

# **Analyzing Consumer behavior on Product Search Engines: Combining Social Media Analytics, Econometric Modeling and Randomized Experiments**

Dissertation Proposal (Short Version)

Beibei Li

Stern School of Business, New York University

*(Last Updated: October 9, 2011)*

## **Abstract**

With the growing pervasiveness of Internet and e-commerce today, online search has been at the very heart of consumer information seeking and economic decision making process. However, due to the size and format of information that is approachable by a search engine, as well as the heterogeneous nature of humans, it becomes very hard to identify *what* information is used and *how* exactly it is used by a consumer via search engines to facilitate the decision making activities in an online market.

In my dissertation, I plan to look into such issues from three perspectives: (i) How do consumers search, evaluate and purchase products based on information acquired from search engines and social media? (ii) How do today's product search mechanisms and social media platforms affect consumers' behavior and the subsequent economic outcome in the market? (iii) How to leverage knowledge created within and across a large diversity of social media channels, in order to improve the product search mechanism to increase market efficiency and social welfare? Moreover, how can I translate my research into real world applications? The summarized interaction and relationship among the detailed research questions are shown in Figure 1.

To examine these questions, I propose to combine methodologies from econometrics, structural modeling and Bayesian modeling together with machine learning techniques such as text mining, sentiment analysis, image classification and predictive modeling, as well as other computer-automated techniques such as social-geo-tagging and geo-mapping tools. My study builds on a unique dataset consisting of detailed information on approximately 1 million online sessions from Travelocity.com, including consumer searches, clicks and conversions that occurred from 2008/11 to 2009/1. Furthermore, to verify my empirical findings, I supplement the archival data analysis with surveys and randomized experiments. This interdisciplinary approach allows me to combine the strengths of economics, IS, marketing and computer science, and to analyze online product search engines from both the technical aspect as well as the social and economic aspects. For better understanding, the summarized methodologies of my research are shown in Table 1.

## 1. Research Questions

With the revolutionary development of information technology, Internet today has become the primary tool for consumer information seeking and commercial activity. In particular, online search has dominated consumers' economic decision making process in that over 90 percent of consumers use search engines before making a purchase decision (e.g., BIA/Kelsey's and ConStat 2010). In the mean time, firm decision makers have also adopted search engines as one of the primary channels to learn customers' online behavior and purchase motivation. For instance, many businesses today have started looking at consumers' online search queries and click logs, in order to understand the cognitive pattern how consumers seek and evaluate relevant information during their economic decision processes. The knowledge created within search engines allows firm owners to customize their business and service in an interactive way to gain and retain customers.

More importantly, with the growing pervasiveness of social media (e.g., online word-of-mouth, social communities, geo-/social-tagging, photo/video sharing and personal blogs), Internet has gradually woven into the fabric of *user experience*. By this year, nearly half of all consumers use social media as an indispensable source supplemental to search during their economic decision making (GroupM Search and ComScore 2011). Consequently, encompassed with the information-intensive age reinforced by the rapid explosion of social media, *how do consumers search, evaluate and purchase products online?* This is one major research question in my dissertation.

In particular, I am interested in examining the following questions: *how do consumers leverage knowledge created within a large variety of social media channels together with knowledge generated by search engines, in order to facilitate their online search and purchase? How do these two types of knowledge influence and interact with each other in the context of consumer economic decision making?* Moreover, consumers are heterogeneous in regards to their preferences towards different product characteristics and their efforts to search online. Thus, *how to dynamically capture consumers' strategic online behavior during the economic decision making process while accounting for their heterogeneous preferences and search costs?* Given the increasingly large amount of choices online, *is there an optimal strategy for consumers to search and identify the best value product?*

In addition to examining the consumer behavior, a second set of important research questions that I would like to focus on in my dissertation is from the search engines' perspective. While connecting consumers with businesses, search engines today have evolved into one of the most important strategic platforms for marketing. Indeed, the generation of online attention from product search engines has led to substantial revenue increase for the economy. Nevertheless, due to the size of the Internet and the varying quality of heterogeneous commodities, one major challenge faced by product search engines today is whether or not the search mechanism can effectively deliver the results to a consumer.

Over the last few years, a tremendous amount of research has focused on how to improve the content quality of the search results, for example, by optimizing retrieval of relevant documents from the Web, mainly as a response to a keyword query (e.g., Lavrenko and Croft 2001, Pang and Lee 2008). However, due to the multi-dimensional preferences of consumers for many products and services, several questions remain unanswered in this space. *How can product search engines present their results in a manner that facilitates efficient information exchange and effective marketing activities? Should product search engines allow consumers to interact with the recommendation algorithm to personalize their search results?* Therefore, two challenges appear to be crucial for product search engines today. First, *what ranking mechanism should be used to effectively present the search results?* Second, *what personalization mechanism should be applied to deliver the search results to the population of heterogeneous consumers?* Finally, *is there an optimal search mechanism for product search engines to serve as a stimulus for users to discover “best fit” items, thereby eliciting efficient consumer search and decision making in electronic market to maximize social welfare?*

## **2. Overview of Research Methodology and Main Results**

To explore these research questions, I plan to conduct my dissertation research in three stages.

- **Stage 1: Consumer Demand and Utility-based Ranking– Structural Model for Demand Estimation + Randomized Experiments.**

In the first stage, I focus on modeling consumers’ economic choices and improving consumers’ search experiences on product search engines, taking in to account knowledge created within and across a large variety of social media channels. In particular, consumers today use social media together with product search engines to facilitate their economic decision making. Consumers try to identify products of high quality with specific desired characteristics, but without compromising on their associated prices. However, given the proliferation of available information from the Web, they normally do not have the sophistication or time to conduct exhaustive searches to seek the quality or price information in order to compare for the “best value” product.

Unfortunately, current product search engines fail to effectively leverage information created across diverse social media platforms. Moreover, current ranking algorithms in these product search engines tend to induce consumers to focus on one single product characteristic dimension (e.g., price, star rating, etc). This largely ignores consumers’ multi-dimensional preferences for products. These drawbacks highly necessitate a new recommendation strategy for product search engines that can better perceive consumers' underlying demand, and facilitate the information seeking activities for consumers to make efficient economic decisions.

Therefore, in this stage I propose to design a ranking system that recommends products providing the best value for money. The key idea is that products that provide consumers with a higher surplus should be ranked higher on the screen in response to consumer queries. To achieve this goal, I propose a random

coefficient hybrid structural model, taking into consideration the two sources of consumer heterogeneity introduced by the different purchase occasions and different product characteristics. Based on the estimates from the model, I infer the economic impact of various characteristics of products. I then propose a new product ranking system based on the average utility gain that a consumer gets by purchasing a particular product. By doing so, I can provide customers with the “best-value” products early on, and thereby improve the quality of searches for such products.

To evaluate the proposed ranking system, I propose to design a set of randomized experiments conducted on Amazon Mechanical Turk, using pair-wise ranking comparisons with the existing benchmark ranking systems that are being used by current product search engines. Moreover, I further test the performance of the proposed ranking mechanism on a real-world hotel search engine that is designed and built by ourselves. Both the AMT experiments and the real-world search engine experiments validate the superiority of the proposed ranking system relative to existing systems on several travel search engines.

On a broader note, the objective of this stage is to illustrate how social media on the Internet can be mined (to learn latent product information) and incorporated into a demand estimation model, and how social media can be leveraged to generate a new ranking system in product search engines to improve the quality of choices available to consumers online. Besides providing consumers with direct economic gains, such a ranking system can lead to non-trivial reduction in consumer search costs and increased usage of product search engines.

- **Stage 2: Impact of Search Engine Ranking and Personalization – Hierarchical Bayesian Model + Randomized Experiments.**

In the second stage, I examine how different ranking and personalization mechanisms on product search engines influence consumer online search and purchase behavior. To investigate these effects, I combine archival data analysis with randomized field experiments. In the archival data analysis, I propose to use a hierarchical Bayesian model to jointly estimate the relationship among consumer click and purchase behavior, and search engine ranking decisions. To evaluate the causal effect of search engine interface on user behavior, I conduct randomized field experiments. The field experiments are based on the real-world hotel search engine application designed and built by us. By manipulating the default ranking method of search results, and by enabling or disabling a variety of personalization features on the hotel search engine website, I am able to empirically identify the causal impact of search engines on consumers’ online click and purchase behavior.

The archival data analysis and the randomized experiments are consistent in demonstrating that ranking has a significant effect on consumer click and purchase behavior. I find that hotels with a higher reputation for providing superior services are more adversely affected by an inferior screen position. In addition, a consumer utility-based ranking mechanism yields the highest click and purchase propensities in comparison

to existing benchmark systems such as ranking based on price or customer ratings. The randomized experiments on the impact of active vs. passive personalization mechanisms on user behavior indicate that although active personalization (wherein users can interact with the recommendation algorithm) can lead to a higher click-through rate compared to passive personalization, it leads to a lower conversion rate when consumers have a planned purchase beforehand. This finding suggests that active personalization strategies should not be adopted ubiquitously by product search engines. On a broader note, this inter-disciplinary approach provides a methodological framework for how econometric modeling, randomized field experiments, and IT-based artifacts can be integrated in the same study towards deriving causal relationships between variables of interest.

- **Stage 3: Consumer Search Dynamics and Social Signal-based Ranking – Dynamic Structural Model + Randomized Experiments.**

In the third stage, I plan to deeply examine how consumers search, evaluate and purchase products online. Especially I am interested in looking into consumers' dynamic behavior and strategic actions on product search engines. In particular, this step allows me to more precisely identify consumer preferences not only from the final choices made through the search engines, but also from the entire online search history that leads to the final decisions. Moreover, I am interested to extend this study to a social network setting, and examine how social dynamics may influence consumer search and how social signals can be incorporated into product search engine design.

To achieve the goal, I plan to use a dynamic structural model considering customers' forward-looking behavior. The dynamic model enables me to understand better how consumers behave during the product search process, in particular, how they search, evaluate and make purchase decisions. Moreover, this dynamic model framework also allows me to *predict* the click-through and conversion rates of a product more precisely. Specifically, it is able to predict the "consumer satisfaction" in a sense that it extracts the *net utility* of a product after taking into account the screen position bias, consumer heterogeneity, search cost and social dynamics.

Currently, my research is still on-going at this stage. In the future, I plan to finalize the consumer dynamic search model and the empirical estimation. Moreover, I plan to explore not only the individual dynamics that arise from consumers themselves, but also the social dynamics that arise from consumers' social networks. To do so, I plan to design a randomized hotel search experiment on Facebook and examine how social network may influence consumers' dynamic search behavior. This future step will in the end allow me to design a social network-based ranking mechanism for product search. This new ranking mechanism can be thereby combined with the utility-based ranking designed in the first stage, shedding light on how social signals can be incorporated into product search engine design to reduce consumer search cost and improve online market efficiency.

### 3. Overview of Data

To empirically examine these questions, I plan to look into the online travel industry and focus on the hotel sector. I plan to use a unique panel data set from 2008/11 to 2009/1 for US hotels from Travelocity.com. This data set contains approximately 1 million online user search sessions including detailed information on consumer searches, clicks, and transactions.

Besides, I supplement the search and transaction data with hotel service-, location- and customer review-based information extracted from a large variety of social media using various machine learning techniques such as image classification and text mining tools. The overall data collection contributes to a final dataset with a total of 29,222 weekly observations for 2117 hotels in the United States. More specifically, this dataset combines four major sources:

#### ***(1) Consumer Search, Click and Conversion Data from Travelocity.com***

This dataset contains complete information on consumer online searching and shopping behavior. A typical online session involves the initialization of the session, the search query, the results (in a particular rank order) returned from that search query, the sorting method, the click(s) on hotel(s) if there exists any, the login and actual transaction(s) if any conversion occurs, and the termination of the online session.

I count a “display” for a hotel if that hotel appears visible to a consumer on the web page in an online search session. Meanwhile, a “click” is counted if the hotel is selected by a consumer, and a “conversion” is counted only if a consumer has finished the payment in that online session. A display often leads to a click, but it may not lead to an actual purchase. Each hotel that counts for a display is associated with a page number and a screen position, which capture the corresponding page order and (within-page) rank order of that hotel in the search results. Notice that when Travelocity displays the hotel search results on a web page, it only shows 25 hotels per page<sup>1</sup>. This restricts the rank order for each hotel within the range from 1 to 25. Meanwhile, to facilitate consumer search, Travelocity provides a sorting criterion called “Travelocity Pick” by default. Besides, it also provides multiple alternative sorting criteria: Price, Hotel Class, Hotel Name, and Customer Review Rating.

In addition, I also have supplemental data collected from the following three sources.

#### ***(2) Hotel Location-Related Characteristics***

I use geo-mapping search tools (in particular the Bing Maps API) and social geo-tags (from geonames.org) to identify the “external amenities” (e.g., shops, bars, etc) and public transportation in the area around the hotel. I also use image classification together with human annotation on Amazon Mechanical Turk to examine whether there is a nearby beach, a nearby lake, a downtown area, and whether the hotel is close to a highway. I extract these characteristics within a local area of 0.25-mile, 0.5 mile, 1-mile, and 2-mile radius.

---

<sup>1</sup> Recently Travelocity has upgraded the webpage design by showing 10 hotels per page. However, during our examination time period, this number was 25.

### ***(3) Hotel Service-Related Characteristics***

There are three broad characteristics in the category of service-based characteristics: hotel class, number of internal amenities and number of rooms. “Hotel class” is an internationally accepted standard ranging from 1-5 stars, representing low to high hotel grades. “Number of internal amenities” is the aggregation of hotel internal amenities, such as “swimming pool,” “free breakfast,” “high speed internet,” “hair drier” and “parking facility.” I extract this information from the Tripadvisor website using fully automated parsing. Since hotel amenities are not directly listed on the Tripadvisor website, I retrieve them by following the link provided on the hotel web page, which randomly directs the user to one of its cooperating partner websites (i.e., Travelocity, Orbitz, or Expedia).

### ***(4) Online Review-Related Characteristics***

Finally, I collect customer reviews from Travelocity.com. The online reviews and reviewers’ information were collected on a daily basis up to January 31, 2009 (the last date of transactions in the database). In addition to the total number of reviews and the numeric reviewer rating, I use text mining tools to extract indicators that measure the stylistic characteristics of the available reviews for robustness checks. I examine two text-style features: “subjectivity” and “readability” of reviews (Ghose and Ipeirotis 2010). Also, since prior research suggested that disclosure of identity information is associated with changes in subsequent online product sales (Forman et al 2008), I measure the percentage of reviewers for each hotel who reveal their real name or location information on their profile pages.

Moreover, to further exploit the information about hotel service characteristics that is embedded in the natural language text of the consumer reviews (e.g., the “helpfulness of the hotel staff” is a service feature that can be assessed by reading the actual consumer opinions), I build on the work of Hu and Liu (2004), Popescu and Etzioni (2005), Archak et al. (2011) and apply text mining and sentiment analysis techniques.

In summary, all the different sources that contribute to the final dataset are shown in Table 2.

**Table 1: Summary of Research Methodologies**

| <b>Methods \ Stage</b>      | <b>Stage 1</b> | <b>Stage 2</b> | <b>Stage 3</b> |
|-----------------------------|----------------|----------------|----------------|
| Hierarchical Bayesian Model |                | √              |                |
| Static Structural Model     | √              |                |                |
| Dynamic Structural Model    |                |                | √              |
| Text Mining                 | √              |                | √              |
| Sentiment Analysis          | √              |                | √              |
| Image Classification        | √              | √              | √              |
| Predictive Modeling         | √              | √              | √              |
| Social-Geo-Tagging          | √              | √              | √              |
| Geo-Mapping                 | √              | √              | √              |
| Survey                      | √              | √              | √              |
| Randomized Experiments      | √              | √              | √              |

**Table 2: Summary of Data Sources**

| Category                                    | Hotel Characteristics   | Data Sources   |
|---|---|--|
| Consumer Search, Click and Transaction Data | Detailed Session-Level Search Logs<br>Transaction Price (per room per night)<br>Number of Rooms sold (per night)                    | Travelocity  |
| Service-based                               | Hotel Class<br>Hotel Amenities  | TripAdvisor  |
|   | Number of Customer Reviews<br>Overall Reviewer Rating<br>Disclosure of Reviewer Identity Information                                | Travelocity and TripAdvisor  |
|   | Subjectivity<br>Mean Probability<br>Std. Dev. Of Probability  |  |
| Review-based                                | Readability<br>Number of Characters<br>Number of Syllables<br>Number of Spelling Errors<br>Average Length of Sentence<br>SMOG Index | Text Analysis  |
|   | Service Features Embedded in the Text<br>Breakfast<br>Hotel Staff<br>Bathroom<br>Bedroom<br>Parking                                 |  |
|   | Near the Beach<br>Near Downtown   | Image Classification,<br>Tags from Geonames.org and<br>Social Annotations from<br>Amazon Mechanical Turk |
| Location-based                              | External Amenities (Number of restaurants/<br>Shopping destinations)  | Microsoft Virtual Earth Geo-<br>Mapping Search SDK   |
|   | Near Public Transportation  | Tags from Geonames.org<br>Social Annotations from<br>Amazon Mechanical Turk                              |
|   | Near the Interstate Highway<br>Near the Lake/River  | Social Annotations from<br>Amazon Mechanical Turk  |
|   | City Annual Crime Rate  | FBI online statistics  |

**Figure 1: Summary of Research Scope**

