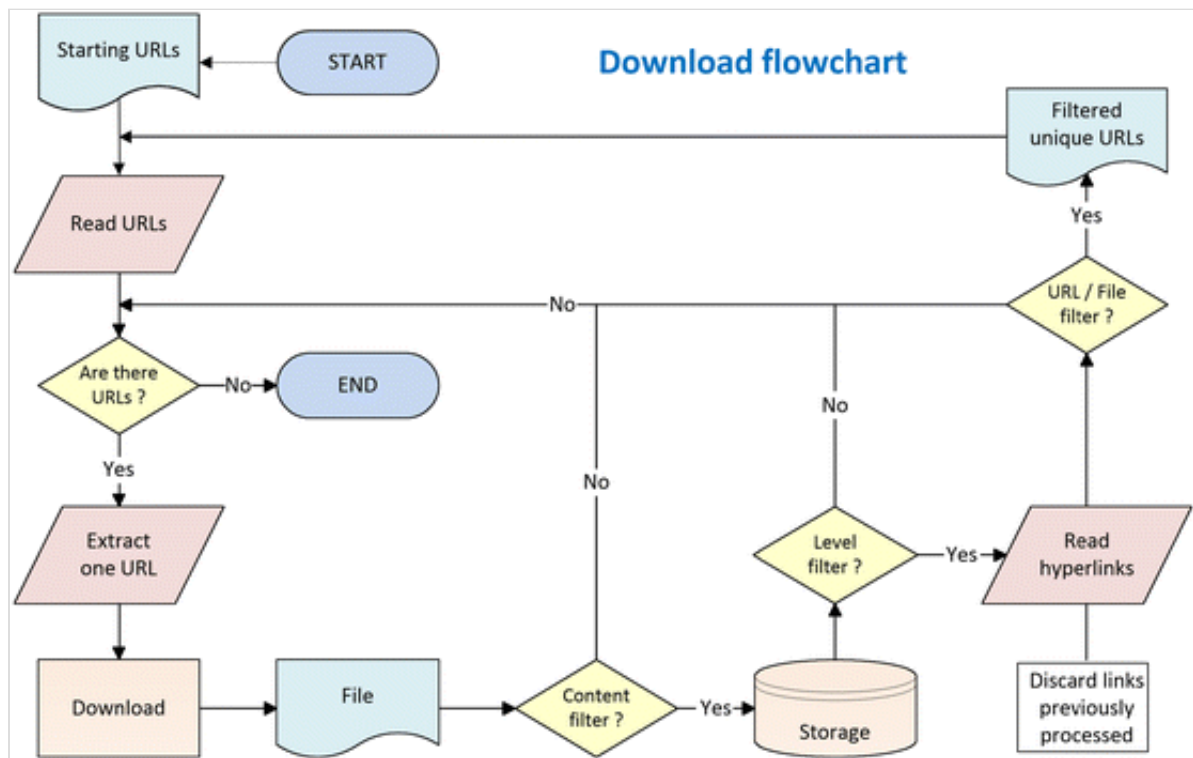
Simplified VT website-map diagram



- *Filter setting* Second, once the information derived from hyperlinks is obtained, a study about which are the most efficient filters (level, file type, URL, and content) to minimize download time and space used on the local disk should be conducted (Marine-Roig and Anton Clave 2015). Filters can be combined and, except for the level filter, they can be inclusive or exclusive: (1) a Level 0 filter only downloads the page indicated by the initial URL, a Level 1 filter, downloads that page and all the resources directly linked to it, etc. (2) The file type filter allows for downloads; for example, only HTML files and the remaining files (multimedia, PDF, etc.) will be visualized if an Internet connection is available. This system is ideal for analysing the textual content of diaries saving space in the local disk. (3) The URL filters allow for action on any part of it (protocol, server, domain, subdirectories or folders, filename and file type); and (4) the content filter is the least efficient, because it is necessary to download the page to assess whether or not it contains the chain of key characters, while with URL filters, only the pages of interest are downloaded (Fig. 3).

Fig. 3

Simplified flow diagram of the process of downloading travel diaries



- *TA* In the case of *TA*, all the files of interest contain the word *Catalonia*. Those which have hyperlinks that lead to OTRs start with *Attraction*, and those of the same OTRs start with *ShowUserReview*; therefore, a couple of inclusive filename filters are enough: *Attraction*Catalonia* and *ShowUserReview*Catalonia*. To understand the importance of the filters in this case, we should bear in mind that *TA* reached more than 225,000,000 reviews and opinions, and all its webpages are linked at different levels by hyperlinks.
- *TB* In this case, it is sufficient to place an inclusive folder filter: */Catalonia/* with no level limit to load files only within the start directory and below, because the server has a hierarchical territorial structure of folders to store the files and some excluded filename filters (keywords: *photo**, *map-**, *hotel**, *flight*...*) to prevent downloading pictures and maps, or information about hotels, flights, etc.
- *TP* This webhost does not have a classification that includes all the destinations in a region, and, therefore, we have to manually locate the start webpages of all Catalan destinations (Barcelona, Lloret de Mar, Salou, etc.).
- *VT* In the case of *VT* the process must be carried out in two phases.

Firstly, the start pages of destinations are downloaded with a “Level 1” filter. In this way, the member’s homepage and the destination’s travel page (Fig. 2) are obtained, which allows for extracting the travelogues and OTR URLs which correspond to said member and destination to download them in a second phase with an included content filter (keyword: *>destination name*). This filter prevents travel pages, travelogues and OTRs from getting mixed up with other destinations the user has visited, because all pages have the destination in the anchor text of the navigation menu bar.

- *Download process* Figure 3 shows schematically how a web copier works. All or some of the filters can be activated; if one of the filters in the diagram is not present, the process advances to the following step. In all cases, HTML files have been downloaded through a type filter (**.htm/*.html*) because they contain all the necessary information for content analysis, and the files associated to the webpage, such as pictures (Fig. 5), can be accessed with Internet connection. The flowchart is more complex when other types of files, such as images or multimedia, are downloaded because the hyperlinks of their respective downloaded webpages must be converted in order to point to their offline locations, and multimedia files occupy a lot of space on the local disk.

Firstly, the starting pages of Catalan destinations must be downloaded. If one downloaded page does not fulfil the content filter, the process downloads the following page. If it does fulfil it, the file is stored on the local hard disk and is processed. After checking the level filter, the hyperlinks on the stored page are read and URLs that have already been processed are excluded. Finally, if hyperlinks fulfil the URL, filename and file type filters, they are placed in a queue of files to be downloaded. The process continues until there is no file left in the queue.

Finally, once the web copier programme has been configured, we proceed with the massive download of the HTML pages of each website. In this research, we used the Offline Explorer Enterprise (OEE) application that has the capacity to download up to 100 million URLs per project, and offers the fastest possible multi-threaded processing of downloaded files

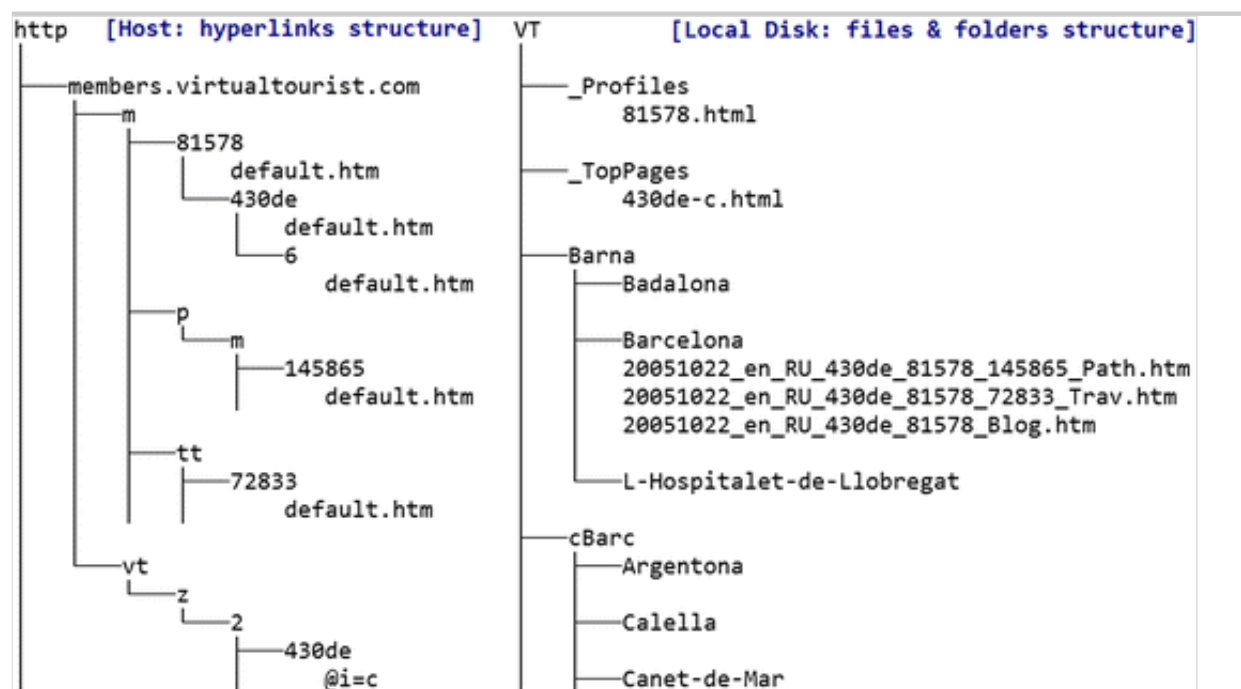
by using all central processing unit (CPU) cores (MetaProducts.com).

3.4. Data arrangement

To facilitate multiple classifications in a region divided into several territorial brands, we adopted the following structure for folders and files (Fig. 4):

Fig. 4

Files and folders of VT before and after the arrangement process



root\website\brand\destination\date_lang_isfrom_pagename_[theme].htm

A *Website* can consist of two-letter initials (Table 2); *Brand*, an abbreviation of five letters (Fig. 1); *Destination*, the name of the destination or if it is a composite name joined by a hyphen (e.g., Lloret-de-Mar); *Date* with the format *YYYYMMDD*, based on the ISO 8061 norm (International Standards Organization) to allow its alphabetical or numerical ordering; *Lang* two-letter code (ISO 639-1; e.g., ca: Catalan, en: English); *IsFrom* two-letter country code (ISO 3166-2; e.g., GB, United Kingdom, US, United States); *PageName* can contain a combination of codes and/or words; *Theme*, an abbreviation of four letters if the website has a thematic classification (e.g., hote: hotel, rest: restaurant, ToDo: things to do).

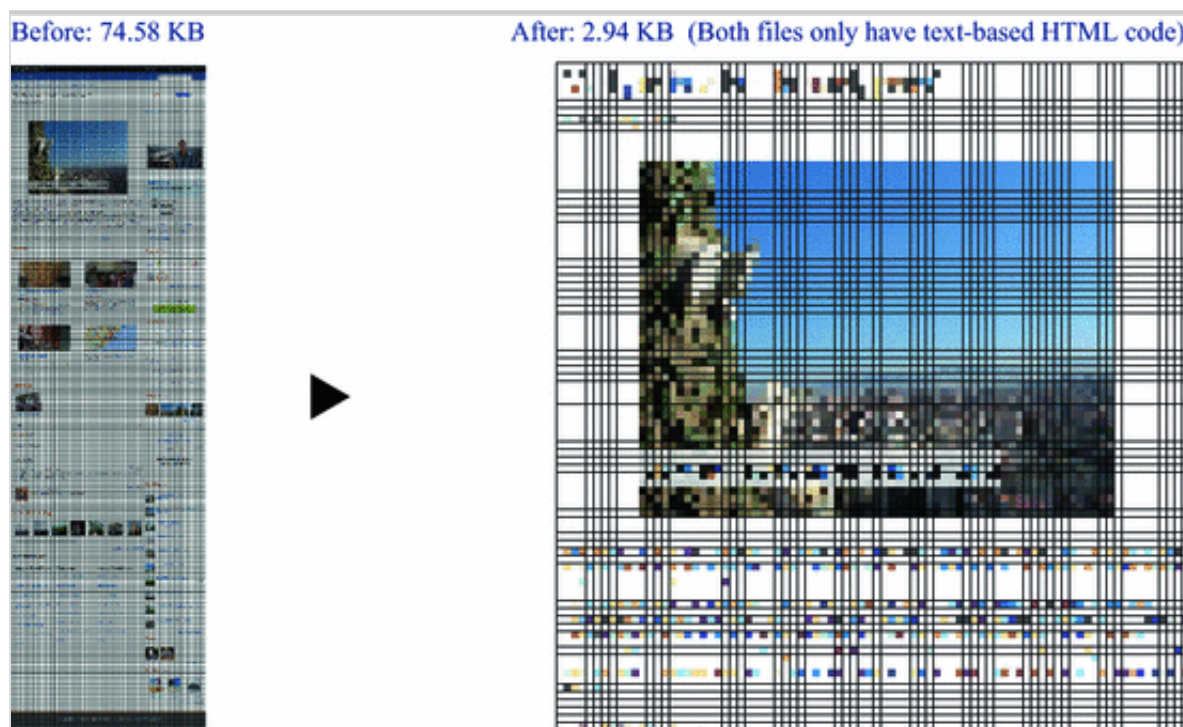
- *Geographic classification* This requires elaborate preparation. It

includes in a comma-separated values (CSV) file all the territorial information about the names and codes (if available) of the destinations. The CSV files are plain text and can be manipulated with a simple text editor or a spreadsheet.

- **Temporal classification**: Based on heterogeneous originals (TA: January 13, 2012; TB: April 24th 2004; TP: Saturday, August 24, 2013; and VT: Dec 6, 2004), dates are converted to the format *YYYYMMDD*.
- *Thematic classification* In TA and VT, there is a general classification of OTRs into hotels, restaurants, and things to do. VT has catalogued the tips: Path (off the beaten path), Shop (shopping), ToDo (things to do), Warn (warnings or dangers), etc. TA sub-classifies the things to do into types and/or categories, but a general thematic classification cannot be made, because the same attraction can belong to multiple categories or types.
- *Language detection*. This process is more complex: (1) it has to be based on textual content without structure (plain text) after the cleaning phase, because what the traveller has written only represents a minimal part of the page content (Fig. 5). (2) Specialized software is needed. In this study a Java programme (Marine-Roig 2013: Annex A3) based on the Language Detection Library (LDL) of N. Shuyo detects 53 languages, which can be extended with an included process. (3) In the case of Catalonia, the Catalan language Wikipedia (more than a million pages) was downloaded, and through a frequency analysis of one character and groups of two and three consecutive characters, the programme deduces the probability that such items appear at the beginning, end or within a word. (4) By means of this system, based on the Naive Bayes classifier, LDL detects each language with a great degree of precision (probability higher than 99 %). To extract textual content a free utility, HTML As Text (NirSoft.net), was used. After that, the language detection programme was run (Marine-Roig, 2013: Annex A3), which returns a CSV file with the code ISO 639-1 and the probability of success. Those diaries with a probability lower than 85 % were considered unclassified, because they usually have an insufficient quantity of text or are multilingual.

Fig. 5

Blurred VT travel page before and after the cleaning stage



Once the CSV files are ready, a batch programme (Marine-Roig 2013: Annex A3) is run for each website, which goes through all files, extracts internal data such as the date of the diary and the name of the destination, eliminates entries without narrative content (more than 100,000 OTRs in the case of TA), changes the format of such dates to *YYYYMMDD*, creates new territorial directories, and transfers the diary to the destination folder previously prepared with its articulated name to facilitate future classifications. Finally, the two-character ISO 639-1 codes are introduced in the name of the files, after the date (Fig. 4).

This organization of data allows obtaining any target subset by means of the utilities in the operating system. For example (see Fig. 4), we can select all trip diaries written in English by Russian tourists during 2005 using a simple expression with wildcards: `2005????_en_RU_*.htm` (*: any character zero or more times, ?: any character once).

3.5. Data cleaning

The data cleaning phase consists of eliminating all the *noise* surrounding what is written and posted by the author (Fig. 5). The original HTML format should be preserved in order to weight keywords and key phrases according to their emphasis or potential impact (Wahsheh et al. 2012).

Programmes such as *Site Content Analyzer* (SCA; CleverStat.com) take this format into account to calculate the frequency, site-wide density and average weight of keywords (Yadav and Yadav 2011).

Considering that the webpages of each site have a homogeneous structure and codification, elements that are not going to be used should be deleted. In general, we can eliminate some tags with their content: *meta* (metadata), *form* (forms), *iframe* (inline frames), *comment* (`<!-- comments in the source code -->`), and *script* (client-side scripts). With a web editor, such as *Microsoft Expression Web 4 (free version)*, all the superfluous HTML elements can be located manually, such as the *header* and *footer* sections (Fig. 5), which do not have a relationship with what *the* user has written, and list opening and closing HTML tags in a TBL file (generic Table file). Both the opening and closing tags as well as the text delimited by them have been recursively removed (i.e., tags and nested tags are removed) with an *ad hoc* programme (Marine-Roig 2013: Annex A3). The programme itself can serve to eliminate the additional OTRs of TA. This website adds some OTRs related with the review contained in the page, which already have their respective pages. To avoid this redundant information, the programme must delete them all except for the first.

AQ2

3.6. Data debugging

Although the analysed travel blogs and OTRs are written in English, there are proper nouns (destination and attraction names) with codified cedillas and accent marks, mistakes and translations. To understand how this step, along with the previous one, affects the quality of collected data, we can consider an example related to the case study: The Basilica (formerly Expiatory Temple/Church) of La *Sagrada Família* (i.e., ‘i’ with an acute accent in Catalan), so-called the “*unfinished cathedral*”, is an outstanding landmark of Catalonia (Marine-Roig 2015b), which is written in multiple ways by encoding problems (due to the acute accent), translations (Spanish: *Sagrada Familia* without accent; and English: Sacred/Holy Family), and misspellings (as a curiosity, we detected and amended more than 100 different ways of writing *Sagrada Família* incorrectly).

- *Character encoding problems* Problems appear with the previously mentioned characters, higher codes than ASCII 127 standard, because the website can codify the ASCII extended characters in three ways:

with an HTML entity (number or name) or with the webpage code (UTF-8 charset). For example, the surname of the Catalan architect Antoni Gaudí ends with an ‘í’ (i.e., ‘i’ with an acute accent), then ‘í’ (ASCII 237) can appear as an HTML number (Gaudí), HTML name (Gaudí), and UTF-8 (GaudÃ–).

- *Misspellings* The most common mistakes (MCM) are found in destination and attraction names (Cakmak and Isaac 2012), when there is no orthographic agreement between the Catalan phonemes and the English graphemes that represent them. Additionally, it is quite common for an English-speaking blogger to omit accents, for example, the surname mentioned in the previous paragraph would be written as Gaudi (without accent). By means of a preliminary frequency analysis we can locate the MCM to correct them in the webpage text or take them into account in the categories, e.g., Gaudi would have two keywords, one with an accent and another without.
- *Translations* It is also common to find proper nouns translated or semi-translated from Catalan to Spanish or English. For example, Parc Güell (‘u’ with dieresis) is an intensely visited attraction due to its World Heritage Site status (UNESCO 2005), and can be found written in multiple forms due to the dieresis (character encoding problem) and the translations (Spanish: Parque, English: Park). On the one hand, the MCM are corrected and on the other, relatively correct keywords are introduced (Parc Guell, Park Guell, Guell Park, etc.) along with the correct expression (Parc Güell) in the composite-words list and in its category (Tangible heritage).

Such codifications and MCMs distort content analysis and should be corrected. The UTF codes and HTML entities can be related in a CSV file with their corresponding Latin-1 character (ISO 8859-15), and the MCM and translations together with the correct word, and transfer it as a parameter to a search-and-replace utility to proceed with their replacement.

3.7. Content analysis

Once the previous phases have been performed, the travel blog and OTR dataset is ready for any kind of content analysis, qualitative or quantitative. In this research, travel diaries written in English have been selected and a

first offline analysis of the frequency, density and weight of keywords has been conducted, with the SCA programme. This software generates a CSV file for each travel blog or OTR conveying all the words appearing in that entry-file, their frequency, density and weight. It parses online and offline for keywords, suggests the most relevant and weighty phrases, and analyses link structure (CleverStat.com). SCA assigns a weight to keywords and key phrases according to their position within the webpage and the HTML tag that defines their format and features (Wahsheh et al. 2012; Yadav and Yadav 2011). The HTML tags in weight order are: first, the “title” (required in all HTML documents), then the headings (“h1”, “h2” and “h3”), followed by “a” (defines a hyperlink), “img alt” (alternate text for an image), then, with the same weight, come “h4” (heading), “b” (bold font), “i” (italic font), “u” (underline font), “strong” (defines important text), “em” (renders emphasized text), etc. Results can serve, for example, to study different tourist modes by grouping keywords into categories (sun, sea, and sand; urban environment; tangible heritage; etc.).

The SCA parser should be configured with the preferences for the case study before it is conducted. The most important ones are the *black list* and the *composite words*. The first prevents meaningless keywords (*stop words*) from being analysed, such as adverbs, articles, conjunctions, prepositions and pronouns (Cakmak and Isaac 2012; Krawczyk and Xiang 2015; Marine-Roig and Anton Clave 2015); moreover, fewer than three-letter words are dismissed. The second list indicates the word groups that form a unit (Cakmak and Isaac 2012; Li et al. 2015) such as *Sagrada Familia*.

4. Results

With the classification system seen in Sect. 3.4 it is easy to analyse the trends followed by the 132,502 travel blogs and OTRs along the 11 years of study (Table 5). In this table, 434 unclassified entries do not appear because the travellers put Catalonia as the destination without specifying brand, city or town.

Table 5

Trends in Catalan tourist brands

	Barna	cBarc	cBrav	cDaur	pBarc	Pyren	tLlei	tEbre	vAra
2004	1177	34	201	61	57	10	6	4	1

2005	1374	42	204	46	45	20	1	3	0
2006	1191	53	163	82	38	12	1	0	7
2007	1309	70	238	117	37	25	3	1	0
2008	1367	79	191	134	45	8	5	1	0
2009	1295	34	134	121	20	10	5	5	3
2010	1742	63	177	288	35	22	11	1	2
2011	5828	115	332	698	89	14	19	2	3
2012	24,211	325	1448	2599	412	62	16	3	9
2013	30,875	560	1707	2498	927	149	16	10	11
2014	40,232	745	1928	3070	517	144	37	24	2

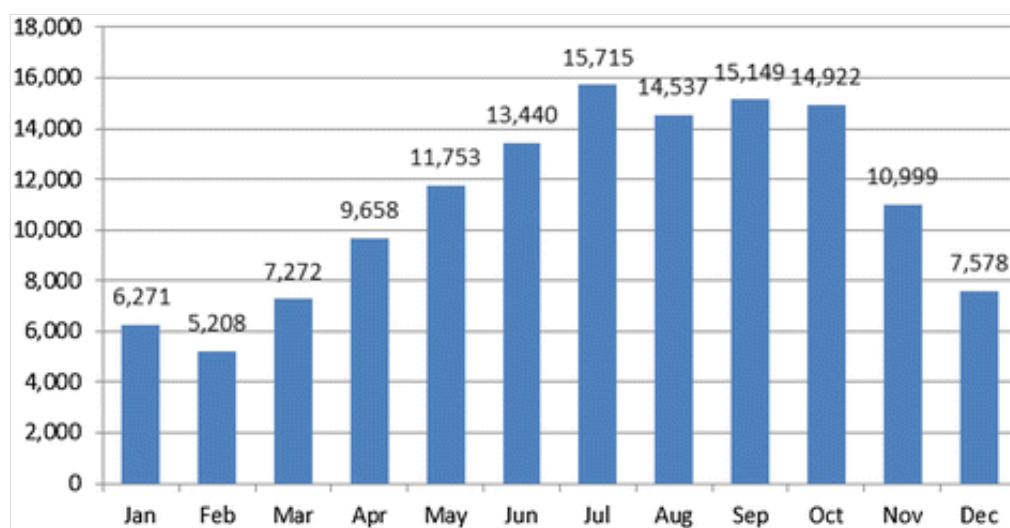


Concerning the distribution by Catalan brands, we observe that the Barcelona brand (*Barna*) far outnumbers all the rest and the number of travel diaries has increased 23 fold in the last 5 years. Additionally, we can observe that the Terres de Lleida (*tLlei*), Terres de l'Ebre (*tEbre*) and Val d'Aran (*vAran*) brands do not have enough entries to sustain reliable content analysis. Comparing both major coastal destinations, it can be seen that Costa Daurada (*cDaur*) has grown considerably in recent years ahead of Costa Brava (*cBrav*). Finally, a decrease was observed during 2014 in travel diaries on Barcelona Landscapes (*pBarc*).

With respect to the problem of the seasonality of the tourist industry in the Spanish coastal regions, consisting of the concentration of *sun, sea and sand* tourists in the months of July and August, Fig. 6 show that high season for the type of tourism studied extends to almost 6 months.

Fig. 6

Monthly distribution of travel blogs and OTRs (132,502 from TA, TB, TP, & VT)



As an example of the validity, usability and capacity of the method, a simple first level of content analysis using SCA was performed with 131,223 travel diaries written in English (99.03 % of the 132,502 gathered). As illustrated in Table 6, very significant results were obtained. Coinciding with the results in Table 5 concerning Barcelona brand (*Barna*), it can be observed in Table 6 that the keyword *Barcelona* has a much higher density than the other ones and a considerable weight. Barcelona is a leading smart city (Marine-Roig and Anton Clave 2015) and the sixth most powerful city brand in the world (Michael 2014). The Catalan Tourist Board is aware of the significance of the Barcelona brand and recently renamed the “Central Catalonia” brand “Barcelona landscapes” (*pBarc*) (Datzira-Masip and Poluzzi 2014). Among the top 25 keywords, we can find eight good feelings, as well as the architect Antoni Gaudi and two of his masterpieces (*Sagrada Familia* and *Parc Guell*) with a highly significant weight. Gaudi’s work is registered in UNESCO’s World Heritage List (UNESCO 2005). A new record of visitors (3260,880) to the *Sagrada Familia* basilica was reached in 2014, of which 12 % were North-American, 10 % French, 6 % Italian, and 6 % English (Marine-Roig 2015b).

Table 6

The 25 most frequent words of 93,765 unique words

Rank	Keyword	Count	Sitewide density (%)	Average weight	Remarks
1	B a rcelona	285,614	2.67	59.66	Capital of Catalonia
2	T o ur	117,845	1.10	33.62	

3	GREATgreat	81,265	0.76	24.17	Good feeling
4	Ssagrada familia	56,133	0.52	65.77	Gaudi's masterpiece
5	Ggaudi	45,420	0.42	19.41	Catalan Architect
6	Vvisit	42,001	0.39	15.16	
7	Pplace	40,329	0.38	16.40	
8	Aamazing	40,076	0.38	24.90	Good feeling
9	Ecity	39,373	0.37	12.12	
10	Ggood	38,753	0.36	15.24	Good feeling
11	Bbeautiful	36,588	0.34	23.84	Good feeling
12	Ppark	36,277	0.34	28.56	
13	Wway	33,614	0.31	15.82	
14	Nnice	29,774	0.28	20.23	Good feeling
15	Gguell park	29,555	0.28	64.07	Gaudi's work
16	Bbest	27,722	0.26	24.98	Good feeling
17	Mmuseum	27,194	0.25	31.06	
18	Eexperience	25,554	0.24	22.69	
19	Gguide	24,183	0.23	10.70	
20	Wwalking	23,573	0.22	36.23	
21	Ppeople	23,502	0.22	4.26	
22	Wwalk	22,169	0.21	10.79	
23	Ffun	21,940	0.21	24.51	Good feeling
24	Ttrip	21,515	0.20	14.61	
25	Iinteresting	19,980	0.19	17.23	Good feeling

These results obtained by SCA can serve for different studies such as grouping keywords into thematic categories, how to ascertain the weight of a certain tourist modality, attraction factors, feelings, dichotomies, etc. in the whole region and in sub-regions. The same studies can be conducted on the Catalan brand or at a city level (Marine-Roig and Anton Clave 2015) and circumscribed to a certain period of time by applying SCA programme

to the corresponding HTML subset.

Furthermore, these initial results give insights into the cognitive/functional and affective components of destination image. It is especially noteworthy that the sum of almost half a million words (492,780) related to good feelings (according to standard lists of positive adjectives in American and British English) present in the top 25 words those users mention. This gives a good indication of the positiveness in general of the affective component, which may have strong implications for tourist satisfaction and destination loyalty. In terms of the cognitive or functional component of image we get insights into the main destinations in the region (Barcelona), the most popular attractions (e.g., *Sagrada Familia*) and the main activities (e.g., tour, city, walk, and trip) conducted.

Moreover, in terms of the relationship of cognitive and affective image components, and their relative weight, this could provide an interesting overview showing that cognitive elements account for about 58 % of the most mentioned words (697,169 mentions), and elements related to the affective component account for approximately 42 % of UGC text in this case.

5. Concluding remarks

The proposed methodology facilitates the massive gathering of online UGC data from the most suitable sources for a specific case study. The hierarchical territorial structure of folders and the articulation of the name of the files which contain individual diaries, enable multiple classifications using utilities to order and manipulate the files of the same operating system. This structure also allows us to focus the analysis on a specific place, period, language or subject (if this is available), in particular or combined, selecting the corresponding subset or a random sample thereof. The cleaning and debugging phases are essential to obtaining quality information, limited to the web content as written and posted by the diary author, and overcoming the most significant errors.

Quantitative analysis results, at the level of territorial brands, may be useful for NTOs to improve their branding and positioning policies; e.g., for the managers of the Costa Daurada brand, it is interesting to know that tourists mention a theme park (PortAventura) ten times more than the Roman amphitheatre in spite of the fact that it is a World Heritage Site

(UNESCO 2000). Metrics (visibility, popularity, audience, and size) of webs hosting UGC data can be useful for the DMOs that want to promote products or services on such websites.

This study aims to contribute to the advancement of research in the field of travel, tourism and hospitality, because it proposes a methodology to extract knowledge from large amounts of online UGC data with a degree of detail that was not covered in previous works. At a first level of analysis, this methodology gives significant insights concerning both the cognitive and the affective components of destination image and may be useful for further analyses in this respect. Most of the proposed phases of this method are useful for the content analysis of other web data sources. The HTML dataset is prepared for any offline content analysis in future work, for instance, a qualitative content analysis using qualitative data analysis software or a sentiment analysis using a computer-aided text-analysis programme.

The main limitation of the application of the method to this case study was the analysis at the territorial boundary brands and, more specifically, a certain influence of the Barcelona brand on the two brands surrounding it: Barcelona coast and Barcelona landscapes; that is, in some trip diaries of *cBarc* and of *pBarc* they find references to Barcelona's attraction factors. Moreover, the implementation of some pre-processing steps requires computer skills that are not available to some DMOs and other stakeholders.

Acknowledgments

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

References

Abburu S, Babu GS (2013) A framework for web information extraction and analysis. *Int J Comp Technol* 7:574–579

Albastroiu I, Felea M (2014) The implications of user-generated content websites for tourism marketing. *Int J Eco Pract Theor* 4:222–229

Amaral F, Tiago T, Tiago F (2014) User-generated content: tourists'

profiles on TripAdvisor. *Int J Strategic Innov Marketing* 1:137–147. doi:10.15556/IJSIM.01.03.002

Anton Clavé S, Gonzalez F (2008) *A proposito del turismo: La construccion social del espacio turistico* [About tourism: The social construction of tourist space]. Editorial UOC, Barcelona

Baggio R, Corigliano MA (2009) On the importance of hyperlinks: A network science approach. In: Hopken W, Gretzel U, Law R (eds), *Information and Communication Technologies in Tourism 2009*, Springer, Austria, pp 309–318. doi: 10.1007/978-3-211-93971-0_26

Banyai M, Glover TD (2012) Evaluating research methods on travel blogs. *J Travel Res* 51:267–277. doi:10.1177/0047287511410323

Cakmak E, Isaac RK (2012) What destination marketers can learn from their visitors' blogs: an image analysis of Bethlehem, Palestine. *J Destination Marketing Manag* 1:124–133. doi:10.1016/j.jdmm.2012.09.004

Chau M, Xu J (2012) Business intelligence in blogs: understanding consumer interactions and communities. *MIS Q* 36:1189–1216

CTB (2015) *Press Pack' 15*. Catalan Tourist Board. <http://www.act.cat/press-pack/?lang=en>. Accessed 31 July 2015

Datzira-Masip J, Poluzzi A (2014) Brand architecture management: the case of four tourist destinations in Catalonia. *J Dest Market Manag* 3:48–58. doi:10.1016/j.jdmm.2013.12.006

Dwyer L, Gill A, Seetaram N (eds) (2014) *Handbook of research methods in tourism: quantitative and qualitative approaches*, Reprint edn. Edward Elgar Publishing, Cheltenham

Eurostat (2014) **45** Tourism. Eurostat regional yearbook 2014**45**. Publications Office of the European Union, Luxembourg, pp 18**79**–2**1006**

Fang B, Ye Q, Kucukusta D, Law R (2016) Analysis of the perceived

value of online tourism reviews: influence of readability and reviewer characteristics. *Tour Manag* 52:498–506.
doi:10.1016/j.tourman.2015.07.018

Gonzalez R (2010) Estudi exploratori de les marques turístiques de Catalunya a la Web 2.0 [Exploratory study of tourist brands of Catalonia to Web 2.0]. Viviential Value, Barcelona, Catalonia

Guo L, Li Z, Sun W (2015) Understanding travel destinations from structured tourism blogs. *WHICEB 2015 Proceedings*, pp 144–151.
<http://aisel.aisnet.org/whiceb2015/80>. Accessed 31 Mai 2015

Gursoy D, Uysal M, Sirakaya-Turk E, Ekinci Y, Baloglu S (2015) Handbook of scales in tourism and hospitality research. CABi, Wallingford

Johnson PA, Sieber RE, Magnien N, Ariwi J (2012) Automated web harvesting to collect and analyse user-generated content for tourism. *Curr Issues Tourism* 15:293–299. doi:10.1080/13683500.2011.555528

Kladou S, Mavragani E (2015) Assessing destination image: an online marketing approach and the case of TripAdvisor. *J Destination Market Manag.* 4:187-193 doi:10.1016/j.jdmm.2015.04.003 (in press)

Koltringer C, Dickinger A (2015) Analyzing destination branding and image from online sources: a web content mining approach. *J Bus Res* 68:1836–1843. doi:10.1016/j.jbusres.2015.01.011

Kotler P, Haider DH, Rein I (1993) Marketing places: Attracting investment, industry and tourism to cities, states and nations. The Free Press, New York

Krawczyk M, Xiang Z (2015) Perceptual mapping of hotel brands using online reviews: a text analytics approach. *Inf Technol Tourism*. doi:10.1007/s40558-015-0033-0 (in press)

Lai LSL, To WM (2015) Content analysis of social media: a grounded theory approach. *J Electronic Commerce Res* 16:138–152.
<http://www.jecr.org/node/466>. Accessed 31 May 2015

Law R, Cheung S (2010) The perceived destination image of Hong Kong as revealed in the travel blogs of mainland Chinese tourists. *Int J Hospitality Tour Administration* 11:303–327.

doi:10.1080/15256480.2010.518521

Li YR, Lin YC, Tsai PH, Wang YY (2015) Traveller-generated contents for destination image formation: mainland China travellers to Taiwan as a case study. *J Travel Tourism Marketing* 32:518–533.

doi:10.1080/10548408.2014.918924

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

Llodra-Riera I, Martinez-Ruiz MP, Jimenez-Zarco AI, Izquierdo-Yusta A (2015) A multidimensional analysis of the information sources construct and its relevance for destination image formation. *Tour Manag* 48:319–328. doi:10.1016/j.tourman.2014.11.012

Lu W, Stepchenkova S (2015) User-generated content as a research mode in tourism and hospitality applications: topics, methods, and software. *J Hospitality Marketing Manag* 24:119–154.

doi:10.1080/19368623.2014.907758

Marchiori E, Onder I (2015) Reframing the image of a destination: A pre-post study on social media exposure. In: Tussyadiah I, Inversini A (eds), *Information and Communication Technologies in Tourism 2015*, Springer, Switzerland, pp 335–347. doi:10.1007/978-3-319-14343-9_25

Marine-Roig E (2013) *From the projected to the transmitted image: the 2.0 construction of tourist destination image and identity in Catalonia*. PhD dissertation. <http://hdl.handle.net/10803/135006>. Accessed 31 May 2015

Marine-Roig E (2014) A webometric analysis of travel blogs and reviews hosting: the case of Catalonia. *J Travel Tourism Marketing* 31:381–396. doi:10.1080/10548408.2013.877413

Marine-Roig E (2015a) Identity and authenticity in destination image construction. *Anatolia Int J Tourism and Hospitality Res* 26:574–587.

doi:10.1080/13032917.2015.1040814

Marine-Roig E (2015b) Religious tourism versus secular pilgrimage: The basilica of La Sagrada Familia. *Int J Religious Tourism Pilgrimage* 3:25–37. <http://arrow.dit.ie/ijrtp/vol3/iss1/5/>. Accessed 31 July 2015

Marine-Roig E, Anton Clave S (2015) Tourism analytics with massive user-generated content: a case study of Barcelona. *J Destination Marketing Manag* 4:162-172. doi:10.1016/j.jdmm.2015.06.004 (**in press**)

Michael C (2014, May 6) From Milan to Mecca: the world's most powerful city brands revealed. *TheGuardian*, News, Cities, City brand. <http://www.theguardian.com/cities/gallery/2014/may/06/from-milan-to-mecca-the-worlds-most-powerful-city-brands-revealed>. Accessed 31 Mai 2015

Minazzi R (2015) *Social media marketing in tourism and hospitality*. Springer, Cham

Moens MF, Li J, Chua TS (eds) (2014) *Mining user generated content*. CRC Press, Boca Raton

Munar AM (2012) Social media strategies and destination management. *Scand J Hospitality Tourism* 12:101–120. doi:10.1080/15022250.2012.679047

Munar AM, Jacobsen JS (2013) Trust and involvement in tourism social media and web-based travel information sources. *Scand J Hospitality and Tourism* 13:1–19. doi:10.1080/15022250.2013.764511

Pan B, MacLaurin T, Crofts JC (2007) Travel blogs and the implications for destination marketing. *J Travel Res* 46:35–45. doi:10.1177/0047287507302378

Schmunk S, Hopken W, Fuchs M, Lexhagen M (2014) Sentiment analysis: Extracting decision-relevant knowledge from UGC. In: Xiang Z, Tussyadiah L (eds), *Information and communication technologies in tourism 2014*, Springer, Switzerland, pp 253–265. doi:10.1007/978-3-

319-03973-2_19

Schuckert M, Liu X, Law R (2015a) Hospitality and tourism online reviews: recent trends and future directions. *J Travel Tourism Marketing* 32:608–621. doi:10.1080/10548408.2014.933154

Schuckert M, Liu X, Law R (2015b) Insights into suspicious online ratings: direct evidence from TripAdvisor. *Asia Pacific J Tourism Res.* doi:10.1080/10941665.2015.1029954 (in press)

Serna A, Marchiori E, Gerrikagoitia JK, Alzua-Sorzabal A, Cantoni L (2015) An auto-coding process for testing the cognitive-affective and conative model of destination image. In: Tussyadiah I, Inversini A (eds), *Information and Communication Technologies in Tourism 2015*, Springer, Switzerland, pp 111–122. doi: 10.1007/978-3-319-14343-9_9

Sigala M, Christou E, Gretzel U (eds) (2012) *Social media in travel, tourism and hospitality: Theory, practice and cases*. Ashgate Publishing, Surrey

Standing C, Tang-Taye JP, Boyer M (2014) The impact of the Internet in travel and tourism: a research review 2001–2010. *J Travel Tourism Marketing* 31:82–113. doi:10.1080/10548408.2014.861724

UNESCO (2000) Archaeological ensemble of tarraco. world heritage list. <http://whc.unesco.org/en/list/875>. Accessed 31 May 2015

UNESCO (2005) Works of Antoni Gaudi. world heritage list. <http://whc.unesco.org/en/list/320>. Accessed 31 May 2015

Wahsheh HA, Alsmadi IM, Al-Kabi MN (2012) Analyzing the popular words to evaluate spam in Arabic web pages. *Res Bull Jordan ACM* 2:22–26

Wang Y, Morais DB (2014) An examination of tourists' identity in tourist weblogs. *Inf Technol Tourism* 14:239–260. doi:10.1007/s40558-014-0016-6

Wang Y, Chan SC, Ngai G, Leong HV (2013) Quantifying reviewer

credibility in online tourism. In: Decker H et al (eds) DEXCA 2013 Proceedings of 24th International Conference: Database and Expert Systems Applications. Czech Republic, Prague, pp 381–395

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: Tussyadiah I, Inversini A (eds), Information and Communication Technologies in Tourism 2015, Springer, Switzerland, pp 33–45. doi:10.1007/978-3-319-14343-9_3

Yadav Y, Yadav PK (2011) Site Content Analyzer in context of keyword density and key phrase. *Int J Comp Technol Appl* 2:860–872

Ying T, Norman WC, Zhou Y (2014) Online networking in the tourism industry: a webometrics and hyperlink network analysis. *J Travel Res*. doi:10.1177/0047287514532371

Zeng B, Gerritsen R (2014) What do we know about social media in tourism? A review. *Tourism Manag Persp* 10:27–36. doi:10.1016/j.tmp.2014.01.001