Semantic Profiling and Destination Recommendation based on Crowd-sourced Tourist Reviews

Fátima Leal[1,4], Horacio González–Vélez[2], Benedita Malheiro[3,4], and Juan Carlos Burguillo[1]

[1] ETSET/UVigo – School of Telecommunications Engineering, University of Vigo, Vigo, Spain
[2] CCC/NCI – Cloud Competency Centre, National College of Ireland, Dublin, Ireland
[3] ISEP/IPP – School of Engineering, Polytechnic Institute of Porto, Porto, Portugal
[4] INESC TEC, Porto, Portugal
fatimaleal2@gmail.com, horacio@ncirl.ie, mbm@isep.ipp.pt, J.C.Burguillo@uvigo.es

Abstract Nowadays tourists rely on technology for inspiration, research, booking, experiencing and sharing. Not only it provides access to endless sources of information, but has become an unbounded source of tourist-related data. In such crowd-sourced data-intensive scenario, we argue that new approaches are required to enrich current and new travelling experiences. This work, which supports the "dreaming stage", proposes the automatic recommendation of personalised destinations based on textual reviews, *i.e.*, a semantic content-based filter of crowd-sourced information. Our approach relies on Topic Modelling – to extract meaningful information from textual reviews – and Semantic Similarity – to identify relevant recommendations. Our main contribution is the processing of crowd-sourced tourism information employing data mining techniques in order to automatically discover untapped destinations on behalf of tourists.

Keywords: Tourism; Crowdsourcing; Topic Modelling; Profiling; Recommendation

1 Introduction

Over the last two decades, the Internet has increased the accessibility of travelling. Tourists not only plan trips based on websites, but, in fact, employ technologies throughout the travel cycle. According to the World Tourism Organization [16], the travel cycle encompasses the dreaming, researching, booking, experiencing and sharing stages. It all starts with the dreaming stage, when the tourist starts to consider travelling, and proceeds with the research stage, when the tourist invests time searching for options. Once the tourist makes his/her mind, the booking stage begins. Finally, in the experiencing stage, the tourist embarks on the trip and relies on context-aware mobile applications to get personalised recommendations. Last, but not least, in the sharing stage, the tourist shares feedback data both in real and deferred time.

This pervasive interaction between tourists and technology consistently generates large volumes of collaborative information on dedicated platforms. Arguably, tourists build their own "crowd-sourced" profile, as their personal information is directly

entered or harvested from tourism websites. Overall, this process comprises experience sharing in the form of likes, posts, images and/or videos (social-network-based); ratings and reviews (evaluation-based); and pages (wiki-based). This valuable feedback information ultimately influences the decisions of both tourists and businesses.

Although tourism crowd-sourced information influences decision making, typically, a tourist cannot monitor or control his/her own crowd-sourced footprint to enhance his/her options, due to the complexity of the diverse platforms and resources. As machine learning and data mining methodologies provide dedicated algorithms for knowledge discovery, we argue that the combination of tourism crowd-sourced information and machine learning methodologies should further enable the personalisation of the tourist travel cycle, eventually proposing *ad hoc* travel stages based on the tourist crowd-sourced footprint.

This paper explores the use of tourism crowd-sourced information to enhance the travel cycle stages. Specifically, we have designed an algorithm to recommend personalised destinations based on Expedia crowd-sourced hotel textual reviews, *i.e.*, for the dreaming stage. Our contribution automatically combines tourism crowd-sourced information and recommender systems to discover untapped destinations for tourists, employing both Topic Modelling (TM) to extract meaningful information from textual reviews; and Semantic Similarity (SS) to recognise relevant recommendations.

Our technique applies content-based filtering to topic-modelled tourists and locations. Topics are clusters of words aggregated according to their meaning as well as frequency in the textual reviews. The content-based filtering provides recommendations according to the semantic similarity among topics. The recommendation engine is assessed through evaluation metrics, *i.e.*, Precision, Recall, and F-Measure.

This paper is organised as follows. Section 2 reviews related work on recommendation supported by crowdsourcing. Section 3 describes the proposed method, including the description of the algorithms used. Section 4 presents the implementation details. Section 5 reports the experiments performed and the results obtained. Finally, Section 6 provides the conclusions and discusses the outcomes of this work.

2 Related Work

There is a significant number of well-known Web-based tourism portals (*e.g.*, TripAdvisor, Expedia, airbnb, Wikivoyage, *etc.*) where a tourist can search, comment, share, and evaluate resources. These collaborative platforms can be envisaged as reputation-based crowdsourcing platforms, as users can increase their reputation by evaluating and making recommendations. Overall, they enable tourists to build their own digital footprint and implement profiling and recommendation mechanisms, i.e. tourists themselves contribute their digital foorprints to form intelligent "crowd-sourced" recommendation systems [4].

On the one hand, recommendation systems for tourism have been extensively studied in the literature. Borras et al. [3], Gavalas et al. [6], and Felfernig et al. [5] provide comprehensive surveys on the recommendation of tourism resources. Some systems just harvest public portal information to suggest destinations or to plan trips [13], while others propose the aggregation of tourism-related information in the context of user

models for personalisation [7]. Of particular relevance to our work is Patil et al. [12] which have surveyed frequent data mining techniques used by tourism recommendation systems. Although some of the tourism recommendation systems documented in the literature use machine learning algorithms to detect tourist preferences, frequent behaviours, new trends, or contexts, most systems do not extensively employ crowd-sourced information.

On the other hand, Tiwari and Kaushik [15], Bachrach et al. [1], and Zhuang et al. [18] use a questionnaire/form-based approach to collect crowd-sourced information. Tiwari and Kaushik specifically rely on the crowd available *in situ* to get updated information and, thus, enrich the list of recommendations. Bachrach et al. ask tourists to rate a set of 20 attractions in order to predict the overall crowd-sourced rating for each attraction. Zhuang et al. provide a form for tourists to suggest each other travel plans. Nonetheless, such questionnaire/form-based crowdsourcing approach requires additional interaction from the tourists, disregarding existing tourism crowdsourcing platforms. Yu et al. [17] collect data from Location-Based Social Networks (LBSN) to model users and, thus, recommend destinations. Finally, Guo et al. [9] combine data from different crowdsourcing sources, but do not provide recommendations. In particular, they get popular routes from travelogues and build scenic spots by matching photo descriptions with attraction reviews.

2.1 Contribution

Scant research has been devoted to the automatic use of heterogeneous crowd-sourced information in tourism recommendation systems fully employing data mining techniques. Table 1 depicts a comparison of the above mentioned related work. While earlier work requires additional interaction by the tourists (questionnaire/forms) or does not provide recommendations (*e.g.*, heterogeneous sources), we are building upon our previous work [10], which relies solely on the available crowd-sourced data. We have explored both tourism crowd-sourced data and recommendation techniques, employing data mining methods, in order to discover untapped destinations for tourists. This research extends this approach by processing a significant amount of heterogeneous crowd-sourced information via tourist reviews to improve tourist recall and destination/location features.

Table 1. Comparison of tourism crowdsourcing and recommendation systems

Systems	Reviews-based Modelling	Crowdsourcing Modality	Data Mining	Recommendation
Tiwari and Kaushik (2014)	No	Questionnaire	No	Context-aware
Bachrach et al. (2014)	No	Questionnaire	No	Collaborative
Zhuang et al. (2014)	No	Form	No	Collaborative
Yu et al. (2016)	No	LBSN	No	Collaborative
Guo et al. (2016)	No	Heterogeneous	Yes	–
Leal et al. (2017)	**Yes**	**Form (Expedia)**	**Yes**	**Content-based**

3 Proposed Method

Our approach automates the dreaming stage by recommending new locations to users based on crowd-sourced information gathered from the Expedia platform. Algorithm 1 summarises the recommendation engine. The algorithm accepts as inputs the required LDA parameters as well as hotel and tourist data, including the crowd-sourced textual reviews, and outputs a list of recommended locations for each tourist.

Algorithm 1 Recommendation algorithm

Inputs	Hotel Data: $H = (<$Hotel h, Location $l>, \ldots)$ User Data: $U = (<$User u, Hotel h, Review $r_{u,h}>, \ldots)$ LDA Parameters: θ; β; n_{iter} and n_{topics}
Outputs	Ordered List of Recommended Locations per User: $L_u = (l_a, \ldots, l_n)$
Step 1	**Review Data Preprocessing**: Tokenising, Stopping & Aggregation
Step 2	**Parallel Topic Modelling** via LDA: **for** $(u = 0; u < users; u$++$)$ **do** $topics_u \leftarrow GetUserTopics()$ **for** $(l = 0; l < locations; l$++$)$ **do** $topics_l \leftarrow GetLocationTopics()$

Step 3	**Recommendations**:
	for $(u = 0, u < users; u$++$)$ **do**
	  for $(l = 0; l < locations; l$++$)$ **do**
	  $SS_{u,l} \leftarrow GetTopicSimilarity(topics_u; topics_l)$
	for $(u = 0, u < users; u$++$)$ **do** return locations sorted by $SS_{u,l}$

Step 4	**Evaluation Procedure**: $Precision$, $Recall$ and $F\text{-}measure$

Data Preprocessing tokenises textual reviews, including individual user as well as the aggregated location reviews, and removes meaningless words (special characters, stop words, and numerical parameters).

Topic Modelling aims to find patterns in unstructured texts, attempting to inject semantic meaning into the vocabulary. Topic modelling algorithms represent a set of computer programs which extract topics from texts. A topic is a list of words which occurs in statistically meaningful ways [8]. We employed Latent Dirichlet Allocation (LDA) topic modelling to extract keywords and use them as implicit features of the respective users and locations. LDA, while a Bayesian generative model for text structures, sees a text corpus (d) as a collection of t topics, where a topic has a probability distribution (θ_d, $d = 1$ to k) over a word dictionary (Equation 1).

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Pi_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \tag{1}$$

The α is k-vector with elements $\alpha_i > 0$ and $\Gamma(x)$ is the gamma function. Now, let β_t be the multinomial distribution over words for topic t. Thus, each word is assigned to a topic via β_t distribution. Blei *et al.* (2003) contains a complete mathematical description regarding LDA algorithm [2]. In our problem, we consider each user review as a document. The underlying topics are the set of possible implicit features.

Semantic Similarity measures the distance between concept meanings. The semantic relatedness is computed using ontologies to measure the distance between terms or concepts. One of the most well-known resources which have been extensively used to compute semantic relatedness is WordNet. This paper uses semantic similarity of Princeton's WordNet to calculate the distance between location and tourist topics identified via topic modelling. WordNet (`http://wordnet.princeton.edu`) is a lexical ontology of English words which organises names, adjectives, verbs, and adverbs according to semantic relations (synonymy, antonymy, hyperonymy and meronymy) [11].

Evaluation Metrics The results were assessed using the Precision Recall and F-Measure classification metrics. On the one hand, the Precision measures the proportion of good recommendations (quality). On the other hand, the Recall measures the proportion of good recommendations which appear among the most important recommendations (quantity). Finally, F-Measure combines both metrics [14].

4 Implementation

Our recommendation system, which is implemented in Java, runs on an Openstack cloud instance with 16 GB RAM, 8 CPU and 160 GB hard-disk. The Java-based application relies on the MAchine Learning for LanguagE Toolkit (`http://mallet.cs.umass.edu/`) (MALLET) for parallel topic modelling and on the WordNet Similarity for Java (WS4J) library (`http://ws4jdemo.appspot.com`) to compute semantic relatedness. In terms of architecture, our content-based filter, which is depicted in Figure 1, includes four modules: (*i*) Data Preprocessing; (*ii*) Topic Modelling; (*iii*) Semantic Similarity; and (*iv*) Recommendation. First, Data Preprocessing removes irrelevant words and aggregates user and location reviews. Then, Topic Modelling applies LDA to identify the most representative topics related with tourists and locations (MALLET). Next, the Semantic Similarity relies on WordNet to compare semantically tourists and location topics. Finally, it orders and recommends, for each user, the locations by descending relatedness.

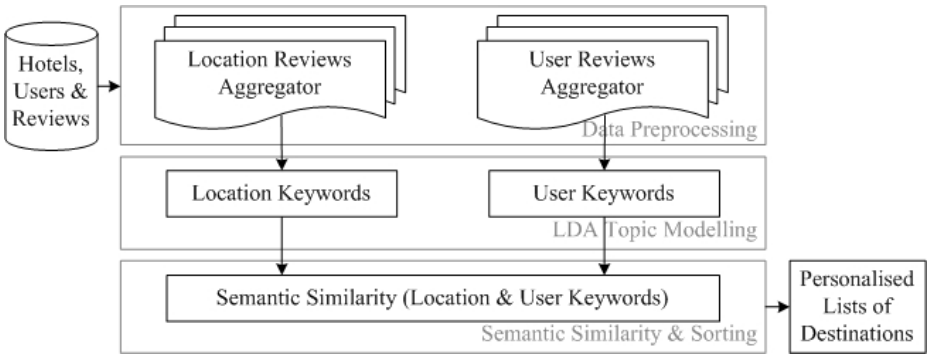

Figure 1. Recommendation engine

5 Experiments and Results

We conducted off-line experiments to evaluate the effectiveness and usefulness of the proposed method with the Expedia crowd-sourced data.

The HotelExpedia data set (`http://ave.dee.isep.ipp.pt/~1080560/ExpediaDataSet.7z`) contains 6030 hotels, 3098 reviewers, including anonymous reviewers, and 381 941 reviews from 11 different hotel locations. Although the data set includes anonymous reviewers and their reviews, we only used in our experiments crowd-sourced data from the 1089 identified reviewers, *i.e.*, we discarded the anonymous users and their inputs. Each user introduced at least 20 reviews. The data set was randomly partitioned into training (75 %) and test (25 %). This data set was built and used as a case study of crowdsourcing in the tourism domain.

5.1 Recommendation Engine

The experiments were focussed on the Parallel Topic Modelling, Recommendation and Evaluation Procedure.

Parallel Topic Modelling is the core of our recommendation engine. It analyses large volumes of textual reviews in order to find topics for describing locations and users. A topic is a cluster of words which, frequently, occur together. Moreover, our topic modelling implementation connects words with similar meanings and distinguishes words with multiple meanings. For each location and user, we select 10 topics using 1000 iterations. The number of topics and the number of iterations were selected according to the F-Measure of the recommendations. The algorithm attributes to each topic a weight based on its relevance in the document. Table 2 and Table 3 present the top five topics together with the corresponding weights (W) for Barcelona and User 15.

Table 2. Barcelona topics

Topic	W	Words
1	0.19	noise night air noisy street
2	0.13	shower bathroom water small door
3	0.34	great close walking clean nice
4	0.17	desk front day time check
5	0.33	stay great friendly clean excellent

Table 3. User 15 topics

Topic	W	Words
1	0.28	check time desk day told
2	0.60	stay good great friendly comfort
3	0.15	airport bus good terminal flight
4	0.46	tube station walk great close
5	0.31	small bathroom shower nice air

Recommendations are based on the WordNet semantic relatedness between the topics provided by the parallel topic modelling module. Considering the example of Table 2 and Table 3, there are similarities between: (*i*) Topic 5 from Barcelona and Topic 2 from User 15; and (*ii*) Topic 2 from Barcelona and Topic 5 from User 15. The algorithm then computes the weighted average between the related topics, using the corresponding topic weight. In the end, the system suggests future travelling destinations to tourists by choosing, for each tourist, the locations with the higher semantic similarity.

Evaluation Procedure adopts Precision, Recall and F-Measure metrics to assess the quality of recommendation. Figure 2 plots, for different number of topics per entity, the F-Measure results. The values grow from 1 to 10 topics and, then, decrease from

10 till 20 topics. The best F-Measure value was achieved with 10 topics per entity, *i.e.*, when the system uses 10 topics to represent tourists and locations. In this case, the recommendation engine presents an F-Measure of 78 %. Based on these results, Figure 3 plots the F-Measure and runtime versus the number of Topic Modelling iterations when using 10 topics per entity. The best F-Measure value was achieved with 1000 iterations and a total runtime of 90 ks (25 h).

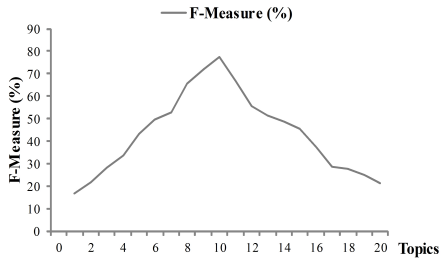

Figure 2. F-Measure *vs.* topics

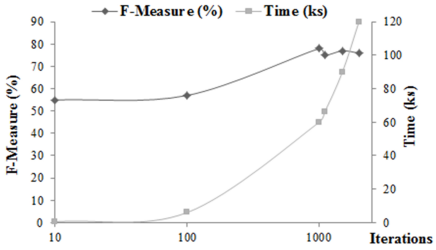

Figure 3. F-Measure and runtime *vs.* iterations

6 Conclusions

Technology has revolutionised both travelling and the tourism industry. In particular, not only it assists tourists in all stages of their travelling, including researching, booking, experiencing as well as sharing, but has transformed tourism business oriented platforms into crowdsourcing platforms, *e.g.*, Expedia, where tourism related knowledge accumulates as tourists leave their digital footprints. These digital footprints are highly influential for businesses and tourists alike.

This paper explores tourism crowd-sourced information to enrich the dreaming stage of the travel cycle. In terms of contributions, we use tourism crowd-sourced textual information and data mining methods to discover and recommend untapped destinations for tourists. Our approach, which uses textual reviews from Expedia, profiles users and locations based on LDA topic modelling and produces personalised recommendations regarding future user destinations based on semantic topic similarity. In order to achieve better recommendations, we tested the topic modelling module with different number of topics and, then, with different number of iterations. The resulting content-based filter was able to recommended future destinations to users solely based on textual reviews with an F-Measure of 78 %.

As future work, we intend to: (*i*) increase the data set dimension with more locations/reviews to explore the tourism Big Data concept in crowdsourcing platforms; (*ii*) introduce hotel recommendations; and (*iii*) design a trust and reputation model for assessing the reliability of the review publishers.

7 Acknowledgements

This work was partially financed by: (*i*) the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness and Internationalisation - COMPETE Programme - within project ≪FCOMP-01-0202-FEDER-023151≫ and

project ≪POCI-01-0145-FEDER-006961≫, and by National Funds through Fundação para a Ciência e a Tecnologia (FCT) - Portuguese Foundation for Science and Technology - as part of project UID/EEA/50014/2013; and (*ii*) ICT COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet).

References

1. Bachrach, Y., Ceppi, S., Kash, I.A., Key, P., Radlinski, F., Porat, E., Armstrong, M., Sharma, V.: Building a personalized tourist attraction recommender system using crowdsourcing. In: AAMAS '14. pp. 1631–1632. International Foundation for Autonomous Agents and Multiagent Systems, Paris (2014)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022 (2003)
3. Borras, J., Moreno, A., Valls, A.: Intelligent tourism recommender systems: A survey. Expert Systems with Applications 41(16), 7370–7389 (2014)
4. Chittilappilly, A.I., Chen, L., Amer-Yahia, S.: A survey of general-purpose crowdsourcing techniques. IEEE Transactions on Knowledge and Data Engineering 28(9), 2246–2266 (2016)
5. Felfernig, A., Gordea, S., Jannach, D., Teppan, E., Zanker, M.: A short survey of recommendation technologies in travel and tourism. OEGAI Journal 25(7), 17–22 (2007)
6. Gavalas, D., Kasapakis, V., Konstantopoulos, C., Mastakas, K., Pantziou, G.: A survey on mobile tourism recommender systems. In: ICCIT 2013. pp. 131–135. IEEE, Beirut (2013)
7. Gonzalez, G., Lopez, B., De la Rosa, J.: Smart user models for tourism: A holistic approach for personaliaed tourism services. Information Technology & Tourism 6(4), 273–286 (2003)
8. Graham, S., Weingart, S., Milligan, I.: Getting started with topic modeling and mallet. The Programming Historian 2, 12 (2012)
9. Guo, T., Guo, B., Zhang, J., Yu, Z., Zhou, X.: Crowdtravel: Leveraging heterogeneous crowdsourced data for scenic spot profiling and recommendation. In: PCM 2016. LNCS, vol. 9917, pp. 617–628. Springer, Xian (2016)
10. Leal, F., Dias, J.M., Malheiro, B., Burguillo, J.C.: Analysis and visualisation of crowdsourced tourism data. In: C3S2E '16. pp. 98–101. ACM, Porto (2016)
11. Miller, G.A.: WordNet: A lexical database for English. Commun. ACM 38(11), 39–41 (Nov 1995)
12. Patil, P., Kolhe, V.: Survey of travel package recommendation system. International Journal of Science and Research 3(12), 1557–1561 (2014)
13. Ricci, F., Werthner, H.: Case base querying for travel planning recommendation. Information Technology & Tourism 4(3-1), 215–226 (2001)
14. Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: Recommender systems handbook, pp. 257–297. Springer (2011)
15. Tiwari, S., Kaushik, S.: Crowdsourcing based fuzzy information enrichment of tourist spot recommender systems. In: ICCSA 2015. LNCS, vol. 9158, pp. 559–574. Springer, Banff (2015)
16. World Tourism Organization (UNWTO) Affiliate Members: Technology in tourism. AM-reports 1, UNWTO (2011)
17. Yu, Z., Xu, H., Yang, Z., Guo, B.: Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints. IEEE Transactions on Human-Machine Systems 46(1), 151–158 (2016)
18. Zhuang, Y., Zhuge, F., Chiu, D., Ju, C., Jiang, B.: A personalized travel system based on crowdsourcing model. In: ADMA 2014. LNCS, vol. 8933, pp. 163–174. Springer, Guilin (2014)