The Identification of Noteworthy Hotel Reviews for Hotel Management

San-Yih Hwang, Chia-Yu Lai, Shanlin Chang, Jia-Jhe Jiang
Department of Information Management
National Sun Yat-sen University
Kaohsiung 80424, Taiwan
syhwang@mis.nsysu.edu.tw

Abstract

User-generated content (UGC) plays an increasingly important role for knowledge sharing among Internet users in recent years. Many travel sites encourage users to share their experiences and express their opinions on attractions, accommodations, restaurants, etc. The UGC about travel provide precious information to the staff in travel industry as well as the end users. Most of previous works in travel recommendation focus on identifying attractions/hotels or reviews that are useful to the customers by analyzing UGC. In this paper, we argue that reviews that are noteworthy for hotel management is critical to the success of hotels in the competitive travel industry. We employ two hotel managers to conduct an examination on Taiwan's hotel reviews in Tripadvisor.com and they report that noteworthy reviews can be characterized by their content features, sentiments, and writing qualities. We propose three methods for representing content features. Through the experiments using tripadvisor.com data, we find that, in terms of content feature representation, LDA method achieves comparable performance to TF-IDF method with higher recall and much fewer features. In addition, all the three types of features are important in identifying noteworthy hotel reviews. Specifically, content features are shown to have the most impact on precisions, whereas sentiments and review qualities impact more on recalls.

Keywords: Review recommendation, Text mining, Topic model, Sentiment analysis, Tourism

Introduction

With the rapid expansion and proliferation of web 2.0 technologies, users have witnessed a profound numbers of platforms for user-generated content (UGC) that are both convenient and ubiquitous. The new style of content sharing enables Internet users to become self-publishing consumers and share their knowledge and experience with others. Online customer reviews are regarded as electronic word of mouth (eWOM) and have been found to have a significant effect on product sales and direct impacts on consumer purchase intention (Duan, Gu, & Whinston, 2008; Zhang, Zhao, Cheung, & Lee, 2014). Furthermore, several studies have demonstrated that reviews shown on UGC platforms are particularly important for experience goods, as they are mostly tangible and their qualities cannot be determined before consumption (L. R. Klein, 1998; Zhu & Zhang, 2006). Therefore, online reviews of experience goods such as traveling, entertainment and professional services due to their variability and intangibility are in high demand. Most services and products offered by hotels are experience goods, and previous studies had shown the empirical evidence that online reviews play a key role in hotel selection and trip planning (Sparks, Perkins, & Buckley, 2013; Ye, Law, & Gu, 2009).

Online customer reviews about a hotel, especially those with critical comments, may significantly impact its reputation, and subsequently sales. Prior studies have shown that searching for travel-related information has become one of the most popular online activities and indicated that more than 74 percent of travelers use the online customer's comments as key information sources when planning their trips (Ye et al., 2009). Within the context of the hotel industry, many hotel staff members now take an active role online by posting their responses to hotel reviews. However, according to a recent survey (Sparks et al., 2013), only 7% of hotels are replying to reviews even though 71% of customers think that a management response has significant influence. Moreover, negative reviews may easily damage the image of the hotels, and hotel staff need to respond to these reviews so as to offset the negative emotions of these customers in the hope to restore the hotel's reputation. Therefore, how to quickly identify reviews that may potentially influence the hotel's business performance is important to hotel staff.

Nevertheless, customer reviews have influenced not just on the hotel sales. Customer reviews, if properly utilized, may prove beneficial for the operations of a business in an ever-changing, competitive environment. In recent years, there is a growing interest to recommend product/service reviews to customers O'Mahony & Smyth, 2009; O'Mahony & Smyth, 2010; Ghose & Ipeirotis 2011; Dong et al. 2013). However, we argue that reviews that are useful to customers may not be noteworthy to managers. In the context of hotel management, hotel staff pay more attention to the comments on services and facilities offered. Reviews on other aspects, such as hotel locations and room sizes, might be important for customers when it comes to choosing hotels, whereas the hotel manager find them less noteworthy from the hotel management point of view because these aspects cannot be easily changed.

In this paper, we propose an approach to automatically identifying customer reviews that are noteworthy for hotel management. To do so, we conduct a preliminary qualitative research that involves an interview with two hotel managers. During this interview, the hotel managers reported that noteworthy reviews for hotel staff are always subject to subsequent actions. For example, reviews with negative comments need to be addressed in a timely manner because their spread may harm the reputations of the hotel. However, not every negative review deserves equal attention. Negative reviews with reasonable writing or written by professional reviewers are particularly harmful. In addition, some

positive reviews may also be noteworthy. For example, reviews that pinpoint some services that are unexpected yet please the customers are particularly welcome. Those reviews could be used to inspire the employees and motivate them to propose innovative services. Finally, some reviews may provide useful suggestions to the hotel, which may help shape future business strategies.

Our study aims to identify the features that are relevant to the noteworthiness of hotel reviews with respect to the hotel management and subsequently propose a method to automatically identifying noteworthy reviews. Prior works in the literature have extensively investigated approaches for recommending hotels or hotel reviews (Chen & Chen, 2014; Ghose, Ipeirotis, & Li, 2012; Levi, Mokryn, Diot, & Taft, 2012; O'Mahony & Smyth, 2009). However, their targets of recommendation are mostly customers, not hotel staff. A hotel review that is helpful to prospective customers may or may not concern the hotel staff. For example, a hotel review that describes how much a customer enjoyed or hated the beach that is close to a hotel could be useful for someone who seeks to relax in the hotel but may not be so much concern for the hotel staff because hotel location cannot be easily changed. Ghose and Ipeirotis (2011) explore the factors that affect the sales of various types of products using data from Amazon.com and conclude that features in subjectivity, readability, and informativeness have different degree of impact on different types of products. However, how to identify reviews that contribute or hinder product sales is not explored in their work. Besides, hotel sales would not be the only concern for the hotel staff.

In this paper, we first conduct interviews with senior managers in travel industry to shed light on the characteristics of noteworthy reviews. The preliminary study leads us to conjecture that three aspects, namely content, sentiment and review quality, may impact the noteworthiness of reviews from the perspective of hotel staff. We subsequently propose several methods to represent the three aspects in the hope to more accurately identify noteworthy hotel reviews for hotel staff. Through the experiments using tripadvisor.com data, we find that all three types of features are important in identifying noteworthy hotel reviews. Specifically, content features are shown to have the most impact on precision of the proposed method, whereas sentiments and writing qualities of reviews impact most on the recall. With respect to the various methods for representing content features, LDA method achieves comparable performance to TF-IDF method with higher recall and much fewer features.

The remainder of this paper is organized as follows. In the next section, we review techniques about content feature identification, sentiment analysis, review quality determination. In the third section, related works on hotel and review recommendation will be described. In the fourth section, we present our approach to identifying noteworthy hotel reviews. In the fifth section, we examine the empirical data to evaluate our approach. Finally, we conclude with the result of our work and give directions for future research.

Background

In this section, we provide information about fundamental techniques used for recommendation based UGC. Specifically, we firstly review the techniques for the identification of content features from a massive set of documents. Then we present the methods about how to determine the sentiment of a review. Finally, the approach for representing the quality of documents is presented.

Content Feature Identification

TF-IDF, term frequency–inverse document frequency, is a typical approach to representing textual features of documents (Chowdhury, 2010). The idea is that if a

word or a phrase appears in a document with high frequency (called term frequency) yet rarely appears in other documents (inverse document frequency), this word or phrase is a good indicator for identifying this document. Let $n_{i,j}$ be the number of times a keyword k_i appears in a document d_j. $TF_{i,j}$ measures the term frequency of keyword k_i in document d_j, as shown below.

$$TF_{i,j} = \frac{n_{ij}}{\sum_{k_l \text{ is a keyword in } d_j} n_{lj}}$$

If a keyword appears in many documents, its importance will decline. $IDF_i = \log\frac{N}{N_i}$ is used to measure the inverse document frequency, where N is the total number of documents and N_i is the number of documents in which keyword k_i appears. The TF-IDF weight for keyword k_i in document d_j, $w_{i,j}$ is then defined as:

$$w_{i,j} = TF_{i,j} \times IDF_i$$

In addition to its basic form shown above, there are several variations about TF-IDF (Chowdhury, 2010). As can be imagined, a large number of TF-IDF features (usually thousands) will be needed for representing a massive set of documents. Another approach for concisely representing documents is to use a small number of latent content features, or called topics. Several topic models have recently been proposed to identify a small number of topics inherent in a set of documents. Each document can be subsequently represented as a topic vector. Latent Dirichlet Allocation (LDA) is the most commonly used approach for deriving the topic model from a set of documents (Blei, Ng, & Jordan, 2003). LDA techniques, such as Gibbs sampling (T. Griffiths, 2002; T. L. Griffiths & Steyvers, 2004), take as input a collection of documents, each represented as a bag of words, and produce two kinds of probability distributions: topic probability distributions, one for each document, and word probability distributions, one for each topic. Two parameters, namely α and β, can be set to adjust the concentration parameters of the Dirichlet prior distributions for topic probabilities and word probabilities respectively.

While TF-IDF and LDA model are both powerful techniques in representing documents by their linguistic forms, semantic information is not considered and problems such as synonyms and homonyms may arise. WordNet is a large lexical database of English that describes the mappings between words and senses (i.e., meanings) and the relationships among senses (e.g., is-a relationships, similar senses, etc). Since a word with multiple meanings can be confusing, a method is needed to identify the particular sense for each word appeared in a sentence. This problem is called the word-sense disambiguation (WSD), which is the ability to identify the meaning of words in context in a computational manner, and there have been many methods proposed for WSD problem (Navigli, 2009).

Sentiment Analysis

Sentiment analysis, also called polarity recognition or opinion mining, aims to determine the sentiment of a document, a sentence, or even an entity, being positive, negative, or neutral. According to Pang and Lee (2008), there are generally two approaches for detecting sentiments: supervised approach and unsupervised approach. The supervised approach for determining the sentiment of a review starts by representing a review as a feature vector, e.g., TF-IDF, and builds a classifier using a training data set. There have been many methods that are devoted to the identification of text features relevant to sentiment (Pang & Lee 2008). On the other hand, several methods have been proposed in the literature for determining the polarity of a review using unsupervised approach (Dave, Lawrence, & Pennock, 2003; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Turney, 2002), which do not require a training data set. These methods generally prepare a domain-specific sentiment lexicon and identify a number of linguistic constructs

commonly used to express sentiments on certain aspects of products. The sentiment of a sentence is determined by looking at its linguistic constructs and the appeared sentiment word(s). Polarity of a review is an aggregation of the sentiments of its constituent sentences. There are also works that try to combine both supervised and unsupervised techniques for sentiment detection. In Carrillo-de-Albornoz et al. (2010), the authors describe a hybrid approach, in which sentences are first converted into senses based on WordNet using Lesk algorithm (Lesk, 1986). Lesk algorithm is based on the assumption that the senses of words in a given "neighborhood" (section of text) will tend to share more common words that explain these senses. By referring to WordNet Affect, which includes senses pertaining to 16 emotions: joy, love, liking, calmness, positive-expectation, hope, fear, sadness, dislike, shame, compassion, despair, anxiety, surprise, ambiguous-agitation and ambiguous-expectation, the emotions pertaining in each review is represented as a 16-tuple, where each element represents the weight of the corresponding emotion (Baccianella, Esuli, & Sebastiani, 2010). Movie reviews in the corpus are tagged positive or negative polarity, and some machine-learning algorithms can be used to train a classifier. The classifier can then be used to classify the polarity (positive or negative) for each incoming review based on its 16 emotional attribute values.

Quality of Reviews

In Liu et al. (2007), the quality of product reviews is determined by several features, namely sentence level informativeness, word level informativeness, and product feature level informativeness. Sentence level informativeness refers to the number of sentences, the average length of each sentence and the number of the sentence with desired product feature. Word level informativeness indicates the number of words, the number of product names, and the number of brand names. In addition, the reputation of the reviewer who wrote a given review is considered as a good indicator about the review quality (Ghose & Ipeirotis, 2011; Ghose et al., 2012; Huang, Shen, Feng, Baudin, & Zhang, 2010). Example features for reputation of a reviewer include the number and the average helpfulness score of the reviews s/he has written. Some previous studies verify the quality of reviews by classifying review features based on the readability of the text, the reputation of the reviewer, the star rating of the review, and various content features based on the review terms (Liu et al., 2007; O'Mahony & Smyth, 2009; Yang, Tang, Wong, & Wei, 2010)

Related Work

There have been quite some works that address recommendation in hotel industry. One line of research is focused on recommending hotels to customers. The other line of research intends to recommend useful hotel reviews to customers. The two types of research works are described below.

Hotel Recommendation

In commercial applications, hotel recommendations are typically based on hotel ratings given by users. In Adomavicius and Kwon (2007), a regression model is adopted to aggregate ratings on various aspects into a single rating, where aspect ratings for unseen items are predicted using collaborative filtering. Similar approach is adopted by Fuchs and Zanker (2012) for recommending hotels using TripAdvisor data, and they further explore the impact of regression models on different customer segments and exploit penalty-reward-performance model. Jannach et al.(2012) extends the model proposed in Adomavicius and Kwon (2007) by incorporating item-based collaborative filtering, more regression models, and aspects selection. However, this line of research does not make use of the textual data of hotel reviews.

Ghose et al.(2012) define a measure called *utility gain* for the economic impact of a hotel by considering consumer heterogeneity, hotel characteristics, as well as UGC pertaining to the hotels. It shows that UGC variables, such as text features, subjectivity, and readability, significantly affect the model's predictive power for utility gain. Levi et al. (2012) propose a context-based method for personalized recommendation of hotels based on hotels' reviews and the reviewers' contextual information. Three types of context are identified, namely travel intent, nationality, and preference. Then for each context group, nouns that frequently appeared in the relevant reviews are collected and form a lexicon. A user who seeks hotel recommendation is asked to provide her intent and nationality, and lexicons of the corresponding context groups are regarded as the traits of the user. Hotels whose reviews contain high positive sentiment on these traits will be recommended. The authors conduct user study using data collected from Tripadvisor.com and Venere.com, and the results show 20% higher satisfaction rate than the ratings-based recommendation. Moreover, effective recommendations often can lead to greater customer loyalty and higher sales.

Hotel Review Recommendation

As shown in previous study, reviews of a hotel play an important role in deciding whether or not to recommend this hotel. However, some reviews are deemed better than the other and more helpful when it comes to decision making. In the past few years, we have seen quite a few works that intend to predict the helpfulness of an incoming review. Product review recommendation was first proposed in the work of (O'Mahony & Smyth, 2009; O'Mahony & Smyth, 2010). They adopt a supervised learning approach by considering four types of features, namely reputation, content, social, and sentiment. However, for each type of features, relatively simple methods are used for defining sub-features. For example, they use user-supplied rating for determining sentiment, and only linguistic sub-features, such as number of terms and the ratio of upper and lower characters, are used to content feature. Experiments using hotel reviews from TripAdvisor shows reasonable recommendation result. This work is further enhanced by incorporating feature selection and exercising various classification schemes (O'Mahony & Smyth, 2010). In Dong et al. (2013), a supervised learning method for identifying helpful reviews is proposed by taking basic features such as age, rating, readability, as well as product features and sentiment features. It is shown based on reviews for various product categories from Amazon.com that both the product features and sentiments expressed in the review are important factors for identifying helpful reviews. In Ghose and Ipeirotis (2011), the authors propose a regression model to predict the helpfulness of a given review by considering readability, subjectivity, and reviewer's reputations. Experiments using Amazon.com's data show that the impact of the three types of features on predicting the usefulness of a review or a product's sale varies across different product types.

Considering the fact that the helpfulness of a review may differ across users, Musat et al. (2013) propose to derive the interest topic profile of a user based on the reviews she wrote. The interest topic profile is subsequently used to filter out less relevant reviews and generate personalized ratings for hotels. Moghaddam et al. (2011) propose a matrix factorization approach for personalized recommendation of product reviews based on review rating data. They have shown from experiments using Epinion.com data that their proposed methods perform better than other non-personalized, textual/social features-based methods. However, all the above mentioned works that intend to recommend either hotels or hotel reviews target at consumers, rather than hotel staff as focused in our work.

The Approach

We follow the design science research framework presented by Hevner et al.(2004). A two-phase process, namely building and evaluating, is adopted. In the build phase, we start by interviewing with hotel managers for the identification of the characteristics of noteworthy online reviews. Subsequently, we explore different methods for representing these characteristics. We then go on to establish a classification model that can be used to identify the noteworthy reviews. The proposed models and methods are implemented and evaluated using real data. Details of the building phase will be described in this section, and those of evaluating phase will be reported in Section 4.

Characteristics Identification of Noteworthy Reviews

The study follows Myers and Newman (2007) qualitative research guideline and adopts semi-structured interviews. Before the interview, the interviewee agreed to accept interviews and have the interview process recorded. The two interviewees are professional hotel management staff members, who have many years working experience in travel industry. The majority of the interview questions might be designed during the interview, allowing both the interviewer and the hotel managers the flexibility to go into details when needed. The purpose of the interview is to identify the important features of reviews for hotel staff. To identify the characteristics of noteworthy reviews for hotel staff, the interviewer has prepared a set of semi-structured questions, such as "*Did you regularly look at online reviews?*", *"Will the online reviews impact on the management of the hotel and how?*", and "*What kinds of reviews are worth noting?*".

After the interviews, we turn the sound files into text files and perform content analysis. The content analysis approach follows Hycner (1985), and characteristics of the noteworthy reviews are identified. These characteristics can be categorized into three types, namely content features, sentiment features, and quality features, as will be illustrated in the following subsections.

Content Features

Content refers to the subjects or topics discussed in a review. Some hotel reviews focus on the trip of the authors and do not express their experiences and opinions on any aspect of the hotel. These reviews are thus considered less important and not noteworthy.

"*Reviews that describes about hotel local, hotel decor, travel planning, or those not directly related to the hotel are less important; we usually ignore them.*" (Interviewee #1)

In addition, some aspects of hotels are considered more important than others from the hotel management point of view. For example, managers usually pay more attention to comments about hotel services than complaints about hotel locations.

"*When a review describes about hotel service, we usually pay more attention to it, even if the writing quality is no good. We attach great importance to reviews about hotel services*." (Interviewee #2)

In this work, we consider three methods for representing content features of a review. The baseline method is TF-IDF, denoted T0, which represents each review as a vector of words. Specifically, each review is parsed, and punctuations and stop words are removed. Stemming is further conducted to identifying synonyms. As a result, each review is regarded as a set of words, and their TF-IDF values are computed. We choose approximately 4000 words with the highest average TF-IDF values as the content features as it yielded the best performance in our preliminary experiments. The second method is a topic model-based method, denoted T1. It first regards each review as a bag of words by excluding stop words and then applies LDA to all the reviews to generate a topic model. The set

of topics are then treated as the content features, and each review is represented as a vector of topics. The third method, denoted T2, uses semantic-based LDA, which utilizes semantic information in the text for determining topics. Specifically, we use Stanford Parser (D. Klein & Manning, 2003), a popular natural language parser, to process each sentence of a review and identify the part-of-speech (POS) of each word. We then extract all the nouns, verbs, adjectives, and adverbs as they may imply important topic information. Because a word may have different meanings (or called senses) in different context, T2 further applies some word sense disambiguation (WSD) technique to identify the senses of the extracted words from a given ontology. In our work, we use the graph based WSD, UKB (Agirre & Soroa, 2009), and the adopted ontology is WordNet.

In WordNet, each sense contains a number of synonyms and is linked to its hypernyms and hyponyms, forming an is-a concept hierarchy. Therefore, each sense can be further extended by including their hypernyms, hyponyms, or similar senses (line 9 in Figure 1). Finally, each review is represented as a bag of senses, and Latent Dirichlet Allocation (LDA) is used for inferring the topic model. The topic model construction algorithm for T2 is shown in Figure 1. For a newly arrived review, we can apply the constructed topic model to infer its topic vector, and the algorithm is shown in Figure 2.

Sentiment Features

Sentiment refers to the strength of a positive (negative) emotion pertaining to a given review. As one can imagine, negative reviews tend to jeopardize the fame of the hotel and may subsequently impact hotel sales. However, not every negative review needs special attention, and some positive reviews are worth noting. Here are some examples for negative reviews:

The service attitude of this hotel staff is bad. The hotel room is not clean and we are not satisfied with breakfast.

Before determining the sentiment of a given review, we need to first identify the sentiments for its sentences. Due to the lack of sentiment training data set, we adopt the unsupervised approach for determining review sentiment. The sentiment lexicon we use in this work is SentiWordNet 3.0 (Baccianella et al., 2010), an extension of WordNet by incorporating emotion values to senses. In SentiWordNet, there are three types of emotions, namely positivity, objectivity, and negativity, and each sense has an emotional value for each type of emotion in the range of [0, 1]. Sentiment terms accompanied by negation cues have to be carefully addressed because their sentiments may become opposite (e.g., “not bad” has a sentiment opposite to “bad”). We use Stanford Parser to find the phrase structure tree for each sentence to delimiting the scope of negation, if any (Carrillo-de-Albornoz, Plaza, Díaz, & Ballesteros, 2012). We interchange positivity with negativity for each sense in the scope of negation, and their values are multiplied by 0.9 by following the work proposed in (Carrillo-de-Albornoz et al., 2010). For example, “not good” is usually considered less negative than “bad”. Finally, we sum the emotional values of every sense in the sentences of a review to determine the overall sentiment score. Note that in our work, the sentiment score of a review is a pair of emotional values for positivity and negativity. The entire algorithm for determining sentiment of a review is shown in Figure 3.

Quality Features

A review of good writing quality usually deserves more attention than those with poor quality as noted by one of our interviewees.

“*Reviews written by professional bloggers are especially noteworthy because they specialize in writing the meaningful and convincing review content, and most importantly, they tend to be*

big shots and influence people." (Interviewee #1)

Writing quality encompasses both lexical quality features, e.g., word choice, grammar, and style, and semantic quality features, e.g., theme relevance and article organization. There have been some measures proposed for automatically evaluating the lexical quality of a review, whereas it is much more difficult to measure the semantic quality of the review. Thus, previous works usually rely on author reputation for determining the semantic quality of a review (Cheung, Luo, Sia, & Chen, 2009; Liu et al., 2007; O'Mahony & Smyth, 2009). By following the previous works, we measure review quality at three levels: sentence level, word level, and user reputation and identify features that are suitable in our context (i.e., hotel reviews). The following lists the quality features used in this work:

- NumSent: the number of the sentences
- LenSent: the average length of sentences
- NumEmoSent: the number of sentences with non-zero sentiment scores
- NumWord: the total number of words
- NumReview: the number of authoring reviews that is the number of reviews authored by the reviewer.
- MeanHelpReview: mean review helpfulness, which is the mean review helpfulness over all reviews authored by the reviewer.
- STDHelpReview: review helpfulness deviation, which is the standard deviation of review helpfulness over all reviews authored by the reviewer.

Classification Model Construction

As a result, each review can be represented by a vector of words/topics, a pair of sentiment scores, and a vector of review qualities. Each review in the training data set is labeled as noteworthy or not-noteworthy. Figure 4 shows the structure of the training data. A classification method can be adopted for training a binary classifier using the training data. Various classification methods will be executed and compared.

Empirical Evaluation

To validate and gain insights about the usefulness of the proposed approach, we perform a set of experiments on various kinds of features used in the classification model. For the purpose of comparison, we choose three content feature extraction methods, namely T0 (the TF-IDF method), T1 (the LDA method), and T2 (the Semantic-LDA method). We firstly implement these three methods to transform the set of reviews into document-term(topic) matrices where terms (topics) are treated as content features of a document vector. Secondly, we apply the sentiment analysis method, denoted as S as shown in Figure 3, to detect the polarity of each review. Finally, with respect to review quality, we use the seven features as shown in Figure 4, which are collectively denoted as Q. We intend to compare the effectiveness of the three types of content features and show how content features, sentiment features, and quality features affect the performance.

In the following, we first describe the data used in our experiments. We then present the experimental design and the performance metrics. Finally we discuss the experimental results.

Data Collection

We collect data from Tripadvisor.com (http://www.tripadvisor.com/), which is a travel site that provides objective and impartial evaluation of hotels, restaurant recommendations, B&B Reviews, membership information, and travel guides. TripAdvisor, established in 2000, pioneers in hosting UGC in tourism. In our work, we focus on hotels and B&Bs reviews. As TripAdvisor is an internationally renowned travel site, many foreign tourists use it to express their opinions and gain feedback. Thus, this study focuses on foreign tourists,

and English reviews for the top ten hotels, up to May, 2013, in each of the following cities in Taiwan: Taipei, Kaohsiung, Taichung, New Taipei City, Hualien, Nantou, and Ilan, are collected. We developed a TripAdvisor web crawler to automatically crawl hotel review pages in tripAdvisor.com and retain only English reviews. We further develop a parser that parses each review page and retrieves the hotel name, review author, review title, overall rating, review date, review URL, accommodation types (e.g., travel accommodations with family or business accommodation), value rating, location rating, sleep rating, comfort rating, cleaning rating, service rating, review respondents, replying date and replying content, in addition to review content. Finally, html tags are removed, resulting in pure textual content.

As a result, we collected 3124 hotel reviews that comprise 28,088 sentences. 2623 authors contributed these 3124 reviews. These authors also wrote totally 54,746 posts.

Algorithms: Semantic Topic Model Construction
Input : A set of reviews S
Output: A Topic Inference Mode I
1. S=Ø
2. For each review r do
3. C(r) = Ø
4. For each sentence s in r do
5. Use Natural Language Parser (NLP) to identify all nouns, verbs, adjectives, and adverbs in s that are not stopwords.
6. Use some Word Sense Disambiguation (WSD) technique to find all senses of these terms.
7. Add theses sense to C(r)
8. S=S $\cup$ {$C(r)$}
9. Extend S
10. Apply LDA on S to find the topic inference model I.

Figure 1 - Algorithm for semantics-based topic model construction

Algorithm: Topic Vector Extraction
Input: A review r and a topic inference model I
Output: A topic vector
1. C(r) = Ø
2. For each sentence s in r do
3. Use Natural Language Parser(NLP) to identify all nouns, verbs, adjectives, and adverbs in s that are not stopwords
4. Use Word Sense Disambiguation(WSD) techniques to find all senses of these terms
5. Add theses senses to C(r)
6. Extend(C(r))
7. Apply I to C(r) and return r's topic vector

Figure 2 - Algorithm for semantics-based topic vector determination

Training Data Set

To construct a review noteworthiness prediction model, we need a training data set. To prepare the training data set, we first retrieved some 500 reviews that are diversified in their emotional polarities. Specifically, we chose 179 reviews for each of the following classes: highest positivity, lowest positivity, highest negativity, and

lowest negativity, as determined by our sentiment detection method shown in Figure 3. After excluding duplicate reviews, we finally obtained 501 hotel reviews. We gave these reviews to two experts, who are senior managers of renowned hotels in Taiwan. The two experts have been in travel industry for more than six years. They manually determine the reviews to be noteworthy or not. To observe the consistency between two labellers, we measured the average Jaccard coefficient (Martin et al., 1995) for the similarity of their labelling results. There are 386 reviews that receive the same label as classified by both experts, resulting in the Jaccard coefficient 0.77, which is moderate. The 386 reviews are used in subsequent experiments.

Algorithm: Sentiment Score Detection
Input: a review r
Output: sentiment score (positivity, negativity)
1. $P(r) = 0$, $N(r) = 0$
2. For each sentence s in r do
3. Use Natural Language Parser(NLP) to identify all nouns, verbs, adjectives and adverbs in s
4. Use word sense Disambiguation(WSD) technique to find the set C f all senses of these terms
5. Find the positive sentiment score P(c) and negative sentiment score $N(c)$ for each $c \in C$ using SentiWordNet3.0 database
6. Use Stanford Parser to find phrase structure tree in s
7. For each negation cue in s do
8. Identify the scope p of negation according to phrase structure tree
9. For each $c \in C$ in p do
10. $tmp = P(c)$
11. $P(c) = 0.9 * N(c)$
12. $N(c) = tmp * 0.9$
13. $P(s) = \sum_{c \in C} P(c)$
14. $N(s) = \sum_{c \in C} N(c)$
15. $P(r) = P(r) + P(s);\ N(r) = N(r) + N(s)$
16. Return $(P(r)/(P(r) + N(r)), N(r)/(P(r) + N(r)))$

Figure 3 - Algorithm for sentiment score detection

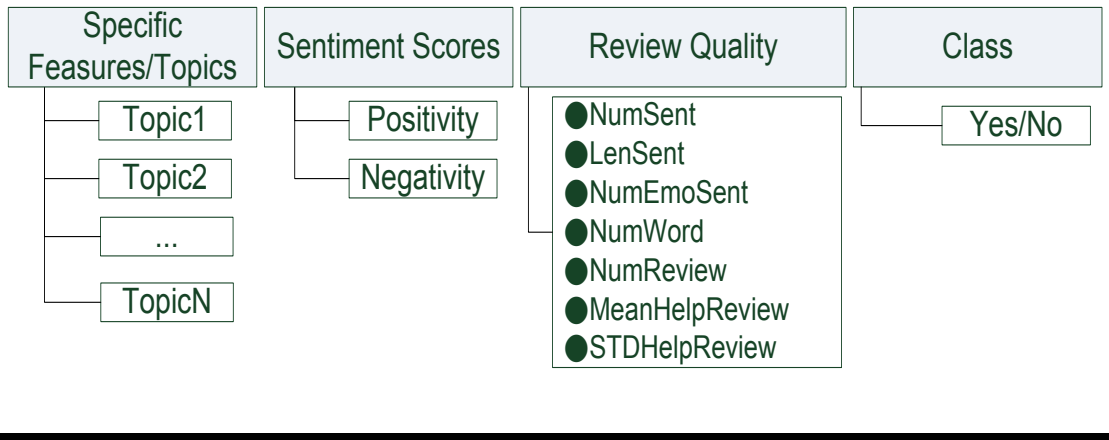


Figure 4 - Representations of reviews

Parameter Settings

In LDA, documents are generated by first picking a Dirichlett probability distribution Dir(α) for generating topics, and then, for each topic, a Dirichlett probability distribution Dir(β) is chosen for generating words. Here α (β) is a hyperparameter specifying the skewness on the topic (word) distribution (T. L. Griffiths & Steyvers, 2004; Hofmann, 2001). Smaller α (β) indicates a bias towards sparsity and results in picking topic (word) distributions favouring just a few topics (words) per document (topic). Based on previous research (T. L. Griffiths & Steyvers, 2004; Steyvers & Griffiths, 2007) , we set the parameters of LDA as follows: α = 1 and β =0.1. α = 1 yields a uniform distribution over a small number of topics. For β =0.1, it is intended that each topic is associated with only a relatively small number of terms out of 4000 terms (Blei, Griffiths, & Jordan, 2010). In addition, we compute the perplexity at different number of topics (Blei et al., 2003), and it was found that with topic number being 25, we are able to achieve the lowest (and best) perplexity.

We exercised several classification techniques, and SVM exhibited the best results. In the following, we present our performance results running using SVM. With respect to the parameter settings of SVM, we keep the default value and use Platt's Sequential Minimal Optimization (SMO) algorithm for training a support vector classifier. We perform 10-fold cross-validation and use average precision and recall on noteworthy reviews as the performance measures. F-measure also serves as a combinational measure. Definitions for Precision, recall and F-measure are listed as follows, where TP, FP, and FN denote true positive, false positive and false negative respectively:

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$$

Preliminary Experiment

Figure 5 shows the F-measure for T1+S+Q under different number of topics, where T1, S and Q denote the incorporations of topic features using method T1, sentiment features and quality features respectively. The result is consistent with our previous experiment using perplexity as the measure in that 25 topics indeed yield the best performance. In our subsequent experiments, we fix the number of topics at 25.

Performance Result

In Section 3.1, we have described three methods for representing content features, namely T0, T1, and T2. T2 is a semantic-based LDA method, where the identified senses may be extended by hypernyms, hyponyms, or similar senses. To evaluate the effect of the sense extension, we name the methods with and without the sense extension as T2 and T2' respectively.

Our next experiment compares the performance of different methods for specifying content features, namely T0, T1, T2, and T2', and the result is shown in Figure 6. It can be seen that T0 and T1 have higher precision and F-measure values. In contrast, the semantic-based methods (T2 and T2') have high recall values but lower precision values. We observed that the semantics-based methods tend to mistakenly predict not-noteworthy reviews as "noteworthy." After looking closely at the high frequency words for topics in T1, we find that quite a few proper names do not show up in WordNet, a general-purpose ontology. For example "Kaohsiung" (place name) or "Hi-Lai"(hotel name) do not appear in WordNet. We thus attribute the poor precision values of the semantic-based methods to the lack to tourism-specific concepts in the general ontology such as WordNet. In addition, T2' is slightly better than T2, which shows that sense extension improves the performance of the semantic-based method, though the

extent of improvement is small. Comparing T0 and T1, T1 has higher recall yet lower precision, and their F-measure values are comparable. For the identification of noteworthy reviews, however, recall is deemed more important than precision as missing a noteworthy review could cause drastic damage to the hotel. Besides, T1 utilizes only 25 content features, in comparison with 4000 TF-IDF features used in T0. Thus, we conclude that T1 is a promising method for representing content features of hotel reviews.

Figure 7 displays different combinations of the content features, T1, the sentiment features, S, and the quality features, Q. As can be seen, the full combination, namely T1+S+Q achieves the best performance. Comparing T1, S, and Q, we find that the content feature (T1) is most important because by excluding T1 (i.e., S+Q), precision drops drastically. This is because by incorporating T1, our approach is able to distinguish between those with and without relevant content. As a result, most noteworthy reviews (which address relevant subjects) will be included. S and Q are both important because by excluding either one, the recall values drop, though to a less degree when comparing to the precision drop by the lack of T1. This is because without quality or sentiment measures, some noteworthy reviews (with good quality and/or negative sentiment) may be mistakenly excluded. In addition, T1 alone achieves the performance comparable to the combination of sentiment features and quality features (S + Q).

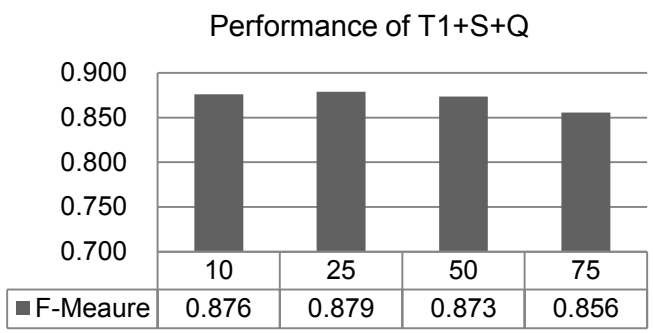
Conclusions

In this paper, we have proposed an effective multi-method approach to identifying hotel reviews that are noteworthy for hotel staff. It starts with a qualitative method for interviewing senior hotel managers to identifying characteristics of noteworthy online reviews. These characteristics can be categorized into three types, namely content features, sentiment features, and quality features, and we developed several methods for deriving these features. Through the experiments using tripadvisor.com data, we found that all the three types of features are important in predicting noteworthy hotel reviews. Specifically, content features have been shown to have most impact on precisions, whereas sentiment and quality features impact recalls. For deriving content features, we have proposed three methods. It has been shown that the LDA method achieves comparable performance to TF-IDF method with higher recall and much fewer features.

One unexpected result of our experiments is that the proposed semantic-based LDA method has lower precision than the word-based LDA method. We attribute the lower precision of the semantic-based method to the general-purpose ontology, namely WordNet, used by our method, which excludes quite a few proper names in tourism domain. For future research, we plan to exercise the semantic-based method by incorporating more domain-specific ontology, in the hope to further increase the performance of review recommendation for hotel managers. In addition, odd numbers of interviewees could be employed and majority voting be adopted to discern the presence of biases when conflicts arise.

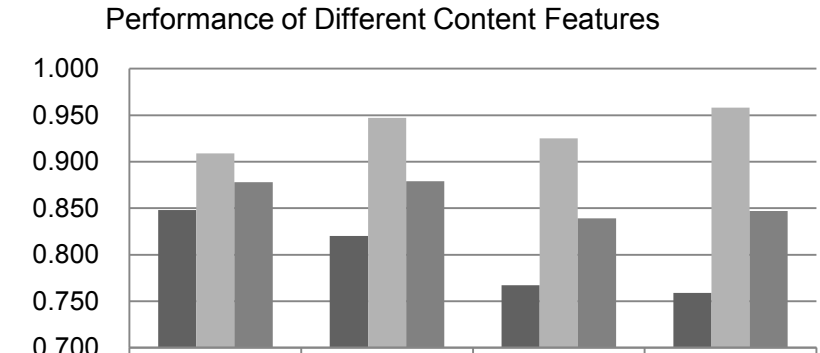
Acknowledgments

The authors would like to thank Ms. Evelyn Kung and Mr. Benson Lin, for their valuable input on the characteristics of noteworthy reviews. This research was supported by NSYSU Aiming for the Top University Project, 2014.



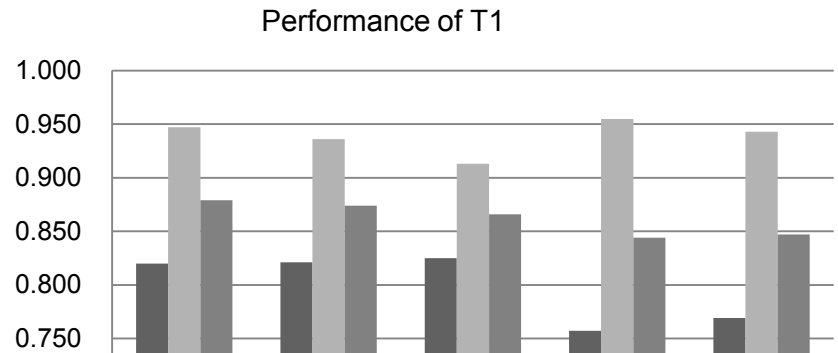
	10	25	50	75
F-Meaure	0.876	0.879	0.873	0.856

Figure 5 - F-measure of T1+S+Q under different numbers of topics



	T0+S+Q	T1+S+Q	T2+S+Q	T2'+S+Q
Precision	0.848	0.820	0.767	0.759
Recall	0.909	0.947	0.925	0.958
F-Measure	0.878	0.879	0.839	0.847

Figure 6 - Performance of different methods for representing content features



	T1+S+Q	T1+Q	T1+S	S+Q	T1
Precision	0.820	0.821	0.825	0.757	0.769
Recall	0.947	0.936	0.913	0.955	0.943
F-Measure	0.879	0.874	0.866	0.844	0.847

Figure 7 - Performance of different combinations of T1, S, and Q

References

Adomavicius, G., & Kwon, Y. (2007). New recommendation techniques for multicriteria rating systems. *Intelligent Systems, IEEE, 22*(3), 48-55.

Agirre, E., & Soroa, A. (2009). *Personalizing pagerank for word sense disambiguation.* Paper presented at the Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.* Paper presented at the LREC.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM), 57*(2), 7.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Carrillo-de-Albornoz, J., Plaza, L., Díaz, A., & Ballesteros, M. (2012). *UCM-I: a rule-based syntactic approach for resolving the scope of negation.* Paper presented at the In proceedings of the *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation.

Carrillo-de-Albornoz, J., Plaza, L., & Gervás, P. (2010). *A hybrid approach to emotional sentence polarity and intensity classification.* Paper presented at the Proceedings of the Fourteenth Conference on Computational Natural Language Learning.

Chen, G., & Chen, L. (2014). Recommendation Based on Contextual Opinions *User Modeling, Adaptation, and Personalization* (pp. 61-73): Springer.

Cheung, M. Y., Luo, C., Sia, C. L., & Chen, H. (2009). Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce, 13*(4), 9-38.

Chowdhury, G. (2010). *Introduction to modern information retrieval*: Facet publishing.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* Paper presented at the Proceedings of the 12th international conference on World Wide Web.

Dong, R., Schaal, M., O'Mahony, M. P., & Smyth, B. (2013). *Topic extraction from online reviews for classification and recommendation.* Paper presented at the Proceedings of the Twenty-Third international joint conference on Artificial Intelligence.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems, 45*(4), 1007-1016.

Fuchs, M., & Zanker, M. (2012). Multi-criteria ratings for recommender systems: An empirical analysis in the tourism domain *E-Commerce and Web Technologies* (pp. 100-111): Springer.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on, 23*(10), 1498-1512.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced

content. *Marketing Science, 31*(3), 493-520.

Griffiths, T. (2002). Gibbs sampling in the generative model of latent dirichlet allocation. *Standford University, 518*(11), 1-3.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proc Natl Acad Sci U S A, 101*(Suppl 1), 5228-5235.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly, 28*(1), 75-105.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning, 42*(1-2), 177-196.

Huang, S., Shen, D., Feng, W., Baudin, C., & Zhang, Y. (2010). Promote product reviews of high quality on e-commerce site. *Pacific Asia Journal of the Association for Information Systems, 2*(3), 51-71.

Hycner, R. H. (1985). Some guidelines for the phenomenological analysis of interview data. *Human studies, 8*(3), 279-303.

Jannach, D., Gedikli, F., Karakaya, Z., & Juwig, O. (2012). Recommending hotels based on multi-dimensional customer ratings *Information and Communication Technologies in Tourism 2012* (pp. 320-331): Springer.

Klein, D., & Manning, C. D. (2003). *Accurate unlexicalized parsing.* Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.

Klein, L. R. (1998). Evaluating the potential of interactive media through a new lens: search versus experience goods. *Journal of business research, 41*(3), 195-203.

Lesk, M. (1986). *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.* Paper presented at the Proceedings of the 5th annual international conference on Systems documentation.

Levi, A., Mokryn, O., Diot, C., & Taft, N. (2012). *Finding a needle in a haystack of reviews: cold start context-based hotel recommender system.* Paper presented at the Proceedings of the sixth ACM conference on Recommender systems.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., & Zhou, M. (2007). *Low-Quality Product Review Detection in Opinion Summarization.* Paper presented at the EMNLP-CoNLL.

Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., & Moos, W. H. (1995). Measuring diversity: experimental design of combinatorial libraries for drug discovery. *Journal of medicinal chemistry, 38*(9), 1431-1436.

Moghaddam, S., Jamali, M., & Ester, M. (2011). *Review recommendation: personalized prediction of the quality of online reviews.* Paper presented at the Proceedings of the 20th ACM international conference on Information and knowledge management.

Musat, C.-C., Liang, Y., & Faltings, B. (2013). *Recommendation using textual opinions.* Paper presented at the Proceedings of the Twenty-Third international joint conference on Artificial Intelligence.

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and organization, 17*(1), 2-26.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR), 41*(2), 10.

O'Mahony, M. P., & Smyth, B. (2009). *Learning to recommend helpful hotel reviews.* Paper presented at the Proceedings of the third ACM conference on Recommender systems.

O'Mahony, M. P., & Smyth, B. (2010). A classification-based review recommender. *Knowledge-Based Systems, 23*(4), 323-329.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval, 2*(1-2), 1-135.

Sparks, B. A., Perkins, H. E., & Buckley, R. (2013). Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. *Tourism Management, 39*, 1-9.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267-307.

Turney, P. D. (2002). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.* Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.

Yang, C. C., Tang, X., Wong, Y., & Wei, C.-P. (2010). Understanding online consumer review opinions with sentiment analysis using machine learning. *Pacific Asia Journal of the Association for Information Systems, 2*(3), 73-89.

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management, 28*(1), 180-182.

Zhang, K. Z. K., Zhao, S. J., Cheung, C. M. K., & Lee, M. K. O. (2014). Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems*.

Zhu, F., & Zhang, X. (2006). *The influence of online consumer reviews on the demand for experience goods: The case of video games.* Paper presented at the International Conference on Information Systems, Milwaukee.

About the Authors

San-Yih Hwang received the B.Sc. and M.Sc. degrees from National Taiwan University, Taiwan, and the Ph.D. degree from the University of Minnesota, Minneapolis in 1994, all in computer science. He joined the Department of Information Management at National Sun Yat-sen University, Taiwan, in 1995 and is presently a professor. His current research interests include text mining, recommender systems, and services computing.

Chiayu Lai is currently a Ph.D. candidate in the department of information Management at National Sun Yet-sen University, Taiwan. Her current research interests include text-mining, machine learning and recommender system.

Shanlin Chang is a Ph.D. candidate in the Department of Information Management at National Sun Yat-sen University, Taiwan. Her research interests include recommender systems, data analysis and text mining. She has published in Pacific Asia Conference on Information Systems.

Jia-Jhe Jiang is currently a project engineer at Darling Life Technical in Taiwan. He received the MBA degree from National Sun Yat-sn University in 2014.