Evaluating Web Search with a Bejeweled Player Model

Fan Zhang
DCST, Tsinghua University
Beijing, China
frankyzf94@gmail.com

Yiqun Liu*
DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Xin Li
DCST, Tsinghua University
Beijing, China
x-l08@163.com

Min Zhang
DCST, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Yinghui Xu
Alibaba Group
Hangzhou, China
renji.xyh@taobao.com

Shaoping Ma
DCST, Tsinghua University
Beijing, China
msp@tsinghua.edu.cn

ABSTRACT

The design of a Web search evaluation metric is closely related with how the user's interaction process is modeled. Each behavioral model results in a different metric used to evaluate search performance. In these models and the user behavior assumptions behind them, when a user ends a search session is one of the prime concerns because it is highly related to both benefit and cost estimation. Existing metric design usually adopts some simplified criteria to decide the stopping time point: (1) upper limit for benefit (e.g. RR, AP); (2) upper limit for cost (e.g. Precision@N, DCG@N). However, in many practical search sessions (e.g. exploratory search), the stopping criterion is more complex than the simplified case. Analyzing benefit and cost of actual users' search sessions, we find that the stopping criteria vary with search tasks and are usually combination effects of both benefit and cost factors. Inspired by a popular computer game named Bejeweled, we propose a Bejeweled Player Model (BPM) to simulate users' search interaction processes and evaluate their search performances. In the BPM, a user stops when he/she either has found sufficient useful information or has no more patience to continue. Given this assumption, a new evaluation framework based on upper limits (either fixed or changeable as search proceeds) for both benefit and cost is proposed. We show how to derive a new metric from the framework and demonstrate that it can be adopted to revise traditional metrics like Discounted Cumulative Gain (DCG), Expected Reciprocal Rank (ERR) and Average Precision (AP). To show effectiveness of the proposed framework, we compare it with a number of existing metrics in terms of correlation between user satisfaction and the metrics based on a dataset that collects users' explicit satisfaction feedbacks and assessors' relevance judgements. Experiment results show that the framework is better correlated with user satisfaction feedbacks.

*Corresponding author

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3077136.3080841

CCS CONCEPTS

•**Information systems → Evaluation of retrieval results;**

KEYWORDS

Benefit and Cost; Evaluation Metrics; User Model

1 INTRODUCTION

System-oriented tests and user-oriented studies are currently two complementary approaches to Web search evaluation. A system-oriented test, which is known as Cranfield approach [15], typically develops a set of relevance judgments to compare the quality of ranked lists returned by different systems in response to a fixed set of queries. On the contrary, a user-oriented study makes use of actual user behaviors during interactive retrieval sessions to measure effectiveness of systems. For instance, *A/B testing* and *interleaving* method [24] are widely used by commercial search engines.

One of the main advantages of user-oriented studies is that they are designed to reflect users' opinions in the realistic search process. However, they are more costly and harder to reproduce than system-oriented tests. On the other hand, although a system-oriented test is inexpensive and reproducible, it cannot capture search activities of users. To tackle this challenge, most Web search evaluation metrics have been built on top of different user models. In these models, when user ends a search session is one of the prime concerns because it is highly related to both benefit and cost estimation.

Benefit, also referred to as *gain* or *utility* in some researches, has been discussed and introduced in a variety of ways. For example, utility is deemed to be associated with relevance [17], in the sense that one can receive benefit or gain from relevant documents. In terms of relevance, what constitutes it is subject to much interpretation [28]. *Cost* is considered as temporal efforts or cognitive efforts in processing, reading and understanding documents in many related works [2, 34, 38, 39]. In this paper, we use the terms *benefit* and *cost*, although many other equivalent terms are also used in existing researches.

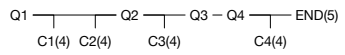
Regarding benefit and cost, underlying user models of existing metrics usually adopt some simplified criteria to decide the stopping time point. For instance, user model of *Reciprocal Rank* (RR) assumes that a user will stop once he/she finds a perfect document. That is, when users stop only depends on when they get the benefit that they expected. We refer to this kind of stopping criterion as

upper limit for benefit. In contrast, *Precision@N* measures the percentage of relevant documents in top-N results, which means that users will scan a ranked list from top to bottom and stop at the N-th document. This kind of stopping criterion is called **upper limit for cost**. The user models of other metrics like *Discounted Cumulative Gain* (DCG) [23], *Expected Reciprocal Rank* (ERR) [12] and *Average Precision* (AP) are more complex, since user variety and stopping probability distribution at different ranked results are considered. Nevertheless, the stopping criteria of these user models still focus on only one aspect of upper limits for benefit or cost, as we show in Section 3.2. However, in many practical search sessions (e.g. exploratory search), the stopping criterion should be usually combination effects of both benefit and cost factors.

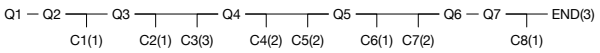
Figure 1 shows two search sessions collected from an experimental user study in [27]. Figure 1(a) shows a session where the user was seeking information on *ice-breaking games*. In this session, the user issued four queries and clicked four results. The usefulness feedbacks provided by the user indicated that the user judged all the clicked results to be "highly useful". The user was also "highly satisfied" with the session according to the satisfaction feedback. Therefore, we suppose that the user ended this session because he/she has received enough benefits from clicked results. It was upper limit for benefit that affected the stopping criterion. On the contrary, in Figure 1(b), we see a session where the user was exploring different aspects of a topic *"Fixed Gear Bicycle"*. In this session, the user issued seven queries and clicked eight results, while only one result is thought to be "fairly useful". The last result he clicked was "useless" and then he ended the session in spite that he was "somewhat satisfied". So we assume that the user stopped with no more patience. It seems to be the upper limit for cost that stopped the user.

Comparing the sessions, we find that the stopping criterion for a search session may be either upper limit for benefit or cost in difference circumstances. To take this a bit complex criterion into consideration, inspired by a popular computer game named *Bejeweled*, we propose a *Bejeweled Player Model* (BPM) to simulate users' search interaction processes and evaluate their search performances. In *Action Mode* on *Bejeweled*[1], the game starts with the timer bar at the bottom half full, which will start to decrease every second. The player must match gems to add more seconds, with bigger moves getting more time. The player will advance to the next level when the bar is full. However, if the bar completely empties, the player lose the game. When no more moves can be made, the game reshuffles the gems. Overall, the stopping criterion of the game may be the bar is either empty (i.e. *Game Over*) or full (i.e. *Level Up*). Similar to the game, frustration and satisfaction are two final states of Web search. Many related works [11, 19, 21] have worked on predicting these two states and tell the difference between them. However, to the best of our knowledge, there is no research that incorporates the difference between frustration and satisfaction into the design of evaluation framework. Inspired by the game, we assume that frustration often means that users have invested too much cost and run out their patience (i.e. the bar is empty) while satisfaction is usually due to the fulfillment of their benefits (i.e. the bar is full), so we propose the BPM to describe the stopping criterion for Web search. To be emphasized, unlike the game where costs and scores are both represented as time in the same bar, we consider users' benefits and costs separately in the BPM. That is to say, benefits and costs are accumulated in two bars and there is an upper limit for each bar. The stopping criterion for a session is either the benefit bar is full or the cost bar is full.

[1]http://bejeweled.wikia.com/wiki/Action_mode



(a) An example session of *upper limit for benefit*



(b) An example session of *upper limit for cost*

Figure 1: Sessions showing different stopping criteria. Q: issuing a query; C: clicking a result. The number in brackets after a click is its usefulness level (4 point scale where 4 means the most useful) and that after the *END* is the session's satisfaction level (5 point scale where 5 means the most satisfied).

As described above, in the BPM, a user stops when he/she either has found sufficient useful information or has no more patience to continue. Given this assumption, we propose a new evaluation framework based on upper limits (either fixed or evolving as search session proceeds) for both benefit and cost. To apply this framework to Web search evaluation, we show how to derive metrics from it. As mentioned previously, the stopping criteria of user models behind some metrics such as DCG, ERR and AP focus on only one aspect of upper limits for benefit or cost. Therefore, we demonstrate that these metrics can be derived from the framework considering one-sided case of the stopping criterion. Finally, to show effectiveness of the proposed framework, we compare it with a number of existing metrics in terms of the correlation between user satisfaction and the metrics.

In summary, we make the following contributions in our work:

- We introduce a *Bejeweled Player Model* to simulate users' search interaction processes and explain the stopping criterion for search sessions.
- Based on the BPM, we propose a new unified framework for Web search evaluation and demonstrate that some existing metrics can be derived from the framework considering one-sided case of the stopping criterion.
- Based on a dataset that collect users' explicit satisfaction feedbacks and assessors' relevance judgements, we show effectiveness of our proposed framework by comparing it with a number of existing metrics in terms of correlation between user satisfaction and the metrics.

The remainder of this paper is organized as follows. In Section 2, we introduce our proposed evaluation framework based on the BPM. Section 3 shows how to instantiate a metric from the framework and how it can be adapted to existing metrics. Then we show effectiveness of the framework by comparing it with existing metrics in terms of correlation between user satisfaction and the metrics in

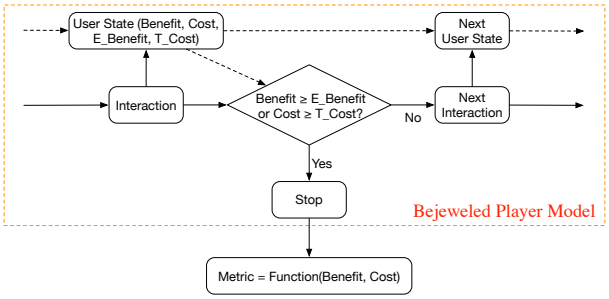


Figure 2: Evaluation framework based on the BPM

Section 4. We review related researches in Section 5 and conclude in Section 6.

2 EVALUATION FRAMEWORK

Figure 2 shows our proposed evaluation framework based on the *Bejeweled Player Model* (BPM). In a session described by the BPM, a user will interact with the system to satisfy her information need. At each round of *interaction*, the user will pay some costs and obtain some benefits simultaneously. As a result, she changes the *user state* with benefit and cost. Benefits and costs are accumulated with interactions and we use *Benefit* and *Cost* to denote them. The BPM supposes that the benefits that the user expects to obtain and the costs that she is willing to pay are limited, which are denoted as *Expected Benefit* (*E_Benefit*) and *Tolerated Cost* (*T_Cost*). Based on upper limits for both benefit and cost, a user stops when either she has found sufficient useful information (reach upper limit for benefit, i.e. $Benefit \geq E\_Benefit$) or she has no more patience to continue searching (reach upper limit for cost, i.e. $Cost \geq T\_Cost$). To be emphasized, *E_Benefit* and *T_Cost* may change with interactions, so we incorporate them into the *user state* with *Benefit* and *Cost* as well.

Note that in the BPM introduced above, we mainly focus on upper limits for both benefit and cost that determine when the user stops, rather than specific interactions. We believe that the BPM is an intrinsic user model which simulates users' search interaction processes and explains the stopping criterion for search sessions. We assume that user satisfaction can be represented by benefit and cost factors. Therefore, based on the BPM, we propose an evaluation framework by combining it with a metric function of *Benefit* and *Cost* at the end of the session. The function is defined to evaluate users' search performances. Given the framework, *interactions*, *Benefit* and *Cost*, *E_Benefit* and *T_Cost*, as well as *metric function* are important components for evaluation. Therefore, before we apply this conceptual framework to Web search evaluation, we should talk more about these components.

2.1 Interactions

In our proposed framework, user behaviors are represented as a sequence of interactions. These interactions are associated with user models. For example, the interactions are examining results one-by-one for user models behind most rank-based retrieval metrics such as DCG [23], ERR [12] and AP. While for click model-based information retrieval metrics [14, 37], the interactions would be examining snippets or click results. Sakai and Dou [33] proposed U-measure to evaluate Web search based on the concept *trailtexts*, where the interactions would be handling *trailtexts* of course. Though we do not focus on how to define interactions in this paper, we believe that interactions that are closer to user behaviors lead to more effective metrics.

2.2 Benefit and Cost

As suggested by Azzopardi et al. [5], benefits and costs are associated with interactions. They provide a summary of different benefits and costs for various interactions. For most traditional metrics that take examining results as interactions, the *benefit* of interaction is considered to be associated with relevance. Binary and graded relevance are two most common ways to model it.

As for the *cost*, most metrics assume that the cost of processing each document is the same. Recently, *cost* has been considered from a variety of angles. For instance, Smucker and Clarke [34] use the time spent by the user as a proxy for *cost* and propose Time Biased Gain (TBG). In addition, the *cost* involved in processing a document in terms of readability and understandability has been explicitly included in other measures (e.g., [2, 38, 39]).

In this paper, we will use $benefit_k$ and $cost_k$ to denote the benefit and the cost of the k-th *interaction*, while *Benefit* and *Cost* denote the benefits and costs accumulated on interactions. Unlike metrics that accumulate benefits or costs with a discount function for different interactions (e.g., DCG [23], TBG [34], U-measure [33] for benefit and ERR [12] for cost), we accumulate them with no discounts. Inspired by Carterette [10] and Moffat et al. [29], we argue that the discount function is the result of user variety and stopping probability distribution at different ranked results. This will be discussed further in Section 3.2.

2.3 E_Benefit and T_Cost

In this paper, we focus on upper limits for benefit and cost that determine when the user stops. *E_Benefit* denotes the benefit that the user expects to obtain (upper limit for benefit), which should be in the same unit with *Benefit*. For example, Moffat et al. [29] use a parameter T to denote the target number of relevant documents the user wishes to identify. *T_Cost* denotes the cost that the user is willing to pay (upper limit for cost), which should be in the same unit with *Cost*. For example, in TBG [34], *T_Cost* can be expressed as the time that the user is willing to use for search.

Motivated by [29], we argue that *E_Benefit* and *T_Cost* should be user-spercific and task-spercific. That is, for different users or different tasks, the values of *E_Benefit* and *T_Cost* are different. For instance, Broder [9] groups Web queries into informational, navigational and transactional categories. Given that informational queries often require more information than navigational or transactional queries to satisfy the information need, we assume that the value of *E_Benefit* will be larger for informational queries. On the other hand, for users who have more patience to search, the value of *T_Cost* will be larger as well, just like the persistence parameter p associated with RBP [30]. In [6], Bailey et al. reveal that searchers display substantial individual variation in the numbers of documents and queries they anticipate needing to issue, and there

are underlying significant differences in these numbers in line with increasing task complexity levels. Therefore, we can consider different $E\_Benefit$ and T_Cost for different users and different tasks to design more sensitive evaluation metrics.

For further thinking, we suggest that *E_Benefit* and *T_Cost* should also be dynamic. Fuhr [20] proposes the Interactive Probability Ranking Principle (iPRP), an extension to the well known Probability Ranking Principle [32]. When developing the iPRP, one of the main requirements is allowing for the information need to change through the course of interaction. Motivated by this point, we assume that *E_Benefit* and *T_Cost* will also involve with interactions, thus incorporated into the *user state* with *Benefit* and *Cost*.

It is also important to note that interactions are affected by *user states*. In [4], Azzopardi et al. examine three theories of Information Seeking and Retrieval. They enumerate a list of hypotheses about search behavior, and show that these theories make similar predictions. Their work indicates that searchers will change their search behavior based on their context. Nevertheless, in this paper, we focus on independent interactions that users scan ranked results one-by-one from top to bottom.

2.4 Metric Function

This function is defined to instantiate a metric from the framework and measure user satisfaction when the user stops. In previous metrics, some focus on *Benefit* (e.g. DCG), some focus on *Cost* (e.g. ERR) while others focus on *Average Utility* which means *Benefit* devided by *Cost* (e.g. AP). Therefore, we assume that the metric should be a function of *Benefit* and *Cost*. In this paper, however, we do not discuss the form what the *metric function* should be. We just compare three choices mentioned above, i.e. *Benefit*, *reciprocal Cost*, and *Benefit* devided by *Cost*.

3 METRICS

In Section 2, we proposed an evaluation framework based on the BPM. Now we will show how to instantiate a metric from the framework and how it can be adapted to existing metrics.

3.1 Metric Based on The BPM

For simplicity, in this paper we take *interactions* as scaning down ranked results one-by-one before the user stops. This simple interaction process is usually regarded as a cascade assumption [18] and accepted by many existing user behavior models. Given the k-th interaction (i.e. examining the k-th result), its benefit and cost will be $benefit_k$ and $cost_k$. In [10], Carterette compares different utility accumulation models that describe how a user accumulates utility in the course of browsing. For simplicity, here we assume *Benefit* or *Cost* accumulated with interactions to be the sum of $benefit_k$ or $cost_k$. Then in the k-th *user state* after the k-th interaction, *Benefit* and *Cost* can be represented as follow:

$$Benefit_k = \sum_{i=1}^{k} benefit_i, \quad Cost_k = \sum_{i=1}^{k} cost_i \tag{1}$$

As mentioned in Section 2.3, *E_Benefit* and *T_Cost* will change with interactions, thus represented as:

$$E\_Benefit_k = E\_Benefit_0 + \sum_{i=1}^{k} \Delta E\_Benefit_i \tag{2}$$

$$T\_Cost_k = T\_Cost_0 + \sum_{i=1}^{k} \Delta T\_Cost_i \tag{3}$$

where $E\_Benefit_0$ and T_Cost_0 are the initial values when the user starts searching. The increments, $\Delta E\_Benefit_i$ and ΔT_Cost_i, may depend on all the interactions and user states up to the i-th *user state*.

Considering the probability that the user stops at rank k:

$$P(k) = P(\geq k) - P(\geq k+1) \tag{4}$$

where $P(\geq k)$ denotes the probability that the stopping rank is not less than k. In our proposed framework, we assume that a user stops only when $Benefit \geq E\_Benefit$ or $Cost \geq T\_Cost$. So $P(\geq k)$ can be represented as follow:

$$P(Benefit_i < E\_Benefit_i, Cost_i < T\_Cost_i : i = 1, ..., k-1) \tag{5}$$

Then we can get the metric:

$$M = \sum_{k=1}^{\infty} Function(Benefit_k,\ Cost_k) * P(k) \tag{6}$$

Note that different users may stop at different ranks. Therefore, considering user variety, the metric is represented as the expectation of the metric function. For a system-oriented test, given definitions of $benefit_i$, $cost_i$, $\Delta E\_Benefit_i$, ΔT_Cost_i and the *metric function*, we can derive a specific metric from the framework.

3.2 Existing Metrics

Equation 6 shows how to instantiate a metric from our proposed framework based on the BPM. Given that the stopping criteria of user models behind some metrics such as DCG, ERR and AP focus on only one aspect of upper limits for benefit or cost, we assume that the framework can be applied to derive these metrics when considering their underlying user models. Therefore, in this part, we take DCG as an example to show how it can be derived from the framework.

Note that DCG is based on an assumption that the user scans down the ranked list one-by-one and the cost of processing each document is the same, so we define $cost_i$ as *one unit* for them, thus $Cost_k$ equals to k *units*. As described in [23], for DCG, $benefit_i$ is defined as $2^{rel_i} - 1$, where rel_i is the relevance level of $document_i$. Since DCG assumes that the probability that the lower-ranked documents are examined is smaller, which leads to a discount function $1/log_2(i+1)$, we explain it with different values of *T_Cost* for different users. To be specific, the proportion of users willing to examining at least i results is $1/log_2(i+1)$. Therefore, the probability that T_Cost_k equals to i is:

$$P(T\_Cost_k = i) = 1/log_2(i+1) - 1/log_2(i+2) \tag{7}$$

Note that T_Cost_k is independent of k, which means that *T_Cost* is static for each user and will not change with interactions. However, for DCG, benefit is not limited, which means *E_Benefit* is infinite.

Therefore, the probability that the stopping rank is not less than k can be written as:

$$\begin{aligned} P(\geq k) =& P(Cost_j < T\_Cost_j : j = 1, ..., k-1) \\ =& \sum_{i=1}^{\infty} P(j < i : j = 1, ..., k-1) \cdot P(T\_Cost_j = i) \\ =& \sum_{i=k}^{\infty} \left(\frac{1}{log_2(i+1)} - \frac{1}{log_2(i+2)} \right) = \frac{1}{log_2(k+1)} \end{aligned} \tag{8}$$

According to the Equation 4, the probability that the user stops at rank k equals to $1/log_2(k+1) - 1/log_2(k+2)$.

Since DCG focuses on cumulative gain, we define the *metric function* as:

$$Function(Benefit_k,\ Cost_k) = Benefit_k = \sum_{i=1}^{k} benefit_i \tag{9}$$

Then the form of DCG is:

$$\begin{aligned} DCG =& \sum_{k=1}^{\infty} Function(Benefit_k,\ Cost_k) * P(k) \\ =& \sum_{k=1}^{\infty} P(k) \sum_{i=1}^{k} benefit_i = \sum_{k=1}^{\infty} benefit_k \sum_{i=k}^{\infty} P(i) \\ =& \sum_{k=1}^{\infty} \left(2^{rel_k} - 1\right) \sum_{i=k}^{\infty} \left(\frac{1}{log_2(i+1)} - \frac{1}{log_2(i+2)} \right) \\ =& \sum_{k=1}^{\infty} \frac{2^{rel_k} - 1}{log_2(k+1)} \end{aligned} \tag{10}$$

Considering a truncate depth K, it reduced to the form that introduced in [23]. Note that the transition at line 2 in Equation 10 shows two ways to compute the expectation. $\sum_{i=1}^{k} benefit_i$ is the benefit cumulated from rank 1 to rank k and $P(k)$ is the probability that the user stops at rank k. Therefore, the left side of the transition should be denoted as *expected cumulated benefit*. On the other hand, $benefit_k$ is the benefit at rank k, and $\sum_{i=k}^{\infty} P(i)$ is equivalent to $P(\geq k)$. It is the probability that the stopping rank is not less than k, which also means the probability that the k-th result will be examined. Consequently, the right side of the transition should be denoted as *cumulated expected benefit*.

Similar to DCG, most existing metrics involve summing over the product of a discount function of ranks and a benefit function mapping relevance assessments to numeric utility values, i.e.

$$M = \sum_{k=1}^{K} benefit(rel_k) \cdot discount(k) \tag{11}$$

Based on the example described above, we can clearly see that the $discount(k)$ can be regarded as the probability that the k-th document is examined when a user scans a ranked list from top to bottom. Note that the k-th document being examined means that the user does not stop before rank k. So the discount function is the result of user variety and stopping probability distribution at different ranked results, as we mentioned in Section 2.2.

Here we should state that besides DCG, most other metrics including ERR, AP, RBP, TBG, U-measure etc. can also be derived from the framework given different definitions of $benefit_i$, $cost_i$, $E\_Benefit_k$, T_Cost_k and the *metric function*. Due to space limitations, we will not discuss all of them in detail. Table 1 shows the definitions for them. For simplicity, here we take *interactions* as examining results before the user stops, thus the benefit and cost are calculated based on rank. Although some metrics such as TBG and U-measure focus on time or trailtext rather than rank, we think the simplified cases can explain underlying user models of these metrics from the perspective of benefit and cost to some extent. However, we find that these metrics focus on only one aspect of upper limits for benefit or cost. That is, either *E_Benefit* or *T_Cost* is infinite. On the other hand, these metrics do not consider the situation where *E_Benefit* and *T_Cost* change with interactions. In this paper, we mainly focus on these two points to instantiate metrics from the framework to show its effectiveness.

3.3 Upper Limits for Both Benefit and Cost

In Section 3.2, we show that many existing metrics can be derived from our proposed framework based on the BPM while these metrics focus on only one aspect of upper limits for benefit or cost. However, according to the examples in Figure 1, we find that the stopping criteria vary with search tasks and are usually combinational effects of both benefit and cost factors, which indicates that upper limits for both benefit and cost exist simultaneously. In this section, at first, we consider a simple case where upper limits for both benefit and cost are independent of users and their interactions. In other words, $E\_Benefit_k$ and T_Cost_k are static values, thus denoted as E_B and T_C in this section. Therefore, Equation 5 can be written as:

$$P(\geq k) = P(Benefit_i < E\_B, Cost_i < T\_C : i = 1, ..., k-1) \tag{12}$$

Following a number of metrics like DCG, we adopt a graded $benefit_i$ associated with relevance level of $document_i$ and a $cost_i$ defined as one unit, which means that:

$$benefit_i = 2^{rel_i} - 1, \quad cost_i = 1 \tag{13}$$

Given E_B and T_C are static, we define them as follows:

$$E\_B = \alpha_B * (2^{rel_{max}} - 1), \quad T\_C = \alpha_C * 1 \tag{14}$$

where α_B and α_C are positive parameters for *Expected Benefit* and *Tolerated Cost* and rel_{max} is the maximum relevance level (e.g. $rel_{max} = 3$ if a 4 point scale is used). Approximately, α_B can be regarded as the number of highly relevant documents that a user expects to find, while α_C is the number of documents that the user is willing to examine. Though we do not know the optimal values of α_B and α_C, we suppose that they would be different for different tasks and compare different values of α_B and α_C. As we mentioned before, a *metric function* should be defined to instantiate a metric for evaluation. In this paper, the metrics defined above are called *Static BPM Metrics* and denoted as $SBPM_f(\alpha_B, \alpha_C)$, where α_B and α_C are parameters, while f is the *metric function*. Referring to the existing metrics, we adopt three forms of *metric function*, which are *Benefit* (B), *reciprocal Cost* ($1/C$) and *Average Benefit* (B/C, i.e. *Benefit* devided by *Cost*), and compare effectiveness of them.

Given the values of α_B and α_C, and the form of *metric function*, we can calculate the value of a *Static BPM Metric* for a ranked list where relevance judgements of top-N results are provided. Algorithm 1 shows the calculation process for *Static BPM Metrics*.

Table 1: Definitions of different components for different metrics. See more details in related researches [12, 30, 33, 34] . Note that here we consider TBG and U-measure as offline metrics given their associated parameters.

	$benefit_i$	$cost_i$	$E\_Benefit$	T_Cost	$P(k)$	$metric\ function$
ERR	$P(benefit_i = 1) = \frac{2^{rel_i}-1}{2^{rel_{max}}}$	1	1	∞	$\prod_{j=1}^{k-1}\left(1-\frac{2^{rel_j}-1}{2^{rel_{max}}}\right)\frac{2^{rel_k}-1}{2^{rel_{max}}}$	$1/Cost_k$
AP	rel_i	1	$P(E\_Benefit = j) = 1/m,\ j \in [1, 2, \ldots, m]$	∞	$\frac{1}{m}I(rel_k)$	$Benefit_k/Cost_k$

	$benefit_i$	$cost_i$	$E\_Benefit$	T_Cost	$P(k)$	$metric\ function$
DCG	$2^{rel_i}-1$	1	∞	$P(T\_Cost = j) = \frac{1}{log_2(j+1)} - \frac{1}{log_2(j+2)},\ j \in [1, 2, \ldots, \infty)$	$\frac{1}{log_2(k+1)} - \frac{1}{log_2(k+2)}$	$Benefit_k$
RBP	rel_i	1	∞	$P(T\_Cost = j) = (1-p)p^{j-1},\ j \in [1, 2, \ldots, \infty)$	$(1-p)p^{k-1}$	$Benefit_k * (1-p)$
TBG	$b_{TBG}(i)$	$c_{TBG}(i)$	∞	$P(T\_Cost \le t) = F(t) = 1 - e^{-t\frac{ln2}{h}},\ t \in [0, \infty)$	$\frac{F(\sum_{i=1}^{k} cost_i) - F(\sum_{i=1}^{k-1} cost_i)}{F(\sum_{i=1}^{L} cost_i)}$	$Benefit_k$
U-measure	$gv_l(s_i)$	$\lvert s_i \rvert$	∞	$P(T\_Cost = j) = 1/L,\ j \in [1, 2, \ldots, L]$	$\frac{min\{pos(s_k), L\} - min\{pos(s_{k-1}), L\}}{L}$	$Benefit_k$

Note: $b_{TBG}(i) = I(rel_i)P(C=1|R=1)P(S=1|R=1)$, $c_{TBG}(i) = T_S + T_D(L_i)P(C=1|R=rel_i)$

Algorithm 1 Calculation process for *Static BPM Metrics*.

Require: $\alpha_B, \alpha_C, f(Benefit, Cost); rel_{max}, rel_i: i \in [1, N]$.
Ensure: $SBPM_f(\alpha_B, \alpha_C)$

```
E_B ← α_B * (2^rel_max − 1), T_C ← α_C * 1;
B ← 0, C ← 0, i ← 1
while B < E_B and C < T_C and i < N do
    i ← i + 1
    B ← B + (2^rel_i − 1), C ← C + 1
end while
SBPM_f(α_B, α_C) = f(B, C)
```

Regarding *Static BPM Metrics*, we examine the following research questions based on a test collection (described in Section 4.1) containing user's explicit satisfaction feedbacks and external assessor's relevance judgments:

- **RQ1** Given proper upper limits for benefit and cost and forms of metric function, will *Static BPM Metrics* have a better correlation with user satisfaction feedbacks than existing metrics?
- **RQ2** Are there differences in optimal upper limits for benefit and cost between different taxonomies of queries (e.g. informational queries and navigational queries)?
- **RQ3** Are there differences in optimal forms of *metric function* between different taxonomies of queries?

3.4 Dynamic E_Benefit and T_Cost

As we mentioned before, upper limits for benefit and cost may evolve with the search interaction processes. Therefore, in this section, we will consider a situation where $E\_Benefit_k$ and T_Cost_k are dynamic values depending on interactions. They are described by Equation 2 and Equation 3.

Regarding $E\_Benefit_k$, we assume that when a user finds a relevant document, she may expect to find more relevant documents because she becomes more interested in this query topic and wants to learn more. On the contrary, if a user finds an irrelevant document, the number of relevant documents she expects to find may decrease. Given this assumption, we define a simple linear function for the increment $\Delta E\_Benefit_i$ in Equation 2 as follows:

$$\Delta E\_Benefit_i = h_B * (benefit_i - benefit_{median}) \quad (15)$$

where h_B is a sensitivity parameter for $E\_Benefit$ increment ($h_B > 0$). The larger h_B is, the more likely it is that a user will stop after finding an irrelevant document. As for $benefit_{median}$, it can be written as:

$$benefit_{median} = 2^{rel_{median}} - 1 \quad (16)$$

where rel_{median} is the median relevance level (e.g. $rel_{median} = 1.5$ if a 4 point scale is used).

On the other hand, in terms of T_Cost_k, our hypothesis is that when a user finds a relevant document, she may be willing to examine more documents since the user is more confident to find useful information in the remaining documents. In contrast, if a user finds an irrelevant document, the number of documents she is willing to examine may decrease. Given this assumption, we also define a simple linear function for the increment ΔT_Cost_i in Equation 3 as follows:

$$\Delta T\_Cost_i = h_C * (benefit_i/benefit_{median} - 1) \quad (17)$$

where h_C is a sensitivity parameter for T_Cost increment ($h_C > 0$). Similar to h_B, the larger h_C is, the more likely it is that a user will stop after finding an irrelevant document.

In addition, the initial values of upper limits for benefit and cost are defined in the same way as E_B and T_C:

$$E\_B_0 = \alpha_B * (2^{rel_{max}} - 1), \quad T\_C_0 = \alpha_C * 1 \quad (18)$$

Therefore, compared with *Static BPM Metrics*, the values of *Dynamic BPM Metrics*, which are denoted as $DBPM_f(h_B, h_C, \alpha_B, \alpha_C)$, based on dynamic upper limits for benefit and cost defined above can be calculated by Algorithm 2.

Regarding *Dynamic BPM Metrics*, we want to answer the following research question:

- **RQ4** Do our hypotheses about dynamic upper limits for benefit and cost hold? In other words, given proper values of h_B and h_C, will *Dynamic BPM Metrics* have a better correlation with user satisfaction feedbacks than *Static BPM Metrics*?

4 EXPERIMENTS

Kelly [25] stated that satisfaction can be understood as the fulfillment of a specified desire or goal. Satisfaction is used to reflect users' actual feelings about the system, thus becoming an important criterion in the user-centric evaluation for Web search engines [1, 22]. To show effectiveness of our proposed framework and answer the

Algorithm 2 A calculation process for *Dynamic BPM Metrics*.

Require: $h_B, h_C; \alpha_B, \alpha_C, f(Benefit, Cost)$;
$rel_{median}, rel_{max}, rel_i: i \in [1, N]$.
Ensure: $DBPM_f(h_B, h_C, \alpha_B, \alpha_C)$

```
E_B ← α_B * (2^rel_max − 1), T_C ← α_C * 1;
B ← 0, C ← 0, i ← 1
while B < E_B and C < T_C and i < N do
    i ← i + 1
    B ← B + (2^rel_i − 1), C ← C + 1
    E_B ← E_B + h_B * [(2^rel_i − 1) − (2^rel_median − 1)]
    T_C ← T_C + h_C * [(2^rel_i − 1)/(2^rel_median − 1) − 1]
end while
DBPM_f(h_B, h_C, α_B, α_C) = f(B, C)
```

Table 2: Statistics of the test collection

#tasks	#SERPs	#participants	#sessions
65	300	98	2685

research questions, we compare *Static BPM Metrics* and *Dynamic BPM Metrics* with a number of existing metrics in terms of correlation between user satisfaction and the metrics based on a test collection containing users' explicit satisfaction feedbacks and assessors' relevance judgements. We will briefly introduce the test collection in Section 4.1 and the results of our experiments will be shown and discussed in Section 4.2.

4.1 Test Collection

In our experiment, the test collection is based on experimental user studies conducted in our previous works [13, 26]. In the user studies, each participant was asked to complete 30 search tasks within about an hour. For each task, after understanding corresponding information need, the participant would be guided to pre-designed search result pages (SERPs) where the query and search results are fixed. The participant was asked to examine the results provided on the SERP and end the search session either if the information need was satisfied or he/she was disappointed with the results. Each time they ended a search session, they were required to provide a five point scaled satisfaction feedback to the session where 5 means the most satisfactory and 1 means the least. Then they would be guided to continue to the next search task. In addition, we invited three professional assessors from a commercial search engine company to label four point scaled relevance scores for all query-result pairs in the experiment. The KAPPA coefficient of their annotation is about 0.7, which can be characterized as a substantial agreement. There are 65 tasks in total, which contains 27 informational queries and 38 navigational queries. The specific statistics of the test collection are shown in Table 2.

4.2 Results

Following Chen et al. [13], considering that satisfaction feedback may be quite subjective, we regularize the satisfaction scores by each participant to Z-scores. For each session, given relevance judgements, we can compute the value of different metrics. In this paper, we use DCG@10, RBP-0.8 (0.8 is the value of the persistence parameter), AP and ERR as representatives of existing metrics. Since we assume that effectiveness of metrics based on the BPM may be affected by different taxonomies of queries, we compare different groups of queries. One group contains 27 informational queries in the test collection, while the other group contains 38 navigational queries.

Table 3: Pearson's Correlations between *Satisfaction Feedbacks* and *Existing Metrics*.

	informational queries	navigational queries
DCG@10	0.493	0.321
RBP-0.8	0.490	0.323
AP	0.400	0.326
ERR	0.393	0.313

4.2.1 Static BPM Metrics. In order to examine **RQ1**, **RQ2** and **RQ3**, first we compute Pearson's correlations between satisfaction feedbacks and existing metrics. The correlations are shown in Table 3. For informational queries, we can see that DCG@10 and RBP-0.8 have better correlations with satisfaction than AP and ERR. However, for navigational queries, all these metrics have similar correlations. Therefore, we use DCG@10 as a baseline to be compared with metrics based on the BPM in the following experiments. Then we compute Pearson's correlations between satisfaction feedbacks and *Static BPM Metrics*. Given that we should tune parameters α_B and α_C for *Static BPM Metrics*, to avoid overfitting and unfair comparison with the baseline, we randomly divide the test collection into two halves. The first half is the training set which containing 620 sessions of informational queries and 723 sessions of navigational queries for tuning parameters, while the other half is the test set which containing 620 sessions of informational queries and 722 sessions of navigational queries for comparing different metrics.

First we try different values of parameters α_B and α_C and different *metric functions* for *Static BPM Metrics* on the training set. The results are shown in Table 4. In order to compare two correlation coefficients (rs), we construct a t-statistic to test the significance of the difference between dependent r's [16]. Note that as long as $\alpha_B > \alpha_C = k$, after examining k documents, $Cost_k$ is equal to T_Cost_k while $Benefit_k$ is smaller than $E\_Benefit_k$. Then users will always stop at rank k, which is the same as the case where $\alpha_B = \alpha_C = k$. Therefore, we omit all the results for the cases where $\alpha_B > \alpha_C$. In addition, if the form of *metric function* is *reciprocal Cost* (i.e. $1/Cost$) and $\alpha_B = \alpha_C = k$, the *metrics* will have the same values (i.e. $1/k$) for all sessions. Then the correlations between satisfaction feedbacks and metrics will make no sense, thus omitted as well.

From the results, we can determine the parameters and the metric function. For informational queries, if we define *metric function* as *Benefit*, when $\alpha_B = 5$ and $\alpha_C = 8$ *or* 9, we can get the best correlation (0.520 or 0.521), which is significantly larger than DCG@10 (0.482) on the training set. For navigational queries, if we define *metric function* as *reciprocal Cost*, when $\alpha_B = 1$ and $\alpha_C = 5$, we can get the best correlation (0.365), which is significantly larger than DCG@10 (0.312) on the training set.

Table 4: Pearson's Correlation between *Satisfaction Feedbacks* and *Static BPM Metrics* for informational and navigitional queries. The first column is the form of *metric function*. Numbers in the second column are values of α_B, while numbers in the second row are values of α_C. * indicates the difference of correlation between the *Static BPM Metric* and *DCG@10* is significant at $p < 0.05$ when the *Static BPM Metric* has a better correlation with *Satisfaction*.

$f(B,C)$		*informational queries*										*navigational queries*									
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
$Benefit$	1	0.345	0.338	0.279	0.242	0.170	0.029	-0.033	-0.068	-0.102	-0.122	0.335	0.248	0.166	0.088	-0.019	-0.100	-0.179	-0.198	-0.204	-0.225
	2	-	0.385	0.398	0.413	0.419	0.374	0.312	0.283	0.214	0.100	-	0.309	0.329	0.296	0.257	0.211	0.086	0.049	0.006	-0.032
	3	-	-	0.409	0.437	0.468	0.490	0.469	0.481	0.450	0.416	-	-	0.321	0.324	0.319	0.289	0.235	0.188	0.157	0.102
	4	-	-	-	0.433	0.462	0.490	0.501	0.521	0.510	0.502	-	-	-	0.308	0.322	0.309	0.277	0.254	0.228	0.204
	5	-	-	-	-	0.454	0.477	0.488	**0.520***	**0.521***	0.523	-	-	-	-	0.312	0.305	0.284	0.264	0.253	0.238
	6	-	-	-	-	-	0.469	0.475	0.497	0.502	0.502	-	-	-	-	-	0.299	0.279	0.266	0.258	0.247
	7	-	-	-	-	-	-	0.475	0.484	0.491	0.488	-	-	-	-	-	-	0.277	0.266	0.258	0.245
	8	-	-	-	-	-	-	-	0.483	0.478	0.477	-	-	-	-	-	-	-	0.264	0.256	0.246
	9	-	-	-	-	-	-	-	-	0.475	0.468	-	-	-	-	-	-	-	-	0.254	0.244
	10	-	-	-	-	-	-	-	-	-	0.465	-	-	-	-	-	-	-	-	-	0.242
$\frac{1}{Cost}$	1	-	0.313	0.332	0.343	0.353	0.360	0.365	0.365	0.366	0.366	-	0.351	0.355	0.362	**0.365***	0.364	0.363	0.363	0.363	0.363
	2	-	-	0.305	0.329	0.358	0.379	0.393	0.403	0.410	0.415	-	-	0.227	0.250	0.273	0.285	0.293	0.295	0.293	0.293
	3	-	-	-	0.303	0.326	0.350	0.369	0.385	0.400	0.410	-	-	-	0.182	0.213	0.234	0.251	0.263	0.270	0.275
	4	-	-	-	-	0.288	0.318	0.332	0.349	0.370	0.386	-	-	-	-	0.134	0.163	0.190	0.206	0.221	0.231
	5	-	-	-	-	-	0.252	0.293	0.319	0.337	0.355	-	-	-	-	-	0.106	0.119	0.169	0.191	0.207
	6	-	-	-	-	-	-	0.218	0.257	0.299	0.326	-	-	-	-	-	-	0.102	0.117	0.143	0.164
	7	-	-	-	-	-	-	-	0.180	0.198	0.222	-	-	-	-	-	-	-	0.061	0.100	0.120
	8	-	-	-	-	-	-	-	-	0.162	0.158	-	-	-	-	-	-	-	-	0.061	0.100
	9	-	-	-	-	-	-	-	-	-	0.162	-	-	-	-	-	-	-	-	-	0.061
	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
$\frac{Benefit}{Cost}$	1	0.345	0.363	0.370	0.378	0.391	0.393	0.392	0.392	0.392	0.391	0.335	0.339	0.353	0.354	0.355	0.355	0.353	0.353	0.353	0.353
	2	-	0.385	0.397	0.419	0.445	0.463	0.463	0.466	0.463	0.458	-	0.309	0.319	0.315	0.316	0.312	0.299	0.297	0.300	0.299
	3	-	-	0.409	0.432	0.450	0.465	0.463	0.471	0.466	0.460	-	-	0.321	0.313	0.313	0.302	0.286	0.281	0.280	0.280
	4	-	-	-	0.433	0.449	0.462	0.459	0.469	0.468	0.465	-	-	-	0.308	0.306	0.292	0.271	0.261	0.257	0.255
	5	-	-	-	-	0.454	0.465	0.465	0.475	0.473	0.468	-	-	-	-	0.312	0.296	0.272	0.259	0.254	0.252
	6	-	-	-	-	-	0.469	0.469	0.476	0.473	0.468	-	-	-	-	-	0.299	0.277	0.262	0.253	0.244
	7	-	-	-	-	-	-	0.475	0.483	0.475	0.464	-	-	-	-	-	-	0.277	0.264	0.254	0.240
	8	-	-	-	-	-	-	-	0.483	0.475	0.463	-	-	-	-	-	-	-	0.264	0.254	0.242
	9	-	-	-	-	-	-	-	-	0.475	0.465	-	-	-	-	-	-	-	-	0.254	0.242
	10	-	-	-	-	-	-	-	-	-	0.465	-	-	-	-	-	-	-	-	-	0.242

To answer **RQ1**, we examine correlations with satisfaction on the test set. For informational queries, we set $\alpha_B = 5$ and $\alpha_C = 8$ and choose *Benefit* as *metric function*. The Pearson's correlation between *Static BPM Metric* and satisfaction is 0.553, which is significantly larger than DCG@10 (0.503, $p < 0.01$). For navigational queries, we set $\alpha_B = 1$ and $\alpha_C = 5$ and choose *reciprocal Cost* as *metric function*. The Pearson's correlation is 0.298, which is smaller than DCG@10 (0.329). Based on these results, we can infer that for informational queries, *Static BPM Metrics* can have a better correlation with user satisfaction feedbacks than existing metrics given proper upper limits for benefit and cost and forms of metric function. However, for navigational queries, *Static BPM Metrics* has poor performance.

In terms of α_B, α_C and forms of *metric function*, we can see that there are differences between informational queries and navigational queries. For informational queries, *Benefit* is a good form of *metric function*, while *reciprocal Cost* is better for navigational queries. We suppose that it is because users are usually willing to pay more costs to get more information for informational queries than navigational queries, thus focusing more on *Benefit* for informational queries and *Cost* for navigational queries. In fact, for informational queries, 4-6 are proper values of α_B while 8-10 are proper values of α_C, which indicates that users may examine the whole result list to find a number of relevant documents for informational queries. However, for navigational queries, 1 is the best value of α_B. It suggests that users want to find just an exactly relevant document for navigational queries, which is consistent with the definition of navigational queries in [9]. Note that values of α_C have little effect on correlations between satisfaction and metrics. Our explanation is that the result lists for navigational queries are usually not bad. As a result, users often stop at lower ranks, which are smaller than upper limit for *Cost*.

Now we can answer **RQ2** and **RQ3**. Based on our results, there are differences in optimal upper limits and forms of *metric function* between informational queries and navigational queries. Users are usually willing to pay more costs and expect to obtain more benefits for informational queries than navigational queries. In other words, users searching for informational queries may have higher upper limits for benefit and cost. On the other hand, since users are willing to pay a lot of costs and expect to obtain a great many benefits for informational queries, they will focus on benefit rather than cost. However, for navigational queries, what users want to find is fixed, so they will focus more on their costs. These differences can guide us to choose proper *metric functions* and upper limits for different queries when we use metrics based on the BPM.

4.2.2 Dynamic BPM Metrics. Regarding **RQ4**, we compare effectiveness of *Dynamic BPM Metrics* with *Static BPM Metrics*. Specifically, we compute Pearson's correlations between satisfaction and *Dynamic BPM Metrics*. Here we focus on informational queries because our results shown in Section 4.2.1 suggest that *Static BPM Metrics* have a significantly better correlation with user satisfaction than existing metrics especially for informational queries. Inspired by the results, we define *metric function* as *Benefit*. Note that here we mainly focus on whether dynamic upper limit for benefit or cost affects effectiveness of *BPM Metrics*, rather than how it works. Therefore, we fixed one of them as zero and try different values of the other one from 0 to 1 with a step of 0.1. Based on our results, Table 5 shows two suboptimal cases for correlations between satisfaction and *Dynamic BPM Metrics*. Further analysis of parameters are not discussed and leaved for future work.

First, we examine effectiveness of dynamic upper limit for benefit based on the case where $h_B = 0.2$ and $h_C = 0$. From the results, we can see that when $\alpha_B = 3$ and $\alpha_C = 9$, we can get the best correlation (0.552), which is significantly larger than the optimal

Table 5: Pearson's Correlation between *Satisfaction Feedbacks* and *Dynamic BPM Metrics* for informational queries. The first column is the form of *metric function*. Numbers in the second column are values of α_B, while numbers in the second row are values of α_C. * (or **) indicates the difference of correlation between the *Dynamic BPM Metric* and the *Static BPM Metric* ($\alpha_B = 5$, $\alpha_C = 8$) is significant at $p < 0.05$ (or $p < 0.01$) when the *Dynamic BPM Metric* has a better correlation with *Satisfaction*.

$f(B, C)$		$h_B = 0.2, h_C = 0$										$h_B = 0, h_C = 0.3$									
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
$Benefit$	1	0.349	0.393	0.384	0.390	0.420	0.400	0.379	0.376	0.372	0.372	0.367	0.330	0.275	0.243	0.160	0.090	0.027	0.012	-0.021	-0.053
	2	-	0.393	0.414	0.441	0.482	0.512	0.502	0.523	0.502	0.476	-	0.409	0.399	0.391	0.358	0.368	0.341	0.333	0.326	0.246
	3	-	-	0.414	0.444	0.477	0.521	0.529	0.550**	0.552**	0.543	-	-	0.423	0.434	0.454	0.480	0.477	0.492	0.481	0.453
	4	-	-	-	0.444	0.469	0.495	0.509	0.542*	0.539	0.541	-	-	-	0.446	0.484	0.498	0.507	0.525	0.526	0.516
	5	-	-	-	-	0.469	0.485	0.492	0.519	0.522	0.518	-	-	-	-	0.489	0.502	0.522	0.543	0.548	0.540
	6	-	-	-	-	-	0.485	0.488	0.497	0.508	0.507	-	-	-	-	-	0.489	0.513	0.532	0.535	0.527
	7	-	-	-	-	-	-	0.488	0.497	0.490	0.485	-	-	-	-	-	-	0.505	0.522	0.523	0.514
	8	-	-	-	-	-	-	-	0.497	0.490	0.479	-	-	-	-	-	-	-	0.511	0.511	0.502
	9	-	-	-	-	-	-	-	-	0.490	0.479	-	-	-	-	-	-	-	-	0.501	0.492
	10	-	-	-	-	-	-	-	-	-	0.479	-	-	-	-	-	-	-	-	-	0.487

Static BPM Metric (0.537) where $\alpha_B = 5$ and $\alpha_C = 8$. This may indicate that dynamic upper limit for benefit is effective to describe some aspects that *Static BPM Metrics* do not consider. When a user start search, she wants to get some useful information. If the user finds a relevant document, she may expect to find more relevant documents. However, if the user find an irrelevant document, the benefits she expects may decrease. We explain it with change of task difficulty perceived by the user. Then, we examine effectiveness of dynamic upper limit for cost based on the case where $h_B = 0$ and $h_C = 0.3$. There are no significant differences of correlations between *Dynamic BPM Metrics* and the optimal *Static BPM Metric*, which suggests that the costs users are willing to pay may not be affected by the relevances of documents they have examined before. We think it is due to the fact that in the experimental user studies, queries and SERPs are fixed and not allowed to change. Participants are usually willing to examine all the results if needed, thus have relatively static upper limits for cost. Based on the results, our answer for **RQ4** is that our hypotheses about dynamic upper limits for benefit and cost partially hold. Considering dynamic upper limit for benefit, effectiveness of *Dynamic BPM Metrics* can be improved. However, participants in experimental user studies usually have static upper limits for cost. We may have to leave further analysis in practical user behavior data set for future work.

5 RELATED WORK

In order to evaluate user satisfaction of Web search, many evaluation metrics are designed with different user models. In these models, when a user ends a search session is one of the prime concerns because it is highly related to both benefit and cost estimation.

The simple model of RBP [30] assumes that users progress from one result in the ranked list to the next with *persistence p* and end their examination with probability *1 - p*. The cascade model proposed by Craswell [18] assumes that a user views search results from top to bottom and has a certain probability of being satisfied at each position. Once the user is satisfied with a document, he/she terminates the search. Based on this model, ERR [12] defines the probability that a user is satisfied with a document to be related with relevance of the document. Considering realistic user behavior, some works [14, 37] combine evaluation metrics with click models, and estimate the probability of leaving a search session given the relevance of the clicked document from click logs. In [29], Moffat et al. explore the link between user models and metrics and use a function $C_M(i)$ to describe the conditional probability that users proceed to depth $i+1$ once they have reached depth i in the ranking. They analyze different forms of $C_M(i)$ for different user models. However, these models lack insights into factors which affect when users stop search sessions.

Some theories of search and search behavior are proposed to describe how users interact with search systems. A well known conceptual model of information seeking is the Berry Picker model proposed by Bates [7], which draws an analogy between a searcher and a forager. Based on this model, Information Foraging Theory (IFT) [31] predicts how long a forager should stay in a patch before moving on to the next patch. IFT assumes that foragers wish to maximize their gain per unit of time. More recently, Fuhr [20] extendes the PRP [32] to consider a series of interactions in the interactive Probability Ranking Principle (iPRP), which accounts for the different costs and benefits associated with particular choices when ranking documents. In [8], Birchler and Butler explain how Stigler's theory [35] can be applied to search in order to predict when a user should stop examining results in a ranked list. However, they did not conduct any empirical study to verify whether the theory was consistent with users' actual behavior. Then Azzopardi suggests that Production Theory [36] could be used to model the search process instead and proposes Search Economic Theory (SET) [3] to model ad-hoc topic retrieval. In [4], Azzopardi et al. examine three theories and show that the models are complementary to each other but operate at different levels. Given these theories, it is possible to explain why users behave the way they do. However, these theories are not applied to developing Web search evaluation.

6 CONCLUSION AND FUTURE WORK

In summary, in this work, we introduce a *Bejeweled Player Model* to simulate users' search interaction processes and explain the complex stopping criterion for search sessions. In the BPM, we suppose that a user stops when he/she either has found sufficient useful information or has no more patience to continue. Given this assumption, we propose a new evaluation framework based on upper limits for both benefit and cost. Then we show how to instantiate a metric from the framework and demonstrate that some existing metrics can be derived from the framework considering one-sided case of the stopping criterion. To show effectiveness of our proposed framework, we compare *Static BPM Metrics* and *Dynamic BPM Metrics* with a number of existing metrics in terms of correlation between user satisfaction and the metrics based on a test collection containing users' explicit satisfaction feedbacks and assessors' relevance judgements. The results show that, given proper upper limits for benefit and cost and forms of metric function, *Static*

BPM Metrics and *Dynamic BPM Metrics* have a better correlation with user satisfaction feedbacks than existing metrics, especially for informational queries. In addition, there are differences in optimal upper limits and forms of metric function between informational queries and navigational queries. We also compare effectiveness of *Static BPM Metrics* and *Dynamic BPM Metrics*, and we find that considering dynamic upper limit for benefit may further improve effectiveness of *BPM Metrics*.

Our work has a few limitations: (1) We make some simplified assumptions for *Static BPM Metrics* and *Dynamic BPM Metrics*. In the future work, we plan to explore more complex situations for them. For instance, we can consider upper limits as latent variables and estimate them with large scale user logs. (2) Our test collection is based on an experimental user study where users examine results on fixed SERPs. In fact, there is a natural upper limit for cost (i.e. 10 results for each SERP), which constrains effectiveness of upper limit for cost in our framework. We would like to use realistic user logs in the future. (3) We only measure correlation with user satisfaction to show effectiveness of the framework, but the usual comparison in IR is to see how well we can determine the quality of one retrieval system versus another. We will apply the framework to this comparison to see the performance of the metrics.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61532011, 61672311) and National Key Basic Research Program (2015CB358700).

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 773–774.
[2] Paavo Arvola, Jaana Kekäläinen, and Marko Junkkari. 2010. Expected reading effort in focused retrieval evaluation. *Information Retrieval* 13, 5 (2010), 460–484.
[3] Leif Azzopardi. 2011. The economics in interactive information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 15–24.
[4] Leif Azzopardi and Guido Zuccon. 2015. An analysis of theories of search and search behavior. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ACM, 81–90.
[5] Leif Azzopardi and Guido Zuccon. 2016. An analysis of the cost and benefit of search interactions. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*. ACM, 59–68.
[6] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 625–634.
[7] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.
[8] Urs Birchler and Monika Bütler. 2007. *Information economics*. Routledge.
[9] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.
[10] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 903–912.
[11] Irina Ceaparu, Jonathan Lazar, Katie Bessiere, John Robinson, and Ben Shneiderman. 2004. Determining causes and severity of end-user frustration. *International journal of human-computer interaction* 17, 3 (2004), 333–356.
[12] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 621–630.
[13] Ye Chen, Yiqun Liu, Ke Zhou, Meng Wang, Min Zhang, and Shaoping Ma. 2015. Does vertical bring more satisfaction?: Predicting search satisfaction in a heterogeneous environment. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1581–1590.
[14] Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. 2013. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 493–502.
[15] Cyril W Cleverdon, Jack Mills, and Michael Keen. 1966. Factors determining the performance of indexing systems. (1966).
[16] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.
[17] William S Cooper. 1973. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science* 24, 2 (1973), 87–100.
[18] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 87–94.
[19] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.
[20] Norbert Fuhr. 2008. A probability ranking principle for interactive information retrieval. *Information Retrieval* 11, 3 (2008), 251–265.
[21] Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. 2014. Struggling or exploring?: disambiguating long search sessions. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 53–62.
[22] Scott B Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction?. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 567–574.
[23] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
[24] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 133–142.
[25] Diane Kelly and others. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.
[26] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 493–502.
[27] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 463–472.
[28] Stefano Mizzaro. 1997. Relevance: The whole history. *JASIS* 48, 9 (1997), 810–832.
[29] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 659–668.
[30] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2.
[31] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.
[32] Stephen E Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* 33, 4 (1977), 294–304.
[33] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 473–482.
[34] Mark D Smucker and Charles LA Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 95–104.
[35] George J Stigler. 1961. The economics of information. *Journal of political economy* 69, 3 (1961), 213–225.
[36] Hal R Varian and Jack Repcheck. 2010. *Intermediate microeconomics: a modern approach*. Vol. 6. WW Norton & Company New York.
[37] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. 2010. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1561–1564.
[38] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 435–444.
[39] Guido Zuccon. 2016. Understandability biased evaluation for information retrieval. In *European Conference on Information Retrieval*. Springer, 280–292.