Incorporating Query Reformulating Behavior into Web Search Evaluation

Jia Chen[1], Yiqun Liu[1]*, Jiaxin Mao[2], Fan Zhang[1], Tetsuya Sakai[3]
Weizhi Ma[4], Min Zhang[1], Shaoping Ma[1]
1 BNRist, Department of Computer Science and Technology, Tsinghua University
2 Gaoling School of Artificial Intelligence, Renmin University of China
3 Department of Computer Science and Engineering, Waseda University
4 Institute for AI Industry Research (AIR), Tsinghua University
yiqunliu@tsinghua.edu.cn

ABSTRACT

While batch evaluation plays a central part in Information Retrieval (IR) research, most evaluation metrics are based on user models which mainly focus on browsing and clicking behaviors. As users' perceived satisfaction may also be impacted by their search intent, constructing different user models across various search intent may help design better evaluation metrics. However, user intents are usually unobservable in practice. As query reformulating behaviors may reflect their search intents to a certain extent and highly correlate with users' perceived satisfaction for a specific query, these observable factors may be beneficial for the design of evaluation metrics. How to incorporate the search intent behind query reformulation into user behavior and satisfaction models remains under-investigated. To investigate the relationships among query reformulations, search intent, and user satisfaction, we explore a publicly available web search dataset and find that query reformulations can be a good proxy for inferring user intent, and therefore, reformulating actions may be beneficial for designing better web search effectiveness metrics. A group of Reformulation-Aware Metrics (RAMs) is then proposed to improve existing click model-based metrics. Experimental results on two public session datasets have shown that RAMs have significantly higher correlations with user satisfaction than existing evaluation metrics. In the robustness test, we have found that RAMs can achieve good performance when only a small proportion of satisfaction training labels are available. We further show that RAMs can be directly applied in a new dataset for offline evaluation once trained. This work shows the possibility of designing better evaluation metrics by incorporating fine-grained search context factors.

CCS CONCEPTS

• **Information systems → Retrieval effectiveness**;

KEYWORDS

Query Reformulation, Web Search, Evaluation Metrics

ACM Reference Format:
Jia Chen[1], Yiqun Liu[1]*, Jiaxin Mao[2], Fan Zhang[1], Tetsuya Sakai[3], Weizhi Ma[4], Min Zhang[1], Shaoping Ma[1]. 2021. Incorporating Query Reformulating Behavior into Web Search Evaluation. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482438

*Corresponding author

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482438

1 INTRODUCTION

As batch evaluation plays an essential role in web search, how to design better evaluation metrics has been a research focus for years. Evaluation metrics usually embed a user model and output the estimated satisfaction scores for users [3, 5, 26, 35, 40, 41]. Generally, a user model simulates the browsing or clicking actions and helps bridge the relationship between user behavior and satisfaction. The estimated scores can be understood as the measurement of user experience of a search process.

Besides browsing and clicking patterns, search intent may also affect users' perceived satisfaction to a certain extent [1]. As users may reformulate their queries for several iterations to strive for helpful information, their reformulating actions are highly correlated with their shifted intents during this process. Therefore, query reformulations can be a good surrogate for inferring user intent and further help model satisfaction. An example is illustrated in Figure 1. A user submits a query "Apple" in the last search round and gets familiar with Apple Inc., and then they may also want to know something about "Apple CEO" with a specialized intent. After browsing the result titled "Tim Cook, Wikipedia", they may get more interested in Apple Inc. thus feels satisfied.

User satisfaction can be affected by both user browsing behavior and their search intent. Therefore, to better estimate users' perceived satisfaction, we should design different evaluation metrics for various search intent, respectively. However, user intents are usually unobservable in realistic scenarios and should be inferred according to observable factors. To this end, we would like to employ query reformulations as the proxy for characterizing user intents to further improve user behavior models and evaluation metrics. Some related work only took query reformulation as an observable behavioral signal but did not consider the impact of user intent on the following query [16, 21]. As there are close relationships between reformulations and user intent, we hypothesize that query reformulations may be beneficial for modeling users'

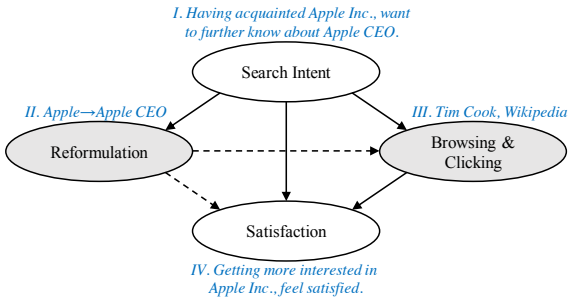


Figure 1: Relationships among query reformulations, search intent, browsing behavior, and query-level satisfaction (gray ellipse: observable factor; white ellipse: latent factor; solid arrow: direct impact; dashed arrow: indirect impact).

perceived satisfaction of the current query. To verify our hypothesis, we attempt to incorporate user reformulations into web search evaluation. Specifically, we aim to address the following research questions in this paper:

- **RQ1**: Can we find evidence indicating that query reformulation is a good surrogate for characterizing user intent and satisfaction?
- **RQ2**: How can we incorporate query reformulations into web search evaluation?
- **RQ3**: How do the proposed Reformulation-Aware Metrics (RAMs) perform compared to the state-of-the-art IR metrics?

To shed light on the above research questions, we first make investigations on a public field study dataset to ascertain the relationships among query reformulations, user intent, and satisfaction. We then introduce a group of Reformulation-Aware Metrics (RAMs) which inherits the framework of click model-based metrics to enhance query-level evaluation and adopts the multi-task learning technique to automatically learn system parameters. Experimental results on two public session datasets show that RAMs perform significantly better than state-of-the-art IR metrics in terms of user satisfaction estimation. Extensive studies also demonstrate the effectiveness of query reformulation information as well as the robustness of RAMs.

In summary, our contributions of this work are listed as follows:

- We are the first to incorporate user reformulation behavior into web search evaluation.
- We propose a novel group of metrics, namely Reformulation-Aware Metrics (RAMs). Constructed on top of click model-based metrics, RAMs model various intents reflected by user reformulations and adopt the multi-task learning technique to automatically learn optimal parameters and calibrate the estimated satisfaction ratings.
- We show that RAMs can better correlate with user satisfaction than existing metrics. Through ablation studies, we find that user reformulation information is essential in satisfaction modeling. We also verify that RAMs can perform well when only a small proportion of satisfaction labels are available or applied in a brand-new dataset for offline evaluation.

2 RELATED WORK

2.1 Web Search Evaluation

Evaluation has always been the research focus in the IR community as it determines whether a search system works well and may help for further improvements. To automatically compare the effectiveness of different search systems, numerous evaluation metrics have been proposed based on the well-established Cranfield evaluation paradigm [12]. Under this paradigm, a test collection-based evaluation with chosen metrics can simulate user behaviors on a search system in a practical setting [32]. With this simulation, each evaluation metric can output the measurement of users' search experience on a given result list. For example, RBP [26] assumes that users will continue examining the search results with a fixed probability from top to bottom, based on the cascade hypothesis [13]. Besides RBP, some other metrics also embed a specific user model, e.g., Expected Reciprocal Rank (ERR) [5], Time-Biased Gain (TBG) [35], Expected Browsing Utility (EBU) [40], U-measure [31], INST [3], Bejeweled Player Model (BPM) [41], etc. To unify various user models, Moffat et al. [25] proposed the C/W/L framework, which describes three related behavioral aspects: Continuation (C) probability, Weight (W) function, and Last examining (L) probability. These metrics have been widely used in different search scenarios and promote the development of IR. However, most of them do not consider the influence of user intent on the perceived satisfaction and estimate the same ratings for a given relevance (or usefulness) list. To better model user satisfaction, we aim to use query reformulation behavior as a proxy to characterize user intent and yield a group of more personalized evaluation metrics.

2.2 User Query Reformulations

As query reformulation is a bottleneck in web search, a broad spectrum of research has focused on understanding user reformulating actions in various search scenarios [7, 8, 17, 19, 28]. Based on a search engine log, Huang and Efthimiadis [18] investigated various reformulation strategies of search users by analyzing content change, including word reorder, remove/add words, URL striping, acronym, substring/superstring, abbreviation, etc. Besides the reformulation content, Chen et al. [7] conducted a field study and investigated the differences in user behavior from more delicate aspects such as the reformulation reason, interface, and the inspiration source. Their work provides valuable insights into understanding users' complex search behavior as well as guidance for designing better query suggestion techniques.

However, as an accessible contextual factor, reformulating actions have seldom been utilized for user modeling or satisfaction estimating. For instance, Hassan et al. [16] found that users' dissatisfaction towards the last query is highly correlated with the similar reformed queries and the short reformulation time. Some existing work also exploited the query change between two consecutive queries to improve session search [15, 23]. To enhance the evaluation of session search, Lipani et al. [21] introduced a new parameter called *balancer* into RBP to quantify the balance between reformulating queries and examining more results. Although previous work has introduced the concept of reformulation, they do not directly bridge the relationship between user behavior and their intents, i.e., to characterize user intent and perceived satisfaction with query

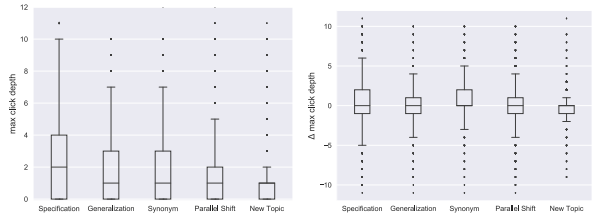
reformulating actions. Therefore, we attempt to incorporate query reformulations into web search evaluation.

3 REFORMULATION, SEARCH INTENT, BROWSING, AND SATISFACTION

To answer **RQ1**, we explore the relationships between query reformulations, search intent, browsing behavior, and satisfaction on a public dataset released by Chen et al. [7]. This dataset contains fine-grained information such as reformulation type, interface, reason, and the corresponding inspiration source. We expect to find evidence showing that users' query reformulating behavior is a relevant contextual factor to their search intent or satisfaction. In the case that there exist differences in browsing patterns or user perception of satisfaction across various reformulation types, we can use query reformulation as a surrogate to better model user intent and satisfaction.

3.1 Reformulation and Browsing & Clicks

Initially, we explore the differences in user browsing behavior with regard to intent-aware reformulation types. Here we adopt the reformulation taxonomy proposed in [7] and condense it into five main types: "Specification", "Generalization", "Synonym", "Parallel Shift", and "New Topic". Some types are merged into one as they present similar intents, e.g., "Specification" vs. "Meronym" and "Generalization" vs. "Holonym". Distributions of 1) the maximum click depth for a query as well as 2) the Δ maximum click depth compared to the former query across reformulation types are presented in Figure 2. As shown in Figure 2(a), we can observe that generally, users will click deeper with a more specialized intent. On the other hand, if their intent has shifted, they may only examine and click the top two results. Users who narrow down the search scope may be more interested in the search topic due to their cumulative gain in previous search rounds; hence, they become more engaged in the search process. From Figure 2(b), we find that there are no significant differences in the averaged Δ maximum click depth after users have taken various reformulating actions. However, the variance is much smaller in the "New Topic" condition. These observations imply that we should focus more on users' intent shift.



(a) Maximum click depth. (b) Δ maximum click depth.

Figure 2: Distribution on user clicks across various intent-level reformulation types.

Besides clicks, the C/W/L framework [24, 25] has been proposed to characterize user models with three related behavioral probabilities: viewing, continuing, and stopping. Following previous work [39], the distributions of continuation $\hat{C}(\cdot)$, weight $\hat{W}(\cdot)$, and last examining $\hat{L}(\cdot)$ vectors can be estimated from observed user behavior as follows:

$$\hat{C}(r) = \frac{\sum_{s\in S} \hat{P}(E_{r+1}^{(s)} = 1)}{\sum_{s\in S} \hat{P}(E_r^{(s)} = 1)} \tag{1}$$

$$\hat{W}(r) = \frac{\sum_{s\in S} \hat{P}(E_r^{(s)} = 1)}{\sum_{s\in S} \sum_{j=1}^{N} \hat{P}(E_j^{(s)} = 1)} \tag{2}$$

$$\hat{L}(r) = \frac{\sum_{s\in S} \hat{P}(E_r^{(s)} = 1) - \hat{P}(E_{r+1}^{(s)} = 1)}{\sum_{s\in S} \hat{P}(E_1^{(s)} = 1)} \tag{3}$$

where $\hat{P}(E_r^{(s)} = 1)$ denotes the estimated probability that a user examines the r-th result in a specific query session s, and S is the set of all query sessions. For simplicity, we use last clicks to estimate this probability. Comparisons of estimated C/W/L vectors for each reformulation type are presented in Figure 3. For all the reformulation types, the continuation probability will first increase and then decrease. This is slightly different from the previous work [38] which found the continuation probability will increase with the rank. Considering that users are not likely to turn the result page, there is a considerable decline for continuation at the bottom of the first page. Moreover, there are subtle differences in all C/W/L vectors among reformulation types, especially between the "New topic" type and others. For the first result, users with a specialized intent may examine the next one with a probability of about 70%, which is almost twice the probability in the "New topic" condition. Similar trends are found in Figure 3(b) and 3(c). Generally speaking, with a more specialized intent, users may engage more on all results within the first page. These findings accord with those in Figure 2.

3.2 Reformulation and Satisfaction

As reformulating behaviors do not directly affect user satisfaction for a specific query, we aim to figure out the relationship between the two factors via several experiments. Here we hold two following assumptions: 1) users will behave in different patterns after they have act various reformulations, which may subsequently affect their satisfaction, and 2) users who commit different reformulations may have various information needs or expectations; therefore, their perception of satisfaction is mainly affected by search intent.

To verify our assumptions, we set three experimental conditions for several IR metrics and then evaluate these metrics by investigating the relationship between their accuracy and correlation with satisfaction [22, 25, 33], e.g., calculating correlation coefficients such as Spearman's ρ [36] and Pearson's r [27]. Following the previous study [42], all the three conditions involve a bootstrapping procedure, i.e., the experiments are conducted on 100 data samples generated from the original data. We consider the following metrics: 1)metrics without parameters such as Precision, U-measure [31], RR, and Average Precision (AP) [37]; 2) metrics with parameters such as RBP [26], DCG [20], INST [3], INSQ [25], and BPM [41]. The three conditions are listed as follows:

- **I**: We tune all metrics with parameters according to the distances between the estimated C/W/L and the observed vectors on the training set by grid search, as done in the previous work [42]. To this end, we utilize the *cwl_eval* [2] tool to generate the $C(\cdot)$, $W(\cdot)$, and $L(\cdot)$ vectors for a specific metric with given parameter(s) and report the expected utility (EU) as the outputs of metric scores. For other details, we use the same settings as previous work [42].

(a) $\hat{C}$ vector. (b) $\hat{W}$ vector. (c) $\hat{L}$ vector.

Figure 3: The observed C/W/L vectors across various intent-level reformulation types (best viewed in color).

- **II**: Upon I, we first tune all metrics based on the reformulation-aware C/W/L losses. As the intents behind the reformulations are unobservable, we use a well-defined syntactic-level taxonomy ("Add", "Delete", "Keep", "Transform", "Others", "First Query") [7, 9] as the surrogate to distinguish various intents. For a specific query, we apply a metric with the best parameter(s) against the corresponding reformulation type and use the context-aware expected utility (CEU, the expected utility given the reformulation action) as the metric score.
- **III**: On top of II, we further calibrate the metric scores with linear regression based on the syntactic query reformulations:

$$Sat = a_\omega \cdot CEU + b_\omega$$

where Sat and ω denote the calibrated score and the observed reformulation type for a query, respectively. a_ω and b_ω are the hyperparameters needed to be estimated from the training set, representing the slope and intercept for the regression of ω.

Experimental results for all metrics in the three conditions are shown in Table 1. By comparing the results in each condition, we have the following findings. Firstly, CEUs correlate better with user satisfaction than EUs in terms of Spearman's ρ for almost all metrics in at least one aspect in the C/W/L framework, especially when using the W (weight) method. However, the improvements in Pearson's r are relatively marginal. This may imply that tuning metrics with reformulations can boost the rank correlation between the predicted satisfaction ratings and the true values. We suggest that this linear correlation depends more on the score distribution of a specific metric. Therefore, it is hard to improve Pearson correlation merely by tuning parameters for a metric. To exemplify this, we plot the distribution of EU and CEU scores for RBP. As shown in Figure 4, CEU yields flatter distributions which have fewer peak values. In this regard, CEU may alleviate the problem that metrics such as RBP have poor *discriminative power* [30]. As for condition III, we surprisingly find considerable improvements for all metrics across the C/W/L vectors in both ρ and r, indicating that users' perceived satisfaction may accord with the result list to different extents across reformulation types. These results verify our two assumptions and may provide guidance for designing better evaluation metrics.

3.3 Reformulation and Intent

All previous efforts merely use the query reformulating behaviors as the surrogates for search intents (or intent shift). To explicitly mine the relationships between the two factors, we calculate the transition probabilities from the syntactic reformulation types to the five aforementioned intent-level ones. The transition probability matrix is presented in Table 2.

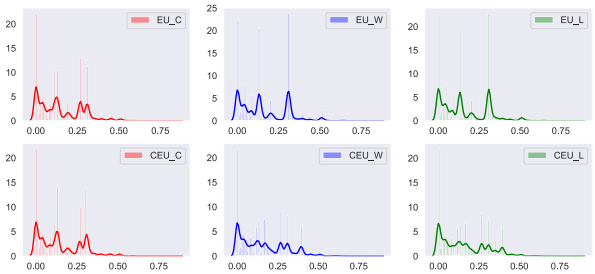


Figure 4: Distribution of Expected Utility (EU) vs. Context-aware Expected Utility (CEU) for RBP in the C/W/L framework, respectively.

As we can observe, mapping relations for "Add", "Delete", and "Repeat" types are more concentrated. Although the "Change" and "Others" types can be mapped into several intent-level types, their most likely intent and overall distribution are also different. The five query change types can be regarded as the coarse-grained intent shift type. This transition relationship may be helpful in conditions where annotations for intent are not available.

3.4 Summary

In this subsection, we summarize the findings for **RQ1** as follows:

- Query reformulation is an accessible yet useful contextual factor that can reflect user intent to a certain extent.
- Users behave differently across various intent-level reformulation types. Employing query reformulations as the surrogate of inferring intent may be beneficial for user modeling.
- It is effective to tune traditional IR metrics and further calibrate them with syntactic reformulation types.
- As Context-aware Expected Utility (CEU) yields a flatter and more dense distribution than Expected Utility (EU), it can be used to improve metrics with poor discriminative power.

4 REFORMULATION-AWARE METRICS

In this section, we would like to answer **RQ2**. As revealed in previous investigations, there are close relationships between query reformulating behavior, search intent, browsing and click behavior, and user satisfaction. To bridge these factors and incorporate query reformulations into evaluation, we design the following three steps:

Table 1: Meta-evaluation of various metrics on TianGong-Qref in three conditions. "△/▽" and "▲/▼" indicate a statistically significant improvement or decline compared to the corresponding value in condition I at $p < 0.05/0.01$ level using a two-tailed pairwise t-test with Bonferroni correction [34]. The significance test is conducted over all bootstrapping samples. Spearman's ρ of metrics without parameters are listed as: Precision@10: 0.3944, U-measure (L=1000): 0.3946, RR: 0.4495, AP: 0.4667.

			I			II			III		
	Metrics	**Parameters**	**C**	**W**	**L**	**C**	**W**	**L**	**C**	**W**	**L**
Spearman's ρ	RBP	p	0.4375	0.4361	0.4361	0.4405▲	0.4447▲	0.4435▲	0.4728▲	0.4731▲	0.4730▲
	DCG	b	0.4416	0.4374	0.4434	0.4464▲	0.4421▲	0.4424▼	0.4744▲	0.4704▲	0.4737▲
	INST	T	0.4413	0.4413	0.4396	0.4464▲	0.4493▲	0.4476▲	0.4725▲	0.4736▲	0.4740▲
	INSQ	T	0.4403	0.4403	0.4389	0.4448▲	0.4472▲	0.4455▲	0.4717▲	0.4732▲	0.4744▲
	BPM-Static	T/K	0.4552	0.4448	0.4180	0.4550	0.4506▲	0.4181	0.4611▲	0.4685▲	0.4546▲
	BPM-Dynamic	T/K	0.4244	0.4210	0.4180	0.4235▼	0.4255▲	0.4181	0.4558▲	0.4599▲	0.4546▲
Pearson's r	RBP	p	0.4180	0.4193	0.4193	0.4200▲	0.4220▲	0.4207▲	0.4745▲	0.4749▲	0.4751▲
	DCG	b	0.4182	0.4174	0.4177	0.4199▲	0.4178	0.4193▲	0.4754▲	0.4739▲	0.4750▲
	INST	T	0.4085	0.4085	0.4085	0.4045▼	0.4030▼	0.4088	0.4637▲	0.4651▲	0.4653▲
	INSQ	T	0.4204	0.4204	0.4184	0.4215△	0.4212△	0.4209▲	0.4757▲	0.4763▲	0.4746▲
	BPM-Static	T/K	0.3635	0.3803	0.3915	0.3633	0.3868▲	0.3915	0.4322▲	0.4445▲	0.4486▲
	BPM-Dynamic	T/K	0.3571	0.3707	0.3915	0.3563▼	0.3703	0.3915	0.4210▲	0.4298▲	0.4486▲

Table 2: Transition probability matrix from syntactic-level reformulation types (i.e., query change) to intent-level reformulation types on TianGong-Qref dataset.

Intent←Syntactic	**Add**	**Delete**	**Change**	**Repeat**	**Others**
Specification	0.8858	0.0806	0.3177	0.1057	0.0843
Generalization	0.0118	0.7630	0.0668	0.0081	0.0427
Synonym	0.0659	0.1185	0.1497	0.8618	0.0444
Parallel Shift	0.0335	0.0237	0.4276	0.0163	0.3383
New Topic	0.0029	0.0142	0.0382	0.0081	0.4903

1) *intent selection*, 2) *click model modification*, and 3) *satisfaction calibration*. Firstly, we select an intent for an observed reformulating action. To model the diversity in user behavior under various intents, we further introduce multiple intents into some click models by adding intent-aware parameters. Following previous work [11], we derive click model-based metrics and then calibrate the metric scores over intents, mapping the normalized values into satisfaction scores with specific scales. Our ultimate goal is to yield a group of Reformulation-Aware Metrics (RAMs) that can be directly applied in any search scenarios wherein query sequences submitted by users are known (e.g., session search). In the following sections, we will introduce the three steps, respectively.

4.1 Intent Selection

User intent is a latent factor which can not be directly observed. To elaborate, we select intents for a given observed reformulation ω, where $\omega \in \Omega$ and $\Omega = \{\mathcal{A}, \mathcal{D}, \mathcal{K}, \mathcal{T}, \mathcal{O}, \mathcal{F}\}$, where $\mathcal{A}, \mathcal{D}, \mathcal{K}, \mathcal{T}, \mathcal{O}$, and $\mathcal{F}$ stand for "Add", "Delete", "Keep" (or Repeat), "Transform" (or Change), "Others", and "First query", respectively.

Given a reformulation action ω, we define the probability that a user is with the k-th intent type as $i_{\omega,k}$:

$$P(I = I_k \mid \omega) = i_{\omega,k}, \text{ where } \sum_{k=1}^{K} i_{\omega,k} = 1 \tag{4}$$

$$i_{\omega,k} = \text{softmax}(\pi_{\omega,k}) = \frac{e^{\pi_{\omega,k}}}{\sum_{j=1}^{K} e^{\pi_{\omega,j}}} \tag{5}$$

Where I_k denotes the k-th intent type and K is the total number of considered intents. Note that to ensure the summation of $i_{\omega,k}$ over k is 1, we make a softmax transformation and introduce a new group of parameters $\pi_{\omega,k}$, where $\pi_{\omega,k} \in \mathbb{R}$.

4.2 Click Model Modification

In this subsection, we would like to introduce multiple intents into user models. Many evaluation metrics encapsulate assumptions about user behavior, e.g., Rank-biased precision (RBP) [26] assumes that users will continue examining the following result with a certain persistence θ. However, user models embedded in these metrics are generally simplified and not necessarily realistic. A more intuitive way is to modify click models by considering different intents and then derive the corresponding click model-based metrics [11].

Take DBN [6] as an example, let $C/E/A/S/R/r/k/u/q$ denote *Click*, *Examination*, *Attractiveness*, *Satisfaction*, *Relevance*, *ranking position*, *intent type*, *url*, and *query*, we modify its assumptions as follows:

$$C_r = 1 \iff E_r = 1 \text{ and } A_r = 1 \tag{6}$$

$$P(A_r = 1) = \alpha_{R_{urq}} \tag{7}$$

$$P(E_1 = 1) = 1 \tag{8}$$

$$P(E_r = 1 \mid E_{r-1} = 0) = 0 \tag{9}$$

$$P(S_r = 1 \mid C_r = 1, I = I_k) = \sigma_{R_{urq}k} \tag{10}$$

$$P(E_r = 1 \mid S_{r-1} = 1) = 0 \tag{11}$$

$$P(E_r = 1 \mid E_{r-1} = 1, S_{r-1} = 0, I = I_k) = \gamma_k \tag{12}$$

The modifications are: 1) we assign the attractiveness $\alpha_{R_{urq}}$ with a fixed value $\frac{2^{R_{urq}}-1}{2^{R_{max}}}$ to support the offline evaluation; 2) we assume that σ, the probability of a user being satisfied with a result given that they have clicked on it, only depends on its relevance and the intent type. This setting can handle unseen query-document pairs and is also more reasonable than the original assumption where σ only depends on the rank r; 3) Finally, we categorize the continuation probability γ with various intents.

Similarly, we can also modify other click models, e.g., SDBN, UBM, and PBM. For SDBN, γ_k is fixed to 1. For UBM and PBM, $\gamma_{rr'}$ and γ_r should be replaced with $\gamma_{rr'k}$ and γ_{rk}, respectively.

4.3 Satisfaction Calibration

To bridge user model and satisfaction, we derive click model-based metrics in the light of the user model defined in the previous subsection. Empirically, we find it effective to fit metric scores across different intent types via linear regression. Therefore, we also calibrate metric scores by learning the linear correlation coefficient β_k and the intercept ψ_k under each intent.

Following previous work [4, 11], we can distinguish *utility-based* (*uMetric*) and *effort-based* satisfaction scores (*rrMetric*) with the following instantiations:

$$uSat = \sum_{k=1}^{K} P(I = I_k) \cdot \beta_k \cdot (\sum_{r=1}^{N} P(C_r = 1 \mid I = I_k) \cdot R_r + \psi_k) \tag{13}$$

$$\begin{aligned} rrSat &= \sum_{k=1}^{K} P(I = I_k) \cdot \beta_k \cdot (\sum_{r=1}^{N} P(S_r = 1 \mid I = I_k) \cdot \frac{1}{r} + \psi_k) \\ &= \sum_{k=1}^{K} i_{\omega,k} \beta_k \cdot (\sum_{r=1}^{N} \sigma_{R u_r q k} \cdot P(C_r = 1 \mid I = I_k) \cdot \frac{1}{r} + \psi_k) \end{aligned} \tag{14}$$

Here *uSat* represents the utility-based satisfaction score, and *rrSat* represents the effort-based one. N is the number of documents we consider for a query. Note that the intent-aware click probability $P(C_r = 1 \mid I = I_k)$ denotes the independent click probability $P(C_r = 1)$ rather than the conditional one $P(C_r = 1 \mid \mathbf{C}_{<r_u})$. These probabilities can be easily calculated based on the variable dependencies of a specific click model [10].

4.4 Model Optimization

To fit our model with both user behavior and satisfaction, we adopt the multi-task learning technique. As for estimating the model parameters, we would like to minimize a loss function $f(\Theta)$:

$$\min_{\Theta} \ f(\Theta), \text{ where } f(\Theta) = (1 - \lambda) \cdot \mathcal{L}_b + \lambda \mathcal{L}_s, \tag{15}$$

$$\mathcal{L}_b = -\frac{1}{|S| \cdot N} \sum_{s \in S} \log(\prod_{r=1}^{N} P(C_r = c_r^{(s)} \mid \mathbf{C}_{<r}^{(s)})) \tag{16}$$

$$\mathcal{L}_s = \frac{1}{|S|} \sum_{s \in S} ||\hat{sat}^{(s)} - sat^{(s)}||^2 \tag{17}$$

Here $\mathcal{L}_b$ and $\mathcal{L}_s$ denote the loss of fitting user behavior and satisfaction, respectively. S is the set of all query sessions, and the superscript (s) denotes the corresponding value in a particular query session s. Note that λ controls the trade-off in fitting the two facets, and Θ represents all parameters involved. For $\mathcal{L}_b$, we formulate it as the negative log-likelihood of user click behavior. Here we factorize the log-likelihood of a click sequence into the product of conditional click probabilities $P(C_r = c_r^{(s)} \mid \mathbf{C}_{<r}^{(s)})$. As shown in Eq. 17, $\mathcal{L}_s$ is written as the Mean Square Error (MSE) between the predicted satisfaction $\hat{sat}$ (e.g., *uSat* or *rrSat*) and the true value *sat*.

5 EXPERIMENTS

In this section, we aim to investigate **RQ3** by conducting a series of experiments. We will first briefly introduce the experimental setups in Section §5.1. Then in Section §5.2, we compare the overall performance of RAMs with several state-of-the-art IR metrics in terms of satisfaction estimation and user behavior prediction. To further investigate the effectiveness of RAMs and the multi-task learning technique, ablation study and robustness test have also been conducted in Section §5.3 and §5.4. Lastly, we analyze the learned parameters to verify the interpretability of RAMs.

Table 3: Basic statistics of two preprocessed datasets.

	TianGong-Qref	TianGong-SS-FSD
# sessions	2,353	664
# queries	7,479	3,342
# results per SERP	10	10
usefulness judgement	4-level	5-level
query-level satisfaction	5-level	5-level

5.1 Experimental Setup

5.1.1 Dataset. There exist several datasets which support the meta-evaluation of IR metrics. Among them, we use TianGong-Qref [7] and TianGong-SS-FSD [42], since they were both collected via field studies and may collect more realistic behavioral information compared to lab-based user study. For simplicity, we denote the two datasets as *Qref* and *FSD* in what follows. To facilitate the application of RAMs, we only consider sessions with at least two queries in the *FSD* dataset. In addition, we adopt the usefulness labels as the relevance scores since only usefulness judgments are available in the *Qref* dataset. As users usually pay more attention to the first result page, we truncate the result lists at a length of 10 and filter the rest of the results for both datasets. Basic statistics of the two datasets after preprocessing are shown in Table 3.

5.1.2 Baselines and meta-evaluation approaches. We compare RAMs with DCG, RBP, and BPM as they outperform other metrics in our previous experiments. For RAMs, we consider six variants: uDBN, rrDBN, uSDBN, rrSDBN, uUBM and uPBM. Considering that UBM and PBM do not involve the concept of satisfaction (S), we only derive their corresponding utility-based metrics (uMetrics). In addition, we also train the variants of RAMs without considering the query reformulations by setting the number of intents $k = 1$. As RAMs need to be tuned with satisfaction labels, we also tune DCG, RBP and BPM according to Spearman's ρ correlation with satisfaction on the training set (denoted as "*w/ sat*"). To ensure a fair and robust comparison, we carefully tune all metrics and report their best performance on the two datasets, respectively.

To evaluate the effectiveness of each metric, we delve into two facets: I) satisfaction estimation and II) behavior prediction. For I, we calculate the correlation coefficients between the predicted satisfaction ratings and the ground truth values, e.g., Spearman's ρ, Pearson's r, and Mean Square Error for satisfaction (SAT MSE). For II, we consider click perplexity (PPL) and the Mean Square Error of the C/W/L vectors (C/W/L MSE). Since we fix the attractiveness in click models to mitigate the position bias, there may exist a discrepancy between the predicted click probability and the real value. This will cause very high PPL values. To this end, we ignore these abnormal cases while calculating the click perplexity (i.e., a query will be filtered if its negative log-likelihood for the click sequence is higher than 50). As the global C/W/L vectors are coarse-grained, we estimate the C/W/L vectors under each syntactic-level reformulation type and then calculate the loss for a query according to the query change. For the $C(\cdot)$ and $L(\cdot)$ vectors, we only consider the top

nine positions because the 11-th result is unknown. Moreover, we ignore the C vector for UBM and PBM since they assume that user examination probabilities $\gamma_{rr'}$ and γ_r do not depend on previous user actions.

5.1.3 Bootstrapping. To obtain fair and robust evaluation results, we generate 100 samples (each sample has a training set and a testing set) for both datasets using the bootstrapping algorithm. For each time, we randomly sample training queries from the whole dataset with replacement until the training set has the same size as the original data. Those queries not included in the training set will then form a corresponding testing set. We will report averaged experimental results and conduct the significance tests over these samples hereinafter.

5.1.4 Implementation details. Since it is non-trivial to calculate the analytical solution for all parameters in RAMs, here we adopt the Stochastic Gradient Descent (SGD) [29] algorithm to learn these parameters. Compared to grid search [42], this approach is more sophisticated and can be easily applied to models with multiple parameters. As for the intent selection, we set k to six and initialize the $\pi_{\omega,k}$ according to the transition probability in Table 2 on both datasets. This technique can stabilize the training process in the case that $\pi_{\omega,k}$ converges to similar values due to the symmetrical characteristic. In addition, we test the performance of RAMs with various λ and find the best value 0.85. The initial learning rate is selected from $\{0.01, 0.005, 0.001\}$ and will be discounted with a rate of 0.99 by each step. We stop the training process if the training loss does not decrease after five iterations. Without loss of generality, we derive the updating formulas for four click models: DBN/SDBN [6], UBM [14], and PBM [14]. To facilitate the reproductivity of our results, we also release the source code for our experiments as well as the derivative process for SGD in the link below [1].

5.2 Overall Performance

Table 4 systematically reports the performance of various metrics on two datasets. Note that to control the Family-Wise Error Rate, we calibrate all p-values with the Bonferroni Correction [34]. From the comparison, we have serveral findings:

- Among traditional metrics, BPM performs the best. We find that when tuning these metrics with satisfaction labels based on ρ, their corresponding linear correlations may slightly decrease (especially for BPM). This indicates that there is a trade-off between the rank and linear correlation in satisfaction prediction.
- RAMs significantly outperform traditional metrics in terms of satisfaction estimation. Among all RAMs, uDBN with reformulation is substantially superior to other variants. Surprisingly, it achieves an improvement of 6.62% and 6.60% on Spearman's ρ and Pearson's r over the best baseline metrics in *Qref* dataset.
- Reformulation information is beneficial for satisfaction estimation since most RAMs perform worse when ignoring query reformulations. Besides the correlation coefficients, the SAT MSEs also reduce a lot when considering query reformulations. This observation verifies the assumptions we mentioned hereinbefore.
- We find that there is no evident relationship between PPL and C/W/L MSE. Comparing the second and the third row groups in Table 4, we find PPLs are slightly improved for uDBN and uSDBN, which demonstrates the consistency of user modeling and satisfaction predicting. However, the C/W/L losses increase in all metrics. We guess PPL can reflect user behavior more accurately than C/W/L vectors. As C/W/L vectors are derived over all cases, they may be too coarse-grained to measure whether a system can predict user behaviors accurately.
- Spearman's ρ for rrMetrics are relatively higher when ignoring query reformulations, which is also reasonable. As we use a linear point-wise loss to fit satisfaction, r and SAT MSE will be directly optimized. If we replace it with a pairwise loss, then the Spearman correlation may also be improved. From another point, RAMs output more balanced scores, which will reduce the number of tie cases. As we have only collected 5-scale satisfaction scores, the Spearman's ρ may decline when the score distribution is more dense or balanced.

5.3 Ablation Study

To further investigate the effectiveness of reformulation information and multi-task learning technique, we conduct an ablation study for the best metric uDBN. Accordingly, we eliminate four factors: 1) the transition probability matrix used to initialize $\pi_{\omega,k}$ (denoted as "w/o fineinit"), 2) $\mathcal{L}_s$, 3) $\mathcal{L}_b$, and 4) reformulation information. For 1), we would use a transition that mainly maps each syntactic reformulation type into one intent. Note that RAMs without $\mathcal{L}_s$ and query reformulations are equivalent to the original version of corresponding click model-based metrics. As revealed in Table 5, we find the effectiveness of this transition matrix on the *Qref* dataset in Spearman's ρ. In contrast, only using a concentrated mapping relationship can also achieve good performance on both datasets. This suggests that RAMs are still superior to traditional metrics by a large margin even without annotations for intent-level reformulation type. We can also observe that $\mathcal{L}_s$ is more important than $\mathcal{L}_b$ when incorporating user query reformulations. However, PPL will increase to a certain extent if we ignore user behavior. $\mathcal{L}_b$ and reformulations can be the regularizers to avoid RAMs overfitting user satisfaction.

5.4 Robustness Test

As satisfaction labels may not be available in all scenarios, it may be hard for RAMs to be directly applied for offline evaluation if they depend heavily on the satisfaction labels. To this end, we test the performance of uDBN when using different sizes of training data. From Figure 5, we find that the satisfaction estimation performance of uDBN is stable when trained on different sizes of data. According to the *central-limit theorem*, the upper bound of the size for a bootstrapping testing sample is about $1/e$ (≈ 0.3679) of the whole data (or about half of the training set). However, uDBN can perform well by using only 20% training queries, which is much smaller than the testing set. Moreover, all boxes are over the average performance of the corresponding best baseline metrics (blue dashed line). This implies that RAMs can still achieve excellent performance by using a small proportion of human satisfaction ratings.

A good metric should also be successfully applied in various datasets with high robustness. To verify this, we train RAMs on the *FSD* dataset and then test their performances on the *Qref* dataset. Note that we can only conduct the transfer application from the *FSD*

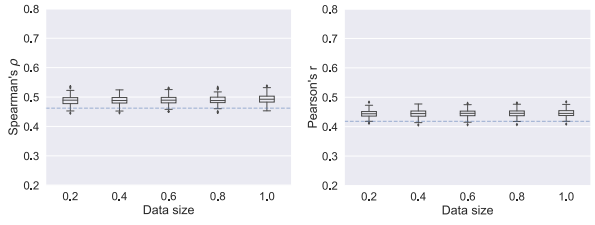
[1] https://github.com/xuanyuan14/Reformulation-Aware-Metrics

Table 4: Comparison of various metrics in terms of overall performance on two datasets. "▲" indicates a statistically significant improvement over the corresponding best baseline in Spearman's ρ and Pearson's r at $p < 0.001$ level using a two-tailed pairwise t-test, respectively. Note that we calibrate all p-values through Bonferroni correction [34].

	TianGong-Qref					TianGong-SS-FSD				
	ρ	r	PPL	C/W/L MSE	SAT MSE	ρ	r	PPL	C/W/L MSE	SAT MSE
RBP	0.4375	0.4180	N/A	N/A	N/A	0.4898	0.5222	N/A	N/A	N/A
DCG	0.4434	0.4182	N/A	N/A	N/A	0.5022	0.5290	N/A	N/A	N/A
BPM	0.4552	0.3915	N/A	N/A	N/A	0.5801	0.6052	N/A	N/A	N/A
RBP w/ sat	0.4389	0.4170	N/A	N/A	N/A	0.5165	0.5527	N/A	N/A	N/A
DCG w/ sat	0.4446	0.4166	N/A	N/A	N/A	0.5047	0.5344	N/A	N/A	N/A
BPM w/ sat	0.4622	0.3674	N/A	N/A	N/A	0.5960	0.6029	N/A	N/A	N/A
rrDBN w/o reform	0.4498	0.3490	1.1150	0.7714/0.0840/0.1882	1.1857	0.6291▲	0.5412	1.1663	0.7350/0.0743/0.1586	1.1023
rrSDBN w/o reform	0.4392	0.3457	**1.1137**	0.9554/0.1013/0.2416	1.1858	0.6289▲	0.5644	1.1731	0.9845/0.0978/0.2293	1.0872
uUBM w/o reform	0.4488	0.3855	1.1481	n.a./0.0977/0.4175	1.1322	0.6198▲	0.5582	1.1536	n.a./0.0424/0.3616	0.9175
uPBM w/o reform	0.4542	0.3954	1.1505	n.a./0.0516/0.1632	1.1091	0.6183▲	0.5696	**1.1506**	n.a./0.097/0.0857	0.8909
uSDBN w/o reform	0.4494	0.4098	1.1161	0.9271/0.0966/0.2252	1.2000	0.6217▲	0.5982	1.1652	0.9483/0.0915/0.2102	0.9002
uDBN w/o reform	0.4521	0.4136	1.1407	**0.1196/0.0079/0.0235**	1.1385	0.6223▲	0.6110▲	1.1689	**0.2711/0.0239/0.0646**	0.8472
rrDBN	0.4123	0.3670	1.1140	0.9473/0.1005/0.2405	1.1508	0.5908	0.5602	1.1667	0.7606/0.0768/0.1649	1.0767
rrSDBN	0.4177	0.3713	1.1141	0.9611/0.1018/0.2456	1.1413	0.5991▲	0.5703	1.1736	0.9836/0.0975/0.2286	1.0524
uUBM	0.4812▲	0.4303▲	1.1663	n.a./0.9613/0.4981	1.0607	0.6242▲	0.5775	1.1597	n.a./0.0462/0.3619	0.8795
uPBM	0.4827▲	0.4369▲	1.1647	n.a./0.0384/0.1471	**1.0524**	0.6210▲	0.5846	1.1550	n.a./0.0095/0.0911	0.8644
uSDBN	0.4837▲	0.4375▲	1.1155	0.9345/0.0976/0.2294	1.1443	0.6290▲	0.6081▲	1.1652	0.9505/0.0921/0.2110	0.8840
uDBN	**0.4928**▲	**0.4458**▲	1.1341	0.1586/0.0093/0.0170	1.0801	**0.6339**▲	**0.6207**▲	1.1686	0.3270/0.0275/0.0638	**0.8322**

Table 5: Ablation study on the uDBN.

	ρ	r	PPL	SAT MSE
TianGong-Qref				
uDBN w/o fineinit	0.4890	0.4460	1.1275	1.0920
uDBN w/o <$\mathcal{L}_s$ + reform>	0.4350	0.3898	1.1141	1.6457
uDBN w/o $\mathcal{L}_b$	0.4940	0.4437	1.1419	1.0816
uDBN w/o <$\mathcal{L}_b$ + reform>	0.4510	0.4120	1.1519	1.1361
uDBN	0.4928	0.4458	1.1341	1.0801
TianGong-SS-FSD				
uDBN w/o fineinit	0.6340	0.6214	1.1686	0.8330
uDBN w/o <$\mathcal{L}_s$ + reform>	0.6175	0.6036	1.1657	0.9504
uDBN w/o $\mathcal{L}_b$	0.6355	0.6202	1.1710	0.8309
uDBN w/o <$\mathcal{L}_b$ + reform>	0.6258	0.6108	1.1717	0.8449
uDBN	0.6339	0.6207	1.1686	0.8322



(a) Spearman's ρ. (b) Pearson's r .

Figure 5: The influence of data size on the performance of uDBN in TianGong-Qref. The blue dashed line denotes the average performance of the best baseline metric.

dataset to the *Qref* dataset because the maximum usefulness rating is higher in the *FSD* dataset. From Table 6, we find that uDBN and uSDBN can still perform well on a brand-new dataset. In contrast, uUBM and uPBM may perform worse when applied in a different dataset. As uDBN/uSDBN can better fit user behavior, it is less likely for them to overfit the satisfaction distribution of a specific dataset. Therefore, it is essential to fit user behavior to ensure the robustness of evaluation metrics. In this regard, uDBN/uSDBN can be easily used for offline evaluation in various search scenarios with a similar scale of relevance or usefulness labels.

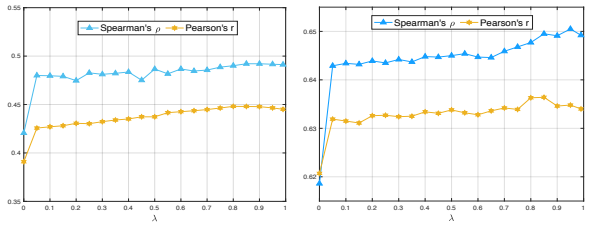
Table 6: Transfer application of RAMs from TianGong-SS-FSD to TianGong-Qref.

Training \ Testing	Qref ρ	Qref r	FSD ρ	FSD r
uDBN-Qref	0.4928	0.4458	N/A	N/A
uDBN-FSD	0.4891	0.4453	0.6339	0.6207
uSDBN-Qref	0.4837	0.4375	N/A	N/A
uSDBN-FSD	0.4837	0.4375	0.6290	0.6081
uUBM-Qref	0.4812	0.4304	N/A	N/A
uUBM-FSD	0.4695	0.4192	0.6242	0.5775
uPBM-Qref	0.4834	0.4220	N/A	N/A
uPBM-FSD	0.4613	0.4251	0.6223	0.5772

5.5 Parameter Sensitivity & Analyses

In this subsection, we would like to analyze the parameters in RAMs. Firstly, we analyze the performance of uDBN with different λ on the first bootstrapping sample for both datasets in Figure 6. There is a considerable improvement in performance when λ increases from 0 to 0.05, which indicates the importance of fitting the satisfaction annotations. Both Spearman's ρ and Pearson's r will gently rise with the increase of λ. We find that when $\lambda = 0.85$, uDBN achieves the best performance on both datasets. If λ approximates 1, the system will ignore the behavior information, which may raise the risk of overfitting the satisfaction ratings. When click PPL keeps increasing, the metrics may be vulnerable and can hardly be applied in other search scenarios where the distribution of user satisfaction is totally different.

To further investigate the interpretability of RAMs, we visualize the distribution of the learned β_k (denotes the linear correlation between metric scores and true satisfaction ratings), γ_k (represents users' patience to continue viewing the next result to some extent), and σ_{Rk} (the satisfaction probability given a result has been clicked) on all bootstrapping samples. As shown in Figure 7(a), there exist differences in the distribution of β_k across various intents. It is clear that the β_k values in the "New Topic" intent are notably higher than



(a) TianGong-Qref. (b) TianGong-SS-FSD.

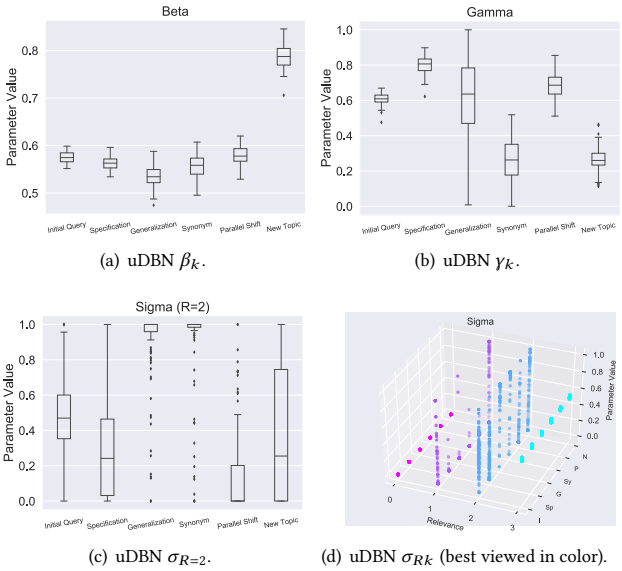
Figure 6: Parameter sensitivity of λ in uDBN on the first bootstrapping sample for both datasets.

those in other intent-level reformulation types. This shows that users' perceived satisfaction is highly correlated with the scores generated by click model-based metrics when their intents have shifted. By contrast, if one generalizes her query, she will perceive a stable level of satisfaction, i.e., with a smaller deviation. Search users who submitted a generalized query may have lower expectations for the SERP and hence can be easily satisfied. As for γ_k, there are huge differences across various intents. We can observe relatively higher values in "Specification", "Initial Query" and "Parallel Shift" types and low values in "Synonym" and "New Topic" conditions. The variance in the "Generalization" type is larger, indicating that users' patience may vary greatly across query cases if their search intent is broader. They may find that the current query is not appropriate and reformulate it very soon, while they may also continue exploiting the page to discover more useful information. Previous work [7] has summarized users' search process into a two-phase process: specialization→intent shift. Combining this rule with the distributions in Figure 7(b), users' patience may first increase at the beginning of a search session and then drop if they want to search for something new. Finally, we find that the averaged σ values are higher for larger relevance R (Figure 7(d)), which accords with our expectation. For $R = 1$, the values are higher in "Initial Query (I)" and "New Topic (N)" conditions where user expectations can be relatively lower. However, users with the same or generalized intents will be highly satisfied with a clicked document when $R = 2$ (as presented in Figure 7(c)).

All in all, the learned parameters are reasonable. The differences of the distribution for each parameter across various intents have also demonstrated the importance of modeling the intents behind each query reformulating action.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel group of Reformulation-Aware Metrics (RAMs) to enhance the evaluation of any search scenarios where query history is available. To the best of our knowledge, this is the first work that directly incorporates users' query reformulation behavior into web search evaluation metrics. For **RQ1**, we investigated a public field study dataset and found that user reformulations are closely coupled with their instant intents and perceived satisfaction. To introduce this factor into evaluation (**RQ2**), we inherited the framework of click model-based metrics and constructed a group of metrics that takes user reformulations as the surrogate to model users' various intents. Through experimental results on two public datasets, we further answer **RQ3**



(a) uDBN β_k. (b) uDBN γ_k.

(c) uDBN $\sigma_{R=2}$. (d) uDBN σ_{Rk} (best viewed in color).

Figure 7: Distribution of learned parameters of uDBN on the TianGong-Qref dataset.

that RAMs can not only learn parameters automatically but also have high robustness in terms of the transfer application and few-label learning. Extensive experiments conducted on two session datasets have shown that RAMs significantly outperform other state-of-the-art metrics in satisfaction estimation.

Our work may provide guidance for further research on designing better effectiveness metrics. Firstly, it may be more appropriate to employ sophisticated approaches such as SGD to better search the parameter space compared to grid search. Secondly, besides query reformulations, there may be other low-cost contextual factors that can expediently characterize various user intents. We can further utilize these factors to model user behavioral patterns or their perceived satisfaction, which is beneficial for constructing evaluation metrics. Last but not least, our experiments reveal that there exist both consistency and contradiction between fitting user behavior and satisfaction. Previous work has found that tuning traditional metrics with C/W/L vectors is effective [39, 42], which shows the consistency between user behavior modeling and satisfaction measurement of evaluation metrics. However, in our study, we have observed the trade-off between fitting user clicks and satisfaction ratings. On the one hand, we conjecture that C/W/L vectors represent the global distributions and may not precisely align with user behaviors. Therefore, the C/W/L framework can be used to tune the metrics but is not appropriate to measure whether a model can predict user behavior accurately. On the other hand, for metrics with high learning power such as RAMs, using a small proportion of human satisfaction labels for model learning will greatly boost the performance. Fitting user behavior is also essential as it guarantees that RAMs will not overfit user satisfaction, ensuring that they can be directly applied in a new dataset once trained.

Our work is only a primary step of considering query reformulating behavior in web search evaluation. In the future, user query

reformulations may be further exploited to construct better session-level evaluation metrics or personalized satisfaction models.

7 ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 61902209, U2001212), Beijing Academy of Artificial Intelligence (BAAI), Tsinghua University Guoqiang Research Institute, Beijing Outstanding Young Scientist Program (NO. BJJWZYJH012019100020098) and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China.

REFERENCES

[1] Azzah Al-Maskari and Mark Sanderson. 2010. A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology* 61, 5 (2010), 859–868.

[2] Leif Azzopardi, Paul Thomas, and Alistair Moffat. 2019. cwl_eval: An evaluation tool for information retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1321–1324.

[3] 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*. 625–634.

[4] 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*. 903–912.

[5] 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*. 621–630.

[6] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*. 1–10.

[7] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a Better Understanding of Query Reformulation Behavior in Web Search. In *Proceedings of The Web Conference 2021*. ACM.

[8] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating query reformulation behavior of search users. In *China Conference on Information Retrieval*. Springer, 39–51.

[9] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. TianGong-ST: a new dataset with large-scale refined real-world web search sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2485–2488.

[10] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services* 7, 3 (2015), 1–115.

[11] 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*. 493–502.

[12] Cyril W Cleverdon, Jack Mills, and E Michael Keen. 1966. Factors determining the performance of indexing systems,(Volume 1: Design). *Cranfield: College of Aeronautics* 28 (1966).

[13] 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*. 87–94.

[14] Georges E Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 331–338.

[15] Dongyi Guan, Sicong Zhang, and Hui Yang. 2013. Utilizing query change for session search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 453–462.

[16] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2019–2028.

[17] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query reformulation in E-commerce search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1319–1328.

[18] Jeff Huang and Efthimis N Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 77–86.

[19] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2009. Patterns of query reformulation during web searching. *Journal of the american society for information science and technology* 60, 7 (2009), 1358–1371.

[20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[21] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2019. From a User Model for Query Sessions to Session Rank Biased Precision (sRBP). In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 109–116.

[22] 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*. 493–502.

[23] Jiyun Luo, Sicong Zhang, and Hui Yang. 2014. Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 587–596.

[24] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 1–38.

[25] 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*. 659–668.

[26] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.

[27] Karl Pearson and Alice Lee. 1900. Mathematical contributions to the theory of evolution. VIII. on the inheritance of characters not capable of exact quantitative measurement. Part I. introductory. Part II. on the inheritance of coat-colour in horses. Part III. on the inheritance of eye-colour in man. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 195 (1900), 79–150.

[28] Filip Radlinski, Martin Szummer, and Nick Craswell. 2010. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*. 1171–1172.

[29] Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics* (1951), 400–407.

[30] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 525–532.

[31] 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*. 473–482.

[32] Mark Sanderson. 2010. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc.

[33] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 555–562.

[34] Philip Sedgwick. 2012. Multiple significance tests: the Bonferroni correction. *Bmj* 344 (2012).

[35] 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*. 95–104.

[36] Charles Spearman. 1961. The proof and measurement of association between two things. (1961).

[37] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge, MA.

[38] Alfan Farizki Wicaksono and Alistair Moffat. 2018. Empirical evidence for search effectiveness models. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1571–1574.

[39] Alfan Farizki Wicaksono and Alistair Moffat. 2020. Metrics, user models, and satisfaction. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 654–662.

[40] 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*. 1561–1564.

[41] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating web search with a bejeweled player model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 425–434.

[42] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models Versus Satisfaction: Towards a Better Understanding of Evaluation Metrics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 379–388.