# A Demo Search Engine for Products*

Beibei Li
bli@stern.nyu.edu

Anindya Ghose
aghose@stern.nyu.edu

Panagiotis G. Ipeirotis
panos@stern.nyu.edu

Department of Information, Operations, and Management Sciences
Leonard N. Stern School of Business, New York University
New York, New York 10012, USA

## ABSTRACT

Most product search engines today build on models of relevance devised for information retrieval. However, the decision mechanism that underlies the process of *buying a product* is different than the process of *locating relevant documents or objects*. We propose a theory model for product search based on expected utility theory from economics. Specifically, we propose a ranking technique in which we rank highest the products that generate the highest *surplus*, after the purchase. We instantiate our research by building a demo search engine for hotels that takes into account consumer heterogeneous preferences, and also accounts for the varying hotel price. Moreover, we achieve this without explicitly asking the preferences or purchasing histories of *individual* consumers but by using aggregate demand data. This new ranking system is able to recommend consumers products with "best value for money" in a privacy-preserving manner. The demo is accessible at `http://nyuhotels.appspot.com/`

**Categories and Subject Descriptors:** H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

**General Terms:** Algorithms, Economics, Experimentation, Measurement

**Keywords:** Consumer Surplus, Economics, Product Search, Ranking, Text Mining, User-Generated Content, Utility Theory

## 1. INTRODUCTION

It is now widely acknowledged that online search for products is increasing in popularity, as more and more users search and purchase products from the Internet. Most search engines for products today are based on *models of relevance* from "classic" information retrieval theory [9] or use variants of faceted search [11] to facilitate browsing. However, the decision mechanism that underlies the process of *buying a product* is different from the process of finding a *relevant* document or object. Customers do not simply seek something relevant to their search, but also try to identify the "best" deal that satisfies their specific criteria. Today's product search engines provide only rudimentary ranking facilities for search results, typically using a single ranking criterion such as price, best selling, or more recently, using customer review ratings. This approach has quite a few shortcomings. First, it ignores the *multidimensional* preferences of consumers. Second, it fails to leverage the information generated by the online communities, going beyond simple numerical ratings. Third, it hardly takes into account the *heterogeneity* of consumers. These drawbacks highly necessitate a recommendation strategy for products that can better model consumers' underlying purchase behavior, to capture their multidimensional preferences and heterogeneous tastes.

Recommender systems [1] could fix some of these problems but, to the best of our knowledge, existing techniques still have limitations: First, most recommendation mechanisms require consumers to log into the system. However, in reality many consumers browse only anonymously. Due to the lack of any meaningful, personalized recommendations, consumers do not feel compelled to login before purchasing. Even when they login, before or after a purchase, consumers are reluctant to give out their individual demographic information due to many reasons (e.g., time constraints, privacy issues, or lack of incentives). Therefore, most context information is missing at the individual consumer level. Second, for goods with a *low purchase frequency* for an individual consumer, such as hotels, cars, or real estate, there are few repeated purchases we could leverage towards building a predictive model (i.e., models based on collaborative filtering). Third, and potentially more importantly, as privacy issues become increasingly noticeable today, marketers may not be able to observe the individual-level purchase history of each consumer (or consumer segment). Instead, the only information available is at an aggregate level (e.g., market share or unit sold). As a consequence, many algorithms that rely on knowing individual-level behavior lack the ability of deriving consumer preferences from such aggregate data.

Alternative techniques try to identify the "Pareto optimal" set of results [2]. Unfortunately, the feasibility of this approach diminishes as the number of product characteristics increases. With more than five or six characteristics, the probability of a point being classified as "Pareto optimal" dramatically increases. As a consequence, the set of Pareto optimal results soon includes *every* product.

In our work, we design a new ranking system for recommendation that leverages economic modeling. We aim at making recommendations based on better perception of the underlying the "*causality*" of consumers' purchase decisions. Our algorithm learns consumer preferences based on the largely anonymous, publicly observed *distributions of consumer demographics* as well as the observed *aggregate-level* purchases (i.e., anonymous purchases and market shares in NYC and

*Supported by NSF grants IIS-0643847 and IIS-0643846.


*WWW 2011*, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

LA), not by learning from the identified behavior or demographics of each individual. We instantiate our research by building a demo search engine for hotels, using a unique data set containing transactions from Nov. 2008 to Jan. 2009 for US hotels from a major travel web site. Our extensive user studies, using more than 15000 user judgments, demonstrate an overwhelming preference for our ranking strategy, compared to a large number of existing strong baselines.

The major contributions of our research are: (1) We present a *causal* model, based on economic theory, to capture consumers' decision-making process, leading to a better understanding of consumer preferences. The causal model relaxes the assumption of "consistent environment" across training and testing data sets: we can now have changes in the environment and can predict what *should* happen under such changes. (2) We infer *personal* preferences from *aggregate* data, in a privacy-preserving manner. (3) We propose a ranking method using the notion of *surplus*, which is derived from a "generative" user behavior model. (4) We present an extensive experimental study: using six hotel markets, and 15000 user evaluations using *blind tests*, we demonstrate that our ranking is significantly better than existing baselines.

# 2. THEORY MODEL

In this section, we first introduce the background of the *expected utility* theory, *characteristics-based* theory, and economic *surplus*. Then we discuss how we leverage these concepts into our setting and empirically estimate our model.

## 2.1 Background

Our model is derived from from *expected utility* and *rational choice* theories. A fundamental notion in utility theory is that each consumer is endowed with an associated utility function $U$, which is "a measure of the satisfaction from consumption of various goods and services." The rationality assumption defines that each person tries to maximize its own utility.

More formally, assume that the consumer has a choice across products $X_1, \ldots, X_n$, and each product has a price $p_j$. Buying a product involves the exchange of money for a product. Therefore, to analyze the purchasing behavior we need two components for the utility function: (1) *Utility of Product:* The utility that the consumer gets by buying the product $X_j$, and (2) *Utility of Money:* The utility that the consumer loses by paying the price $p_j$ for product $X_j$.

On one hand, the decision to purchase product $X_j$ generates a product utility $U(X_j)$. According to Lancaster's *characteristics theory* [6] and Rosen's *hedonic price model* [10], differentiated products are described by vectors of objectively measured characteristics. Let $x_j^k$ denote the $k$th observed characteristics of $X_j$. Thus, the utility of product is defined as the aggregation of weighted utilities of observed characteristics and an unobserved characteristic, $\xi_j$, as follows

$$U(X_j) = U(x_j^1, \ldots, x_j^k) = \sum_k \beta_j^k \cdot x_j^k + \xi_j. \quad (1)$$

On the other hand, assume that the consumer has some disposable income $I$ that generates a money utility $U(I)$. Paying the price $p_j$ decreases the money utility to $U(I - p_j)$. We typically assume that $p_j$ is relatively small compared to the disposable income $I$, and the *marginal utility* of money remains constant in the interval $I - p_j$ to $I$ [8]. In this case,

$$U(I) - U(I - p_j) = \alpha I - \alpha(I - p_j) = \alpha p_j. \quad (2)$$

With the assumption of rationality, a consumer purchases product $X_j$ if and only if it provides him with the highest increase in utility. Let *consumer surplus* denote the "increase" in utility after purchasing a product. This idea naturally generates a ranking order: The products that generate the highest consumer surplus should be ranked on top.

## 2.2 The BLP Model

The key for our model is to identify the different product characteristics and estimate the corresponding weights assigned by consumers towards the product characteristics. However, different consumers hold different evaluations towards the product characteristics and towards the money. To capture the consumer heterogeneity, we use the Random-Coefficient Logit Model [3] (also known as BLP). This model assumes that consumers have idiosyncratic tastes towards product characteristics. In other words, the coefficients $\beta$ and $\alpha$ in equation 1 and 2 are different for each consumer. Based on this, we define the *utility surplus* for consumer $i$ to buy product $X_j$ as

$$US_j^i = U_h(X_j) - [U_m(I^i) - U_m(I^i - p_j)] + \varepsilon_j^i \quad (3)$$

$$= \underbrace{\sum_k \beta^{ik} \cdot x_j^k + \xi_j}_{\text{Utility of product}} - \underbrace{\alpha^i p_j}_{\text{Utility of money}} + \underbrace{\varepsilon_j^i.}_{\text{Stochastic error}}$$

Here, $I^i$ is the income of consumer $i$, $p_j$ is the price of product $X_j$, $U_m$ is the utility of money, and $U_h$ is the utility of product purchased. Note that $\xi$ is a *product-specific* disturbance scalar summarizing unobserved characteristics of product $X_j$, whereas $\varepsilon_j^i$ is a stochastic choice error term that is assumed to be i.i.d. across *products and consumers* in the selection process. The parameters to be estimated are $\alpha^i$ and $\beta^i$, which represent the weights that consumer $i$ assigns towards "money" and towards different observed product characteristics, respectively. The technical details for the model estimation are in [7]. To better understand our model, let's consider an example.

EXAMPLE 1. *Suppose that we have two cities, A and B and two types of consumers: business trip travelers and family trip travelers. City A is a business destination (e.g., New York City) with 80% of the travelers being business travelers and 20% families. City B is mainly a family destination (e.g., Orlando) with 10% business travelers and 90% family travelers. In city A, we have two hotels: Hilton ($A_1$) and Doubletree ($A_2$). In city B, we have again two hotels: Hilton ($B_1$) and Doubletree ($B_2$). Hilton hotels ($A_1$ and $B_1$) have a conference center but not a pool, and Doubletree hotels ($A_2$ and $B_2$) have a pool but not a conference center. To keep the example simple, we assume that preferences of consumers do not change when they travel in different cities and that prices are the same.*

*By observing demand, we see that demand in city A (business destination) is 820 bookings per day for Hilton and 120 bookings for Doubletree. In city B (family destination) the demand is 540 bookings per day for Hilton and 460 bookings for Doubletree. Since the hotels are identical in the two cities, the changes in demand must be the result of different traveler demographics.*

*More specifically, for business traveler, the utility surplus from hotel $A_1$ (conference center, no pool) is $US^B(A_1) = \delta_{A_1} + (\beta^B_{conf} \cdot 1 + \beta^B_{pool} \cdot 0) + \epsilon$, and for family travelers, the corresponding utility surplus is $US^F(A_1) = \delta_{A_1} + (\beta^F_{conf} \cdot 1 + \beta^F_{pool} \cdot 0) + \epsilon$.*

*By $\beta^B_\bullet$ we denote the deviations from the population mean for business travelers towards "conference center" and "pool" and by $\beta^F_\bullet$ we denote the respective deviations for family travelers. Similarly, we can write down the utilities for hotels $A_2$, $B_1$ and $B_2$. Following the estimation steps, we discover that family travelers have $\beta^F_{conf} = \beta^F_{pool} = 0.5$. In other words, they have the same preferences regarding a pool and conference center. On the other hand, for business travelers, their preference towards "conference center" is much higher than towards "pool," with $\beta^P_{conf} = 0.9$ and $\beta^F_{pool} = 0.1$, respectively.*

*This estimation result can be further interpreted with monetary meanings. For instance, we can infer that a business trip traveler is willing to pay \$54 for the conference center and \$6 for the pool, whereas a family trip traveler is willing to pay equally \$30 for each of the two features.*

# 3. SURPLUS-BASED RANKING

So far, we have described the economic model used for inferring the preferences of consumers using a utility model and aggregate demand data. This model uses the concept of surplus mainly as a conceptual tool to infer consumer preferences towards different product characteristics. In our work, the concept of surplus is directly used to find the product that is the "*best value for money*" for a given consumer.

We define *Consumer Surplus* for consumer $i$ from product $j$ as the "normalized utility surplus," the surplus $\bar{US}_j^{(i)}$ divided by the mean marginal utility of money $\bar{\alpha}$.

$$CS_j = Normalized\_US_j = \sum_t \frac{1}{\bar{\alpha}} \bar{US}_j^{(i)}. \quad (4)$$

We thereby use the estimated surplus for each product and rank the products in decreasing order of surplus. So, products at the top are the "best value" for consumers, for a given price. Furthermore, we extend our ranking to include a personalization component. To compute the personalized surplus, we ask the consumer to give the appropriate demographic characteristics and purchase context (e.g., 25-34 years old, \$100K income, business traveler) and then use the corresponding deviation matrices $\beta_T$ and $\alpha_I$. It is then easy to compute the personalized "value for money" for this consumer, and rank products accordingly. Notice that the consumer *has the incentive to reveal demographics* in this scenario.

EXAMPLE 2. *For better understanding, let's re-consider the setting of the two hotels $A_1$ and $A_2$ for city A from Examples 1. Suppose that two consumers are traveling to city A on the same day: $C_1$, **a 25-34 years old business traveler, with an income \$50,000-100,000**, and $C_2$, **a 35-64 years old family traveler, with an income less than \$50,000**. Since these two travelers belong to different demographic groups and travel with different purposes, their preferences towards "conference center" and "pool" are different. Thus, the surplus they obtain from $A_1$ and $A_2$ varies. For example, the business traveler gets higher utility from $A_1$ due to the specialized conference center services, whereas the family traveler find $A_2$ more valuable due to the pool and price.*

# 4. A DEMO SEARCH ENGINE FOR HOTELS

We instantiated our product search framework using as target application the area of *hotel search*. The demo is accessible at `http://nyuhotels.appspot.com/`.

## 4.1 Data

First, to simulate the online search environment, we created one exhaustive data set using multiple data sources.

**Demand data**: Travelocity, a large hotel booking system, provided us with the set of all hotel booking transactions, for 2117 randomly selected hotels over the United States. The transactions covered the period from November 2008 until January 2009.

**Consumer demographics**: To measure the demographics of consumers in each market, we used data from the TripAdvisor web site: The consumers that write reviews about hotels on TripAdvisor also identify their *travel purpose* (*business, romance, family, friend, other*) and their *age group* (*13-17, 18-24, 25-34, 35-49, 50-64, 65+*). Based on the data, we were able to identify the demographic distribution of travelers for each destination.

**Hotel location characteristics**: We used geo-mapping search tools (in particular the Bing Maps API) and social geo-tags (from geonames.org) to identify the "external amenities" (such as shops, restaurants, etc) and available public transportation in the area around the hotel. We also used examined whether there is a nearby beach, a nearby lake, a downtown area, and whether the hotel is close to a highway. We extracted these characteristics within an area of 0.25-mile, 0.5 mile, 1-mile, and 2-mile radius.

**Hotel service characteristics**: We extracted the service-based characteristics from the reviews on TripAdvisor. We also used the hotel description information from Travelocity, Orbitz, and Expedia, to identify the "internal amenities" of the hotels (e.g., pool, spa.)

**Characteristics of online reviews**: Finally, we extracted indicators that measure stylistic characteristics of the reviews. We examined the "subjectivity" and "readability" of reviews [5] and measured the percentage of reviewers for each hotel who reveal their real name or location information on their profile web pages.

## 4.2 An Example: Personalized Hotel Search

Using the data described above, we are able to construct our economic model and create a system that generates hotel rankings. We estimate the mean and variance of the weights that consumers assign to each hotel characteristic. Using these estimates, we can derive the consumer surplus from each hotel, for a given customer.

We developed a prototype hotel search and ranking system and deployed it on Google App Engine. It consists of three basic components: a user search interface, a summary result page with the ranked hotels, and a (set of) explanatory web pages with details of each individual hotel listed in the results. First, a customer is required to select the location of the trip destination, the type of the trip (e.g., business, family, romance, friend.), and his/her income level via the search interface. Given the input search criteria and the demographic information, the system computes the personalized consumer surplus for each hotel in the specified location and ranks the search results in descending order of consumer surplus (i.e., best value on top). The customer can review the list of search results and can click on the hotel to get more information. In the detailed explanatory page of each hotel, we list the breakdown of the surplus computation, showing the value of each individual hotel characteristic. Moreover, to help customers interpret the meaning of those surplus values, the system provides not only the personalized surplus tailored

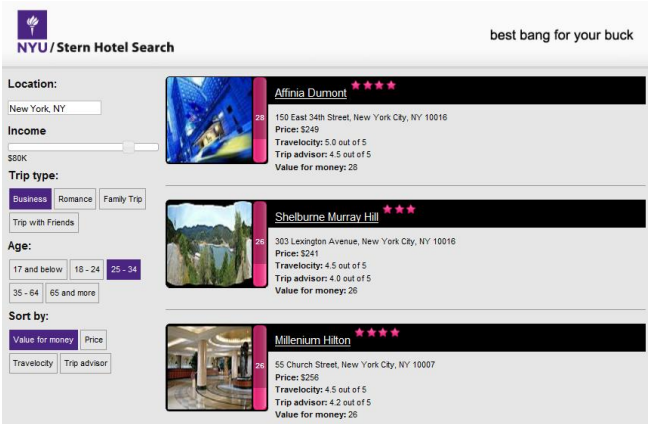


**Figure 1: Ranking results for $C_1$ (Business, $80,000, 25-34)**

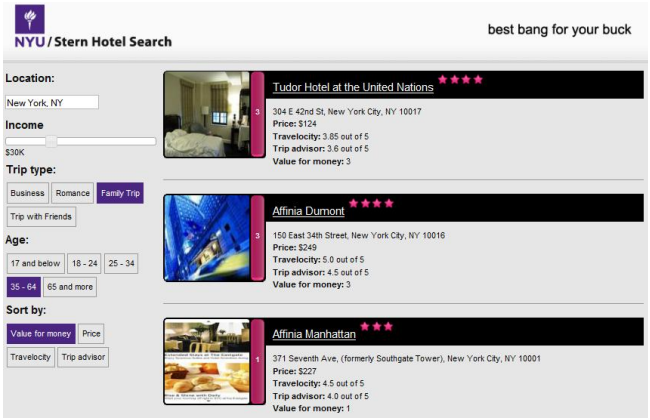


**Figure 2: Ranking results for $C_2$ (Family, $30,000, 35-64)**

for each customer, but also provides the population average surplus as a baseline for comparison. This gives customers a better idea of the *relative, personalized* value they get from each hotel characteristic.

To better illustrate this, let's look at an example.

EXAMPLE 3. *We have the same setting as in Examples 1 and 2. To find the best-value hotel, customer $C_1$ specifies the search criteria as "Location: New York, NY; Trip type: business; Income $80,000; Age group: 25-34." Similarly, customer $C_2$ specifies "Location: New York, NY; Trip type: family; Income $30,000; Age group: 35-64." Figure 1 and 2 shows the top three hotels in response to the two customized searches by $C_1$ and $C_2$. As we can see, "Affinia Dumont," a 4-star hotel with an price of $249, appears on top of the ranking list for customer $C_1$, providing a "Value for Money" of $28. On the other hand, "Tudor Hotel at the United Nations," a 4-star hotel with an lower price of $124, is ranked the first to customer $C_2$. The ranking results are dynamically justified based on the demographic of the customers (e.g., For $C_2$ with lower income, the top-ranked hotels have mainly within lower class and price range, compared to the ones for $C_1$.).*

*Customers can click each hotel for details on how each individual hotel characteristic contributes to the total value for money of that hotel. Figure 3 illustrates as an example the breakdown personalized scores of "Affinia Dumont" for*

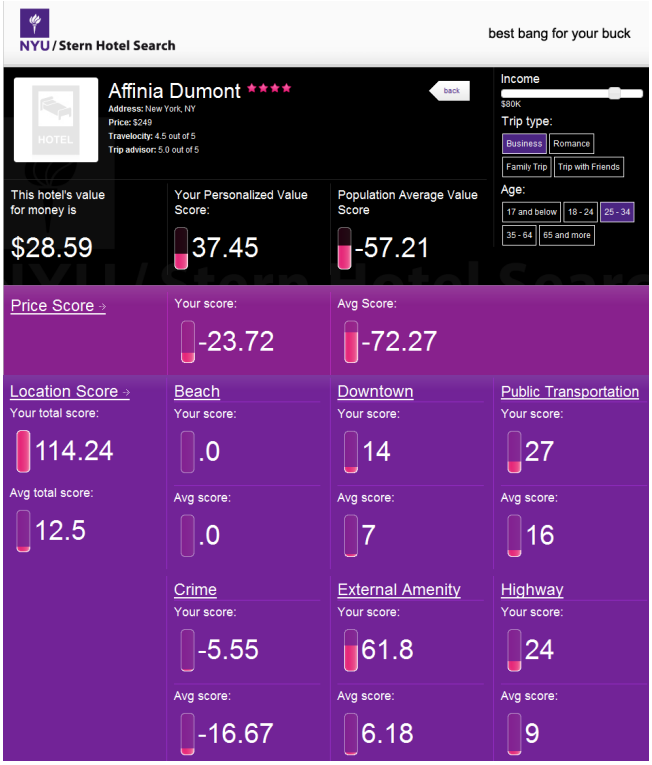


**Figure 3: Hotel overall score and breakdown across individual hotel characteristics**

*$C_1$, paired with the population average scores. For instance, we found this hotel has a personalized score (27) for "public transportation," higher than the overall population score (16). This result demonstrates that business travelers have a stronger preference towards "public transportation" than the overall population.*

## References


[1] ADOMAVICIUS, G., AND TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TKDE 17* (2005), 734–749.

[2] BALKE, W.-T., AND GÜNTZER, U. Multi-objective query processing for database systems. In *VLDB* (2004), pp. 936–947.

[3] BERRY, S., LEVINSOHN, J., AND PAKES, A. Automobile prices in market equilibrium. *Econometrica 63* (1995), 841–890.

[4] FORMAN, C., GHOSE, A., AND WIESENFELD, B. Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. *ISR 19*, 3 (2008), 291–313.

[5] GHOSE, A., , AND IPEIROTIS, P. G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE TKDE* (2010).

[6] LANCASTER, K. *Consumer Demand: A New Approach*. Columbia University Press, New York, 1971.

[7] LI, B., GHOSE, A., AND IPEIROTIS, P. G. Towards a theory model for product search. In *WWW* (2011).

[8] MARSHALL, A. *Principles of Economics*, Eighth ed. Macmillan and Co., London, 1926.

[9] NIE, Z., WEN, J.-R., AND MA, W.-Y. Webpage understanding: beyond page-level search. *SIGMOD Record 37*, 4 (2008), 48–54.

[10] ROSEN, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *J. of Political Econ. 82*, 1 (1974), 34–55.

[11] YEE, K.-P., SWEARINGEN, K., LI, K., AND HEARST, M. Faceted metadata for image search and browsing. In *CHI* (2003), pp. 401–408.