TOWARDS DESIGNING RANKING SYSTEMS FOR HOTELS ON TRAVEL SEARCH ENGINES: COMBINING TEXT MINING AND IMAGE CLASSIFICATION WITH ECONOMETRICS

Anindya Ghose, Panagiotis G. Ipeirotis, Beibei Li
Department of Information, Operations, and Management Sciences
Stern School of Business, New York University
{aghose, panos, bli}@stern.nyu.edu

Abstract

In this paper, we empirically estimate the economic value of different hotel characteristics, especially the location-based and service-based characteristics given the associated local infrastructure. We build a random coefficients-based structural model taking into consideration the multiple-levels of consumer heterogeneity introduced by different travel contexts and different hotel characteristics. We estimate this econometric model with a unique dataset of hotel reservations located in the US over 3 months and user-generated content data that was processed based on techniques from text mining, image classification, and on-demand annotations. This enables us to infer the economic significance of various hotel characteristics. We then propose to design a new hotel ranking system based on the empirical estimates that take into account the multi-dimensional preferences of customers and imputes consumer surplus from transactions for a given hotel. By doing so, we are able to provide customers with the "best value for money" hotels. Based on blind tests of users from Amazon Mechanical Turk, we test our ranking system with some benchmark hotel ranking systems. We find that our system performs significantly better than existing ones. This suggests that our inter-disciplinary approach has the potential to improve the quality of hotel search.

Keywords: Structural modeling, text mining, image classification, hotel search engine, user-generated content, ecommerce.

1. Introduction

It is now widely acknowledged that local search for hotel accommodations is a component of general Web searches that is increasing in popularity as more and more users search and reserve their trips online. Online travel search engines provide only rudimentary ranking facilities, typically using a single ranking criterion such as distance from the city center, star ratings, price per night, etc. This approach has quite a few shortcomings. First, it ignores the multidimensional preferences of the consumer in that a customer's ideal choice may consist of several hotel-specific attributes. Second, it largely ignores characteristics related to the location of the hotel, for instance, in terms of proximity to a "beach" or proximity to a "downtown shopping area". These location-based features represent important characteristics that can influence the desirability of a particular hotel. In this paper, using demand estimation techniques, we propose to estimate the weight that consumers place on different internal (service) and external (locational) characteristics of hotels. Thereafter, based on the estimation of consumer surplus we compute the "best value for money" of a particular hotel. The eventual outcome of our analysis is to design a new ranking system for hotels based on this concept. Such a ranking can be displayed in response to a user query on hotel search engines.

We undertake this study in the context of demand for hotel rooms using a unique dataset consisting of actual transactions and different kinds of user-generated content such as product reviews describing hotel service characteristics as well social geo-tags and on-demand annotations describing location characteristics. The theory that product reviews affect product sales has received support in prior empirical studies (for example, Chevalier and Mayzlin 2006). However, these studies have only used the numeric review ratings (e.g., the valence and the volume of reviews) in their empirical analysis. An emerging stream of work has begun to examine whether the textual information embedded in online user-generated content can have an economic impact (Ghose et al 2006, Das & Chen 2007, Archak et al. 2008, Ghose and Ipeirotis 2008, Ghose 2009). But these studies do not focus on estimating the impact of reviews in influencing real transactions nor do they aim to design new IT systems based on economic variables. Hence, another research objective in this paper is to analyze the extent to which user-generated content can help us learn consumer preferences for different hotel attributes.

Our work involves three stages. First, we need to identify the important hotel characteristics, both internal (service) and external (locational) that influence demand. Second, we need to empirically quantify the extent to which these characteristics influence demand. Finally, we aim to improve local search for hotels by incorporating the economic impact of these characteristics on consumer surplus from hotel transactions and designing a ranking system that incorporates "value for money" as a criterion for ranking. Any successful attempt to address these issues needs to answer the following questions: How can we automatically extract information about hotel attributes captured in textual content of product reviews and social tags, and visual content of satellite images? How can we incorporate extracted variables in a structural demand estimation model so as to be able to precisely identify parameter estimates?

A key challenge is in bridging the gap between the essentially textual and qualitative nature of review and image content and the quantitative nature of structural demand estimation models. With the rapid growth and popularity of user-generated content on the Web, a new area of research applying text mining techniques product reviews has emerged (for example, Hu & Liu 2004, Pang & Lee 2004, Das & Chen 2007). Similarly, advances in image classification have been made using non-parametric classifiers such as decision trees and support vector machines (Lu and Weng 2007). We use techniques from both these streams of work in finalizing our dataset (see below).

2. Data Description

We have complete information on all transactions conducted over a 3 month period from 2008/11–2009/1 for 2117 hotels in the US via Travelocity. These hotels were randomly selected by Travelocity. Further, we have data on hotel attributes from three sources: (i) service descriptions based on mining users' hotel reviews from Travelocity, (ii) location descriptions based on social geo-tags identifying different "external amenities" (such as shopping malls, restaurants, tourist attractions, etc) from Geonames.org, and (iii) user-contributed opinions on important hotel characteristics from Amazon Mechanical Turk such as whether a hotel is located "near the interstate highway", "near public transportation", "near the beach," etc.

Since some location-based characteristics, such as "proximity to the beach" and "distance from downtown", are not directly measurable based on reviews, tags or opinions, we use image classification techniques to infer such features from the satellite images of the area. We extracted hybrid satellite images (sized 256 × 256 pixels) using the Visual Earth Tile System, for each of the hotel venues, with 4 different zoom levels for each. These images were then used to extract information about the surroundings of the hotels. We performed a study to examine the performance of the classifiers. To perform the classification, we classified the out-of-sample images using Amazon Mechanical Turk; our results show that our SVM classifier has an accuracy of 91.2% for "Beach" image classification and 80.7% for "Downtown" image classification.

With regard to the service-based hotel characteristics, we extracted them from the website of TripAdvisor using fully automated JavaScript parsing engines. Since hotel amenities are not directly listed on TripAdvisor website, we retrieved them by following the link provided on the hotel web page, which randomly directs to one of its cooperative partner websites (i.e., Travelocity, Orbitz, Expedia, etc.).

We looked into two text-style features: "subjectivity" and "readability" of reviews. In order to capture more objectively the review text-style, we used a multiple-item method for subjectivity and readability. We included 2 sub-features for subjectivity and 5 sub-features for readability, each of which measures the review text-style from an independent point of view. In order to decide the probability of subjectivity for review text, we trained a classifier using as "objective" documents the hotel descriptions of each of the hotels in our data set. We randomly retrieved 1000 reviews to construct the "subjective" examples in the training set. We conducted the training process by using a 4-gram Dynamic Language Model classifier provided by the LingPipe toolkit. Thus, we were able to acquire a subjectivity confidence score for each sentence in a review, thereby deriving the mean and standard deviation of this score, which represent the probability of the review being subjective.

Finally, previous research suggested that the prevalence of reviewer disclosure of identity information is associated with changes in subsequent online product sales (Forman et al. 2008). Therefore, we decide to include one particular characteristic capturing the level of reviewers' disclosure of their identity information on these websites – "real name or location." These different data sources are then merged to create one comprehensive dataset (Table 1).

Table 1. Summary of Different Sources for Extracting Hotel Characteristics

Category	**Variables - Hotel Characteristics**		**Methods**
Transaction Data	• Transaction Price (per room per night)		Travelocity
Service-based	• Hotel Class • Internal Amenities ("ice machine," "pets allowed," "fitness center," "free breakfast", "wheelchair accessible", etc)		TripAdvisor
Review-based	• Number of Customer Reviews • Overall Reviewer Rating • Disclosure of Reviewer Identity Information		Travelocity and TripAdvisor
	Subjectivity	• Mean Probability • Std. Dev. of Probability	Text Mining Analysis
	Readability	• Number of Characters • Number of Syllables • Number of Spelling Errors • Average Length of Sentence • SMOG Readability Index	
Location-based	• Near the Beach • Near Downtown		Image Classification
	• External Amenities (Number of restaurants, shopping malls, historical sites, etc) • Number of Local Competitors Within 2 miles		Geonames and Virtual Earth Interactive SDK
	• Near the Interstate Highway • Near Public Transportation		Amazon Mechanical Turk (AMT)
	• City Annual Crime Rate		FBI online statistics

3. Model

In this section, we discuss how we develop our structural model and how we apply it to empirically estimate the distribution of consumer preferences for different hotel characteristics.

3.1 Random Coefficients-Based Structural Model

We define a consumer's decision-making behavior in the hotel market to be in accordance with the following two-step procedure. In the first step, the consumer aims to find a subset of hotels that best matches her travel context. For instance, if a consumer wants to go on a business trip, he would be more interested in a subset of hotels that specialize in business services; while if he plans to take his four-year kid for a family fun trip, he would be more likely to look for those hotels which are regarded as being kid-friendly. We have eight such unique category types in our data (Family Trip, Business Trip, Romantic Trip, Tourist Trip, Trip with Kids, Trip with Seniors, Pets Friendly and Disabilities Friendly). Then, in the second step, once the consumer has picked a corresponding subset of hotels which satisfy his travel requirement, he makes a further decision based on his evaluation of the value provided by the hotels.

Let the utility $u_{ij^k t}$ for consumer *i* from choosing hotel *j* with category type *k* in market *t* be as shown in equation (1),

$$u_{ij^k t} = X_{j^k t}\beta_i - \alpha_i P_{j^k t} + \xi_{j^k t} + \varepsilon_{it}^k, \tag{1}$$

where, *i* represents a consumer, j^k represents hotel *j* with category type *k* ($1 \le k \le 7$), and *t* represents a hotel market which in our case is defined as a "city-night" combination. In this model, β_i and α_i are random coefficients that capture consumers' heterogeneous tastes towards different observed hotel characteristics, *X*, and towards the average price per night, *P*, respectively. ξ represents the set of hotel characteristics that are unobservable to the econometrician. ε_{it}^k with a superscript *k* represents a travel context level "taste shock".

Consistent with prior research (Berry and Pakes 2007, Song 2008), we assume that β_i and α_i are distributed among consumers per some known statistical distribution, i.e., $\beta_i \sim (\beta_i \mid \bar{\beta}, \sigma_\beta)$ and $\alpha_i \sim (\alpha_i \mid \bar{\alpha}, \sigma_\alpha)$. Our goal is then to estimate the means (β_i , α_i) and the standard deviations (σ_β , σ_α) of these two distributions. The means correspond to the set of coefficients on hotel characteristics and on hotel price, which measure the average weight placed by consumers; while the standard deviations provide a measure of the consumer heterogeneity in those weights. Furthermore, these heterogeneities result from some particular demographic attributes of consumers. Hence, we assume that $\sigma_\alpha \sim \alpha_I I_i$, where I_i represents the income whose distribution can be inferred from consumer demographics; $\sigma_\beta \sim \beta_v v_i$, where $v_i \sim N(0,1)$ represents some random factor that will influences people's preferences towards individual hotel characteristics. Therefore, we rewrite our model in the following form:

$$u_{ij^k t} = \delta_{j^k t} + X_{j^k t}\beta_v v_i - \alpha_I I_i P_{j^k t} + \varepsilon_{it}^k, \tag{2}$$

where $\delta_{j^k t} = X_{j^k t}\bar{\beta} - \bar{\alpha} P_{j^k t} + \xi_{j^k t}$, represents the mean utility of hotel *j* with category type *k* in market *t*. β_v and α_I are the set of parameters to be estimated.

3.2 Estimtaion

Due to lack of space, we describe the estimation procedure very briefly. As mentioned before, our goal here is to estimate the mean and variance of β_i and α_i. We apply estimation methods similar to those used in Berry and Pakes (2007) and Song (2008). This problem can be essentially reduced to a procedure of solving a system of nonlinear equations. In general, with a given starting value of $\theta_0 = (\alpha_I^0, \beta_v^0)$, we look for the mean utility δ such that the model predicted market share equates the observed market share. From there, we form a GMM objective function using the moment conditions such that the mean of unobserved characteristics is uncorrelated with instrumental variables. Based on this, we identify a new value of $\theta_1 = (\alpha_I^1, \beta_v^1)$, which is used as the starting point for the next round iteration. This procedure is repeated until the algorithm finds the optimal value of θ that minimizes the GMM objective function. To find a solution, we applied the contraction mapping method suggested by Berry et al. (1995).

4. Empirical Analysis

4.1 Results

The estimation results are shown in Table 2. An important factor in influencing demand is the textual content and style of customer reviews. We see that "Complexity", "Syllables" and "Spelling Errors" have a negative sign. This implies that most consumers would prefer to read reviews with shorter sentences, less syllables and fewer spelling errors in total. On the other hand, variables “Characters” and “SMOG Readability index” present a positive influence. This implies that consumers appreciate longer reviews with more characters, and with a more professional writing style. For the subjectivity features, both “Mean Subjectivity” and “Subjectivity standard deviation” turn out to be negative. Therefore, consumers prefer to obtain as much objective information as possible from others’ experiences.

There are at least five location-based characteristics which have a positive impact on hotel demand: “External Amenities,” “Near Public transportation,” “Near Highway”, “Near Downtown” and “Near Beach” showing that consumers prefer to stay in hotels with these features. A few location-based characteristics have a negative impact on hotel demand. Not surprisingly, one of them is the “Annual Crime Rate.” The higher the average crime rate reported in a local area, the lower the desirability of consumers for staying in a hotel located in that area. Another factor that has a negative impact is the number of ``Local Competitors” within 2 miles. These estimates imply that controlling for price and content of user reviews, the geographical and other location attributes of a hotel can make a big difference in attracting consumers.

Table 2. Estimation Results

Variable	Coefficient (Std. Err)^{I}	Coefficient (Std. Err)^{II}
Room Price Per Night	-0.1768*** (.0289)	-0.0080 (.0144)
Number of Characters in Review	0.0155*** (.0020)	0.0108*** (.0015)
Review Complexity	-0.0121*** (.0026)	-0.0070*** (.0020)
Number of Syllables in Review	-0.0482*** (.0063)	-0.0331*** (.0048)
SMOG Readability Index	0.1137*** (.0280)	0.0650*** (.0195)
Number of Spelling Errors in Review	-0.1575*** (.0416)	-0.1250*** (.0318)
Mean Subjectivity of Review	-0.8268* (.3322)	-0.2265† (.1317)
Subjectivity Deviation of Review	-0.2298** (.0758)	-0.2221*** (.0576)
Hotel Class	0.0421*** (.0128)	-0.0049 (.0055)
Number of Competitors of Hotel	-0.0853*** (.0118)	-0.1435*** (.0147)
City Annual Crime Rate	-0.1523*** (.0174)	-0.0598*** (.0095)
Number of Internal Amenities	0.0022 (.0020)	0.0023* (.0010)
Number of External Amenities	0.0066*** (.0019)	0.0052*** (.0011)
Near Beach or Not	0.0693* (.0335)	0.1035*** (.0178)
Near Public Transportation or Not	0.01495*** (.0290)	0.00003** (9.61e-06)
Near Interstate Highway or Not	0.1332*** (.0272)	0.0848*** (.0153)
Near Downtown or Not	0.0275 (.0287)	0.0713*** (.0160)

***, **, *, and † denote significance at 0.1%, 1%, 5% and 10% levels, respectively. Control variables include volume and valence of reviews and whether reviewer disclosed his identity or not.

The above results are based on the dataset of hotels from Travelocity, which may or may not have online customer reviews on its website. As a robustness check, we also collected reviews from a third party site - TripAdvisor, which is regarded as the world’s largest online travel search engine. We therefore narrowed down the sample to consist of those hotels that have at least one review from either Travelocity or TripAdvisor. The estimation results from this filtered dataset (II) are shown in column 2 of Table 2. Further, we conducted the similar estimations after incorporating the textual content of reviews from TripAdvisor. All the results were qualitatively very consistent with our findings above.

4.2 Consumer Surplus-Based Hotel Ranking

After we have estimated the parameters in the model, we can derive the consumer surplus from our model. The mean utility provides us a good estimation of how much consumers in general can benefit from choosing this particular hotel, and the standard deviation of utility describes the variance of this benefits from different consumers. In our case, we are interested to know what the excess utility, or consumer surplus, is for consumers on an aggregate level to choose a certain hotel. We thereby propose a new ranking approach for hotels based on the consumer surplus of each hotel for consumers on an aggregate level. This ranking idea is based on how much “extra value” consumers can obtain after paying for that hotel, which is what consumers really care about. If a hotel provides a comparably higher surplus for consumers on an aggregate level, then it should appear on the top of our ranking list for that city.

4.3 Evaluation With User Study

To evaluate the quality of our ranking technique, we conducted a user study using Amazon Mechanical Turk (AMT). First, we generated different rankings for the top-20 hotels, in various areas, according to a set of baseline criteria: price low to high, price high to low, maximum online review count, hotel class, hotel size (number of rooms), number of internal amenities, and popularity rank (generated by TripAdvisor). We then computed the consumer surplus for each hotel, and ranked the hotels in each city according to their surplus. Then, we performed blind tests, presenting various lists to 100 anonymous AMT users and asking them which ranking list they prefer. Further, we asked users to compare pairs of lists and tell us which of the hotel ranking lists they prefer the most. We tested the results for a few large cities like New York, Chicago, Dallas, Atlanta, Los Angeles, San Francisco and Washington DC. The results were highly encouraging. For example, in New York city, more than 80% of the customers preferred our ranking when listed side-by-side with the other, competing baseline techniques (p = 0:001, sign test).

We also asked consumers why they chose a particular ranking, to understand better how users interpret the surplus-based ranking. In our NYC experiment, the majority of the users indicated that they preferred the diversity of the returned results given that the list consisted of a mix of hotels cutting across several price ranges. In contrast, the other ranking approaches tend to list hotels of only one type (e.g., very expensive hotels). We found that a ranking system generated with “value for the money” returns a better variety of hotels, covering 30% 5-star, 40% 4-star, and 30% 3-star hotels in a given city. It generally starts out with lower class hotels and increases to 5-star hotels, providing a logical way to present the information on the screen which will help customers in their decision-making procedure. Based on the qualitative opinions of the users, it appears that diversity in hotel choices is indeed an important factor that improves the satisfaction of consumers, and an economic approach for ranking introduces diversity naturally. This result seems intuitive: if a specific segment of the market systematically appeared to be underpriced, then market forces would move the prices for the whole segment accordingly. However, this effect may be less pronounced with individual hotels, especially under a personalized consumer surplus calculation.

5. Conclusion

To summarize we show information about hotel characteristics captured from different sources of data can be incorporated in a demand estimation model to empirically estimate the economic value of different hotel characteristics, including both service based and location-based characteristics. Our research allows us to not only quantify the economic impact of hotel characteristics, but also by reversing the logic of this analysis, allows us to identify the characteristics that most influence the demand for a particular hotel. After inferring the economic significance of each characteristic, we incorporate the economic value of hotels characteristics into a local ranking function based on estimation of consumer surplus from transactions of that hotel. The key idea is that hotels that provide consumers with a higher surplus would be placed higher on the screen in response to consumer queries. We then conduct blind tests using users from AMT to examine how well our ranking system performs and find that our system performs significantly better than existing benchmark ones. We are currently working on extending our system into a more personalized ranking system that incorporates surplus for each individual consumer of a hotel and then displays hotel recommendations in a personalized manner for each consumer. We hope that our inter-disciplinary methods and approach can improve the quality of results displayed for hotel search engines on the Internet.

References

Archak, N., Ghose, A., Ipeirotis, G. 2008. Deriving the pricing power of product features by mining consumer reviews, Working Paper, SSRN.

Berry, S., Levinsohn, J., Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* (63), 841-890.

Berry, S. and Pakes, A. 2007. The pure characteristics demand model. *International Economic Review* (48), 1193-1225.

Chevalier, J., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Res*. 43(3), 345-354.

Das, S., M. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53(9). 1375–1388.

Ghose, A., P. Ipeirotis. 2008. Estimating the socio-economic impact of product reviews: Mining text and reviewer characteristics, Working paper, SSRN.

Ghose, A., P. Ipeirotis, A. Sundararajan. 2006. The dimensions of reputation in electronic markets, Working paper, SSRN.

Ghose, A. 2009. Internet exchanges for used goods: An empirical analysis of trade patterns and adverse selection, *MIS Quarterly*, 33(2), 263-291.

Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Info Sys Res* 19(3),291-313.

Hu, M., B. Liu. 2004. Mining and summarizing customer reviews. *Proceedings of the Tenth ACM International Conference on Knowledge Discovery and Data Mining*, 168–177.

Lu, D.,Weng, Q. 2007. A survey of image classification methods and techniques for improving classification performance, *Int. Journal of Remote Sensing* 28 (5), 823-870.

Pang, B., Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd ACL*, 271-278.

Song, M. 2008. A hybrid discrete choice model of differentiated product demand with an application to personal computers. SSRN Working Paper.