Analysis and Visualisation of Crowd-sourced Tourism Data

Fátima Leal
EET, University of Vigo, Spain
INESC TEC, Porto, Portugal
fatimaleal2@gmail.com

Joana Matos Dias
FEUC, University of Coimbra
INESC Coimbra, Portugal
joana@fe.uc.pt

Benedita Malheiro
ISEP, Porto, Portugal
INESC TEC, Porto, Portugal
mbm@isep.ipp.pt

Juan Carlos Burguillo
EET, University of Vigo, Spain
AtlanTIC, Vigo, Spain
J.C.Burguillo@uvigo.es

ABSTRACT

The tourist behaviour has changed significantly over the last decades due to technological advancement (*e.g.*, ubiquitous access to the Web) and Web 2.0 approaches (*e.g.*, Crowdsourcing). Tourism Crowdsourcing includes experience sharing in the form of ratings and reviews (evaluation-based), pages (wiki-based), likes, posts, images or videos (social-network-based). The main contribution of this paper is a tourist-centred off-line and on-line analysis, using hotel ratings and reviews, to discover and present relevant trends and patterns to tourists and businesses. On the one hand, on-line, we provide a list of the top ten hotels, according to the user query, ordered by the overall rating, price and the ratio between the positive and negative Word Clouds reviews. On the other hand, off-line, we apply Multiple Linear Regression to identify the most relevant ratings that influence the hotel overall rating, and generate hotel clusters based on these ratings.

Keywords

Data Mining, Crowdsourcing Analysis, Travel Planning

1. INTRODUCTION

Travelling changed dramatically in the last decades due to the evolution and popularisation of information and communication technologies as well as mobile devices. In particular, tourists constantly share on-line information regarding their travel experiences through ratings, reviews, comments, photos or videos. Therefore, this paper presents a tourist-centred analysis of crowd-sourced data concerning hotels, including both on-line and off-line processing.

The on-line module identifies, based on the user queries, the most relevant hotels using the hotel stars, `Overall` rating and text reviews provided by other tourists. First, for each filtered hotel, we build and scale the positive and negative reviews by means of Word Clouds. Then, we provide to the user a list with the top ten hotels ordered by decreasing hotel `Overall` rating, `Price` and ratio between the positive and negative review Word Clouds.

The off-line module, first performs a correlation analysis between variables, and applies a Multiple Linear Regression (MLR) to the data set in order to find relations among ratings, while retaining the existing data trends (regarding the ratings and reviews). Second it uses clustering to analyse the trend between the hotel ratings and the `Price`.

The main contribution of this work is a crowd-sourced data processing methodology to support travel planning which forecasts relevant rating trends (off-line analysis of TripAdvisor hotel ratings), and to visualise tourist reviews (on-line processing and visualisation of Expedia hotel reviews).

This paper is organised as follows. Section 2 reviews related work on analysis and visualisation of crowd-sourced data. Section 3 introduces tourism data analytics, describing current techniques and trends. Section 4 describes the implemented on-line visualisation of tourism crowd-sourced data. Section 5 the adopted off-line processing. Finally, Section 6 provides the conclusions and discusses the results.

2. RELATED WORK

Crowdsourcing, which was introduced in 2006 by Jeff Howe [10], is a process of getting work done by a crowd of people, *i.e.*, corresponds to any collective and collaborative activity performed by a large number of volunteers with the support of information and communication technologies. Individuals can play two different roles within Crowdsourcing: *task requester* or *task worker* [1]. The crowd wisdom is very influential in the tourism domain. While tourists make decisions based on crowd-sourced reviews and ratings, tourism businesses, according to Sigala [14], regard the tourist crowd know-how as a valuable contribution for personalised marketing. In addition, Crowdsourcing, while a continuous source of tourist-generated, shared and maintained data, promotes intangible tourism experiences [15].

In terms of tourism data visualisation, there are maps, opinion wheels, tag clouds, bubble trees, tree maps, *etc.* Bjørkelund *et al.* [2] and Marchetti *et al.* [12] present a review opinion mining analysis to provide a map-based visualisation of sentiments expressed in the reviews and, thus, allow tourists to identify promising areas. Wu *et al.* [22] propose *OpinionSeer* – a system which enables an interactive radial visualisation, including a tag cloud and opinion wheel, of hotel reviews. Carvalho & Chaves [5] [4] follow an

C3S2E '16, July 20-22, 2016, Porto, Portugal

 ISBN 978-1-4503-4075-5/16/07$15.00

DOI: http://dx.doi.org/10.1145/2948992.2949008

identical approach, *i.e.*, analyse hotel reviews using relevant adjectives through a concept ontology and provide three visualisation techniques: a bubble tree, tree map and plot visualisation. Colantonio *et al.* [6] propose a matrix-based visualisation approach reusing the Access Data visualiseR (ADVISER) algorithm [7] in order to identify singularities or trends.

The processing of reviews includes multiple techniques, *e.g.*, Natural Language Processing (NPL), Clustering, Regression, *etc.* Suzuki *et al.* [17] use NPL to analyse word-of-mouth communication regarding tourist hotel experiences. Fang *et al.* [8] and [9] Han *et al.* perform a regression analysis concerning textual reviews.

These works focus on the analysis and the visualisation of on-line reviews or ratings. However, this related work does not combine the analysis of both crowd-sourced data (ratings and reviews) in order to find trends and patterns. We, in this paper, combine both analyses (ratings and reviews) with an on-line and off-line approach aiming to support tourists as well as businesses.

3. TOURISM DATA ANALYTICS

The tourist behaviour has changed dramatically in recent years. On the one hand, technology provides ubiquitous access to endless collections of tourism-related Web Services and, on the other hand, tourists generate through Crowdsourcing systems large volumes of tourism-related data in the form of ratings and reviews. The number of tourism Crowdsourcing systems on Web has been growing, including dedicated map-based (OpenStreetMap), wiki-based (Wikivoyage), evaluation-based (TripAdvisor or Yelp) tourism platforms or general purpose social networks (Facebook).

For tourism businesses, data analytics identifies important travel patterns and, thus, empowers businesses with the ability to enhance and personalise the customer travel experience. The power to analyse, find and visualise the highlights underlying tourism-related crowd-sourced data offers businesses and tourists an insight into existing market opportunities. Crowdsourcing has marketing impact since tourists act both as clients and marketers [14], allowing businesses to use the user-generated content to promote and re-adapt their tourism offers [15].

In the tourism domain, Data Mining has been used to detect tourist preferences, frequent behaviours, new trends or contexts from tourism-related datasets. Data Mining is the process of discovering unidentified patterns and properties in large data repositories, *i.e.*, corresponds to the discovery stage of data analysis. It often incorporates and uses machine learning techniques in the process, *e.g.*, unsupervised machine learning. Cabena *et al.* [3] defines Data Mining as an interdisciplinary field, which gathers mechanisms from machine learning, pattern recognition, statistics, databases and visualization to extract useful information from large repositories. The information needs to be gathered, cleaned, analysed, interpreted and evaluated in order to find useful data. The most popular Data Mining techniques found in the literature [11] include Classification, Clustering as well as Forecasting [13] algorithms. In this work we use Clustering and Regression together with Word Cloud visualisation. The entire data processing was implemented in Python, using the *scikit-learn*[1] and the *word cloud*[2] library.

4. ON-LINE ANALYSIS

The majority of tourism Crowdsourcing platforms enables the introduction of textual reviews regarding tourism resources and tourist experiences. Since reviews influence the tourist decision making process and, thus, have an impact on tourism businesses, it is important to gather, analyse and visualise crowd-sourced tourism data. However, the dimension of the data available, which is beyond human processing capabilities, leads to the application of Data Mining (text mining) and Big Data visualisation techniques. This section analyses crowd-sourced information from Expedia platform since TripAdvisor does not provide a free Application Programming Interface (API). Therefore, we present an on-line module for helping the tourist to visualise the third party experiences and evaluations of hotels from Expedia, using Data Mining techniques. This on-line module provides the most influential hotels of a location according to the following crowd-sourced information: (*i*) hotel stars; (*ii*) `Overall` rating of hotel; (*iii*) positive Word Cloud; and (*iv*) negative Word Cloud.

Expedia API. Expedia provides a set of public API[3] which allows the access to Expedia real-time information regarding accommodations and attractions, comprising `location`, `name`, `reviews`, `ratings`, `price`, *etc.* Our on-line module uses the *Natural Language Hotel Search*, *Hotel Reviews* and *Hotel Search, Offers and Info* API in order to obtain on-line information from Expedia and thus, perform real time data analysis.

Processing. The on-line module – governed by Algorithm 1 – returns the top 10 hotel based on the hotel price, ratings and reviews according to the specified user query. It creates: (*i*) the scaled positive (+) and negative (-) review Word Clouds for each hotel, where the scale represents their relative dimension; (*ii*) the interactive `Overall` *versus* `Price` chart regarding the filtered hotels; and (*iii*) a list of the top 10 hotel ordered by `Overall` rating, `Price` and the ratio between the scales of the + and - review Word Clouds.

Algorithm 1 On-line Processing

Inputs	Hotel stars `Overall` rating Textual reviews
Outputs	Top 10 hotels ordered by decreasing `Overall` rating, `Price` and ratio of + and - reviews Interactive `Overall` rating *vs* `Price` hotel chart
	Individual hotel analysis
Step 1	Visualise stars and `Overall` rating
Step 2	Remove irrelevant words from reviews
Step 3	Generate and scale the + and - review clouds
Step 4	Visualise the scaled + and - review clouds
	Interactive hotel chart
Step 5	Build the interactive `Overall` rating *vs* `Price` chart
	Top 10 hotel list
Step 6	Order by decreasing `Overall` rating, `Price` and ratio of + and - reviews

[1] http://scikit-learn.org
[2] https://pypi.python.org/pypi/wordcloud
[3] http://hackathon.expedia.com/directory

5. OFF-LINE ANALYSIS

The off-line processing was performed with TripAdvisor data and involved the processing of hotel ratings and reviews using: (*i*) MLR for rating analysis; and (*ii*) Clustering to compare hotel ratings and price. The goal of this off-line analysis is to verify how the price and the overall rating are related with the crowd-sourced partial ratings.

TripAdvisor Data Set. TripAdvisor is a powerful platform, containing a huge volume of crowd-sourced opinions in the form of hotel ratings and reviews. The platform enables the assessment of hotel resources according to the described features. The selected data set, which was retrieved from University of Illinois at Urbana-Champaign [21], contains metadata regarding reviewers, ratings and hotels from TripAdvisor. It was chosen due to its lower data sparsity and big size. The reviewers data encompass `AuthorLocation`, review `Title`, `Author`, `ReviewID`, review textual `Content` and review `Date`. In terms of ratings, the data set provides information about nine features in a scale from 0 to 5: `Overall` (O), `Value` (V), `Rooms` (R), `Location` (L), `Cleanliness` (C), `Check in/front desk` (CI), `Service` (S), `Sleep Quality` (SQ) and `Business Service` (BS). Finally, the hotel data includes the `Name`, `HotelURL`, `Price`, `Address`, `HotelID` and `ImgURL`. All text reviews, aspect segmented reviews and vectors are organised by `HotelID` in different JavaScript Object Notation (JSON) files. The data set contains 235 793 hotel reviews from February 14, 2009 to March 15, 2009.

Rating Analysis. The off-line module – summarised by Algorithm 2 – implements the off-line rating analysis. First, to analyse the ratings relations, it performs a correlation analysis between the different rating features and the `Overall` rating. Then, applies the MLR to validate the results of the first step and to identify the ratings with more influence in the `Overall` hotel rating. Finally, it builds clusters to group the results of MLR analysis based on the identified hotel ratings.

Algorithm 2 Off-line Processing

Inputs	`Overall` (dependent variable), `Value`, `Rooms`, `CheckIn`, `Cleanliness`, `Location`, `Service`, `Business` and `SleepQuality` ratings (independent variables)
Outputs	MLR of the `Overall` rating Clusters based on ratings (`O`, `C`, `R`, `S`, `V`) and `Price`
Step 1	Calculate the correlation between independent variables
Step 2	Apply MLR regression to the variables more correlated with `Overall` (`Cleanliness`, `Rooms`, `Service` and `Value`)
Step 3	Calculate the ranting and `Price` average per hotel
Step 4	Create clusters using k-means based on `O`, `C`, `R`, `S`, `V` ratings *vs* `Price`

Multiple Linear Regression. A Linear Regression analysis predicts one or more continuous variables based on other attributes, identifying dependence relationships among variables [18]. MLR is typically applied in multi-variable scenarios to estimate the value of a dependent variable based on a set of other explicative independent variables. Equation 1 displays the model of the MLR with k regression variables. The parameters β_i ($i = 1$ to k) are the partial regression coefficients [20].

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i \quad (1)$$

In this work, MLR is used to verify if the `Overall` rating can be explained by other hotel ratings, *i.e.*, which ratings are more relevant to the `Overall` rating. Firstly, we calculate the correlation between the different ratings to select the most relevant variables for the MLR. The results show that the partial ratings included in an user review are capable of explaining 78 % of the `Overall` rating of the hotel. The regression uses the Ordinary Least Squares (OLS), which estimates the unknown parameters using a linear regression model, and minimises the differences between the observed responses, and the responses predicted by a linear approximation of the data [16]. Based on the correlation results, we used as independent variables `Cleanliness`, `Rooms`, `Service` and `Value`. In terms of evaluation, this MLR presents a Root Mean Square Error (RMSE) of 0.78. These results show that, since the `Overall` rating includes the clients' appreciation regarding the different hotel services, we can use the `Overall` rating as the global hotel appreciation.

Clustering. High price is not necessarily a synonym of high quality or high value. In Crowdsourcing platforms, tourists provide feedback regarding essentially the service quality. However, typically these platforms do not offer comparisons between the crowd-sourced information and relevant features such as the hotel price. Based on the MLR results, we perform a k-means Clustering analysis, *i.e.*, using the hotel `Overall`, `Cleanliness`, `Rooms`, `Service` and `Value` ratings, and, finally, plot the clusters against the hotel `Price`. Using the Elbow Method [19], we concluded that the optimum k is 5. This analysis shows that high prices usually lead to ratings with less dispersion and higher values. It seems that expensive hotels are more prone to have high ratings, while cheaper hotels have both clients who love and clients who loath the hotel. We also concluded that, in the case of the TripAdvisor data set, high price hotels have a high crowd-sourced rating.

Reviews Visualisation. The analysis of textual reviews plays an important role in the processing of crowd-sourced data since tourists tend to make their decisions based on the opinions of others. Word Clouds are a Big Data Visualisation technique, which highlights the most frequent relevant words present in a text. In the context of hotel reviews, it provides a means to visualise both positive and negative opinions regarding any given tourism resource. In our case, we analyse the positive and negative reviews provided by users, using reference words, removing stop words, and building the corresponding scaled Word Clouds.

6. CONCLUSIONS

The emergence of tourism Crowdsourcing platforms, *e.g.*, TripAdvisor or Expedia, allow tourists to evaluate and share opinions regarding a tourism resource. This crowd-sourced information influences the planning of new tourists. However, in face of this Big Data scenario, tourist are unable to process, relate and visualise the available volumes of crowd-sourced information adequately. To address this problem, *i.e.*, to process tourism crowd-sourced data, and to provide the tourist with relevant information for travel planning;

this paper exploits essentially tourism crowd-sourced information, and provides meaningful information to a potential tourist. As a result, the main contribution of the paper is a methodology to process crowd-sourced hotel data, both off-line and on-line, to analyse rating trends and to visualise reviews. Our methodology[4] relies on Data Mining (Regression and Clustering) and Big Data Visualisation (Word Clouds). Therefore, while the goal of the off-line processing is to identify relevant trends using Crowdsourcing ratings, the goal of the on-line processing is to provide travel planners with the top hotels, together with an instant and intuitive visualisation of their reviews.

As future work, we intend to take into account the reputation of data publishers, in order to determine more precisely the quality of crowd-sourced contents, and use data from Expedia both for on-line and for off-line analysis.

7. ACKNOWLEDGEMENTS

This work was partially financed by: (*i*) ICT COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet); and (*ii*) Mobility grant for doctoral students at the University of Vigo in adapted programs to the European Higher Education in 2016.

8. REFERENCES

[1] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, S. Beheshti, E. Bertino, and N. Foo. Reputation management in crowdsourcing systems. In *Collaborative Computing: Networking, Applications and Worksharing*, pages 664–671. IEEE, 2012.

[2] E. Bjørkelund, T. H. Burnett, and K. Nørvåg. A study of opinion mining and visualization of hotel reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, pages 229–238. ACM, 2012.

[3] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi. *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998.

[4] E. Carvalho and M. S. Chaves. Detecting end-user's visual model to build a visualization tool based on online reviews. *Parsons Journal for Information Mapping (PJIM)*, 5(4):1–11, 2013.

[5] E. S. Carvalho and M. S. Chaves. Exploring user-generated data visualization in the accommodation sector. In *Information Visualisation (IV), 2012 16th International Conference on*, pages 198–203. IEEE, 2012.

[6] A. Colantonio, R. Di Pietro, M. Petrocchi, and A. Spognardi. Visual detection of singularities in review platforms. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 1294–1295. ACM, 2015.

[7] A. Colantonio, R. D. Pietro, A. Ocello, and N. V. Verde. Visual role mining: A picture is worth a thousand roles. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6):1120–1133, 2012.

[4]This work is under progress, and comprises preliminary results regarding the analysis of crowd-sourced tourism data. Due to that, the approach has some limitations, for instance, the on-line analysis was done with Expedia data (as TripAdvisor API was closed recently), and unabled us to compare adequately on-line and off-line results.

[8] B. Fang, Q. Ye, D. Kucukusta, and R. Law. Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics. *Tourism Management*, 52:498–506, 2016.

[9] H. J. Han, S. Mankad, N. Gavirneni, R. Verma, et al. What guests really think of your hotel: Text analytics of online customer reviews. 2016.

[10] J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[11] M. Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. Wiley-IEEE Press, 2 edition, 2011.

[12] A. Marchetti, M. Tesconi, S. Abbate, A. Lo Duca, A. D'Errico, F. Frontini, and M. Monachini. Tour-pedia: A web application for the analysis and visualization of opinions for tourism domain. In *The 6th Language & Technology Conference on Human Language Technology*, pages 594–595, 2013.

[13] I. Olmeda and P. J. Sheldon. Data mining techniques and applications for tourism internet marketing. *Journal of Travel & Tourism Marketing*, 11(2-3):1–20, 2002.

[14] M. Sigala. Gamification for crowdsourcing marketing practices: Applications and benefits in tourism. In *Advances in Crowdsourcing*, pages 129–145. Springer, 2015.

[15] M. Sigala, E. Christou, and U. Gretzel. *Social media in travel, tourism and hospitality: Theory, practice and cases*. Ashgate Publishing, Ltd., 2012.

[16] M. Stone and R. J. Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society*, pages 237–269, 1990.

[17] T. Suzuki, K. Gemba, and A. Aoyama. Hotel classification visualization using natural language processing of user reviews. In *Industrial Engineering and Engineering Management (IEEM)*, pages 892–895. IEEE, 2013.

[18] A. O. Sykes. An introduction to regression analysis. 1993.

[19] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, 63(2):411–423, 2001.

[20] M. Tranmer and M. Elliot. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 2008.

[21] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACM, 2010.

[22] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. Opinionseer: interactive visualization of hotel customer feedback. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1109–1118, 2010.