A Reference-Dependent Model of Search Evaluation

Jiqun Liu

School of Communication and Information, Rutgers University
New Brunswick, New Jersey
jl2033@scarletmail.rutgers.edu

ABSTRACT

Most of the existing IR studies employed final values of search behavior measures in building evaluation metrics. However, according to the theories and empirical evidences from Behavioral Economics studies, in people's evaluations of actions and outcomes, the carriers of the values of different actions are gains and losses defined relative to a reference point, rather than the absolute final assets. Based on this idea, I will first explore how users' levels of search satisfaction are affected by the gains and losses defined relative to the pre-search expectations of system performance or reference levels in a controlled lab study. Then, based on the data collected from a field study, I will test the predicative power of my reference-dependent models (built upon Δ -value-based behavioral features given the corresponding reference points) in predicting user satisfaction in naturalistic settings, aiming to examine the extent to which the reference-dependent approach can approximate real users' search evaluations. The findings of this work can help us better understand the subjectivity, bias, and variation in users' evaluation of search experience and thus have implications for user modeling and system recommendations design.

CCS CONCEPTS

• Information systems → Users and interactive retrieval.

KEYWORDS

Reference dependence; search evaluation; interactive IR

ACM Reference Format:

Jiqun Liu. 2019. A Reference-Dependent Model of Search Evaluation. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*, March 10–14, 2019, Glasgow, United Kingdom. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3295750.3298970>

1 INTRODUCTION

How users evaluate their search experience when interacting with information retrieval (IR) systems is a fundamental research problem in the area of Interactive Information Retrieval (IIR). In particular, *user satisfaction* as a major aspect of search experience can significantly affect search actions and judgments in online information seeking episodes and thus has been widely used as a ground truth for evaluating IR evaluation metrics of different types (including both behavior-based online metrics and annotation-based

offline metrics) [5, 12]. Therefore, to embrace a deeper understanding of user characteristics in information interaction, it is critical to investigate how user satisfaction is affected by their interactions with IR systems and the associated outcomes.

Most of the existing IR studies employed final values of search behavior measures in building evaluation metrics. However, according to the theories and empirical evidences from Behavioral Economics studies, in real people (instead of "rational man")'s evaluations of actions and outcomes, the carriers of the values of different actions are gains and losses defined relative to a reference point, rather than the absolute final assets [18]. Given these findings, it is critical to explore (1) *how, if at all, are users' levels of search satisfaction affected by the gains and losses defined relative to the corresponding reference levels*, and (2) *to what extent can we infer user satisfaction based on a reference-dependent approach*.

2 MOTIVATION

A promising development in the modeling of search interaction and evaluation came from the application of classical economic theories. Specifically, for example, Azzopardi and his colleagues adopted models (e.g., Cobb-Douglas Production Function) from Microeconomics Theory to develop Search Economic Theory and empirically tested some of the connections between behavioral factors (e.g., number of queries issued, number of documents assessed) [2, 4]. With respect to evaluation, [3] designed and applied information-foraging-based measure in evaluating the utility of search engine result pages (SERPs). Built upon classical economic theories, these models use the final values of search behavior measures (e.g., total number of documents assessed within a query segment, number of queries issued within a task session) in developing behavioral features and evaluation metrics. Behind these measurements is the implicit assumption that individuals evaluate utility based on the final objective assets or outcomes associated with different actions and decisions (cf. expected utility theory [9]).

However, as one of the cornerstone of classical/rational economic models, the final-outcome-based objective preference and outcome evaluation assumption has been demonstrated to be problematic in a variety of *Behavioral Economics* studies (e.g., [10, 16–18]). Based on the evidences collected from a series of experiments, Tversky and Kahneman [17] rejected the expected utility approach which assumes that people evaluate the utility of an option based on the final outcome. Instead, they proposed a *Reference-Dependent Model* and argued that the carriers of the values of different actions are gains and losses defined relative to a reference point, rather than the absolute final assets. Therefore, people's preferences and result evaluations are essentially dynamic and relative rather than static and objective as they depend on their corresponding reference levels (which is often represented by status quo). This theory also helps explain another relevant phenomenon: *Framing Effect*.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6025-8/19/03.

<https://doi.org/10.1145/3295750.3298970>

Specifically, according to [18], the framing of options in evaluation can yield systematically different preferences. An individual's perception of a given option and the associated decision can be easily reversed with merely adjusting the reference point. Therefore, it is not appropriate to define and represent people's preferences using predefined indifference curves (cf. ordinal utility theory [6]). Also, Tversky and Kahneman [18] found that people are more sensitive to the gains and losses which are close to the reference point.

Furthermore, based on the idea of reference dependence, Tversky and Kahneman [18] also demonstrated that people tend to prefer avoiding losses to acquiring equivalent gains: it is better to not loss \$5 than to find \$5 (*Loss Aversion*). Consequently, when the outcome of an option is perceived as a loss, the decision maker's satisfaction level will drop quickly. Also, when an option consists of different facets and dimensions, people are more sensitive to the dimension in which they are losing relative to the reference point.

As one of the fundamental principle of behavioral economics, reference dependence (1) helps explain other associated "anomalies" (e.g., framing effects, loss aversion) which clearly violate the assumption of final-outcome-based evaluation and (2) speaks to the human bias in evaluation and decision making, which are not covered by the classical rational economic models. Inspired by the reference-dependent approach, in my dissertation work I will build a reference-dependent model of search evaluation and explore how the dynamic reference points affect users' evaluations of their search experience and performance. It is likely that the reference-dependent approach can offer search evaluation model a more realistic behavioral and psychological foundation.

3 RESEARCH QUESTIONS

In the light of the relevant theories and empirical evidences from behavioral economics research [10, 16–18], I will study the roles of reference level and the associated tendency, loss aversion, in users' search evaluation. Specifically, my dissertation study seeks to address the following research questions (RQs):

- **RQ1:** *To what extent, if at all, does the difference in reference level lead to the difference in user satisfaction?*
- **RQ2:** *Are users' levels of satisfaction more sensitive to losses than to gains given a reference level?*

RQ1 and RQ2 examine the effects of reference level and the tendency of loss aversion on user satisfaction. Given the existing evidences on the phenomenon of reference dependence from various behavioral economic experiments, we may find similar effects of reference level and loss aversion in the context of user search evaluation and IIR in general.

- **RQ3:** *To what extent, if at all, do different amounts of loss and gain cause different levels of cognitive load and stress?*

To explain the effect of reference dependence at cognitive level, Gonzalez et al. [8] ran fMRI tests and their neuro evidences demonstrate that without changing the nature of the options, the cognitive efforts or loads increased when a participant needed to select a option that was framed as a loss. I will take a step forward by investigating the effect of relevance dependence at both cognitive level (i.e. cognitive load) and emotional level (i.e. perceived stress), aiming to help explain how reference level affect user evaluation.

- **RQ4:** *To what extent can we predict a user's expectation disconfirmation state using the reference-dependent model?*
- **RQ5:** *To what extent can we predict a user's level of satisfaction in search interaction using the reference-dependent model?*

According to the Expectation Confirmation Theory (ECT) [15], when a system outperforms a user's expectations (which means the expectation disconfirmation state is *positive*), the user is very likely to be satisfied with the system performance. In contrast, when the expectation disconfirmation state is *negative*, the user is more likely to be dissatisfied and thus may discontinue the usage of the information system. In the reference-dependent framework, a user's pre-search expectation of system performance can be approximated by the reference level or status quo for search evaluation. Accordingly, the expectation disconfirmation state can be considered as the perceived result of search interaction (i.e. gain or loss). Predicting the expectation disconfirmation state from search behavioral signals can help explain the effect of reference level on users' evaluations. As a natural extension from RQ4, RQ5 directly speaks to the predicative power of search-behavioral-based reference-dependent model. When building reference-dependent models, I will use the Δ values of different search behaviors (i.e. the difference between each behavioral measurement and the corresponding reference point) as features and build prediction models based on several possible reference levels. Given the Δ -value-based measure, RQ5 essentially pushes us to investigate the extent to which we can infer the implicit user satisfaction based on a user's gains and losses over various dimensions of explicit search interaction.

4 METHODOLOGY

Given the nature of different research questions, my dissertation work consists of two different user studies. The first user study is a *controlled lab study* where I control task topic and cost-benefit scenarios (i.e. number of relevant results on predefined SERPs) in search evaluation, aiming to accurately examine the "clean effects" of reference level and loss aversion. The controlled lab study is designed to answer RQ1-3. The second study is a *field study* where I collect data on participants' search behavior and evaluation in naturalistic settings and test the predicative power of behavior-based reference-dependent model in predicting participants' expectation disconfirmation states and levels of search satisfaction (RQ4-5).

4.1 Controlled Lab Study

SERPs and cost-benefit scenarios. I will represent different levels of benefits by varying the amount of relevant results on a SERP consisting of 12 results in total (see the 12 scenarios in Figure 1). I include 12 text-only snippets (instead of 10 results) to make sure there is enough room for detecting the variations on both the gain (above) and loss (below) sides of the reference levels in search evaluation.

For each participant, each scenario has two different SERPs. Thus, each participant needs to evaluate 26 SERPs in total. The queries and search result snippets will be extracted from the results retrieved from Microsoft Bing API. I will prepare 26 different tasks with different topics to contextualize the 26 different SERPs in order to avoid learning effects. The relevance (highly relevant, somewhat relevant, not relevant) of each result snippet to the given task will be evaluated by at least two external assessors separately. To avoid

Figure 1: 13 SERPs (blue: relevant, white: non-relevant).

the potential ambiguity in relevance judgment, among the search result snippets which receive the same relevance rating from two assessors, I will only use the ones which are rated as either “highly relevant” or “not relevant”. Also, I only use factual task here in order to make it clearer for relevance judgment. Since query length may affect users’ judgments [19], all queries here will be the same in length (for example, 4 words). In Figure 1, for each condition, I mix the relevant and irrelevant results in SERPs (except the 0 condition) instead of aggregating relevant results altogether on the top as I want participants to be sensitive all the time to the relevant results. I control the ranking of the results instead of randomizing them in order to clear out the effect of ranking when analyzing the role of reference level. To make sure each SERP can create a unique, well-controlled cost-benefit scenario, I will ask participants to evaluate all result snippets on each SERP (so the total cost is controlled and the total benefit is manipulated by the scenarios).

Note that the search result snippets here are not clickable. The participants will be asked to evaluate the relevance of each result based on the snippet only. I make this decision in order to control the length of study within a reasonable amount of time. The sequence of the SERP scenarios will be rotated for each participant.

Participant and study procedure. I design two different reference levels here, *4 relevant results out of 12 results* and *8 relevant results out of 12 results*, aiming to make sure there are enough data points on both gain and loss sides of each reference level. The reference levels will be introduced as the average level of performance of the current system both at the beginning of study and in the side bar of the system during the study (see Figure 2). The number of relevant documents found will also be shown in the side bar. Participant can indicate their level of satisfaction on a continuum by dragging the slider. I adopt a *between-subjects design* here to avoid possible learning effect. This study only recruits undergraduate students who are native English speakers in order to control the possible effect of language and education level on snippet evaluation. I will recruit 60 participants and randomly assign them into the two reference level groups (30 participants for each group).

With respect to the study procedure, the study will start with a short survey asking questions about their backgrounds and demographic information. Then I will ask the participants to start the first session where they need to evaluate the first set of 13 SERPs (about 26 minutes). Then participants will take 2 minutes break (to mitigate fatigue) and continue with evaluating the other 13 SERPs in the second session. Participants will be asked to find as many relevant result snippets as they can. To answer RQ3, I will collect *Electroencephalogram (EEG)* data (using EEG headset) [1] and *Heart Rate* data (using BioPac) [7] to infer participants’ cognitive load and perceived stress respectively.

Figure 2: Mockup of the study system (Note that the real system will include 12 result snippets).

Data analysis. Before answering RQ1-3, I will test the correlation between the user-reported expectation disconfirmation state (below or above expectation) and the perceived result of SERP evaluation (loss or gain). A high correlation value would indicate that our manipulation of participants’ reference levels is successful.

To answer RQ1, I will employ t-test or Mann-Whitney U test (depending on the data distribution) to compare the satisfaction levels of the participants from the two reference level groups under same SERP conditions. If there is a reference-dependent effect, then the differences in satisfaction rating between the groups should be statistically significant in all conditions (with the “4 out 12” group always higher than the “8 out 12” group). To answer the RQ2, in each reference level group, I will divide the dataset into two parts according to the result state (above or below the reference level) and run regression models (regress user satisfaction on evaluation result) on them separately. If there is a tendency of loss aversion, then the slope of the loss-side model should be greater than that of the gain-side model. To answer RQ3, I will test the difference in the electrical activity of the brain and heart rate between the two different reference-level groups within the same SERP scenarios.

4.2 Field Study

Participant and study procedure. To answer RQ4-5, in the field study, I will recruit 30 native English speakers to do seven-day Web search in naturalistic settings. Participants’ search behavioral data will be collected via a Chrome browser plugin.

The field study will start with a pre-search interview session where I will ask questions about participants’ daily work tasks and search tasks and also give participants a tutorial on installing and using the browser plugin. Then, the participants will do their regular Web search with the plugin turned on and annotate their search tasks at the end of each day (annotate the name of search task for the corresponding segments of search logs and offer task descriptions). To collect ground truth data for evaluating prediction models, when a participant searches on the Web with the browser plugin turned on, a pop-up window will show up at the end of each

query segment. In this work, query segment is defined as *a search session segment which starts from one issued query and ends at the next query*. In the pop-up window, the participant will be asked to indicate their expectation disconfirmation state and satisfaction level using the sliders shown in Figure 2. To reduce participants' cognitive load in study participation, I will enable them to skip one questionnaire/pop-up window for every three query segments.

After the seven-day Web search in the wild, participants will be invited to a post-search individual interview where I will ask detailed questions on how they decide their pre-search expectations and search strategies and evaluate their search experiences.

Search behavioral measurements and data analysis. To measure the gains and losses in search interaction, I employ a variety of search behavioral measures here: (1) the benefit side of interaction, such as number of content pages that are read for more than 30 seconds (i.e. number of relevant pages), total dwell time on relevant pages. (2) the cost side of interaction, such as dwell time on SERP, dwell time on content page, number of search result snippets examined, number of content pages clicked and hovered by the mouse cursor, distance of cursor movement, time spent on formulating query. Given the query-level user satisfaction evaluation, all behavioral measures are calculated within query segments. To apply the reference-dependent approach, I design six “candidate reference levels” here and compute six sets of Δ -value-based features:

- average value of each search behavioral measure within the current search task (local reference level1);
- average value of each measure within the most recent search task + the current task (local reference level2);
- average value of each measure within the most recent search task of same type + the current task (local reference level3);
- average value of each measure within all search tasks from the same day (meso reference level);
- average value of each measure within the search tasks of same type from all past days (global reference level1);
- average value of each measure within the search tasks from all past days (global reference level2);

To answer RQ4-5, I will build six regression prediction models based on the six sets of Δ -value-based behavioral features to predict participants' perceived system performance and levels of search satisfaction within query segments. I will compare the prediction performances of different models in order to determine the best model which can approximate the reference level employed by participants in real-time search evaluation. Furthermore, this study will also compare the prediction performances of the models across different groups of participants (e.g., male and female) and tasks of different types (e.g., fact finding task and exploratory task), aiming to obtain a deeper understanding of the reference-dependent evaluation in different contexts. In addition, I will code and extract qualitative evidences from the transcripts of post-search interviews in order to better interpret the prediction models and results.

5 PROGRESS AND FUTURE PLAN

Previously, I have conducted a series of studies at both conceptual level and empirical level to explore the task and information seeking intention aspects of search interaction (e.g., [13, 14]) as well as the applicability of economic approach in modeling information

seeking and search behaviors [11]. The search tasks, prediction models, and study instruments (e.g., questionnaire, browser plugin) developed in my previous studies can be modified and applied in the proposed study. My next step will be finalizing the study design, preparing the recruitment information and study instruments, and conducting pilot tests for the controlled lab study. A first round of pilot tests with five participants will be conducted to test the study instruments, procedure, and the predefined SERPs.

6 ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Chirag Shah. This work is supported by National Science Foundation grant IIS-1717488.

REFERENCES

- [1] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara Van Gog. 2010. Using electroencephalography to measure cognitive load. *Educational Psychology Review* 22, 4 (2010), 425–438.
- [2] Leif Azzopardi. 2014. Modelling interaction with economic models of search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 3–12.
- [3] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- [4] Leif Azzopardi and Guido Zuccon. 2016. An analysis of the cost and benefit of search interactions. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ACM, 59–68.
- [5] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of Online and Offline Web Search Evaluation Metrics. *SIGIR*.
- [6] James S Coleman and Thomas J Fararo. 1992. Rational choice theory. *Nueva York: Sage* (1992).
- [7] Ashlee Edwards and Diane Kelly. 2017. Engaged or Frustrated?: Disambiguating Emotional State in Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 125–134.
- [8] Cleotilde Gonzalez, Jason Dana, Hideya Koshino, and Marcel Just. 2005. The framing effect and risky decisions: Examining cognitive functions with fMRI. *Journal of economic psychology* 26, 1 (2005), 1–20.
- [9] John D Hey and Chris Orme. 1994. Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society* (1994), 1291–1326.
- [10] Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* 5, 1 (1991), 193–206.
- [11] Jiqun Liu. 2017. Toward a unified model of human information behavior: an equilibrium perspective. *Journal of Documentation* 73, 4 (2017), 666–688.
- [12] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. “Satisfaction with Failure” or “Unsatisfied Success”: Investigating the Relationship between Search Success and User Satisfaction. In *Proceedings of the 2018 World Wide Web Conference*. Geneva, Switzerland. 1533–1542.
- [13] Matthew Mitsui, Jiqun Liu, Nicholas J Belkin, and Chirag Shah. 2017. Predicting Information Seeking Intentions from Search Behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1121–1124.
- [14] Matthew Mitsui, Jiqun Liu, and Chirag Shah. 2018. How Much is Too Much?: Whole Session vs. First Query Behaviors in Task Type Prediction. In *Proceedings of ACM SIGIR*. ACM, 1141–1144.
- [15] Richard L Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research* (1980), 460–469.
- [16] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science* 211, 4481 (1981), 453–458.
- [17] Amos Tversky and Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics* 106, 4 (1991), 1039–1061.
- [18] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5, 4 (1992), 297–323.
- [19] Wan-Ching Wu, Diane Kelly, and Kun Huang. 2012. User evaluation of query quality. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 215–224.