Knowledge-Based Systems 23 (2010) 323–329

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A classification-based review recommender

M.P. O'Mahony *, B. Smyth

CLARITY: Centre for Sensor Web Technologies, School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland

ARTICLE INFO

Article history:
Available online 20 November 2009

Keywords:
User-generated reviews
Classification
Helpful
TripAdvisor

ABSTRACT

Many online stores encourage their users to submit product or service reviews in order to guide future purchasing decisions. These reviews are often listed alongside product recommendations but, to date, limited attention has been paid as to how best to present these reviews to the end-user. In this paper, we describe a supervised classification approach that is designed to identify and recommend the most helpful product reviews. Using the TripAdvisor service as a case study, we compare the performance of several classification techniques using a range of features derived from hotel reviews. We then describe how these classifiers can be used as the basis for a practical recommender that automatically suggests the most-helpful contrasting reviews to end-users. We present an empirical evaluation which shows that our approach achieves a statistically significant improvement over alternative review ranking schemes.

1. Introduction

Recommendations are now an established part of online life. In the so-called *Social Web*, we receive recommendations everyday from friends and colleagues, as well as from more distant connections in our growing social graphs. Recommender systems have played a key role in automating the generation of high-quality recommendations based on our online histories and/or purchasing preferences. For example, music services such as Pandora and Last.fm are distinguished by their ability to suggest interesting music based on our short-term and long-term listening habits. Indeed, online stores such as Amazon, iTunes, and BestBuy have long established the critical role of recommender systems when it comes to turning browsers into buyers.

Recently, information in the form of *user-generated reviews* has become increasingly important when it comes to helping users make the sort of buying decisions that recommender systems hope to influence. Many sites, such as Amazon, TripAdvisor and Yelp, complement their product descriptions with a rich collection of user reviews. Indeed, many of us use sites like Amazon and TripAdvisor primarily for their review information, even when we make our purchases elsewhere. In the world of recommender systems, reviews serve as a form of *recommendation explanation* [2,5,12], helping users to evaluate the quality of suggestions.

The availability of user-generated reviews introduces a new type of recommendation problem. While reviews are becoming increasingly more common, they can vary greatly in their quality and helpfulness. For example, reviews can be biased or poorly authored, while others can be very balanced and insightful. For this reason, the ability to accurately identify helpful reviews would be a useful, albeit challenging, feature to automate. While some services are addressing this by allowing users to rate the helpfulness of each review, this type of feedback can be sparse and varied, with many reviews, particularly the more recent ones, failing to attract any feedback.

In this paper, we describe a system that is designed to recommend the most helpful product reviews to users. In the next section, we motivate the task in the context of the TripAdvisor service, which we use as a test domain. In Section 3, we adopt a classification approach to harness available review feedback to learn a classifier that identifies helpful and non-helpful reviews. We then describe how this classifier can be used as the basis for a practical recommender that automatically suggests the most-helpful contrasting reviews to end-users. In Section 4, we describe a comprehensive evaluation that is based on a large set of TripAdvisor hotel reviews. We show that our recommender system is capable of suggesting superior reviews compared to benchmark approaches, and highlight an interesting performance asymmetry that is biased in favour of reviews expressing negative sentiment.

2. Towards recommending helpful reviews

Insightful product reviews can be extremely helpful in guiding purchasing decisions. As reviews accumulate, however, it can become difficult for users to identify those that are helpful, thereby introducing yet another information overload problem. This signals a new and challenging recommendation task – to recommend reviews based on *helpfulness* – which complements the more

* Corresponding author.
E-mail addresses: michael.p.omahony@ucd.ie (M.P. O'Mahony), barry.smyth@ucd.ie (B. Smyth).

0950-7051/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2009.11.004

traditional product recommendation scenarios. Thus the job of the *product recommender* is to suggest a shortlist of relevant products to users, and the role of the *review recommender* is to suggest a small number of helpful reviews for each of these products. We address review recommendation in Section 2.3, but first we consider user-generated reviews and review helpfulness in respect of TripAdvisor reviews, which form the basis of our study.

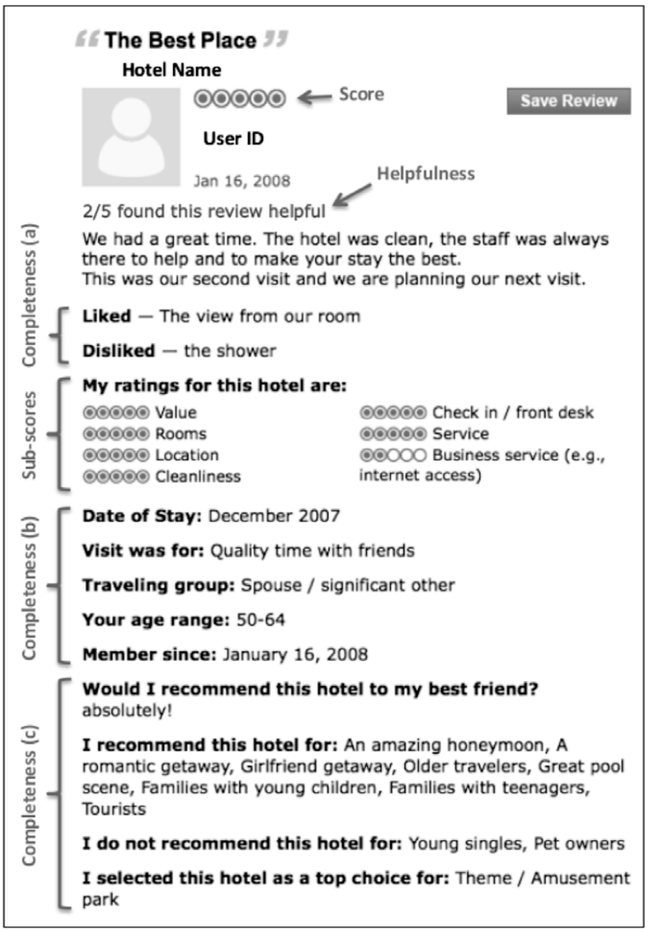

Fig. 1. A TripAdvisor review.

2.1. TripAdvisor reviews

Fig. 1 shows a typical TripAdvisor review. In addition to the *hotel-ID* and the *user-ID*, each review includes an overall score (in this example, 5 out of 5 stars), a title ("The Best Place") and the review-text (in this case, just three lines of text).

Optionally, users can specify what they *liked* and *disliked* about the hotel, and can provide *sub-scores* in relation to certain aspects of the hotel (e.g. *Value*, *Rooms*, *Location* etc.). Further, users can provide some personal information (*Your age range* and *Member since*) and details relating to the date and purpose of visit (*Date of Stay*, *Visit was for* and *Traveling group*). Finally, users can respond to set review-template questions such as *Would I recommend this hotel to my best friend?* and *I recommend this hotel for*.

For the study described in this paper, we created two large datasets by extracting all TripAdvisor reviews prior to April 2009 for users who had reviewed any hotel in either of two US cities, Chicago or Las Vegas. In total, there are approximately 225,000 reviews by 45,000 users on 70,000 hotels (Table 1). For both datasets, the median number of reviews per user and per hotel is 7 and 1, respectively. These distributions are, however, significantly skewed; for example, the most reviewed hotel in the Chicago and Las Vegas datasets has 575 and 2205 reviews, respectively, while the greatest number of reviews written by any user is 165 and 134, respectively.

Table 1
TripAdvisor dataset statistics.

Dataset	# Users	# Hotels	# Reviews
Chicago	13,473	28,840	77,863
Las Vegas	32,002	41,154	146,409

2.2. Review helpfulness

Importantly for our case study, TripAdvisor allows users to provide feedback on review helpfulness. We define *helpfulness* as the percentage of positive opinions that a review has received. For example, the review shown in Fig. 1 has received 2 positive and three negative opinions and thus it has a helpfulness of 0.4.

Fig. 2a shows the number of reviews in the Las Vegas dataset versus user score. It is clear that the majority of reviews attracted high scores, with more than 95,000 4-star and 5-star reviews submitted, compared to less than 10,000 1-star reviews. This suggests that users are far more likely to review hotels that they have liked.

In addition, Fig. 2a indicates that many reviews attracted very few opinions; for example, approximately 20% of reviews received no feedback and, while some 80% of reviews received $\geqslant 1$ opinion, only 38% of reviews received $\geqslant 5$ opinions. Excluding reviews with no feedback, Fig. 2b shows that the most poorly-scored reviews attracted on average the highest number of opinions (almost 11), while reviews with scores of $\geqslant$ 2-stars received on average between 6 and 8 opinions.

Interestingly, reviews with lower scores were perceived as being less helpful by users (Fig. 2b). For example, on average 63%

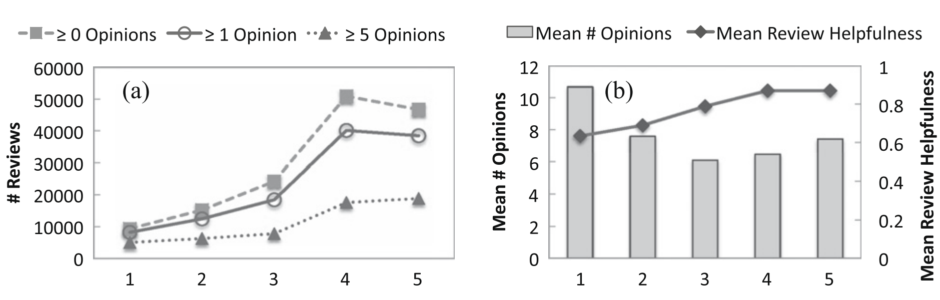

Fig. 2. Las Vegas dataset trends. (a) Number of reviews versus score. (b) Mean number of opinions per review and mean review helpfulness versus score. Similar trends applied for the Chicago dataset.