Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content

Beibei Li
(with Anindya Ghose and Panos Ipeirotis)

Stern School of Business
New York University

IDEI – Toulouse, France
Jan-14-2011

I Want to Find…

*“**How** can I find the* *best hotel* in *Long Beach? ”*

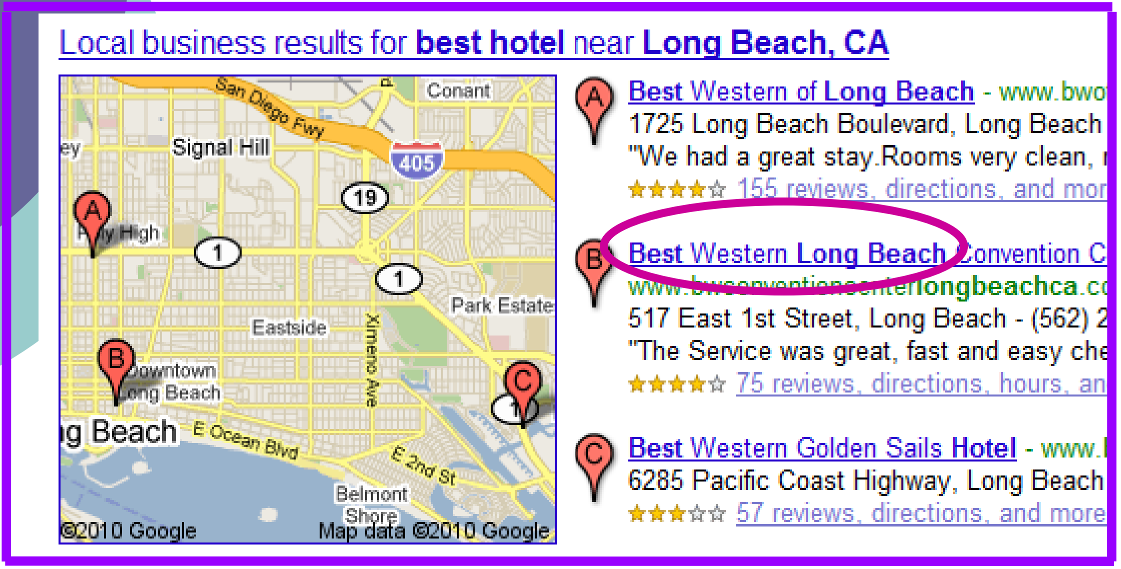
Downtown

Beach

Nightlife, Restaurants, Bars…

A Great Price for what it offers!

Limitations of Online Search Engines



Local business results for **best hotel** near **Long Beach, CA**

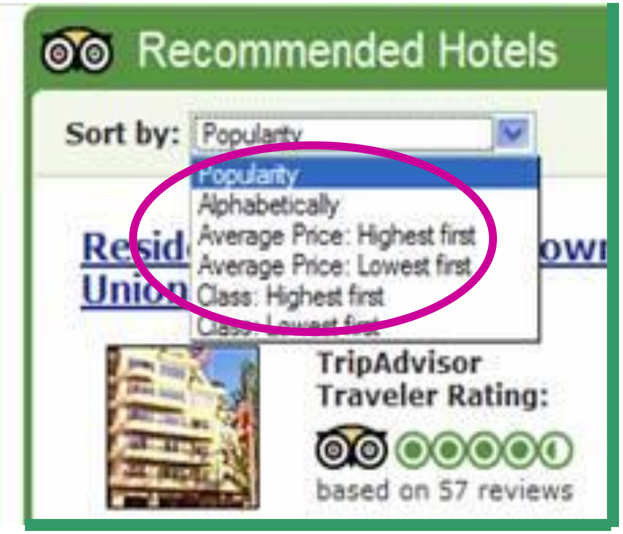
A **Best Western** of **Long Beach** - www.bwo
1725 Long Beach Boulevard, Long Beach
"We had a great stay.Rooms very clean, r
★★★★☆ 155 reviews, directions, and mor

B **Best Western Long Beach** Convention C
www.bwconventioncenterlongbeachca.co
517 East 1st Street, Long Beach - (562) 2
"The Service was great, fast and easy che
★★★★☆ 75 reviews, directions, hours, an

C **Best Western** Golden Sails **Hotel** - www.l
6285 Pacific Coast Highway, Long Beach
★★★☆☆ 57 reviews, directions, and more

***Problem*:**

-Decision of buying a product ≠ judging a document as relevant.

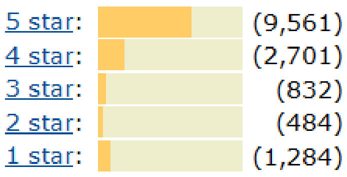


***Problem*:**

-Largely ignore multidimensional preferences;

-No consumer heterogeneity.

Limitations of Customer Information Seeking



5 star:	▬	(9,561)
4 star:	▬	(2,701)
3 star:	▬	(832)
2 star:	▬	(484)
1 star:	▬	(1,284)

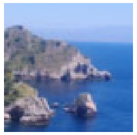
If you see the correct category among

○ Baby & Kids' Furniture
○ Breakroom Supplies
○ Arm Chairs & Recliners
○ Kitchen Furniture
○ Home Organization
○ Ottomans & Benches

amazonmechanical turk
Artificial Artificial Intelligence

“ A Perfect Stay ”

The Mandala Hotel



Mutleygirl ▾ 3 contributions
Devon

Apr 1, 2010 | Trip type: Couples

We stayed in the Mandala Berlin for 5 nights and loved it. The hotel is very well situated at Potsdamer Platz and is across the road from the Sony Centre, so it was convenient for bars, restaurants and shops as well as public transport. It is also close enough to walk to many of the attractions of this vibrant city.
The hotel is very smart and stylish and our room on the 10th floor was very spacious with a large comfortable bed, walk in dressing area and beautiful bathroom. There was also a good desk if you are there on business We had an excellent view of the modern Sony Centre, but there was no traffic noise to disturb our sleep. It was definitely in the luxury class. The hotel restaurant is Michelin starred.

How can we effectively learn consumer preferences from various types of social media data (e.g., UGC and Crowd-Sourced Data) to improve user search experience?

How to improve consumer search experiences with various types of social media data?

Main Idea: Identify latent demand pattern by analyzing social media data → Ranking hotels based on consumer surplus → Reduce consumer search costs of evaluating offers online.

Key Challenge: Bridge the gap between the qualitative nature of social media content and the quantitative nature of economic choice models.

Method: Combine Text Mining, Image Classification, Social Geo-Tagging, Crowd-Sourced User Survey with Structural Modeling for Demand Estimation.

Validation: User experiments to validate our ranking system with existing benchmark systems offered by travel search engines.

Agenda

- Data
- Hybrid Structural Demand Model
- Results
- Conclusion

Data

Transaction data: *Travelocity.com,* 1497 US hotels, 2008/11-2009/1

Location Characteristics:

- Social geo-tags: *Geonames.org,* "Public transportation"
- GeoMapping Search Tools: *Microsoft Virtual Earth SDK,* "Restaurants"
- Image Classification: "Beach", "Downtown"
- On-Demand Survey: *Amazon Mechanical Turk (AMT),* "Highway"

Service Characteristics:

- JavaScript parsing engines: *TripAdvisor & Travelocity,*

"# of Internal amenities", "Reviewer Rating", "# of online reviews"

Additional Review Characteristics:

- Text Mining: *Review-based content from TripAdvisor & Travelocity,*

Text features (e.g., "Breakfast", "Staff"), "Subjectivity", "Readability", "Disclosure of Reviewer Identity"

Structural Model Framework

- Each hotel exclusively belongs to 1 of the 8 Travel categories:

Family Trip, Business Trip, Romantic Trip, Tourist Trip, Trip with Kids, Trip with Seniors, Pet Friendly, and Disability Friendly.

- A consumer makes a choice decision by looking for the hotel with specific service/location characteristics that best match her travel purpose:

e.g., A romantic trip traveler prefers:

1) characteristics: downtown, beach, nightlife, restaurants…

2) hotels in romance category are best candidates to satisfy her needs.

Our Model: Hybrid random coefficient demand model, capturing two levels of consumer heterogeneity during the decision process.

Structural Model

We propose a hybrid random coefficient-based structural model:

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

hotel *j* with category type ***k* (1≤k≤8)**

ε with a superscript *k* represents a travel category level “taste shock”.

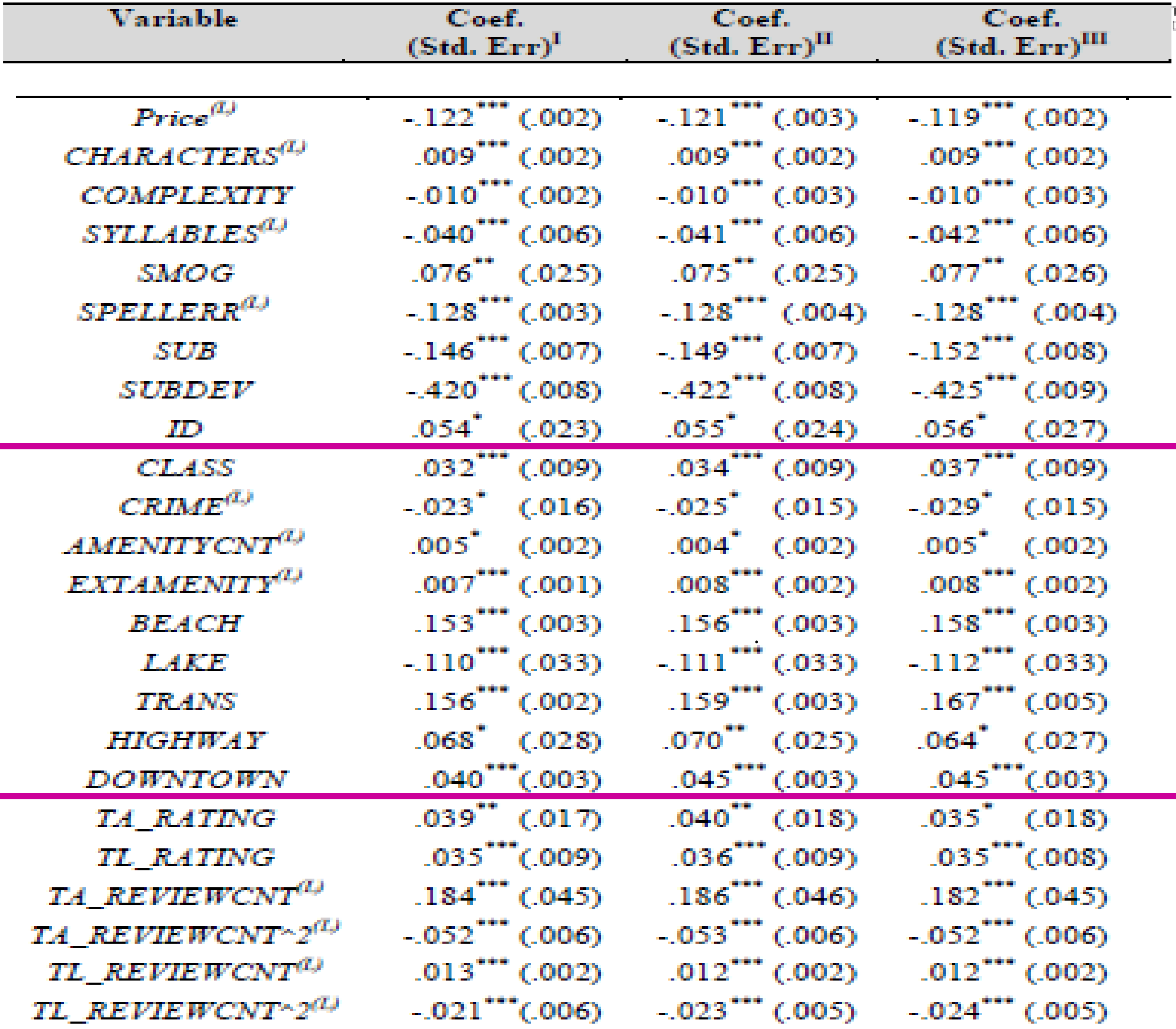
consumer-specific random coefficients

$$\beta_i \sim (\bar{\beta}, \sigma_\beta), \quad \alpha_i \sim (\bar{\alpha}, \sigma_\alpha)$$

- Choice of travel category
- Choice of hotel within a travel category

Summary: Combine the BLP (1995) and PCM (Berry &Pakes 2007).

Estimation Results



Variable	Coef. (Std. Err)^{I}	Coef. (Std. Err)^{II}	Coef. (Std. Err)^{III}
Price$^{(L)}$	-.122*** (.002)	-.121*** (.003)	-.119*** (.002)
CHARACTERS$^{(L)}$	.009*** (.002)	.009*** (.002)	.009*** (.002)
COMPLEXITY	-.010*** (.002)	-.010*** (.003)	-.010*** (.003)
SYLLABLES$^{(L)}$	-.040*** (.006)	-.041*** (.006)	-.042*** (.006)
SMOG	.076** (.025)	.075** (.025)	.077** (.026)
SPELLERR$^{(L)}$	-.128*** (.003)	-.128*** (.004)	-.128*** (.004)
SUB	-.146*** (.007)	-.149*** (.007)	-.152*** (.008)
SUBDEV	-.420*** (.008)	-.422*** (.008)	-.425*** (.009)
ID	.054* (.023)	.055* (.024)	.056* (.027)
CLASS	.032*** (.009)	.034*** (.009)	.037*** (.009)
CRIME$^{(L)}$	-.023* (.016)	-.025* (.015)	-.029* (.015)
AMENITYCNT$^{(L)}$	.005* (.002)	.004* (.002)	.005* (.002)
EXTAMENITY$^{(L)}$	.007*** (.001)	.008*** (.002)	.008*** (.002)
BEACH	.153*** (.003)	.156*** (.003)	.158*** (.003)
LAKE	-.110*** (.033)	-.111*** (.033)	-.112*** (.033)
TRANS	.156*** (.002)	.159*** (.003)	.167*** (.005)
HIGHWAY	.068* (.028)	.070** (.025)	.064* (.027)
DOWNTOWN	.040*** (.003)	.045*** (.003)	.045*** (.003)
TA_RATING	.039** (.017)	.040** (.018)	.035* (.018)
TL_RATING	.035*** (.009)	.036*** (.009)	.035*** (.008)
TA_REVIEWCNT$^{(L)}$	.184*** (.045)	.186*** (.046)	.182*** (.045)
TA_REVIEWCNT^2$^{(L)}$	-.052*** (.006)	-.053*** (.006)	-.052*** (.006)
TL_REVIEWCNT$^{(L)}$	.013*** (.002)	.012*** (.002)	.012*** (.002)
TL_REVIEWCNT^2$^{(L)}$	-.021*** (.006)	-.023*** (.005)	-.024*** (.005)

I. >= 1 review from either TA or TL.

II. Reviews >= 5.

III. Review >= 10.

Marginal Effects

Characteristics	Marginal Effect
Public transportation	18.09%
Beach	18.00%
Interstate highway	7.99%
Downtown	4.70%
Hotel class (Star rating)	3.77%
External amenities	0.08%
Internal amenities	0.06%

Robustness Checks

- Estimation using alternative sample split:

  Samples consisting of those hotels that have at least one review from only Travelocity or only TripAdvisor or both.

- Estimation using alternative models:

  Use BLP model, PCM, Nested Logit, OLS with random effect.

- Estimation using additional features:

  "*Airport*", "*Convention center*", and 7 individual service ratings from TripAdvisor, *eg, Value, Room, Location, Cleanliness, Service, Check-in, Business service.*

- Estimation using alternative IVs:

  Lag Price with Google Trend, Employee wage, Region dummies, BLP style IVs (Average characteristics of the same star-rating hotels in the other markets).

- All the results are qualitatively very consistent with our findings above.

Consumer Surplus Based Ranking

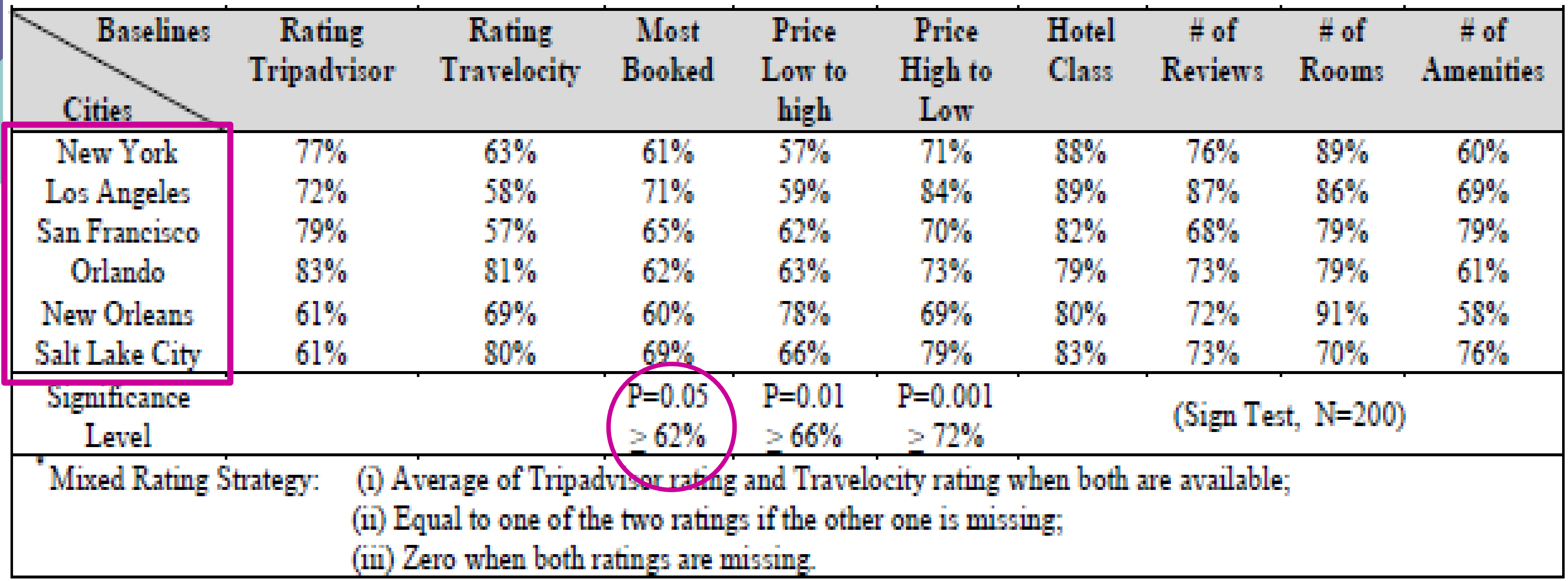
- Ranking hotels based on the consumer surplus of each hotel for consumers on an aggregate level.
- - How much "extra value" consumers can obtain after purchase?

User Study: Blind pair-wise comparisons, 200 anonymous AMT users;
Compare with 13 existing benchmarks.

Result: CS-based ranking is overwhelmingly preferred.
More than 80% customers (p=0.001, sign test)

Reasoning: - Diversity;
- Price & Quality;
- Multi-dimensional preferences.

Ranking Evaluation - User Study (1)



Baselines / Cities	Rating Tripadvisor	Rating Travelocity	Most Booked	Price Low to high	Price High to Low	Hotel Class	# of Reviews	# of Rooms	# of Amenities
New York	77%	63%	61%	57%	71%	88%	76%	89%	60%
Los Angeles	72%	58%	71%	59%	84%	89%	87%	86%	69%
San Francisco	79%	57%	65%	62%	70%	82%	68%	79%	79%
Orlando	83%	81%	62%	63%	73%	79%	73%	79%	61%
New Orleans	61%	69%	60%	78%	69%	80%	72%	91%	58%
Salt Lake City	61%	80%	69%	66%	79%	83%	73%	70%	76%
Significance Level			P=0.05 ≥ 62%	P=0.01 ≥ 66%	P=0.001 ≥ 72%		(Sign Test, N=200)		

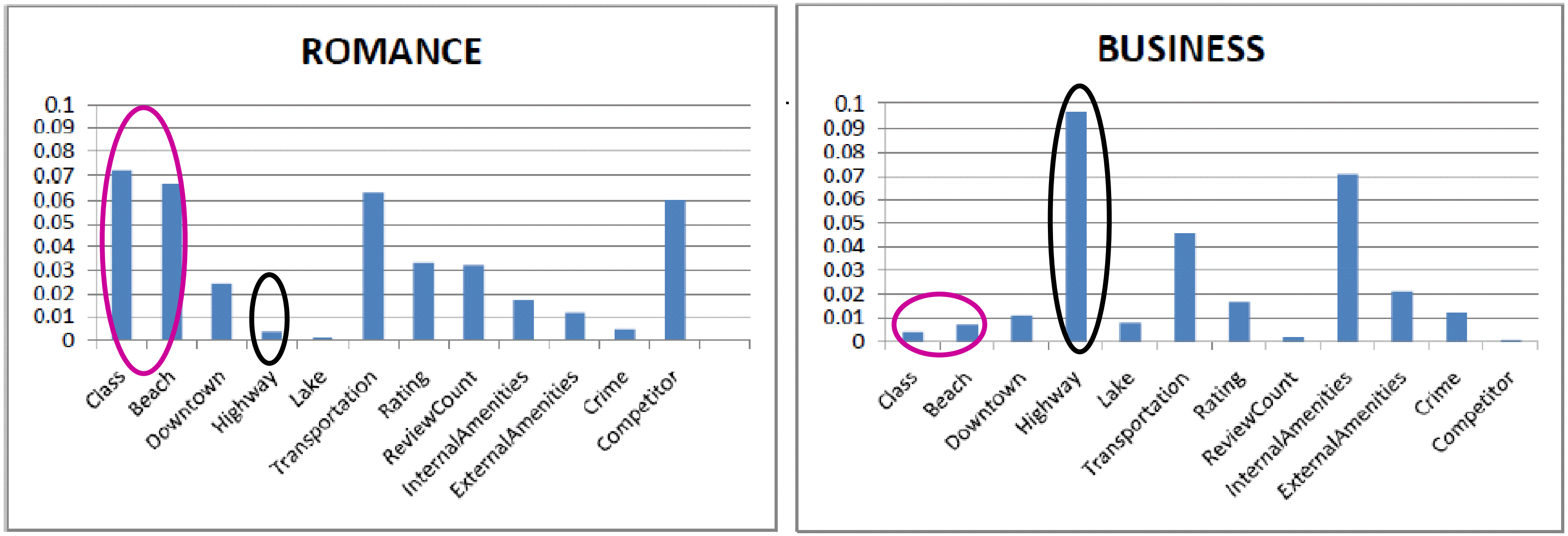
*Mixed Rating Strategy: (i) Average of Tripadvisor rating and Travelocity rating when both are available;
(ii) Equal to one of the two ratings if the other one is missing;
(iii) Zero when both ratings are missing.

Can we do better?

- Personalized ranking

- Incorporate consumer demographics;
  - i.e., age group, travel purpose
- Examine the interaction effect between consumer demographics and hotel characteristics; for example,
  - *Travel Purpose* and *Price;*
  - *Travel Purpose* and *Hotel Characteristics* (e.g., location, service, etc.);
  - *Travel Purpose* and *Brands.*
- Derive personalized consumer surplus for ranking.

Weights of Hotel Characteristics Based on Travel Purposes



(a)

(b)

Consumers with different travel purposes assign different weight distributions on the same set of hotel characteristics.

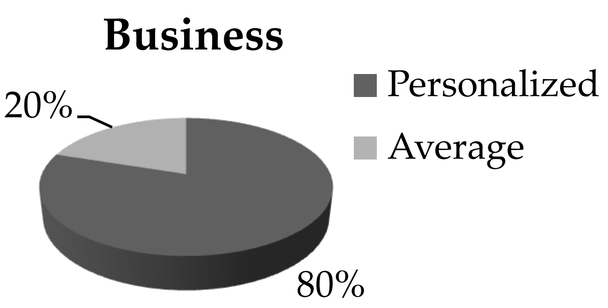
Selected Findings from Extended Model

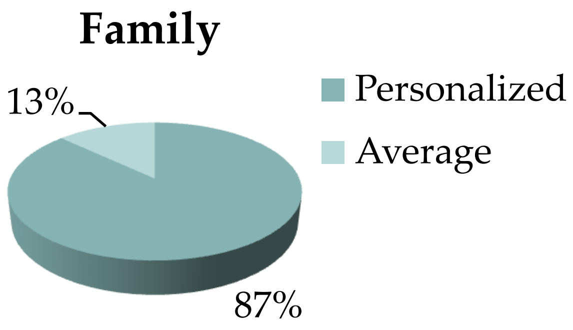
- Business travelers have the highest marginal valuation for "highway" and "transportation."
- Romance travelers have the highest marginal valuation for "hotel class" and "beach."
- Tourists have stronger preferences towards "Hilton" and "Intercontinental." Senior citizens have stronger preference towards ``Best Western"
- etc.

User Study (2)

Experiment 2: Blind pair-wise comparisons, 200 anonymous AMT users; baseline: generalized CS-based ranking (for an average consumer).

E.g., Business trip and family trip AMT user study results in the NYC experiment.





Conclusion: Personalized CS-based ranking is overwhelmingly preferred.

Reasoning: The personalized ranking model can capture consumers' specific expectations, dovetail with their actual purchase motivation in the real world.

Conclusions

✓ Economic impact of hotel characteristics using user-generated and crowd-sourced data.

✓ Structural model for demand estimation, image classification, text mining & on-demand social annotations.

✓ Model extension to capture interactions of hotel features with travel purpose and brands.

✓ New ranking system for hotels on travel search engines and validate it with field experiments.

- → Any product search engines

Demo: http://nyuhotels.appspot.com/

Future Work

In the future, we plan to look into:

- How does the "rank" of hotel affect clicks and conversions for that hotel? What other factors drive "search", "clicks", and "conversions" of a hotel?

  - ~1 M online sessions, >40 M search events;
  - **Hierarchical Bayesian modeling;**

- Why do some people search more, while others search less?

  - Empirically estimate consumer search cost in travel industry;
  - **Dynamic structural modeling.**