Inf Syst Front (2018) 20:617–625
DOI 10.1007/s10796-016-9697-z

Product recommendation with latent review topics

Juheng Zhang[1] · Selwyn Piramuthu[2]

Published online: 6 September 2016

Abstract Online customer reviews complement information from product and service providers. While the latter is directly from the source of the product and/or service, the former is generally from users of these products and/or services. Clearly, these two information sets are generated from different perspectives with possibly different sets of intentions. For a prospective customer, both these perspectives together provide a complementary set of information and support their purchase decisions. Given the different perspective and incentive structure, the information from these two source sets tends to be necessarily biased, clearly with the high probability of negative information omission from that provided by the product/service providers. Moreover, customers oftentimes face information overload during their attempts at deciphering existing online customer reviews. We attempt to alleviate this through mining hidden information in online customer reviews. We use a variant of the Latent Dirichlet Allocation (LDA) model and clustering to generate equivalent options that the customer could then use in their purchase decisions. We illustrate this using online hotel review data.

Keywords Latent Dirichlet Allocation · Hidden information · Online reviews

1 Introduction

Most online and some brick & mortar (B&M) sellers of products and services provide information that is generally biased, with the ultimate goal of nudging prospective customers to decide in their favor. However, the widespread availability of online and offline customer reviews provides some (albeit imperfect) balance to information from sellers. While seller-generated information generally tends to be positive in terms of support to the seller's products and services, customer reviews cover the entire positive–negative spectrum as appropriate and warranted (e.g., Hu et al. 2011, 2012). Since customer reviews are based on their experiences with the products/services, the possibility for such reviews to contain specific information that is not available directly from the seller is high. A drawback associated with customer reviews is that useful information nuggets tend to hide among other not so useful information. Given the availability of a large number of customer reviews, it is a challenge for an average customer to sift through large volumes of reviews in order to gather actionable knowledge.

From picking a hotel to choosing a dress, customer purchase experiences more often than not attest to the fact that the products they see online and their formed perceptions do not necessarily translate to what they actually experience with the actual product. For example, dresses may not be true to size or color. Hotels may look like a completely different property from what is seen online at the hotel's Web site. A recent collection of "photo fake-outs" released on Oyster.com (Zeveloff 2013) shows swimming pools that are small in reality and are "upgraded" to beautiful skyline ones on hotel

✉ Selwyn Piramuthu
selwyn@ufl.edu

Juheng Zhang
Juheng_Zhang@uml.edu

[1] Operations and Information Systems, University of Massachusetts, Lowell, MA 01854, USA

[2] Information Systems and Operations Management, University of Florida, Gainesville, FL 32611, USA

websites; hotel rooms with a view of concrete buildings pictured as river-view rooms. Clearly, hotels often accentuate the positive or display altered perspectives to increase their appeal to prospective customers. Similarly, online sellers may only release favorable information on their products or services to attract more potential buyers. For instance, on eBay, an online seller can easily hide bad transaction record from potential buyers by registering a new ID (Ba 2001; Baron 2002). Companies often strategically present financial news to attract more investors (Healy and Palepu 2001; Hirshleifer and Teoh 2003). The problem of information asymmetry (Akerlof 1970) is widely observed in electronic markets.

If they solely rely on the information disclosed on corporate or seller Web sites, customers have a high probability of making decisions with less desirable outcomes. For example, customers who rely completely on the information that is provided on a hotel's Web site may encounter an unpleasant surprise during their stay at the hotel. A customer can easily overestimate an investment opportunity when complete trust is placed on the sale materials presented in corporate documents. Sufficient information of high quality is inherently the foundation of good decisions, and such information can be intentionally or unintentionally distorted at the source.

When faced with the availability of only limited and possibly biased information, prospective customers turn to alternative information sources such as customer reviews or User Generated Content (UGC) for more information to help them become better informed. UGC refers to content that is generated by open collaboration of users and is available through a variety of media that include product reviews, blogs, among others. According to the Nielsen report in the year of 2012, customers consider reviews on UGC Web sites to be more trustworthy compared to information listed by marketers in general. For example, when choosing a local restaurant, diners consider ratings and reviews on yelp.com; travelers check the popularity index of hotels on TripAdvisor.com or other travel Web sites to help with their decision to choose and book a hotel. Online reviews are helpful for customers to discover more information on a product or service.

Customers often use information discovered through reviews in their decision making process. For example, it is not uncommon for customers to check whether the rooms are located in close proximity to a noisy environment when booking hotel rooms; words such as "train tracks nearby the hotel" in reviews can certainly have a high impact during the customers' decision-making process, and ultimately in the elimination of that hotel from the considered set of hotels. When reading reviews, customers identify information based on their decision criteria. We refer the information to be discovered in the reviews as "hidden information" or "hidden topics", since such information is normally not available through common sources such as the seller/company official Web sites. Given that such information is generally not available from the seller, customers search for possible hidden topics in online product reviews and incorporate them along with other related and relevant information when making purchase decisions. Note that hidden information discovery is different from feature selection. The hidden information or hidden topics discovered are the underlying topics in the collection of reviews, whereas in feature selection, the features are generally specifically selected words in text. For example, given a short text including words such as "train tracks nearby the hotel", feature selection extracts words "train", "track", "hotel" as features, while hidden information may group these words into a hidden topic such as "noise". The hidden information is defined in this study as the information of a product or service embedded in user reviews but unavailable in the sources provided by the company, e.g., its official website.

While the general idea of the existence of customer reviews as a source for complementary information is encouraging, the challenge is in finding usable hidden information among the huge volume of unstructured text comprising such reviews. Oftentimes, it is relatively resource (e.g., time) intensive and not realistic for anyone to peruse all existing related reviews on a product or service. For example, on TripAdvisor.com, the Flamingo Las Vegas Hotel & Casino has more than 12,000 reviews. Customer reviews are often in the form of unstructured text, and lack the consistency of the features of products to be covered. Taking the reviews of a hotel as an example, some reviewers talk about how crowded the swimming pool is; some reviewers focus on the extra fees charged to access the fitness room; others are concerned about the construction noise due to ongoing renovation activities. Customer reviews are based on personal experience, and are written based on context-specific experiences and different perspectives. Seldom does one single customer review cover all possible facets and aspects of a product or service. Nevertheless, to be useful to a prospective customer, essential implicit and explicit information from all related reviews need to be somehow distilled into a compact, understandable, and usable format.

Due to the sheer volume and the unstructured format of customer reviews and associated resource constraints that are imposed on anyone perusing such reviews, a tool that automatically discovers and extracts hidden topics in reviews is needed. In this paper, we illustrate how to accomplish automated discovery of hidden topics in online reviews with the use of hotel reviews downloaded from TripAdvisor.com. The hidden topics are underneath the threads of online reviews for any product. For a given set of potentially hidden characteristics, we identify the products/services that provide better values and generate associated product/service recommendations to customers based on the topics of interest. The contributions of this study are summarized as follows.

First, we identify and incorporate hidden topics that are discovered through customer reviews into the customers' decision-making process. While existing studies on UGC have studied the value of customer reviews as well as UGC in general, few of them consider hidden topics present in reviews in customer decision models. We use a variant of the Latent Dirichlet Allocation (LDA) model to automatically extract the hidden topic structure from customer reviews. For each review, the hidden topics and the probability of being assigned to each such topic are inferred. The inferred topic probabilities are then incorporated into the decision-making process.

Second, after the preprocessing (Farquad and Bose 2012) step, we conduct clustering analysis (Bose and Chen 2015) to identify products with better values based on the topics of interest. Similar products in the same cluster are considered to be the ones that offer the same level of utility to customers in terms of related topics. The products that offer the same level of utility but with lower prices are identified as those with better values.

The rest of the paper is organized as follows. We review papers relevant to this study in Section 2. We discuss the concepts with an illustrative example in Section 3. In Section 4, we propose an algorithm that discovers hidden topics in product reviews with the LDA model, and identify the products with better values based on the inferred topics. In Section 5, we include description of the data and discuss results from our experiments with the downloaded hotel review data from TripAdvisor.com. We provide detailed discussion on the hidden topics of the reviews. In addition, we include results for the reviews with a specific rating. We conclude the paper and discuss future extensions in Section 6.

2 Relevant background

There is an extensive set of extant published research on online customer reviews that consider various facets of this interesting area. Given that there are also several papers that survey related research in this general area (e.g., Cantallops and Salvi 2014; Dolnicar and Otter 2003; Leung et al. 2013), we do not attempt to replicate the process of reviewing existing literature here. Also related to this study are papers that mine user-generated product/service reviews in the hotel context as well as study the general dynamics associated with user-generated product/service reviews (e.g., Burgess et al. 2011; Litvin et al. 2008; McCarthy et al. 2010; Sparks and Browning 2011; Yan et al. 2015; Zhang 2015). It should be noted that a majority of these studies consider the star ratings provided by reviewers, the general sentiment (positive or negative) associated with such reviews, as well as explicitly present attributes from the provider's (i.e., Hotel Web sites) own description of their services (e.g., physical location of the hotel, price, available services).

We are interested in *implicit* knowledge that is *hidden* in customer-generated reviews, specifically those that are associated with hotel reviews. Such implicit or hidden knowledge most likely is not present in the explicitly available information from hotel Web sites. However, through first-hand experiences during their hotel stays, past customers have the necessary knowledge to write reviews from a customer-based perspective. We believe that such a perspective provides a complementary set of information for a prospective customer to obtain well-rounded knowledge on hotels before making an informed stay decision. Clearly, reviews by different customers will be biased based on their unique requirements. For example, while the absence of non-vegetarian options at the hotel restaurant breakfast offering might be irrelevant for a vegan, it may not necessarily be so for someone expecting a non-vegetarian option. Similarly, while the exact physical location of the hotel (e.g., near a touristy area) may not be important for a budget-minded customer, it may be relevant for a customer whose stay focus is on proximity to touristy areas. The reviews from these two `types' of customers naturally will tend to focus on different aspects of the hotel. To get a complete picture of hidden knowledge, and to avoid idiosyncrasies associated with a small fraction of customers, it is therefore necessary to simultaneously consider reviews from several different customers and to identify elements that are repeated by several customers.

A related set of literature is on feature selection (e.g., Piramuthu 1999, 2004; Piramuthu et al. 2012), where features that are relevant to a concept of interest are extracted for further analysis. Another relevant area includes topic models (Blei 2012; Blei et al. 2003; Hofmann 2001). In topic models, probabilistic Latent Semantic Analysis (pLSA) uncovers the hidden thematic structure in the collection of documents (Hofmann 2001). Each document is considered as a bag-of-words and a mixture of latent topics. The pLSA method connects words with similar meanings and distinguishes between uses of words with multiple meanings. LDA (Blei 2012; Blei et al. 2003) further improves pLSA and offers solutions to the problems of pLSA. In pLSA, the topic mixture weights are only learned for trained documents. It cannot assign topic probabilities to unseen document. It does not provide generative probabilities of topic proportions. The pLSA method is based on point estimation while the LDA method uses a full topic simplex. The LDA method provides a generative probabilistic model for collection of documents. The Bayesian approach used in the LDA model resolves the problem of overfitting which we may see in the pLSA. The LDA method also can provide reasonable estimation when sample data is small. Related to this study also includes research on strategically hidden information (Zhang and Aytug 2016; Zhang et al. 2014), in which information is strategically hidden by information providers and decision makers deal with strategically hidden information in data analytics.

3 An illustrative example with hotel reviews

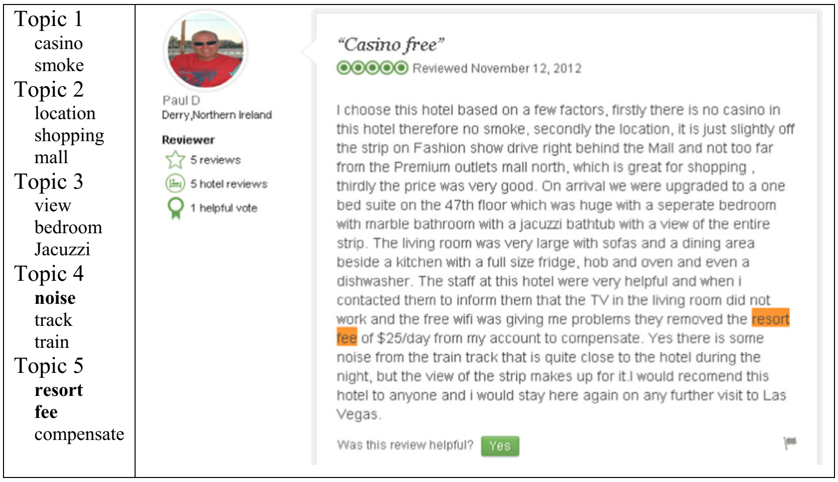
We illustrate the general idea of this study with the use of a hotel review on TripAdvisor.com. The hotel review is listed in Fig. 1. It is about the Flamingo Las Vegas Hotel & Casino. As we read the review, we see that the reviewer commented on the casino and the hotel's location, price, bathroom, and free Wi-Fi. More interestingly, he mentioned in the review that there was originally "the resort fee of $25 per day" and "some noise from the train track that is quite close". The information of the resort fee and the noise from nearby train track is not available on the hotel's website, but can be quite valuable for travelers. The idea of this study is to find such hidden information in reviews and use it for product recommendation.

We list topics that are embedded in this review on the left side of Fig. 1. Topics one, two, and three are about casino, location, and view respectively. Topic four is about train track noise and topic five is on resort fee.

We plot the Flamingo hotel in a simplex based on its probabilities of the topics. For ease of visualization, here we use three of the topics, i.e., location, view, and resort fee, to plot the simplex. In practice, consumers or recommendation systems can choose the topics of interest and the number of the topics of interest. We also download the reviews of two other hotels in the same city, enVision Hotel and Haborside Inn, and plot the two hotels in the same simplex based on their probabilities of the three topics in the reviews. The simplex with the three hotels is shown in Fig. 2. As can be seen in Fig. 2, enVision Hotel is located next to Flamingo Hotel in the simplex, which suggests that these two hotels have similar topic assignments and offer the same level of utility in terms of location, view, and resort fee to consumers. The Haborside Inn, on the other side, is located away from the Flamingo Hotel. Its probability of topic "view" is zero, and it is located on the edge connecting the topic location and resort fee. The Haborside Inn is therefore considered different from the Flamingo Hotel in these three aspects.

Given that the 3.5-star hotel Flamingo charges $190 per night, the 3-star hotel enVision with a price of $130 per night offers a better value to travelers since it provides a similar level of utility to consumers but at a lower price. The Haborside Inn, the 3.5-star hotel with a price of $200 per night is not better than Flamingo. Flamingo and enVision hotels are grouped together based on their similarity in terms of the considered topics. The simplex in Fig. 2 illustrates how we can identify the hotels that provide better value to consumers in terms of the topics of interest.

In the example, we considered only one review for each hotel for illustration purposes. In our experiments discussed later, each hotel has thousands of consumer reviews and our proposed method automatically discovers the hidden topics in the collection of the reviews and the topic probabilities of hotels. Similar hotels are grouped together based on the topic probabilities. The hotels clustered within the same group are considered as providing the same level of utility to consumers. Among the hotels in the same cluster, the ones with lower prices can be recommended to consumers. Following this methodology, customers can either (a) name the topics of interest or (b) choose from the identified topics. The proposed method can recommend the hotels with high values to consumers. Please note that the identified/chosen topics of interest need not necessarily be from the hidden set – the hidden set complements the evoked set of characteristics. As for standard product characteristics, companies often make the information available to consumer, e.g., weight or size of a product, so it is unnecessary to discover the characteristics using data mining tools. The topics of interest can include hidden product features and/or standard characteristics.



Topic 1
casino
smoke
Topic 2
location
shopping
mall
Topic 3
view
bedroom
Jacuzzi
Topic 4
noise
track
train
Topic 5
resort
fee
compensate

Paul D
Derry,Northern Ireland

Reviewer
5 reviews
5 hotel reviews
1 helpful vote

"Casino free"

Reviewed November 12, 2012

I choose this hotel based on a few factors, firstly there is no casino in this hotel therefore no smoke, secondly the location, it is just slightly off the strip on Fashion show drive right behind the Mall and not too far from the Premium outlets mall north, which is great for shopping , thirdly the price was very good. On arrival we were upgraded to a one bed suite on the 47th floor which was huge with a seperate bedroom with marble bathroom with a jacuzzi bathtub with a view of the entire strip. The living room was very large with sofas and a dining area beside a kitchen with a full size fridge, hob and oven and even a dishwasher. The staff at this hotel were very helpful and when i contacted them to inform them that the TV in the living room did not work and the free wifi was giving me problems they removed the resort fee of $25/day from my account to compensate. Yes there is some noise from the train track that is quite close to the hotel during the night, but the view of the strip makes up for it.I would recomend this hotel to anyone and i would stay here again on any further visit to Las Vegas.

Fig. 1 Information hidden in online reviews

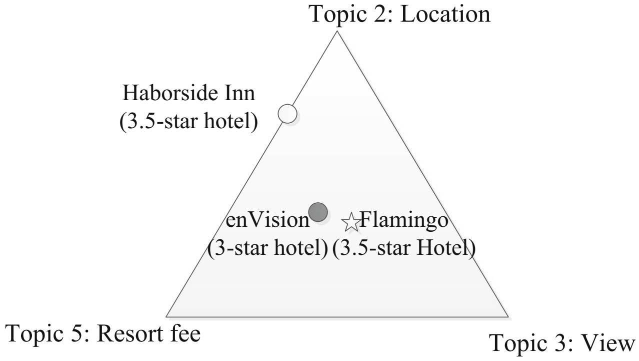


Fig. 2 Simplex plot with three topics

4 Algorithm based on Latent Dirichlet allocation model

4.1 Latent Dirichlet allocation model

From a text mining perspective, the latent Dirichlet allocation (LDA) model (Blei et al. 2003) is based on the assumption of bag-of-words and the exchangeability of words. Entities in this model form a hierarchical structure. We have a text *corpus* which comprises a collection of *documents*. Each such document comprises a number of words that are the basic building blocks among the modeled entities.

In the LDA model, each document is considered as a distribution over latent *topics*, and each topic is a distribution over words. Each word is assigned to a topic with a certain probability, and words are chosen from the corresponding topic. The observed variables are the words in all documents. The variables to estimate are the topics, topic distribution of each document, and topic assignment of each word in each document. These variables consist of the hidden structure of topics in the corpus.

Each word w is drawn from a vocabulary that is indexed by vectors with index values of $\{1, \ldots, V\}$. A word w is represented by a vector in which only a single element equals one and all other elements are set to zero. The element with the value of one corresponds to the index where the word is represented in the vocabulary. The corpus is the collection of documents, $D = (d_1, d_2, \ldots, d_m)$. In this study, each document is the collection of online reviews of each hotel. The corpus is the collection of all of the hotel reviews. The LDA method comprises a procedure that operates as follows.

(1) For each document d, choose the number of words, $N \sim Poisson(\xi)$.
(2) For each document d, the topic proportions θ_d are drawn from the Dirichlet distribution, $\theta_d \sim Dirichlet(\alpha)$, where α is the vector of parameters of the Dirichlet distribution, and each element of $\alpha > 0$.
(3) For each word in the document, draw a topic assignment, $z_{d,n} \sim Multinomial(\theta_d)$.
(4) Draw the specific word from the word distribution over vocabulary with condition on the topic assignment $z_{d,n}$, $w_{d,n} \sim Multinomial(\beta, z_{d,n})$, where β is the word distribution, $\beta \sim Dirichlet(\eta)$.

The Dirichlet distribution is a multivariate generalization of the beta distribution. Dirichlet distributions are often used for prior distributions in Bayesian statistics. To infer the hidden structure given the observed documents, we compute the following posterior distribution,

$$p\left(\beta, \theta, z \middle| w\right) = \frac{p(\beta, \theta, z, w)}{p(w)}.$$

While the numerator in the above equation, $p(\beta, \theta, z, w) = p(\beta)p(\theta)p(z|\theta)p(w|\beta, z)$, can be computed, the denominator is the evidence or the probability of observing the corpus under any topic model, which can be intractable to compute and is generally approximated. The two types of approximation methods of topic modeling include sampling based algorithms and variational algorithms (Blei et al. 2003; Jordan et al. 1999).

4.2 The proposed algorithm

We propose an algorithm to recommend a product (a hotel in this study) with a better value in terms of the product features of interest. The algorithm discovers the hidden topics and topic proportions in the reviews based on the LDA model. Similar products based on the topics of interest are identified through cluster analysis. Among the products in the same cluster, the products with lower price are chosen for recommendation to prospective customers. These products are considered to be the ones with better value since they have a lower price but with similar characteristics as relatively expensive products. The proposed algorithm is given in Fig. 3.

The input to this algorithm is a collection of documents – online customer reviews for a set of selected hotels. We randomly chose this set of hotels to illustrate the concepts of interest. Since the purpose here is purely illustrational, any bias due to the representativeness of this chosen set to the universe of hotels or otherwise is irrelevant to this study. Step 1 (Fig. 3) comprises the following: A vocabulary that comprises a set of words that are deemed to be significant is generated from the collection of documents; the number of occurrences of each of these words in the corpus (i.e., the collection of all reviews from all hotels considered) are then counted; this frequency count is normalized, and then compared with the inverse document frequency, which is a normalized count of the word frequency in the entire corpus; the resulting matrix is the term-by-document (tdm) matrix, and it

Fig. 3 The proposed algorithm based on a variant of LDA

```
1. Given a collection of documents, $D=(d_1,\ldots,d_i,\ldots d_N)$, generate the term-by-document-matrix (tdm).
2. Based on the tdm generated in (1), discover the top $K$ hidden topics, $T=(t_1,\ldots,t_k,\ldots,t_K)$ with the LDA model. For $d_i$, $i=1,\ldots,N$, compute the topic proportions, $P_i=(p_{i1},\ldots,p_{ik},\ldots,p_{iK})$.
3. Specify the topics of interest, $Ts\subset T$.
4. Cluster the documents based on Euclidean distance, $\mathbf{min}\sum_{j}\sum_{i\in C_j,k\in Ts}(p_{ik}-u_{jk})^2$, where $\mu_{jk}$ is the topic $k$ of the center of the cluster $C_j$.
5. $\forall i\in C_j$, find: $\arg\max_i\{utility_i/price_i\}$, where $utility_i=\sum_{k\in Ts}weight_{ik}*p_{ik}$.
```

comprises the tf-idf values of each of the documents as its columns. The tf-idf (term frequency – inverse document frequency) represents the importance of a word in a document that is a part of a corpus. The tf-idf value increases with the word frequency in a document, and is offset by the same word's frequency in the entire corpus in order to address the fact that some words are inherently more frequent than others.

Based on the top K hidden topics that are generated in Step 2 as well as the topic proportions, we obtain generic specifications for the set of topics. We use a Euclidean distance based clustering (e.g., Seret et al. 2014) method in Step 4. The purpose here is not to introduce a new clustering method, but rather use one to show the utility of the proposed algorithm in being able to generate a recommended set of hotels based on the utility per unit price (Step 5).

The proposed algorithm can therefore be used to extract the hidden topics from documents and generate customer recommendations based purely on (hidden) attributes that are similar for the different products considered, but with better value propositions due to lower price. The value is measured by the ratio of utility and price, which may be replaced by other measurements based on the context of interest.

5 Experiments

We used customer review data on 2000 hotels that we randomly selected from TripAdvisor.com. The corpus for this study is the collection of all customer review text of the 2000 hotels. Each hotel is a unit in the corpus. The experiments include topic inference and cluster analysis for product recommendation.

5.1 Data description

We downloaded the hotel reviews from TripAdvisor.com with a Java-based crawler. We then parsed the reviews so only the text reviews for each hotel are kept to be processed further. We filtered out numeric rating of hotels, reviewer information, the date of the review, and other attributes that are not relevant for this study. In addition, we accessed an open database of hotels at http://api.hotelsbase.org/ for information on hotel star classes and prices. This is a database with thousands of hotels listed with their year around price, star rating, hotel name, location, facilities, etc. We matched the hotels listed in the database with the ones downloaded from TripAdvisor.com based on the hotel name and location. We used this as our data set for further analysis.

5.2 Results

5.2.1 Top topics in hotel reviews

The topics discovered from the hotel reviews through LDA are listed in Table 1. As shown in Table 1, we list ten topics that include "Beach", "Service", "Location", "Check", "Cleanliness", "Parking", "Shuttle", "Pool", "Distance" and "Time." We also list the top terms under each topic in decreasing order of frequency from top to bottom in Table 1. For example, under "Service" topic, we have as the top frequent terms, "service", "staff", "love", "great", "excel", and so on. Under "Location" topic, the top words are "great", "location", "love", "place", etc. Each hotel is assigned to these top topics with associated probabilities.

The topic distribution graph is shown in Fig. 4. The histogram plot in Fig. 4 illustrates the topic probabilities. The most likely topics in all of the hotels' reviews are no.5 and no.9, which are "cleanliness" and "distance" topics respectively as shown in Table 1. The least likely topics are no.1 and no.10, which are respectively the topics "beach" and "time".

A plot of cloud of words that are commonly shared across all reviews is shown in Fig. 5. The word cloud allows for quick visualization to highlight the most frequent words in the reviews. As can be seen in Fig. 5, "great", "good", "time", "location", "staff", "service", "clean", and "place" are the most frequent words in the collection of reviews. Some words in Fig. 5 are truncated, such as "restaur" for "restaurant" because we preprocess the words with a standard stemming document. Stemming is commonly used in text mining to reduce inflected words to their root form (e.g., Feldman and Sanger

Table 1 Topics identified from hotel reviews

"Beach"	"Service"	"Location"	"Check"	"Cleanness"	"Parking"	"Shuttle"	"Pool"	"Distance"	"Time"
resort	service	great	nice	clean	parking	good	Pool	location	Great
beach	staff	location	great	place	free	airport	Beach	good	time
food	love	love	desk	bed	car	shuttle	View	great	location
Time	great	place	check	look	staff	service	Great	walk	staff
good	excel	help	good	time	walk	shop	Place	breakfast	good
great	restaurant	breakfast	bed	bathroom	good	breakfast	Nice	staff	new
pool	bar	friend	floor	staff	nice	food	Ocean	clean	floor
people	best	wonder	service	book	clean	free	Area	nice	bed
restaurant	view	staff	park	check	location	area	restaurant	station	clean
Nice	beautiful	recommend	location	door	great	taxi	Time	help	friend

2007). For example, "location" and "located" are mapped to the stem "locat."

5.2.2 Recommendation of hotels with better value

The hotels are clustered together using the discovered topics and the topic probability assignments. We use a sample of 5-star, 4.5-star, and 4-star hotels to illustrate how the better-valued hotels within the same star level are identified.

Figure 6 is a plot of the hotel samples in the three dimensional cube for illustration purposes, although the actual cluster analysis is conducted on all of the topics of interest. The top three topics in the reviews of these hotels are topic 2, topic 8, and topic 10. The 5-star hotels are plotted as red solid dots, 4.5-star hotels are represented by blue triangles, and 4 star hotels are denoted with green circles. Depending on their topic assignments, the hotels are located in different positions in the cube. Also, note that the scale of axis is not necessarily between zero and one. The scale of each axis is between zero and the maximum probability assignment of document.

As can be seen in Fig. 6, there is an overlap between 5-star hotels and 4/4.5-star hotels. The hotels close to each other are considered similar in terms of these topics because the distance between two hotels is based on the topic assignments of hotels. The overlap shows that certain hotels are considered to be very similar by reviewers even though they are from different hotel star level categories. Thus, hotels in the overlapped area provide a similar level of utility in terms of topics of interest. Generally speaking, hotels with lower star levels charge lower prices. Therefore, the hotels that provide the same level of service and accommodation but charge lower

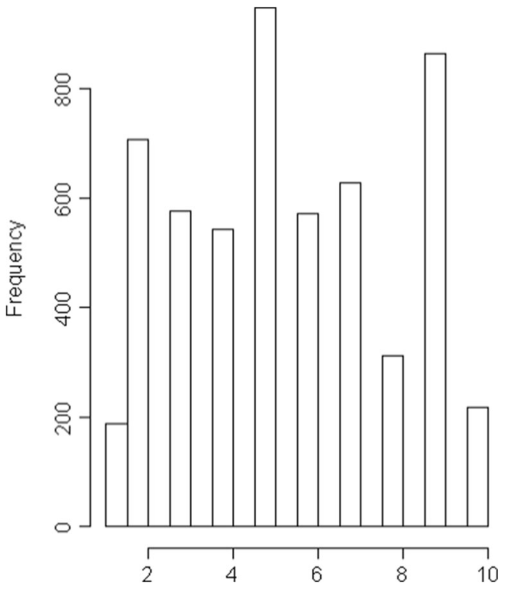


Fig. 4 Histogram of topics

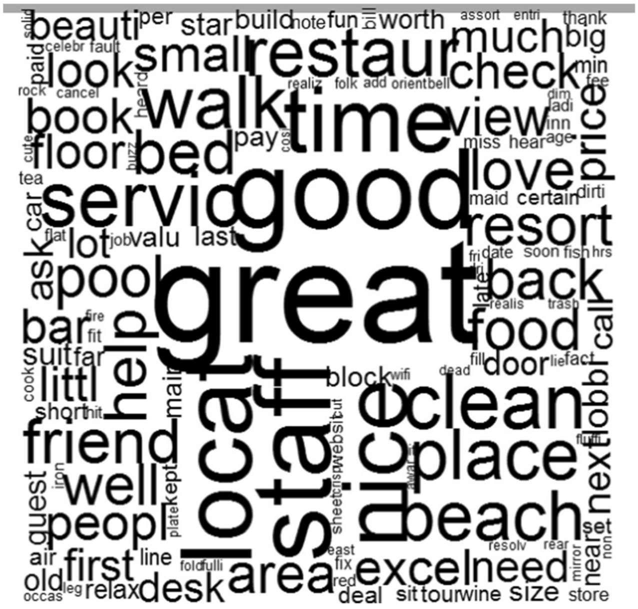


Fig. 5 Word cloud of hotel reviews

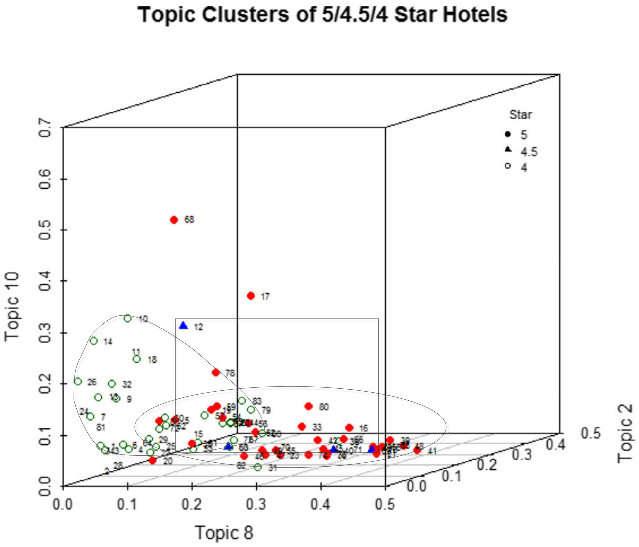


Fig. 6 Clusters of 5/4/4.5 star hotels

price can be recommended as good alternatives to hotels in the high price level.

5.2.3 Topics of hotel reviews with specific ratings

The above analysis is conducted on all the reviews of the hotel samples. We can discover hidden topics in the reviews and recommend hotels with better values. In practice, it can be also helpful for hotels to know which area they can improve to meet consumers' satisfaction. To do that, we now focus only on the negative reviews and consider what aspects of hotels made consumers leave negative comments.

Fig. 7 shows the word clouds for the reviews with numeric rating value “1”, “2”, or “3”. We use the word “rating" referring the hotel level rated by reviewers, different from the word “star” which level is rated by an official organization. We observe that the most frequent words in the reviews with overall rating value of “1” are different from those in the reviews with overall rating value of “2” or “3”.

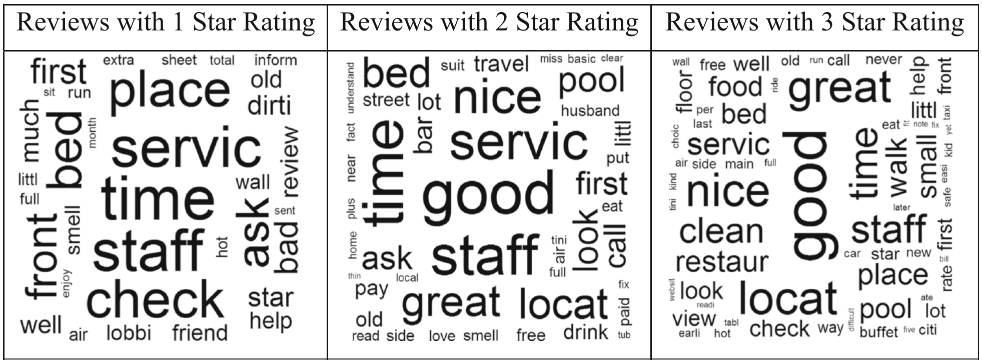
As can be seen in Fig. 7, for the negative reviews with rating value “1”, the most frequent words include the ones such as “bad”, “dirty”, “smell”, “hot”, and “old.” In addition, the word “never” is also in the list of top 50 frequent words, and, interestingly, the most associated word with “never” is the word “back.” The most frequent words in the reviews with rating value “2” include some nice words, “good”, “staff”, etc. But the words such as “old” and “smell” are frequently used.

6 Discussion and conclusion

We considered the hidden valuable information that is embedded in customer reviews and their incorporation in the customer decision-making process. Product/service information released on company/seller Web sites is often limited. Customer reviews are a valuable complementary source for prospective customers to find hidden information on products/services. Given the large amount of information in customer review repositories, it is a challenge for prospective customers to sift through and obtain usable knowledge. We developed a method and illustrated how to automatically discover the hidden topics in online reviews with a variant of the LDA model. Based on the discovered topics, we cluster the products that offer similar level of consumer utility in terms of topics of interest. Among the products in the same cluster, the ones with lower prices are considered to be those that offer better values since they are at the same utility level. For any prospective customer, given his or her topics or characteristics of interest, the proposed method generates a recommendation of products/services with better value based on the automatically extracted hidden information.

This study has some limitations. We assume that the hidden topics in the online reviews remain static over time and reviews are written independently of one another. In reality, those assumptions may not necessarily be true. Existing reviews may affect new reviews (e.g., Piramuthu et al. 2012), and the topics themselves may change over time. It would be interesting to study whether the topics of online reviews have evolved over the years as well as the existence of herding behavior (Dellarocas 2006) among online reviewers.

Fig. 7 Word cloud of reviews of 1/2/3 star hotels



References

Akerlof, G. A. (1970). The market for "Lemons": quality uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*(3), 488–500.

Ba, S. (2001). Establishing online trust through a community responsibility system. *Decision Support Systems, 31*, 323–336.

Baron, D. P. (2002). Private Ordering on the Internet: The Ebay Community of Traders. *Business and Politics* (4:3).

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*(2003), 993–1022.

Bose, I., & Chen, X. (2015). Detecting the migration of mobile service customers using fuzzy clustering. *Information & Management, 52*(2), 227–238.

Burgess, S., Sellitto, C., Cox, C., & Buultjens, J. (2011). Trust perceptions of online travel information by different content creators: some social and legal implications. *Information Systems Frontiers, 13*(2), 221–235.

Cantallops, A. S., & Salvi, F. (2014). New consumer behavior: a review of research on ewom and hotels. *International Journal of Hospitality Management, 36*, 41–51.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: implications for consumers and firms. *Management Science, 52*(10), 1577–1593.

Dolnicar, S., & Otter, T. (2003). Which Hotel Attributes Matter? A Review of Previous and a Framework for Future Research. *Proceedings of the 9th Annual Conference of the Asia Pacific Tourism Association (APTA)* (pp. 176–188). University of Technology Sydney.

Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems, 53*(1), 226–233.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. *Journal of Accounting and Economics, 31*(1–3), 405–440.

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics, 36*(1–3), 337–386.

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

Hu, N., Bose, I., Gao, Y., & Liu, L. (2011). Manipulation in digital word-of-mouth: a reality check for book reviews. *Decision Support Systems, 50*(3), 627–635.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: an analysis of ratings, readability, and sentiments. *Decision Support Systems, 52*(3), 674–684.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*(2), 183–233.

Leung, D., Law, R., Van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: a literature review. *Journal of Travel & Tourism Marketing, 30*(1–2), 3–22.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29*(3), 458–468.

McCarthy, L., Stock, D., & Verma, R. (2010). How travelers use online and social media channels to make hotel-choice decisions. *Cornell Hospitality Report, 10*(18), 4–18.

Piramuthu, S. (1999). Feature selection for financial credit-risk evaluation decisions. *INFORMS Journal on Computing, 11*(3), 258–266.

Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research, 156*(2), 483–494.

Piramuthu, S., Kapoor, G., Zhou, W., & Mauw, S. (2012). Input online review data and related bias in recommender systems. *Decision Support Systems, 53*(3), 418–424.

Seret, A., vanden Broucke, S. K., Baesens, B., & Vanthienen, J. (2014). A dynamic understanding of customer behavior processes based on clustering and sequence mining. *Expert Systems with Applications, 41*(10), 4648–4657.

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*(6), 1310–1323.

Yan, X., Wang, J., & Chau, M. (2015). Customer revisit intention to restaurants: evidence from online reviews. *Information Systems Frontiers, 17*(3), 645–657.

Zeveloff, J. (2013). Why You Should Never Trust the Photos Hotels Post Online. In *Business Insider*.

Zhang, J. (2015). Voluntary information disclosure on social media. *Decision Support Systems, 73*(2015), 28–36.

Zhang, J., & Aytug, H. (2016). Comparison of imputation methods for discriminant analysis with strategically hidden data. *European Journal of Operational Research, 255*(2), 522–530.

Zhang, J., Aytug, H., & Koehler, G. J. (2014). Discriminant analysis with strategically manipulated data. *Information Systems Research, 25*(3), 654–662.

Juheng Zhang is an Assistant Professor of Information Systems at University of Massachusetts, Lowell. She earned a Ph.D. in Management Information Systems from University of Florida in August 2011. Her research focuses on data analytics and data manipulation. Juheng Zhang has published in Information Systems Research, Decision Support Systems, European Journal of Operational Research, and other academic journals.

Selwyn Piramuthu is a Professor of Information Systems at the University of Florida. His research interests include recommender systems.