# Chapter 5
# Back to the Fundamentals: Extend the Rational Assumptions


**Abstract** In this chapter, we revisit the fundamental formal models of IR and associated simplified assumptions, with the goal of exploring and introducing actionable directions toward which the assumptions can be extended to at least partially cover the triggers and characteristics of bounded rationality. To this end, we first categorize different types of explicit and implicit assumptions into three groups, pre-search, within-search, and post-search, and discuss their conflicts with empirical findings on bounded rationality. Within each group, we discuss possible ways to extend and revise existing rational assumptions, as a key preparation for enhancing formal user models and IR evaluation techniques. When explaining the methods for extending rational assumptions, we also discuss related boundaries and explain the implications for user modeling and evaluation and how these potential boundaries are related to IIR-specific factors.

## 5.1 Introduction

The main goal of our book is to develop a behavioral economics framework that can characterize the role of human biases and boundedly rational decisions, especially in the context of search interaction and user-centered system evaluation. Also, we hope that the knowledge shared in our work can motivate students and future researchers to broadly explore critical, understudied research problems and hidden research paths that would enhance bounded rational or bias-aware user modeling and evaluation. In this chapter, we revisit the fundamental formal models of IR and associated simplified assumptions, with the goal of exploring and introducing actionable directions toward which the assumptions can be extended to at least partially cover the triggers and characteristics of bounded rationality. Our analysis is built upon the in-depth reviews on formal models (e.g., click models, user models of offline metrics, formal models of search sessions) and human biases offered in Chaps. 2 and 4, respectively. This chapter also takes a step forward from the identified gaps and preliminary framework introduced in Chap. 3 by discussing ways to extend rational assumptions and the components of ideal models (e.g., static costs and



J. Liu, *A Behavioral Economics Approach to Interactive Information Retrieval*, The Information Retrieval Series 48, https://doi.org/10.1007/978-3-031-23229-9_5

rewards, optimization functions, unbiased judgments) based on relevant concepts, theories, and empirical evidences prepared in previous chapters.

Although researchers could keep adding new parameters, representations, and components to accommodate the impacts of various cognitive and perceptual factors in user models, the increasingly complex models may not be practically applicable, especially for model training and testing purposes. Also, as discussed in Chap. 4, different human biases may operate within their respective boundaries and limits and may involve significant individual differences in actual behavioral impacts. Therefore, when explaining possible approaches to extending rational assumptions, we will also discuss related boundaries and explain the implications for user modeling and evaluation and how these potential boundaries are related to IIR-specific factors, such as dimensions of search tasks, labels from document judgments (e.g., query-document relevance, or *qrel*), as well as the characteristics of individual searchers (Liu, 2021).

Specifically, this chapter will categorize different types of explicit and implicit assumptions into three groups, pre-search, within-search and post-search, and discuss their conflicts with empirical findings on bounded rationality (e.g., Azzopardi, 2021; Kahneman, 2003; Simon, 1955; Thaler, 2016). Most of the existing formal models and assumptions are proposed to characterize and simulate the activities during search, especially in ad hoc retrieval scenarios, so that each search iteration and query-based retrieval evaluation be analyzed and evaluated individually. However, as indicated in Chap. 4, there are factors associated with bounded rationality that could affect people's in situ preferences, expectations, and retrospective evaluations in pre-search estimation and post-search stages as well.

Based on the identified gaps and conflicts, we will discuss possible ways to extend and revise existing rational assumptions, as a key preparation for enhancing formal user models and IR evaluation tools and methods. This chapter will be built upon the gaps and three main problems introduced in Chap. 3 and discuss more details regarding each category or phase of search modeling and the implications of research progresses on human biases and bounded rationality for updating rational assumptions. We believe that extending and revising existing assumptions based on rich theoretical and empirical basis would be an appropriate initial step toward building an actionable research agenda on bias-aware IR modeling and implementing next-generation intelligent search systems that can mitigate the negative effects of human biases.

## 5.2 Pre-search Stage

In most formal models of search and implicit assumptions underpinning evaluation metrics, factors emerging in *pre-search* stage, such as existing beliefs, initial preferences, pre-search expectations, and motivating tasks, are not represented or examined. Although there are offline evaluation metrics that include individual characteristics in underlying user models (e.g., patient and impatient users in

rank-biased precision measure; Moffat & Zobel, 2008), there are still a wide range of user features and contextual factors, especially the ones involving human biases and bounded rationality, that are not considered in user simulation and system evaluation. The implicit assumption behind this general model setup is that users' search behavior and strategies of search evaluation are not affected by the factors beyond topics, queries (and associated search intents), and characteristics of retrieved documents (e.g., relevance, rank position).

A straightforward approach to revising the general assumption and enhancing existing user models is incorporating representations of the key pre-search factors and their associations with search interactions into models and metrics. For instance, instead of starting with no prior preferences or expectations, different users may have different in situ search expectations and search strategies due to their varying past experiences under similar *cases*, especially in situations where the solution space of current task is complex and uncertain. According to case-based decision theory (CBDT) (Gilboa & Schmeidler, 1995), the extent to which past search experience and actions affect current behaviors of decision makers depends on the perceived *similarity* of the past case(s) to the current task. In addition, this basic setup of CBDT also naturally connects to the principle of satisficing and aspiration levels in boundedly rational decision-making processes (cf. Schwartz et al., 2002; Simon, 1955).

From the reference-dependence perspective introduced in Chap. 4 (Kahneman, 2003; Tversky & Kahneman, 1991), the effects of past similar cases on current decision-making strategies and thresholds in judgments can also be framed as a type of reference effects. Under this circumstance, a representation of *pre-search reference* in user-related assumptions may need to crystallize one case or a set of multiple similar and recent cases that are *mentally accessible* to the decision maker at the moment. The case, in this context, can be considered as a motivating task that happens or is assigned to a person within a particular problematic situation. Different task facets and characteristics of problematic situations (e.g., available social and technical support, urgency of the problem) may have varying impacts or weights in similarity assessment. For instance, a past task (and the associated information search experience) of learning Python data analytics may be considered as similar to the current task of studying Python text analysis. However, a past task of learning a deep learning package for completing a self-designed project and the current task of learning deep learning functions for preparing a computer science final exam might be considered as separate cases with low level of similarity, due to the difference in underlying motivations and requirements in information seeking and use. In addition, people's judgments on case similarity may also be affected by users' familiarity with the involved topics and domains (Liu et al., 2019; White et al., 2009). Higher levels of knowledge and familiarity on involved tasks, topics, or cases may increase the accuracy of similarity estimation and enable users to bring truly relevant past experiences into current decision-making scenarios.

Figure 5.1 summarizes the structure of pre-search user preferences and expectations under the CBDT framework. The pre-search factors are affected by the cases, actions, and outcomes a user experienced before, and their respective weights in

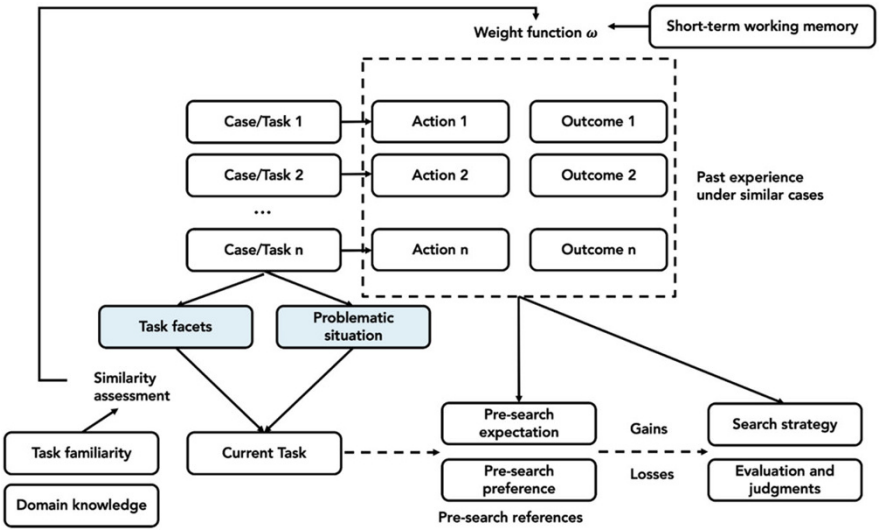


**Fig. 5.1** Pre-search factors and search interaction: a CBDT perspective

current search strategies are affected by the implicit similarity assessment. These pre-search factors, from reference-dependence perspective, could serve as initial references based on which users evaluate their current information gains and search efforts. For example, based on past search experience on similar factual tasks, a user who is working on the task of "finding the Asian supermarket near me" may expect to see the most relevant results ranked on the top of the first SERP and would be dissatisfied or perceive a loss if this was not the case. In particular, from the expectation disconfirmation perspective (Oliver, 1980; Venkatesh & Goyal, 2010), people may have initial expectations and preferences regarding multiple aspects of search interactions, such as quality of retrieved results, effectiveness of search recommendations, and interface layouts. The ways in which these expectations are confirmed or disconfirmed may affect users' in situ search tactics, especially query reformulation types, and the thresholds for evaluating retrieved documents. For instance, a previous frustrating search experience under similar task (e.g., poor search results, irrelevant recommendations) may lower the user's expectation regarding document quality under their current similar task, which may result in a relatively lower threshold for relevance judgment and overestimation of document relevance.

As discussed above, estimating pre-search preferences and expectations from a traditional reference-dependence or CBDT perspective can enrich the pre-search component of user models and may facilitate subsequent search behavior predictions. At application level, however, achieving this representation and conducting model training would be challenging, as it would require sufficient information and knowledge about the individual users and their past relevant experiences beforehand. It might also be difficult to infer or simulate these pre-search factors merely based on

previous search logs or document features. In addition, regarding possible perceptual biases, we need to take into consideration the gap between actual set of similar cases and mentally accessible set of similar cases at the moment. This gap can be traced back to even broader problems of investigating perception-outcome differences (under the influence of in situ relative gains and losses) and examining the limits of individuals' divergent limits in working memory. We will discuss more details regarding these problems and their implications at the *within-search* stage. Overall, although it is challenging to represent and estimate the impacts of all components discussed above, it would be helpful to utilize the framework illustrated in Fig. 5.1 to locate the progresses and limitations of our current models on related research problems and identify potential research themes for future efforts.

Compared to modeling the full structure of pre-search factors presented in Fig. 5.1 and estimating all initial references, it might be more feasible in most cases for researchers to estimate people's initial beliefs and preferences from a *confirmation bias* perspective (cf. Nickerson, 1998). Specifically, for instance, although it would be difficult to infer all accessible past cases, researchers might be able to infer users' preferences over different subtopics, opinions, and sentiments based on their past search logs and initial couple of queries and the associated search interactions with the SERPs and social media contents (e.g., examination and clicking, dwell time on content pages that covering certain opinions) (Knobloch-Westerwick et al., 2015; Rieger et al., 2021; Workman, 2018). Retrieved results that disconfirm these initial opinions and beliefs may receive less attention and underestimated relevance score (even though they may actually be topically relevant to the queries).

Thus, when developing user models for simulation and evaluation purposes, it would be reasonable to assume that individual users have one or multiple initial beliefs over a set of subtopics, opinions, and sentiments before the search session starts, and the search result snippets and documents that confirm the belief(s) would receive more attention from users and might also reinforce users' existing beliefs and biases. From the *loss aversion* perspective (cf. Tversky & Kahneman, 1991), examining and accepting the results that confirm existing beliefs and expectations could be perceived of a gain or at least an avoidance of possible loss, as the user would not need to give up the existing beliefs that cost previous cognitive resources to establish. At the implementation level, the initial beliefs and preferences can be considered as variables affecting browsing patterns and relevance judgments, apart from several widely examined factors, such as queries, rank positions, and externally labelled relevance scores. Identifying the implicit initial preferences and beliefs waiting to be confirmed can help researchers better predict users' clicking and evaluation behaviors and design effective low-cost search interventions for mitigating the potential negative impacts of confirmation bias and algorithmically debiasing relevance and credibility judgments (Draws et al., 2021; Rieger et al., 2021).

Apart from pre-search references, users' behaviors are also affected by human biases and heuristics that operate *within* search sessions. In the following section, we will discuss the ways in which we could extend the components of assumptions

regarding people's actions and decision-making during information search processes.

## 5.3 Within-Search Stage

Compared to pre-search stage, within-search stage is more complicated as it involves multiple aspects of ongoing search interactions and changing user experience. Meanwhile, however, researchers can collect more diverse signals based on which user models and evaluation metrics can be constructed. In this section, we discuss a series of widely discussed intuitive assumptions applied in a broad range of formal user models and explain the ways in which we can (and should) revise and extend them to better predict real-world user behaviors and explain why they search and evaluate in such ways.

The first set of assumptions is related to the *Problem 1* discussed in the Chap. 3. When modeling users and search interactions, building pre-defined rules and assumptions based on actual *costs* and *rewards* tend to be a natural starting point for developing user models and evaluating system performances in a simulated environment. In simulation-based experiments, cost and reward measures are usually linked to user behavior and relevance-based scores, respectively. Based on these measures, researchers often assume the following:

1) Users' behaviors and implicit optimization goals are defined based on the actual experienced costs and rewards during search processes.
2) The costs associated with different actions (e.g., query formulation, search result snippet examination, clicks, dwell time on content pages) and the *relevance-based* reward functions remain the same across different queries, topics, and task types.

Based upon these two main assumptions, researchers can model user behavior and optimize retrieval algorithms on the same ground across different search states and problematic situations. In IR experiments, the two assumptions and their similar variants about costs and rewards have been widely applied in various formal models of search interactions (e.g., Moffat et al., 2012; Zhang et al., 2017; Zhang & Zhai, 2016), including the models that integrate economic theories into search cost modeling (e.g., Azzopardi, 2011, 2014). These assumptions largely simplify the process of estimating costs and rewards associated with different components of searching. They also allow researchers to turn the complex problem of improving search interactions and experiences into straightforward, mathematically solvable optimization problems that involve minimizing behavior-based costs and maximizing rewards and utilities measured by relevance or other judgment labels. In addition, from replicability and reproducibility perspective, assuming fixed connections between action types and costs also facilitates more flexible reuse and replication of user models and IR evaluation experiments. With these assumptions as the basis, many of the potential challenges related to changes of task nature, individual

differences, and within-session cognitive variations in IR evaluations (cf. Gäde et al., 2021; Liu, 2022) could be at least temporarily bypassed in standardized experiments. In sum, before we bring in the behavioral economics perspective for extending the assumptions, we believe that it is almost equally important to acknowledge the value and contribution of the simplified assumptions to the field of IR and computing in general.

As discussed in Chap. 4, when users perceive and evaluate costs and rewards, their perceptions and search decisions are usually developed based on *gains* and *losses* relative to certain reference points, rather than the actual absolute values. This reference dependence perspective casts doubt on the fundamentals of a broad range of formal models and user-oriented evaluation techniques. When the reference points change before or during search sessions, it would lead to the variations in perceived costs and rewards associated with the same type of actions. In addition, depending on the nature of perceived changes (i.e., as gains or losses), the same size of changes in search behavior and result quality across queries and sessions may have different impacts on subsequent search interactions and retrospective evaluations within search sessions (i.e., *loss aversion*, Tversky & Kahneman, 1992). Perceived losses, such as increased dwell time or search actions, less relevant search result snippets or documents, lower readability of retrieved documents, and higher difficulty in formulating effective search queries, could generate larger impacts on users' following search tactics and levels of satisfaction, compared to the same or similar sizes of perceived search gains.

Therefore, to extend the original cost-reward analytical framework from a bounded rationality perspective, researchers need to identify the reference points in effect and compute the in situ perceived gains and losses relative to the reference points. Previous behavioral economics research introduced in Chap. 4 on related topics (e.g., reference dependence and prospect theory, confirmation bias, anchoring bias) have demonstrated that people's decision-making under uncertainty could be influenced by varying types of potential references that emerge at varying stages and are associated with different internal and external factors (Caputo, 2014; Gneezy et al., 2017; Kahneman, 2003; Nickerson, 1998; Tversky & Kahneman, 1991). In the context of information seeking and retrieval, different dimensions of search interactions may have different reference levels and are associated with divergent contextual factors, such as different task facets, user characteristics, as well as in situ search dynamics. Also, the co-exist references may also interact with each other and jointly affect users' in situ search evaluations.

Based on the discussion above, the users' references in search can be written as:

$$R = f_R(\omega_1 r_1, \omega_1 r_2, \ldots \omega_n r_n) \tag{5.1}$$

where $R$ represents the *integrated reference point* for a certain dimension of current search session (e.g., cost of query reformulation and SERP browsing, gain from ranked result list). The integrated reference point is formulated based on a variety of active potential reference points that may have different weights in the user's search decision-making and evaluation. For instance, if a user is at *m-th* query

segment, the user's reference point in terms of the cost of browsing may be affected by the pre-search expectations, beliefs, and preferences. These pre-search references may emerge from past search experience under similar motivating tasks or cases or other people's search interactions that the user observed. In addition to pre-search references, according to the studies on *anchoring bias* (e.g., Tversky & Kahneman, 1974; Caputo, 2014), the user's reference point could also be significantly affected by the initially encountered information objects (e.g., top ranked results presented on the SERPs under first or second queries). The content and quality of the first set of examined documents may heavily influence the user's understanding and threshold of document relevance (Scholer et al., 2013) and thereby affect the perceived gains and losses in following search iterations. The weight distribution on different original reference points, $W = \{\omega_1, \omega_2, \omega_3 \ldots \omega_n\}$, may vary significantly across different dimensions of search interactions and thus may need to be estimated separately. In addition, changes in reference points and their associated weights may also be associated with the transitions of local information seeking intentions and task states (Jansen et al., 2007; Liu et al., 2020). Under different intentions, users often search and evaluate documents differently and may also be affected by distinct reference points.

Based on the identified reference points, the perceived gains or losses for each dimension can be written as:

$$C = \sum_{i=1}^{m} f_c(c_i, R_{c_i}) * \mid c_i - R_{c_i} \mid \tag{5.2}$$

$$Re = \sum_{i=1}^{n} f_{Re}(Re_i, R_{Re_i}) * \mid Re_i - R_{Re_i} \mid \tag{5.3}$$

$$U = u(C, Re) \tag{5.4}$$

where $C$ and $Re$ measure the total *perceived gains* and *losses* in terms of search cost and search reward, respectively. Both *Cost* and *Reward* are multidimensional search components and could be deconstructed into $m$ and $n$ dimensions, respectively. $R$ represents the in situ reference point corresponding to each specific dimension. $U$ refers to the overall perceived utility, which is a function of $C$ and $Re$. Depending on the nature of perceived changes (as gains or losses), the corresponding weights $f_c(c_i, R_{c_i})$ and $f_{Re}(Re_i, R_{Re_i})$ may vary. Users' search tactics and evaluation are more sensitive to perceived losses than to gains. Also, the dimensions where relative losses are perceived are more likely to attract users' attentions and thus receive higher weights in decisions. Figure 5.2 summarizes the process of gain- and loss-based search decision-making in sessions.

As presented, users' current search interactions and outcomes could be evaluated based upon a diverse set of potential reference points. The perceived losses relative to the integrated reference point, such as lower levels of relevance and usefulness of the retrieved SERP and increased dwell time on retrieved pages, may lead to significant changes in following query reformulation behaviors (e.g., formulating a

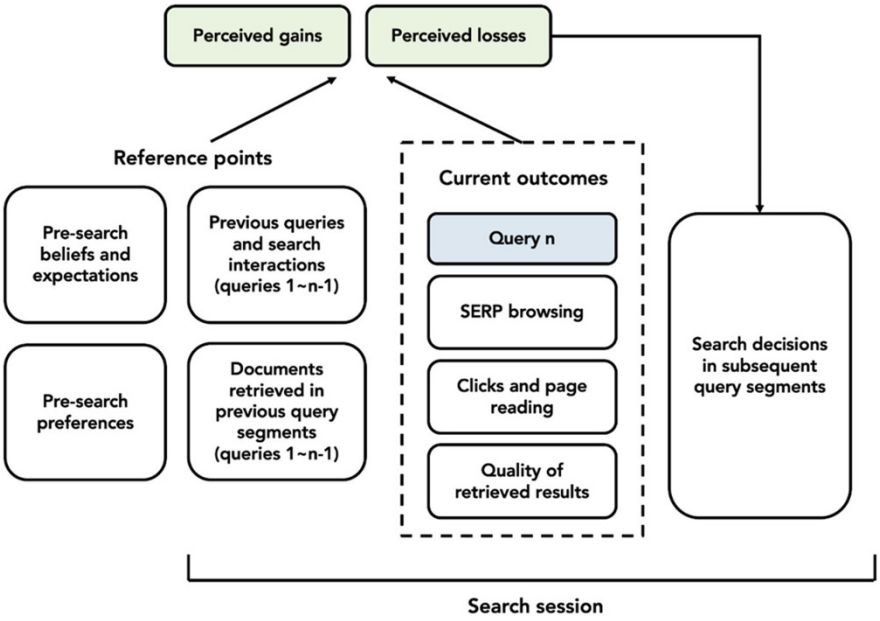


**Fig. 5.2** Gain- and loss-based search decision-making process

new exploratory query, instead of slightly adjusting current query) and in situ search expectations. These changes of search outcomes and experiences in ongoing sessions may also lead to minor or major changes to the reference points in the user's mind, especially for subsequent query segments under similar search intentions.

Apart from the explicit, observable aspects of search behaviors and costs, many of the implicit components of search introduced in Chap. 2 may also be affected by reference dependence effects. Regarding clicking behavior, the perceived *attractiveness* and *probability of examination* on retrieved documents may also be affected by the user's previous experience (e.g., similar documents or subtopics encountered on similar rank positions). Previous examinations of search result snippets, clicking on documents, and click depth on SERPs as potential reference points may affect the thresholds of attractiveness and examination decisions for the documents encountered in following search iterations. When users examine result snippets on current SERP, the search result snippet that have a clearly lower quality than the previously encountered ones (e.g., lower perceived relevance or readability, higher level of ambiguity) may receive significantly less attention from the user. However, if the users start the session with encountering a set of poor-quality search results and documents, the relatively low thresholds or reference points may increase the attractiveness and probability of examination on subsequently retrieved documents that have an intermediate level of quality and are ranked at similar rank positions.

Note that some of the relative, reference-dependence aspects may be partially captured and characterized by some of the existing click models (e.g., graph-based

session click models that take document- and query-based edges into account, cf. Lin et al., 2021; click model that includes users' click and examination preferences, cf. Xing et al., 2013). However, it would still be useful to explicitly investigate the relative changes in search result features and incorporate gain- and loss-based parameters into the estimation of the attractiveness, examination probability, and in situ preferences on clicking. Our multidimensional reference-dependence approach, which also considers the impacts of other related biases and heuristics (e.g., loss aversion, confirmation bias, anchoring bias), could also better explain the individual differences in click actions and extend existing personalized click models (PCMs) that seek to integrate user factors into click prediction and relevance estimation algorithms (e.g., Cheng & Cantú-Paz, 2010; Shen et al., 2012).

With respect to search browsing, apart from the explicit observable dimensions, such as dwell time on different pages and SERP components, scrolling patterns, and eye movements, users' implicit *cost budgets* in evaluation (cf. Zhang et al., 2017), if any, may also be affected by previously established reference points. Thus, the existing fixed cost budget setup or assumption in evaluation metrics could be extended by including the reference points extracted from or simulated based on empirical evidences. For instance, a lower in situ search expectation or threshold of relevance may result in higher tolerance of irrelevant documents and extended browsing sessions. Although these actual interactions can lead to linear increases of search costs as it is assumed in classic cost-reward models, due to the influence of relatively low references, users' perceived costs may increase slowly (e.g., as characterized by a logarithmic function). As a result, the perceived cost and subject cost budget may systematically deviate from the actual or simulated cost budgets. On the contrary, users may perceive a relative quick accumulation of costs under a high reference point (e.g., high-quality SERPs and documents encountered in previous query segments and sessions). Once the perceived costs hit the cost budget in mind, users may become increasingly sensitive to the relative changes in search due to the effect of loss aversion. Apart from quantifiable references, users' cost budgets may also vary across different search intentions and task states. For example, under exploratory search states, users may be more open to examine more search results and click deeper results on SERPs. In contrast, under factual known-item searches, users may have a very limited cost budgets in mind and expect to see the correct answers being ranked on the top positions of the SERPs.

Regarding search result evaluation, similar to other dimensions of search sessions, users' perceived gains and rewards obtained from each clicked relevant document are not fixed. Also, the thresholds of relevance and usefulness judgment may not be static or predefined as it is often implicitly assumed in user models, underpinning a variety of offline evaluation metrics. Instead, the in situ gains and underlying thresholds of relevance and usefulness evaluation could be affected by threshold priming effects (cf. Scholer et al., 2013) and are related to the document evaluation experience under previous queries or other similar search tasks. The threshold priming effects may also be moderated by other user characteristics, such as topic and task familiarity, domain knowledge, and search skills, in real-life search scenarios. Compared to topical relevance, document usefulness tends to be

more subjective and diverse as documents and results could be useful in different ways during search sessions. From a behavioral economics perspective, estimating and simulating the reference points of usefulness judgment might be more challenging and involve a broader range of situational factors, such as search intentions, task progresses, distractions on SERPs, and information encountering and serendipity events (e.g., André et al., 2009; Mao et al., 2017; Mitsui et al., 2017; Rahman & Wilson, 2015).

Related to the idea and assumptions regarding cost budget and evaluation thresholds, users' *stopping rate* and utility *discounting factor* in browsing and evaluation (e.g., Chapelle et al., 2009; Zhang et al., 2017) may also be gain- and loss-based and be affected by both pre-search and in situ references. Specifically, a higher perceived gain (e.g., increased number of relevant documents ranked on top positions; reduced amount of dwell time needed before collecting useful information from clicked documents) may result in lower stopping rate and discounting rate in current SERP browsing. However, if an in situ search loss is encountered (e.g., dropping in search result quality, increased dwell time), the user's stopping rate in following rank positions may increase quickly and result in early search stopping or even query abandonment behavior. Thus, in the context of interactive search sessions, the assumptions of fixed or rank-based stopping rate and discounting factor by ranks could be extended by including reference dependence features or parameters and connecting to previously encountered search results and search costs within the same session. For instance, following the stopping rate function adapted from cascade model and expected reciprocal rank (ERR) measure (Chapelle et al., 2009), a bounded rational stopping rate function can be written as follows:

$$R_i = \frac{2^{r_i - Re_{i-1}} - 1}{2^{r_{max}}} \tag{5.5}$$

$$P_j = \prod_{i=1}^{j} (1 - R_i) R_j \tag{5.6}$$

where $r_i$ measures the graded relevance score of the current document $i$. $Re_{i\ -\ 1}$ refers to the total perceived reward or accumulated gains up to the rank position $i - 1$. During SERP browsing and document examination, a document satisfies the user with the probability $R_i$. $P_j$ represents the probability that the user is satisfied and stops at document $j$. In this simple initial setup, we change the absolute graded relevance score to the relative gain-based score as the basis for calculating the probability that the user is satisfied with the current document $i$. Note that in ERR measure, it is assumed that users will stop searching once they find the one document that satisfies their information needs. However, in exploratory searches, people may not stop at just one satisfactory document and be open to broader explorations and deeper clicking behavior. Under this circumstance, it is critical to extend evaluation metrics and consider the systematic impacts of perceived gains and losses at different levels. Incorporating potential reference points and biases into the estimation and simulation of stopping rate and discounting factor may be a viable approach to

extending user-centered IR metrics, especially in the context of whole-session IR evaluation.

Regarding within-SERP examination and evaluation, the interdependence between different search results may also affect the user's overall perception of the SERP and the associated examination behavior. In existing research, *Document interdependence* has been examined in a series of offline IR evaluation studies as a factor or constraint in document relevance estimation (e.g., Montazeralghaem et al., 2018; Radlinski et al., 2009; Zhai et al., 2015). Taking *decoy effect* into consideration would offer a new perspective for examining different aspects and forms of document interdependence (Kahneman, 2003; Tversky & Kahneman, 1985; Wedell & Pettibone, 1996; Wu et al., 2020). Specifically, for example, under a short exploratory search query, the SERP may present documents involving different opinions, information sources, and subtopics, for which users may not have prior preferences. However, a potential decoy search result associated with a subtopic may increase the user's examination and clickthrough probability on results with relatively higher quality or level of informativeness under the same subtopic. Therefore, in information searches that involve a diverse set of results, users' in situ preferences over different types of contents may be shaped by implicit decoy options.

This decoy effect may also interact with the existing impacts of search snippet features and rank position biases and could be included and represented in both click models for attractiveness and examination probability estimation and offline evaluation metrics. Note that the decoy effect could happen at multiple dimensions, such as relevance and informativeness of search result snippets, document readability (e.g., Collins-Thompson et al., 2011), and perceived credibility of the search results (e.g., Hilligoss & Rieh, 2008), which may cause different reactions from users.

$$\alpha_i = \frac{d(sr_1, sr_2 \ldots sr_n)}{d(i,\ \ t)} \tag{5.7}$$

$$d(sr_1, sr_2 \ldots sr_n) = \sum_{r=1}^{n} w_r \mid Sr_{ri} - Sr_{rt} \mid \tag{5.8}$$

$$P\left(C_t = 1 | d_{ipresent} = 1\right) - P\left(C_t = 1 | d_{ipresent} = 0\right) = \alpha_i + \delta * max\,(i, t) \tag{5.9}$$

At the implementation level, researchers may need to start with identifying potential decoy options at multiple levels among the retrieved results and estimate possible decoy effects based on the distance between decoy options and the search results associated with similar subtopics, opinions, and information sources. This distance measure should consider both the superficial-level distance (i.e., distance in rank positions on SERPs) and the distance or difference in search result quality and presentation. As shown in Formula (5.7), $\alpha_i$ measures the potential decoy effect generated by the decoy option ranked at the rank position $i$. In Formula (5.8), the function $d$ measures the aggregated differences between the regular/target result and decoy result, which plays an essential role in triggering potential decoy effect. As presented in Formula (5.9), the decoy effect can be represented by the probability

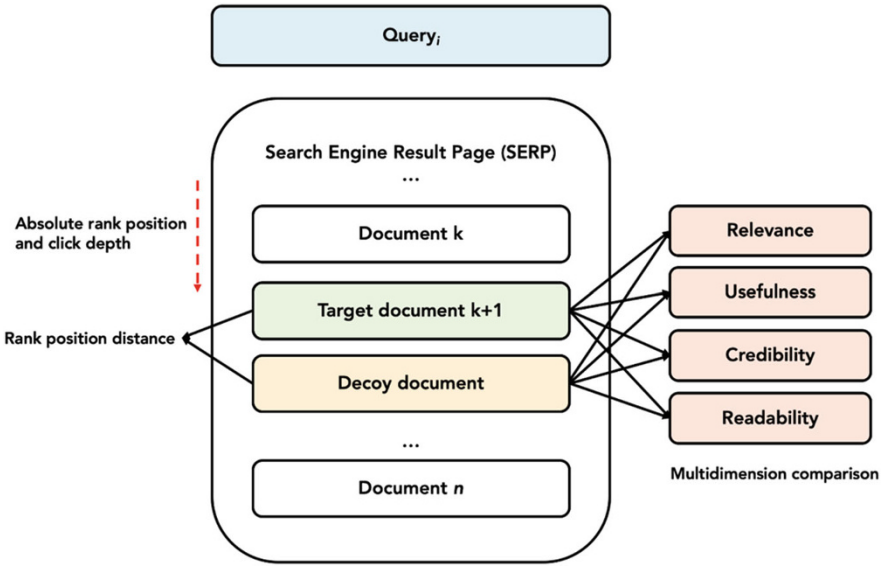


**Fig. 5.3** The structure of decoy effect in IR evaluation

difference in user's clickthrough rate on the target document (with similar subtopics, opinions, themes, or sentiments) ranked at position $t$ between two scenarios: (1) the decoy result is present, and (2) the decoy result is absent. $sr_n$ represents different dimensions of search result snippets and corresponding documents, which may cause perceived quality difference and trigger decoy effect in user judgments.

$\delta$ measures the potential discounts on decoy impacts due to rank positions. Based on previous research on the effects of *rank position bias* (e.g., Agarwal et al., 2019; Wang et al., 2018), we assume that with all other conditions remaining the same, the decoy effect is more likely to occur when the decoy and target results are ranked higher as they would be more likely to receive the user's attention during browsing in top rank positions. In contrast, if both results are ranked in relatively low position, users may be less sensitive to the potential differences between the two results and thus are less likely to be influenced by the decoy result in evaluation. Also, given the potential effect of *diminishing sensitivity* (Trautmann & Kuilen, 2012; Tversky & Kahneman, 1991; Wakker & Tversky, 1993), users are less likely to completely change their click preferences at lower ranked results, especially in situations where the corresponding level of click depth is already way beyond their expected cost or in situ cost budget.

Figure 5.3 illustrates the structure of decoy effect discussed above and highlights the factors that may influence the effect size of decoy results, such as absolute rank position (where the target search result and decoy result are ranked on the SERP), rank position distance between target result and decoy result, and, more importantly, the multidimensional differences *perceived* by users. Note that different dimensions involved in comparison may have significantly different levels of saliency to users in

SERP browsing. For instance, the presentation quality of search result snippet (e.g., readability, informativeness, color, and length) may be more salient than some other implicit factors (e.g., overall usefulness and credibility of the documents) that are difficult to judge at the first glance, especially for users who are not quite familiar with the topic and domain involved. In addition, users may be more sensitive to the dimensions where a clear relative loss is perceived in the comparison between decoy and target results, such as a significant drop in the quality of search result snippets and decreased relevance or quality of images (i.e., loss aversion bias). As a result, different dimensions perceived in comparison may contribute differently to the final behavioral impacts caused by the decoy result.

As shown in the formulas, when the potential decoy result and regular search results are close to each other in rank position but significantly differ from each other in other dimensions, the decoy effect is more likely to occur as the contrast between the decoy and regular results would be more noticeable to the user. Investigating decoy effect in SERP evaluation can extend the implicit assumption of document or search result independence in a wide range of click models and evaluation metrics, especially in terms of estimating the probability of examination, predicting clicks, and modeling users' perceptions of SERP utility. Studying decoy effect in IR can also pave the path toward a new layer of document interdependence studies. The structure of decoy effect presented in Fig. 5.3 can serve as a theoretical framework for characterizing the behavioral impacts of decoy results in information seeking and retrieval and may also inform the design of controlled user studies (e.g., SERP-based crowdsourcing evaluation study), focusing on decoy options in SERPs across different information-intensive decision-making scenarios.

The following section will move on to the *post-search* stage and discuss the possible extensions of user model assumptions in light of the knowledge regarding bounded rationality, especially in terms of whole-session retrospective evaluation.

## 5.4 Post-search Stage

In addition to modeling user behaviors and system performance during search sessions, how users retrospectively evaluate the performance of search systems and their overall search experiences, especially in whole-session IR, is also one of the central themes in IR research. In post-search retrospective evaluation, researchers usually evaluate search system performances based on the average value and total value of each measure or dimension (e.g., sum dwell time on SERPs, average number of clicks and pages visited, average precision, reciprocal rank and nDCG scores of SERPs) (e.g., Chen et al., 2017; Liu et al., 2012). In offline Cranfield experiments, a series of average-value-based evaluation metrics have also been proposed to evaluate search systems in a set of diverse queries and topics (Voorhees, 2001). Differing from common average-value- and total-value-based metrics, *session-based* DCG (sDCG) metric takes query order into consideration when evaluating search sessions and discounts relevant search results retrieved from later queries

within a session (Järvelin et al., 2008). Compared to other retrospective evaluation metrics, sDCG takes a step forward by simulating the discounted weights of each search iteration based on the linear query order in the session.

Estimating the weight distribution of different query segments is a critical aspect of whole-session search evaluation. In light of the empirical findings on *peak-end rule* and *recency effects* (Kahneman, 2003; Redelmeier et al., 2003), researchers can adjust the average-value-based metrics and linear query weight functions and assign higher weights on peak experience and most recent search iterations. As discussed in related behavioral economic experiments, people's in situ peak experience and most recent experience could generate relatively higher impacts on the retrospective remembered utility. This *remembered utility*, rather than actual *experienced utility*, serves as the basis for people's intuitive judgments and subsequent decision-making, especially under the operations of System 1 (Kahneman, 2003). Researchers also found that when retrospectively evaluating an extended episode, people are not sensitive to the actual duration of the entire session (i.e., *duration neglect*, Fredrickson & Kahneman, 1993; Hands & Avons, 2001). In addition, the initial queries in a session may also be associated with relatively higher weights compared to other following queries as they may serve the anchoring points in evaluation (i.e., anchoring bias, cf. Nickerson, 1998). Also, at the application level, extracting the anchoring references from initial queries may also facilitate early prediction of whole-session search effectiveness and thereby offer opportunities for proactive search intervention and recommendation (e.g., Koskela et al., 2018; Mitsui et al., 2018; Shah, 2018), especially in cases where the current user is on a potentially poor-performing or high-loss search path predicted by the search system.

The user biases and heuristics discussed above can help extend the implicit assumptions regarding the extent to which average-value-based and total-value-based metrics can approximate whole-session experience. Specifically, knowledge regarding these biases highlight several key moments in search sessions and can inform the design of a more behaviorally realistic weight distribution for connecting query-level evaluation to whole-session-based evaluation of system performance. In constructing session-level evaluation models and metrics, researchers should consider assigning relatively higher weights to these key points and examining their respective impacts on user evaluation under different tasks and search scenarios. In addition, given the impact of *duration neglect*, researchers may not be able to rely on session duration time as a main predictor in estimating a user's *remembered* whole-session experience.

Figure 5.4 illustrates and contrasts the key factors involved in the whole-session evaluations characterized by classical rational approach and boundedly rational approach, respectively. This figure highlights the difference in weight distributions of different factors and query positions and explains the role of each related human bias and heuristics in different aspects of the session evaluation process. Future researchers can use the framework presented here as a guideline in variable and model design for predicting whole-session search experience (e.g., levels of search satisfaction and cognitive loads) and evaluating the performance of bounded rational prediction models against that of the classic rational models as baselines. In addition

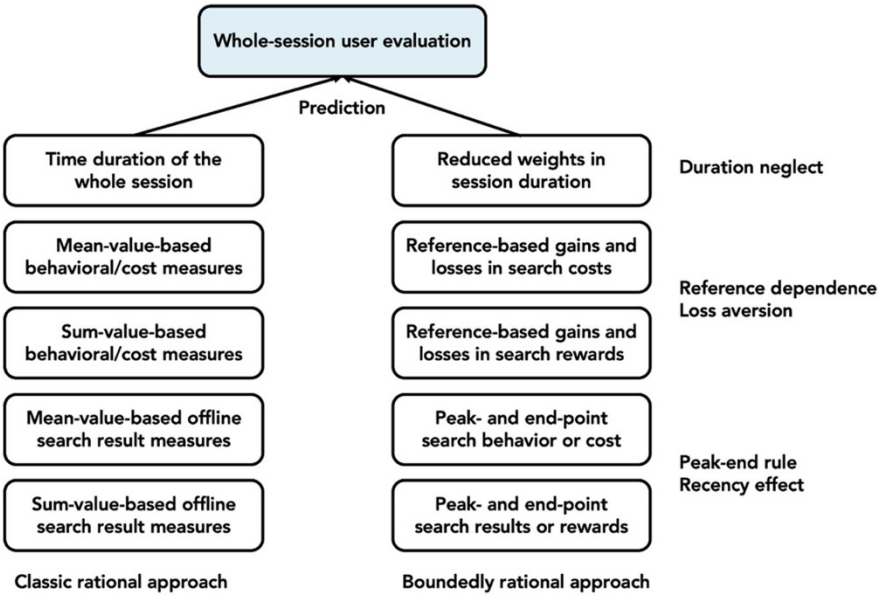


**Fig. 5.4** Boundedly rational whole-session evaluation

to the findings from behavioral experiments introduced in the Chap. 4, in the following chapter, we will introduce empirical evidences from IR research that confirms the impacts of peak-end rule on retrospective evaluations and discuss how we can leverage the knowledge regarding related biases in better answering critical research questions in information seeking and retrieval communities.

It is also worth noting that peak-end effect usually has its boundaries and limits across different dimensions of decision-making problems (Langer et al., 2005; Schneider et al., 2011). The extent to which peak-end rule affects whole-session evaluation may vary significantly across different facets of search interactions and thus may have different weights and limits in affecting overall search interaction experience and judgments of the search system, such as levels of user engagement, task and cognitive loads, as well as perceived levels of success and satisfaction. Therefore, the actual weight distributions across different search dimensions may need be estimated individually based on corresponding search interaction signals. Given existing research on the divergent effects of peak-end rule across varying behaviors and decision-making sessions, IR researchers may also find similar variations in weight distributions and effect sizes of different query moments (particularly the initial, peak, and end queries that may trigger boundedly rational search decisions) in retrospective session evaluation across different behavioral measures, offline evaluation metrics, as well as types of search tasks.

In addition to the findings on peak-end rule and recency effects, behavioral economics research also casts doubt on the fundamental assumption regarding users' intents of *always pursuing maximized utility*. In contrast to the assumption

of optimization, behavioral science and decision-making researchers found that people's decision-making often follows the principle of *satisficing* and aims for satisfactory or adequate options in all accessible options, rather than the optimal solution predicted by normative models (Simon, 1955). The satisficing strategy can by triggered by different reasons. For instance, people may find it difficult to deliberately compare and calculate the possible utility from different options as they are often restricted by limited time, cognitive resources, and computing capability. Also, although the optimal option can be identified through rational mathematical analysis and simulations, it may not always be actually accessible to users or decision makers due to the individual differences in varying aspects (e.g., domain knowledge, information search skills, existing beliefs, and cognitive biases) and situational limits. Moreover, the threshold of satisficing in each specific decision-making scenario may be linked to a prior reference point, perceived gains or losses that a person has in mind before current search iteration or session, and thus may vary significantly across different individuals and problematic situations (Kahneman, 2003; Schwartz et al., 2002).

In IR evaluation, the satisficing strategy in decision-making may be partially captured by some of the existing offline metrics, assuming that users will stop searches once they find the *first relevant document* that satisfies their information needs (e.g., reciprocal rank, expected reciprocal rank). However, current relevance-based metrics may not be able to fully characterize the nature of in situ satisficing moment or threshold. The threshold may also involve other features and dimensions of search results, such as usefulness for completing the task, topical diversity, interestingness, as well as unexpectedness (or information serendipity). Also, the evaluation of relevance itself may also be affected by other related human biases. For instance, the first couple of documents encountered in a session are often considered as relevant ones and may significantly affect users' evaluations of subsequently retrieved documents. Thus, focusing on available relevance labels in test collections only may lead to biased search ranking and inaccurate estimation of levels of user satisfaction. Consequently, it might be difficult for researchers and system designers to adaptively optimize ranking and search recommendation algorithms toward the goal of satisficing in real-world information seeking and search settings.

## 5.5 Summary

This chapter brings together the insights regarding formal IR models and human bounded rationality discussed in previous chapters and discusses the ways in which we can extend several widely adopted (explicitly or implicitly) assumptions in user modeling and make them more behaviorally realistic. Specifically, based on the nature of human biases and heuristics discussed in Chap. 4, we explain their possible impacts on users' search interaction and evaluation at pre-search, within-search, and post-search retrospective evaluation stages and suggest revised forms of existing

assumptions that can incorporate the knowledge about human biases into formal user models and system evaluation metrics.

Nevertheless, it is worth noting that we as a research community still has a long way to go before achieving reliable, intelligent bias-aware user modeling and personalized recommendation. Extending rational assumptions and discussing potential research problems is the first step toward reaching the ultimate goal of bias-aware IR that addresses both human biases and algorithmic biases. Beyond this initial step, researchers also need to develop and further enhance bias-aware user models and incorporate the knowledge regarding bounded rationality into search and ranking algorithms, evaluation metrics, and standardized IR evaluation experiments (Liu, 2022). In addition to the open research problems discussed above, methodologically, how to accurately capture and estimate the real impacts of human biases and heuristics in naturalistic complex information seeking and retrieval settings, rather than the (over)simplified decision-making experiments employed in a series of classic behavioral economics research, also remains an open challenge to the research community. Also, given the high cost of user studies in both lab and naturalistic environments, we also need to develop and evaluate the methods through which we can reliably reuse the study materials (e.g., study design and instruments, collected data, statistical and machine learning models built and tested) and replicate the completed experiments in different settings. Researchers may need to both explore existing user study designs and techniques (e.g., Kelly, 2009; Kelly & Sugimoto, 2013; Liu & Shah, 2019) and also employ additional signals and design new study settings where boundedly rational decision-making processes could be better observed and identified.

In the following chapters, we will discuss the recent research progress on modeling and simulating human biases in information seeking, retrieval, and recommendation and identify more specific research questions, directions, as well as challenges that may require more attention and research efforts from future studies.

## References

Agarwal, A., Zaitsev, I., Wang, X., Li, C., Najork, M., & Joachims, T. (2019). Estimating position bias without intrusive interventions. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 474–482). ACM. https://doi.org/10.1145/3289600.3291017

André, P., Teevan, J., & Dumais, S. T. (2009). From x-rays to silly putty via Uranus: serendipity and its role in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2033–2036). ACM. https://doi.org/10.1145/1518701.1519009

Azzopardi, L. (2011). The economics in interactive information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 15–24). ACM. https://doi.org/10.1145/2009916.2009923

Azzopardi, L. (2014). Modelling interaction with economic models of search. In *Proceedings of the 37th ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 3–12). ACM. https://doi.org/10.1145/2600428.2609574

Azzopardi, L. (2021). Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 27–37). ACM. https://doi.org/10.1145/3406522.3446023

Caputo, A. (2014). Relevant information, personality traits and anchoring effect. *International Journal of Management and Decision Making, 13*(1), 62–76.

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 621–630). ACM. https://doi.org/10.1145/1645953.1646033

Chen, Y., Zhou, K., Liu, Y., Zhang, M., & Ma, S. (2017). Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 15–24). ACM. https://doi.org/10.1145/3077136.3080804

Cheng, H., & Cantú-Paz, E. (2010). Personalized click prediction in sponsored search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 351–360). ACM. https://doi.org/10.1145/1718487.1718531

Collins-Thompson, K., Bennett, P. N., White, R. W., De La Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 403–412). ACM. https://doi.org/10.1145/2063576.2063639

Draws, T., Rieger, A., Inel, O., Gadiraju, U., & Tintarev, N. (2021). A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 9, pp. 48–59). https://ojs.aaai.org/index.php/HCOMP/article/view/18939

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*(1), 45–55. https://doi.org/10.1037/0022-3514.65.1.45

Gäde, M., Koolen, M., Hall, M., Bogers, T., & Petras, V. (2021). A manifesto on resource re-use in interactive information retrieval. In *Proceedings of the 2021 International ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 141–149). ACM. https://doi.org/10.1145/3406522.3446056

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics, 110*(3), 605–639. https://doi.org/10.2307/2946694

Gneezy, U., Goette, L., Sprenger, C., & Zimmermann, F. (2017). The limits of expectations-based reference dependence. *Journal of the European Economic Association, 15*(4), 861–876. https://doi.org/10.1093/jeea/jvw020

Hands, D. S., & Avons, S. E. (2001). Recency and duration neglect in subjective assessment of television picture quality. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 15*(6), 639–657. https://doi.org/10.1002/acp.731

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management, 44*(4), 1467–1484. https://doi.org/10.1016/j.ipm.2007.10.001

Jansen, B. J., Booth, D. L., & Spink, A. (2007). Determining the user intent of web search engine queries. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 1149–1150). https://doi.org/10.1145/1242572.1242739

Järvelin, K., Price, S. L., Delcambre, L. M., & Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query IR sessions. In *European Conference on Information Retrieval* (pp. 4–15). Springer.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1–2), 1–224. https://doi.org/10.1561/1500000012

Kelly, D., & Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology, 64*(4), 745–770. https://doi.org/10.1002/asi.22799

Knobloch-Westerwick, S., Johnson, B. K., & Westerwick, A. (2015). Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. *Journal of Computer-Mediated Communication, 20*(2), 171–187. https://doi.org/10.1111/jcc4.12105

Koskela, M., Luukkonen, P., Ruotsalo, T., Sjöberg, M., & Floréen, P. (2018). Proactive information retrieval by capturing search intent from primary task context. *ACM Transactions on Interactive Intelligent Systems (TiiS), 8*(3), 1–25. https://doi.org/10.1145/3150975

Langer, T., Sarin, R., & Weber, M. (2005). The retrospective evaluation of payment sequences: Duration neglect and peak-and-end effects. *Journal of Economic Behavior & Organization, 58*(1), 157–175. https://doi.org/10.1016/j.jebo.2004.01.001

Lin, J., Liu, W., Dai, X., Zhang, W., Li, S., Tang, R., He, X., Hao, J., & Yu, Y. (2021). A graph-enhanced click model for web search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1259–1268). ACM. https://doi.org/10.1145/3404835.3462895

Liu, J. (2021). Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management, 58*(3), 102522. https://doi.org/10.1016/j.ipm.2021.102522

Liu, J. (2022). Toward Cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing & Management, 59*(5), 103007. https://doi.org/10.1016/j.ipm.2022.103007

Liu, C., Belkin, N. J., & Cole, M. J. (2012). Personalization of search results using interaction behaviors in search sessions. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 205–214). ACM. https://doi.org/10.1145/2348283.2348314

Liu, J., Mitsui, M., Belkin, N. J., & Shah, C. (2019). Task, information seeking intentions, and user behavior: Toward a multi-level understanding of Web search. In *Proceedings of the 2019 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 123–132). ACM. https://doi.org/10.1145/3295750.3298922

Liu, J., Sarkar, S., & Shah, C. (2020). Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 193–202). ACM. https://doi.org/10.1145/3343413.3377976

Liu, J., & Shah, C. (2019). Interactive IR user study design, evaluation, and reporting. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 11*(2), i–93. https://doi.org/10.2200/S00923ED1V01Y201905ICR067

Mao, J., Liu, Y., Luan, H., Zhang, M., Ma, S., Luo, H., & Zhang, Y. (2017). Understanding and predicting usefulness judgment in web search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1169–1172). ACM. https://doi.org/10.1145/3077136.3080750

Mitsui, M., Liu, J., Belkin, N. J., & Shah, C. (2017). Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1121–1124). ACM. https://doi.org/10.1145/3077136.3080737

Mitsui, M., Liu, J., & Shah, C. (2018). How much is too much? Whole session vs. first query behaviors in task type prediction. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1141–1144). ACM. https://doi.org/10.1145/3209978.3210105

Moffat, A., Scholer, F., & Thomas, P. (2012). Models and metrics: IR evaluation as a user process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium* (pp. 47–54). https://doi.org/10.1145/2407085.2407092

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS), 27*(1), 1–27. https://doi.org/10.1145/1416950.1416952

Montazeralghaem, A., Zamani, H., & Shakery, A. (2018). Theoretical analysis of interdependent constraints in pseudo-relevance feedback. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1249–1252). ACM. https://doi.org/10.1145/3209978.3210156

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research, 17*(4), 460–469. https://doi.org/10.1177/002224378001700405

Radlinski, F., Bennett, P. N., Carterette, B., & Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. In *ACM SIGIR Forum* (Vol. 43, No. 2, pp. 46–52). ACM. https://doi.org/10.1145/1670564.1670572

Rahman, A., & Wilson, M. L. (2015). Exploring opportunities to facilitate serendipity in search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 939–942). ACM. https://doi.org/10.1145/2766462.2767783

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: A randomized trial. *Pain, 104*(1–2), 187–194. https://doi.org/10.1016/S0304-3959(03)00003-4

Rieger, A., Draws, T., Theune, M., & Tintarev, N. (2021). This item might reinforce your opinion: Obfuscation and labeling of search results to mitigate confirmation bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (pp. 189–199). ACM. https://doi.org/10.1145/3465336.3475101

Schneider, S., Stone, A. A., Schwartz, J. E., & Broderick, J. E. (2011). Peak and end effects in patients' daily recall of pain and fatigue: A within-subjects analysis. *The Journal of Pain, 12*(2), 228–235. https://doi.org/10.1016/j.jpain.2010.07.001

Scholer, F., Kelly, D., Wu, W. C., Lee, H. S., & Webber, W. (2013). The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 623–632). ACM. https://doi.org/10.1145/2484028.2484090

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology, 83*(5), 1178. https://doi.org/10.1037/0022-3514.83.5.1178

Shah, C. (2018). Information fostering-being proactive with information seeking and retrieval: Perspective paper. In *Proceedings of the 2018 International ACM SIGIR Conference on Human Information Interaction & Retrieval* (pp. 62–71). ACM. https://doi.org/10.1145/3176349.3176389

Shen, S., Hu, B., Chen, W., & Yang, Q. (2012). Personalized click model through collaborative filtering. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 323–332). ACM. https://doi.org/10.1145/2124295.2124336

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118. https://doi.org/10.2307/1884852

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review, 106*(7), 1577–1600. https://doi.org/10.1257/aer.106.7.1577

Trautmann, S. T., & van de Kuilen, G. (2012). Prospect theory or construal level theory? Diminishing sensitivity vs. psychological distance in risky decisions. *Acta Psychologica, 139*(1), 254–260. https://doi.org/10.1016/j.actpsy.2011.08.006

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Behavioral Decision Making* (pp. 25–41). Springer.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039–1061. https://doi.org/10.2307/2937956

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297–323. https://doi.org/10.1007/BF00122574

Venkatesh, V., & Goyal, S. (2010). Expectation disconfirmation and technology adoption: Polynomial modeling and response surface analysis. *MIS Quarterly, 34*(2), 281–303. https://doi.org/10.2307/20721428

Voorhees, E. M. (2001). The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 355–370). Springer.

Wakker, P., & Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty, 7*(2), 147–175. https://doi.org/10.1007/BF01065812

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., & Najork, M. (2018). Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 610–618). ACM. https://doi.org/10.1145/3159652.3159732

Wedell, D. H., & Pettibone, J. C. (1996). Using judgments to understand decoy effects in choice. *Organizational Behavior and Human Decision Processes, 67*(3), 326–344. https://doi.org/10.1006/obhd.1996.0083

White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM International Conference on Web Search and Data Mining* (pp. 132–141). ACM. https://doi.org/10.1145/1498759.1498819

Workman, M. (2018). An empirical study of social media exchanges about a controversial topic: Confirmation bias and participant characteristics. *The Journal of Social Media in Society, 7*(1), 381–400.

Wu, L., Liu, P., Chen, X., Hu, W., Fan, X., & Chen, Y. (2020). Decoy effect in food appearance, traceability, and price: Case of consumer preference for pork hindquarters. *Journal of Behavioral and Experimental Economics, 87*, 101553. https://doi.org/10.1016/j.socec.2020.101553

Xing, Q., Liu, Y., Nie, J. Y., Zhang, M., Ma, S., & Zhang, K. (2013). Incorporating user preferences into click models. In *Proceedings of the 22nd ACM international Conference on Information & Knowledge Management* (pp. 1301–1310). ACM. https://doi.org/10.1145/2505515.2505704

Zhai, C., Cohen, W. W., & Lafferty, J. (2015). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *ACM SIGIR Forum* (Vol. 49, No. 1, pp. 2–9). ACM. https://doi.org/10.1145/2795403.2795405

Zhang, Y., Liu, X., & Zhai, C. (2017). Information retrieval evaluation as search simulation: A general formal framework for IR evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 193–200). ACM. https://doi.org/10.1145/3121050.3121070

Zhang, Y., & Zhai, C. (2016). A sequential decision formulation of the interface card model for interactive IR. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 85–94). ACM. https://doi.org/10.1145/2911451.2911543