Opinion Extraction from Hotel Reviews

F. Javier Ortega[1], José A. Troyano[1], Fermín Cruz[1], and Fernando Enríquez[1]

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Av. Reina Mercedes s/n 41012, Sevilla (Spain)
{javierortega,troyano,fcruz,fenros}@us.es

Abstract. In this work we carry out a study of the usefulness of social-based online systems on the tourism industry from the point of view of being aware of the perception expressed by users about the services enjoyed in their travels. To this end, we have compiled and analyzed opinions expressed and shared by users on TripAdvisor, one of the most relevant UGC-based websites in the field of tourism. This work shows the method used for the compilation of the dataset and discusses its characteristics focusing on two aspects of the information gathered: the structured and the unstructured data. Regarding the last one, we show the results of applying TOES, an Opinion Mining domain-adaptable tool for the extraction and classification of opinions from user reviews, comparing them to the scores given by users to the features of the hotels.

1 Introduction

Tourism has become the main industry of many countries over the world, such is its relevance in the economy nowadays. This fact explains the appearance and growing interest on the e-Tourism research field. e-Tourism is mainly focused on the opportunities offered by the application of the information technologies to the field of tourism, in addition to the analysis of the effects that these technologies provoke to the industry. e-Tourism has evolved together with the technologies that support it adopting the Web 2.0 technologies, with the inclusion of social capabilities in their systems for example, as a way of leveraging the user experience. In this work, we discuss the usefulness of a dataset compiled from one of the most important web sites on tourism: TripAdvisor. We have retrieved the opinions written in Spanish about hotels located in the Canary Islands. Furthermore, we show an Opinion Mining approach to e-Tourism, unlike most of the works carried out in this area that are based on personal surveys and interviews. Our work aims to study the user perception about tourism services regarding the on-line opinions that those users express through tourism-related on-line social platforms.
The work is structured as follows. In order to carry out our study, we have compiled a dataset from on-line reviews about hotels, generated by users of a well-known tourism social platform, TripAdvisor. This resource is briefly discussed in Section 2. Then, in Section 3 we apply an Opinion Mining tool in order to extract useful knowledge from our dataset about

the user perception of the services provided. We discuss the results obtained in Section 4. Finally, in Section 5 we point out the ideas of future works that we plan, following the conclusions derived from this work.

2 Dataset Compilation

Since we intend to determine the reliability of UGC-based websites in determining the tourist perception about a destination or a tourist service, we have compiled a dataset to support our experiments. Our aim is to analyze the opinions expressed by the users of websites focused on tourism on one of the most important UGC-based websites in this area: TripAdvisor. Finally, we decide to work with user-generated reviews about hotels in a specific location, such as the Canary Islands, due to their particularities as the unique subtropical area in Europe and the importance of the tourism industry in their economy, which assures a huge amount of hotels and user-generated reviews of them, with a high variety of tourists with different needs and perceptions.

The structured data retrieved about the hotels consists in: name of the hotel, category (in the range of 0-5 stars), location (the island where the hotel is located) and the average of the scores provided by the users. About the opinions, we have gathered the following information: the user who wrote the opinion, the origin of the user, the profile (whether the user has traveled solo, with friends or with family), the textual opinion, and a set of detailed scores given by the users to six specific features: location, service, comfort, cleanliness, rooms and quality of the hotel. TripAdvisor allows their users to assign a numerical value to each of those features, computing the overall score of a hotel as the average of the feature scores.

In Table 1 we show some metrics related to the size of the dataset focusing just on those reviews originally written in Spanish.we can see that we have collected a high number of hotels and user opinions. Although the average number of reviews per hotel is about 34, the standard deviation is high (50), so there are a lot of hotels with just one review in Spanish, while others have more than 300. These reviews are usually of a medium size (more than 170 words in average) which could provide valuable information if properly processed.

Hotels with reviews	382
Reviews in Total	12,950
Reviews per hotel	33.99
Std. Dev. of Reviews	50.01
Number of Words	2,224,301
Words per review	171.76

Table 1. Statistics of the reviews in Spanish of the dataset.

3 TOES

TOES is a domain-adaptable Opinion Mining tool intended to detect and classify the opinions in a text. The underlying idea is to capture knowledge about a particular product class and the way people write their reviews on it. To that end, TOES performs two phases: first, it detects the features that are opinionated; in the second step, it computes the polarity of each opinion and the intensity of the polarity, and assigns a score in the range [-1,1], representing -1 the most negative polarity and 1 the most positive.
TOES needs a training phase where a set of resources adapted to the application domain are built. Some resources are automatically induced from a corpus of annotated reviews, while others are manually generated by an expert with some computational assessment. The training corpus is tagged by an expert aided by TOES, indicating the words that refers to features of the domain, *feature words*, and also the phrases that constitute opinions about those features, *opinions words*. A taxonomy on the given domain is built from the *feature words*, defining the characteristics that users are expected to write about. Using the taxonomy and the annotations of the expert, TOES builds a set of domain-dependent resources which are used for the detection and classification of opinions.

4 Evaluation

We discuss in this section the method followed to study whether the information extracted by TOES about users opinions is related to the structured information provided by those users in the form of numerical values. One of the major advantages of using TOES is its capability to identify specific opinions about each feature of the object being opinionated. It is accomplished by using a taxonomy of features defined by an expert for the application domain. On the other hand, TripAdvisor defines a taxonomy of features as well, in order to allow their users to assign specific scores to each feature. Regarding the evaluation of the system, we need to define the relation between the categories in both taxonomies in order to properly assess the reliability of TOES in the hotel domain. We show the proposed mapping in Table 2, where we relate the features in TOES to each feature in TripAdvisor.

TripAdvisor	**TOES**
Quality	Building, Hotel, Price
Comfort	Bed
Rooms	Rooms, Television, Bathroom, Facilities
Cleanliness	Cleanliness
Location	Location, Views
Services	Services, Staff, Internet, Food/Drink

Table 2. Mapping between the feature taxonomies of TOES and TripAdvisor.

Once we map the taxonomies, we can compare the information extracted by TOES to the structured information provided by users in the website. Regarding the data in Table 3, we observe that the overall results

obtained by TOES from the textual opinions are fairly close to those expressed by users through the scores.

	TripAdvisor		**TOES**	
Features	**Pos.**	**Neg.**	**Pos.**	**Neg.**
Cleanliness	79.07%	20.93%	78.98%	21.02%
Comfort	63.95%	36.05%	66.41%	33.59%
Location	75.79%	24.21%	86.51%	13.49%
Quality	79.39%	20.61%	80.19%	19.81%
Rooms	78.72%	21.28%	81.93%	18.07%
Services	82.90%	17.10%	84.83%	15.17%
Average	**76.64%**	**23.36%**	**79.81%**	**20.19%**

Table 3. Percentage of positive and negative opinions per feature according to the scores in TripAdvisor(columns 2 and 3) and TOES (columns 4 and 5).

5 Conclusions and Future Work

In this work we discuss a study of a new tourism dataset compiled from TripAdvisor, one of the most used eTourism websites. We have explained the process followed to build the dataset. Then, we have performed a set of experiments in order to study the applicability of TOES (0) in the hotel domain. To that end, we have compared the results obtained by TOES from the textual opinion of users to the numerical scores given by those users to the hotels in the dataset. According to our observations, TOES is able to infer from unstructured data very similar information to that expressed by users in the form of structured data. This is really useful in the cases where the websites do not offer such information to the users.

We plan to further our work by studying the expansion of our dataset with information extracted from other sources, and the possibilities offered by TOES in the extraction of implicit information not provided explicitly by the websites. On the other hand, it is very interesting the adaptation of these resources to the multilingual environment where our research is focused on. Other important lines of research are the computation of hotel rankings in terms of the user opinions, the inclusion of the temporal aspect of the reviews of users in the computation of the relevance of opinions, and also the influence of the user profiles (origin, type of traveller, age, etc.) in the opinions.

Bibliografía

Fermín L. Cruz, José A. Troyano, Fernando Enríquez, F. Javier Ortega, and Carlos G. Vallejo. 'long autonomy or long delay?' the importance of domain in opinion mining. *Expert Systems with Applications*, 40:3174–3184, 2013.