ICIC Express Letters
Volume **12**, Number **10**, October **2018**

ICIC International ©2018 ISSN 1881-803X
pp. **1047–1052**

IMPROVING FLIGHT SEARCH ENGINE BY LEARNING CONSUMER PREFERENCE FUNCTION FROM CHOICE DATA

Jihwan Lee[1] and Yoo Suk Hong[2]

[1]Division of Systems Management and Engineering
Pukyong National University
45, Yongso-ro, Nam-gu, Busan 48513, Korea
jihwan@pknu.ac.kr

[2]Department of Industrial Engineering
Seoul National University
1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
yhong@snu.ac.kr

Received March 2018; accepted June 2018

Abstract. *Online Travel Agency (OTA) provides the flight search and the booking capability throughout an online web site. Searching the flight ticket is provided by the flight search engine. If the user specifies his destination with the departure date, then the search engine returns available tickets that meet the user's query. As a result, several competing tickets are usually found from the query. Since there are usually a large number of competing itineraries in the same route, filtering or recommending itineraries that are likely to be selected by customers becomes an important feature for the search engine. In this study, discrete choice model which can predict the probability of customer's choice among several product alternatives is used to recommend the favorable itineraries. In learning the choice model, conjoint experiment was conducted with 500 actual customers. As a result, a utility function that can measure the customer's preference on itinerary alternatives was identified. We show that the elicited utility function then can be used to rank the search result in the flight search engine.*

Keywords: Flight search, Search engine optimization, Preference learning, Discrete choice, Experimental design

1. **Introduction.** Online Travel Agency (OTA) provides the flight search and the booking capability throughout an online web site. Recently OTA has become a major player in the airline and travel market. In 2015, the gross booking of Expedia.com which is the market leader in OTA accounted for $50.4 billion.

The main service of OTA is provided by the flight search engine. Given the destination and departure date, the search engine usually returns hundreds of competing itineraries that differ in attributes such as airline operators, price, stopovers, and departure time. Because the number of itineraries is so large, navigating throughout them usually becomes a daunting task for the user. Therefore, filtering itineraries that are likely to be selected by the user is an important issue for the search engine.

In this study, discrete choice model (DCM) was applied to filter itineraries based on user's preference. DCM, which was originally proposed by McFadden [1], explains the individual's choice among multiple alternatives with the simple utility function. Due to its good performance and theoretical simplicity, it has been applied to various areas including marketing [2], economics [3], and artificial intelligence [4].

DCM also has wide applications in numerous industry areas including retail [5] and Transportation [6]. In an airline industry, DCM has been applied to forecast the demand of airlines that compete over the same flight route. For example, Coldren and Koppelman

DOI: 10.24507/icicel.12.10.1047

[7] predict the itinerary market share of airlines from the historical airport reservation database. Garrow and Koppleman [8] apply DCM to estimating the price sensitivity of flight itineraries that can be used in revenue management. For more applications of DCM in airline industry, please refer to Garrow [9]. Nevertheless, DCM has not been extensively used in search engine. Smith and Brynjolfsson [10] developed the model to predict the choice of items in online bookstore. The preference coefficient of utility model was then examined to reveal the difference of customer's preference between online and offline store. However, their model was not used in filtering alternatives in search engine.

To the author's knowledge, this work is the first attempt to adopt the DCM in filtering itineraries in the flight search engine. In this study, DCM was used to learn the preference of flight search engine users. To systematically learn the preference model, a choice experiment was conducted with artificially generated itineraries with 500 users. As a result, a preference model that can predict the choice probability of each attribute can be obtained from the aggregated choice data. This model then can be used to filter itineraries that are likely to be selected by the search engine user.

The reminder of this paper is organized as follows. Section 2 introduces the DCM briefly. Section 3 explains the choice experiments. Section 4 explains the experiment result and shows its application in recommendation of the search result. Finally, Section 5 draws conclusions and anticipates the future work.

2. Discrete Choice Model.

The discrete choice model assumes that the individual makes choices based on his/her latent utility function. Consider the person n who selects the flight i among the set of alternative flight set $A = \{1, 2, \ldots, n\}$. Let U_{ni} be the latent utility function of person n that obtains from the choice of flight i. In discrete choice model, U_{ni} is decomposed into the systematic part (V_{ni}) and stochastic component (ε_{ni}):

$$U_{ni} = V_{ni} + \varepsilon_{ni} \tag{1}$$

The systematic part V_{ni} depends on the observed attributes of alternative i, while the stochastic part ε_{ni} captures the impact of all unobserved factors that affect the person's choice. In our problem, possible attributes may include price, journey time, or the number of stops that may influence the individual choice on flight. The systematic part V_{ni} can be further decomposed into the following equation:

$$U_{ni} = V_{ni} + \varepsilon_{ni} = x_{ni}\beta + \varepsilon_{ni} \tag{2}$$

where x_{ni} is a vector of observed variables relating to alternative i and β is a person n's preference vector with respect to the attributes of the alternative.

The person would choose the flight i if and only if it would provide the highest utility compared with other flight alternatives, which is expressed as the following:

$$U_{ni} > U_{nj}, \quad \forall j \neq i \tag{3}$$

Due to the stochastic part ε_{ni}, which is a random variable, the individual's choice only is expressed with probabilistic measure. There are several types of discrete choice models depending on the distributional assumption of ε_{ni}. The most widely used model is multinomial logit model. The multinomial logit assumes that ε_{ni} belongs to the type I extreme distribution (Gumble distribution) and also holds independence and identical distribution (IID). In this study, we also assume multinomial logit model. Under the MNL model, the probability of choosing alternative i is expressed as follows:

$$P_{ij} = \frac{e^{x_{ni}\beta}}{\sum_{k=1}^{J} e^{x_{nk}\beta}} \tag{4}$$

where J is the total number of alternatives that individual n considers. Based on the above model structure, our goal is to estimate β by aggregating multiple individual's

choice data. In MNL, we can easily estimate the β using maximum likelihood method. For the details of the estimation method, please refer to Train [11].

3. **Experimental Design.** The discrete choice model builds upon the experimental design and data gathering procedure. The experiment procedure requires respondents to make choice among a set of artificially generated itinerary alternatives that may exist in a same flight segment. Each flight is represented by the combination of their attributes, and the flight alternatives are constructed by varying each attribute over a range of levels. A sample of alternatives are then selected by the experimental design principle. The consumer's utility function is then obtained by aggregating respondent's choice.

Firstly, itinerary attributes were determined. The list of attributes and its corresponding value range was listed in Table 1. The value range of itinerary price was from 0.8 to 1.3. We adopted normalized scale because the average itinerary price may differ across the flight route. The value range of itinerary price was determined by examining more than 50,000 itineraries over 500 randomly chosen routes.

Table 1. Attribute values for the choice experiment

Attributes	**Values**	**Unit**
Price (normalized scale factor)	0.8, 0.9, 1, 1.1, 1.2, 1.3	Ratio
Number of stops	0 (direct), 1, 2	Number
Transfer time (1 connection)	1, 3, 5 hours	Hours
Transfer time (2 connection)	1, 3, 5 hours	Hours
Departure time	Dawn (0), Morning (1), Afternoon (2), Evening (3), Night (4)	

Table 2. Flight segment information (average price and flight time)

Route types	Routes	Avg. price (USD)	Flight time	Time lag from Seoul
Overseas trip	Seoul-LA	1184	600	−14
	Seoul-Washington DC	1566	898	−17
	Seoul-Rome	1334	745	−8
	Seoul-Paris	1357	725	−8
	Seoul-Madrid	1345	820	−8
	Seoul-Budapest	1543	730	−8
Short trip	London-Munich	341	140	0
	Barcelona-Vienna	300	170	0
	Paris-Frankfrut	279	115	0

Another attribute is total traveling time. The traveling time is the actual time taken to travel from the departure city to the arrival city. The traveling time is equal to the sum of actual flight time and additional time taken to the stopover. Since the flight time is almost the same within a flight segment, the difference of traveling time mainly depends on the additional time taken in the stopover. As shown in the table, the maximum stopover that each itinerary can have was set to 2 because the itinerary with more than 3 stops may be rarely chosen. The possible additional time for each stopover was from 1 to 5 hours.

Finally, departure or arrival time may affect the consumer's choice. Since consumer usually tries to avoid too early or late time zone, binary indicator for the early morning for both arrival and departure time was chosen for the attribute.

After attribute levels are identified, experimental design principle was applied to generate the choice set. Attributes in the choice set should be orthogonal, which means that the attributes presented to the respondent vary independently from one another. However, the complete orthogonal design requires respondent to compare total $5*3*3*3*2*2 = 640$ profiles of flights, which become too many surveys to handle. To reduce the size of the choice set, fractional factorial design is applied. Fractional factorial design arranges attribute levels of each alternative with fewer runs by ignoring some of the interactions except for main effects. For the details of the fractional factorial design is referred to Louviere et al. [12].

Another issue in experiment design is to ensure that there is no dominating alternative in the choice set. The alternative is dominant if its attribute values are obviously preferable than the others. For example, consider three flight alternatives, where flight A is \$500 with traveling time 6 hours, flight B is \$600 with 7 hours and flight C is \$700 with 6 hours. Obviously a respondent would choose flight A because it has cheapest price as well as short traveling times. To avoid the dominant alternative in the choice set, the Bayesian optimal design algorithm was applied. For the details of the Bayesian optimal design, please refer to DuMouchel and Jones [13].

In the survey, each respondent was asked to choose itinerary alternatives over 9 different flight segments. The base price and traveling time of each flight segment are summarized in Table 3. As shown in the table, six segments were overseas flights, and three segments were domestic flights. The survey was conducted with 500 respondents, who were recruited by a market research company in South Korea. All respondents were in 20-30 age group, who consider overseas trip in near future. The survey was conducted throughout the mobile phone interface. The screenshot of the choice survey is shown in Figure 1.

Table 3. Estimated coefficients of the utility function

	Model 1		Model 2		Model 3	
Coefficients	Estimate	t-value	Estimate	t-value	Estimate	t-value
Price (\$)	−0.0031	−23.14**	–	–	−0.0031	−25.31**
Price (ratio)	–	–	−4.457	−23.44**	–	–
Stops (no.)	−0.474	−9.48**	−0.481	−9.73**	−0.4267	−10.19**
Duration (hr)	−0.211	−13.94**	–	–	−0.2087	−16.31**
Duration (ratio)	–	–	−2.728	−14.26**	–	–
Departure at Dawn $(0, 1)$	0.203	4.42**	0.209	4.578*	–	–
Arrival at Dawn $(0, 1)$	−0.115	−2.34*	−0.091	−1.860	–	–
Log-likelihood	−3119.3		−3119.8		−4487.7	
McFadden R^2	0.1875		0.1874		0.1819	

4. **Result.** Learning the DCM from the data was conducted with R. Especially, 'mlogit' package was used to estimate the multinomial logit model from the aggregated choice data. Basically, our preference model includes the price, total flight hour, the number of stops, and the indicator of early flight time (whether the arrival and departure time is from 0 am to 7 am).

Because each itinerary attribute can have different unit measures, several attempts have been conducted to find the best combination of attributes. Table 3 reveals three preference models that have highest model fitness measure (McFadden R^2). Although each model has different attribute combination, their predictive power was almost the same. As expected, price, stops and flight time have negative coefficient in the utility function. Among these models, only the 'Departure at Dawn' was associated with positive coefficient. This states

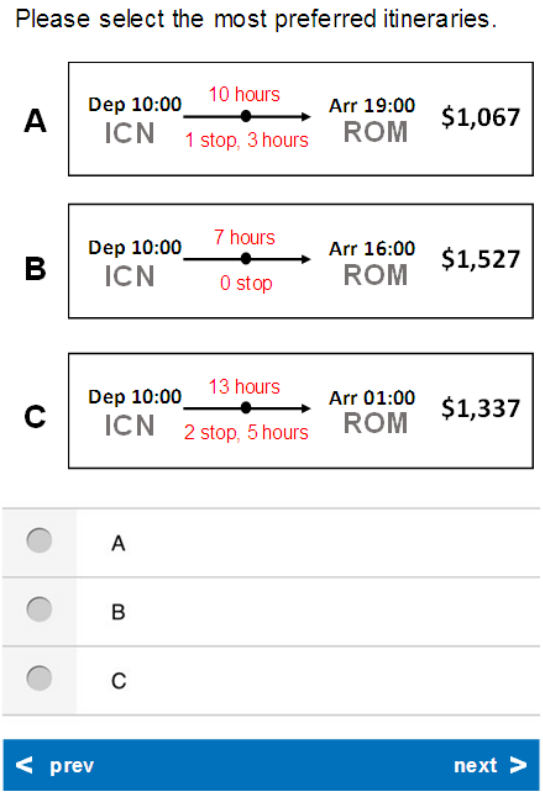

FIGURE 1. Screenshot of the choice survey

that the user prefers the departure at early morning. The predictive power of obtained model was examined by dividing the raw data into the training and the validation data. Among total 4500 response data, 3000 were used in training and rest of 1500 choice data were used to validate the accuracy of the model. The average accuracy was 60.73% which shows moderate performance.

The trained model was also applied to actual search engine data. Table 3 shows itineraries recommended by the preference model. Among over 200 itineraries that were obtained from the flight search engine website, top 20 itineraries with highest utility score were shown in the table. It is noteworthy that the most preferred ticket does not have the cheapest price. Instead the most preferred ticket was a little bit expensive but it saves about 1 hours compared with the cheapest one. This result indicates that the preference model can recommend the ticket that may not be ranked highly when sorting by the single criterion.

5. **Conclusion and Future Works.** This work suggests ticket filtering system by measuring the customer preference on itineraries. To systematically learn the preference model from the choice data, discrete choice model was applied. Several experiment designs were conducted to obtain the reliable and unbiased models. The utility function was obtained from the large choice data, and it is then implemented in the actual flight search engine.

Although our model shows the promising result, it also has some limitations that require further studies. Currently, preference model was obtained by aggregating the choice data over 9 different flight segments. Since value range of price and flight hour differ across different flight segments, the aggregated model may result in the moderate performance. One way to deal with this issue is to separate the preference models according to several subgroups. Clustering analysis may be used to identify such subgroups. Such refined model would result in much better prediction performance.

Another limitation is that the current model only considers linear combination of attributes. Although the linear model is simple and intuitive, it cannot address the complex

interactions that lie between attributes or alternatives. For example, negative interaction usually exists between itinerary price and the stopover. Similarly, some itinerary may interact with another itinerary. For example, itineraries from the same airline operator may show similar price pattern and flight hours. Authors are currently working on to deal with such complex interaction by applying the non-linear model such as neural network or support vector machines.

Acknowledgement. This work was supported by Pukyong National University Research Fund of 2017.

REFERENCES

[1] D. McFadden, Econometric models of probabilistic choice, *Structural Analysis of Discrete Data with Econometric Applications*, pp.198-272, 1981.
[2] S. T. Berry, Estimating discrete-choice models of product differentiation, *The RAND Journal of Economics*, vol.25, no.2, pp.242-262, 1994.
[3] K. A. Small and H. S. Rosen, Applied welfare economics with discrete choice models, *Econometrica: Journal of the Econometric Society*, vol.49, no.1, pp.105-130, 1981.
[4] B. Eric, N. D. Freitas and A. Ghosh, Active preference learning with discrete choice data, *Advances in Neural Information Processing Systems*, pp.409-416, 2008.
[5] C. P. Lamberton and K. Diehl, Retail choice architecture: The effects of benefit- and attribute-based assortment organization on consumer perceptions and choice, *Journal of Consumer Research*, vol.40, no.3, pp.393-411, 2013.
[6] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, 1985.
[7] G. M. Coldren and F. S. Koppelman, Modeling the competition among air-travel itinerary shares: GEV model development, *Transportation Research Part A: Policy and Practice*, vol.39, no.4, pp.345-365, 2005.
[8] L. A. Garrow and F. S. Koppelman, Multinomial and nested logit models of airline passengers' no-show and standby behaviour, *Journal of Revenue and Pricing Management*, vol.3, no.3, pp.237-253, 2004.
[9] L. A. Garrow, *Discrete Choice Modelling and Air Travel Demand: Theory and Applications*, Routledge, 2016.
[10] M. D. Smith and E. Brynjolfsson, Consumer decision-making at an Internet shopbot: Brand still matters, *The Journal of Industrial Economics*, vol.49, no.4, pp.541-558, 2001.
[11] K. E. Train, *Discrete Choice Methods with Simulation*, Cambridge University Press, 2009.
[12] J. J. Louviere, D. A. Hensher and J. D. Swait, *Stated Choice Methods: Analysis and Applications*, Cambridge University Press, 2000.
[13] W. DuMouchel and B. A. Jones, A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model, *Technometrics*, vol.36, no.1, pp.37-47, 1994.