Discover Computing

Research

Improving searcher struggle detection via the reversal theory

Jiyun Luo[1,4] · **Yan Yang**[2,4] · **Valerie Nayak**[3] · **Grace Hui Yang**[4]

Received: 21 June 2024 / Accepted: 2 December 2024
Published online: 19 December 2024

Abstract
Searcher struggle is important feedback to Web search engines. Existing Web search struggle detection methods rely on effort-based features to identify the struggling moments. Their underlying assumption is that the more effort a user spends, the more struggling the user may be. However, studies have shown that this simple association might be incorrect. This paper proposes a new feature modulation method for struggle detection and refers to the Reversal Theory in psychology. Reversal Theory points out that instead of having a static personality trait, people constantly switch between opposite psychological states, complicating the relationship between the efforts they spend and the level of frustration they feel. Supported by the theory, our method modulates the effort-based features based on Reversal Theory's bi-modal arousal model. After modification, the users' effort level is better aligned with their struggling experience. Evaluations on Pinterest search logs confirm that the proposed method can statistically significantly improve searcher struggle detection methods.

Keywords Web search · Searcher struggle detection · Reversal theory · Information retrieval

1 Introduction

Web searchers can face difficulties ranging from vague search results to the sheer volume of information available online. These struggles in a search session can be attributed to poor query formulation, the challenge of filtering through irrelevant data, and identifying trustworthy sources amidst a sea of content. As web content continues to evolve, being adept in search methodologies is essential for efficient and effective information retrieval. Detecting instances of searcher struggle is an important task for web search engines, as it enables them to modify their algorithms promptly, thereby mitigating user difficulties in subsequent searches. Detecting instances of user struggle in sessions that involve multiple rounds of user-machine interactions is also important to intelligent assistants such as ChatGPT [1–3] and other advanced dialogue systems. [4]

In this paper, we define *searcher struggle* as an event in which a web searcher exerts significant effort to overcome challenges while trying to fulfill a search task within a session. Such a struggle is characterized by the user's experience of negative emotions, such as frustration, upset, or annoyance, during the search activity. We treat the detection of searcher struggle as a binary classification problem, distinguishing between *struggling* and *non-struggling* based on the presence or absence of such stressful events in a search session.

✉ Jiyun Luo, jluo@pinterest.com; ✉ Grace Hui Yang, grace.yang@georgetown.edu; Yan Yang, yy490@georgetown.edu; Valerie Nayak, vjn@andrew.cmu.edu | [1]Pinterest Inc, 651 Brannan St, San Francisco 94107, CA, USA. [2]Department of Computer Science, University of Nevada, Reno, 1664 N Virginia St, Reno 89557, NV, USA. [3]School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh 15213, PA, USA. [4]InfoSense, Department of Computer Science, Georgetown University, 37 and O Streets Northwest, Washington 20057, DC, USA.

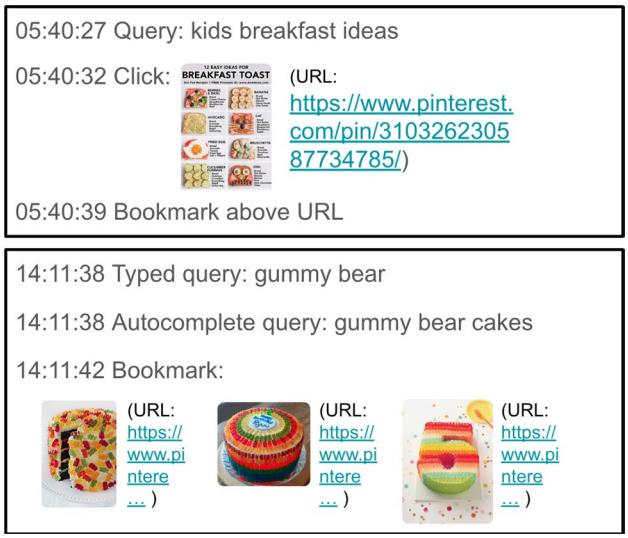

Fig. 1 Two example search sessions with no searcher struggle

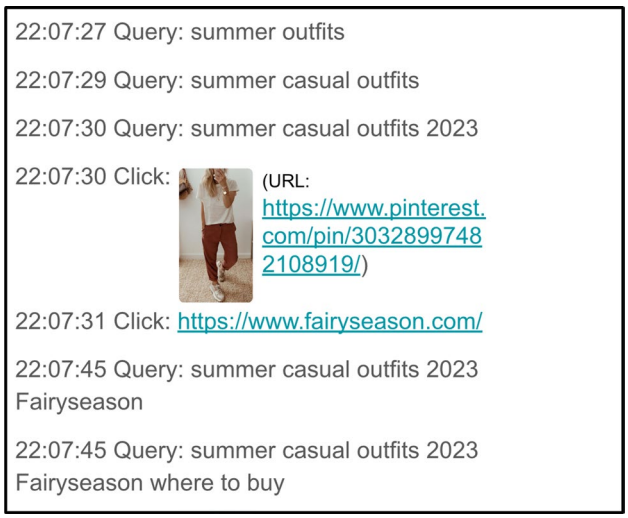

Fig. 2 One example search session with searcher struggles

Examples of non-struggling and struggling search sessions are illustrated in Figs. 1 and 2, respectively. Figure 1 depicts two seamless search sessions. In the first, the user efficiently searches for kids' breakfast recipes, quickly finding and bookmarking a desirable image. In the second session, an auto-complete suggestion "gummy bear cakes" perfectly meets the user's needs, leading to the discovery and bookmarking of three relevant recipes without any hassle. These examples demonstrate smooth search experiences free from struggle.

Conversely, Fig. 2 presents a session fraught with difficulty. The user's initial search for "summer outfits" requires multiple query reformulations before landing on an appealing image. The subsequent realization that the image is linked to the Fairyseason brand does not ease the user's struggle, as further efforts to find and purchase the product on the brand's website are unsuccessful. This session exemplifies the kind of struggle that users may encounter, characterized by repeated attempts and unmet information needs.

The majority of current methods for detecting searcher struggle employ supervised classification techniques, utilizing features indicative of user effort [5, 6]. This approach is based on the understanding that struggle is frequently manifested by an excessive repetition of actions by the user. For example, in an analysis of Pinterest search logs, it was noted that a user executed over twenty queries in the pursuit of "curly hair dye" without arriving at a satisfying outcome. This behavior typifies the kind of repetitive action patterns that are commonly interpreted as indicators of struggle.

While it might appear logical to presume that an extensive amount of user effort indicates struggle, this assumption does not consistently hold true in actual practice. It has been observed that even sessions without struggle can involve a significant number of queries and clicks by the user. For example, in a session where a user looked up

"guy proposing ideas" on Pinterest, the search log recorded over 900 results reviewed and 27 items bookmarked. Despite the high level of activity, the user reported enjoying the search experience throughout the session via a pulse survey. This indicates that a large amount of effort does not necessarily correlate with the presence of struggle, as both struggling and non-struggling sessions may exhibit high user engagement.

Laboratory studies have echoed these observations, indicating that high user engagement does not necessarily equate to struggle. Edwards and Kelly [7] have highlighted the complexity in interpreting increased user effort during web searches. They argue that while increased activity could indicate struggle, it might also represent positive engagement or exploratory behavior. An engaged user, similar to a frustrated one, is likely to issue multiple queries and click on numerous links related to the same topic. This overlap suggests that without a nuanced method of interpretation, effort-based metrics alone may not sufficiently differentiate between negative experiences of struggle and positive experiences of engagement, presenting a challenge in enhancing web user experiences.

This paper turns to the psychology literature to adopt a more nuanced approach to the problem. We propose a novel feature modulation technique for detecting searcher struggle, drawing inspiration from Reversal Theory. [8–11] This theoretical framework informs our method, suggesting that the same user behaviors can signify different emotional states in different contexts, thus providing a more refined analysis of user engagement and struggle. Reversal Theory is a *mode-based* psychological framework that questions established beliefs in motivation and personality studies. It emphasizes the complexity and variability of human behavior, suggesting that personality traits and motivations are dynamic and can shift in response to different situations. [11] A key proposition of this theory is that individuals' behaviors are fluid, and motivations can *reverse* in the everyday flow of life. [10] This can lead to surprising behavioral changes, such as a typically selfish person acting unselfishly in certain contexts. The theory posits that factors like conforming to rules or rebelling have less influence on whether a person struggles during a search task than the nature of the task itself, whether it's approached with seriousness or playfulness. By applying statistical hypothesis testing to Pinterest search logs, our research has validated these claims, confirming the relevance of task nature to the experience of struggle.

Reversal Theory posits that human motivations are not linear but multidimensional, with dimensions such as "means-ends," "rules," "transactions," and "relationships," each containing two diametrically opposed states. At any given moment, an individual operates within one state of a dimension, suggesting motivations shift in a bi-modal rather than a uni-modal pattern. Our paper harnesses this theoretical insight to develop an innovative feature modulation approach for detecting struggle within search sessions.

In applying this method, we first isolate features that correspond to the dimensions of Reversal Theory. We then adjust these features to address the inherent bias between the two states in the bi-modal arousal model. Through this refinement, we better correlate the user's level of effort with their actual experience of struggle. The final step involves feeding these adjusted features into classification models to discern the presence of struggle within a search session. This process not only enhances the accuracy of struggle detection but also aligns with the dynamic nature of user motivation as described by the Reversal Theory. Our approach is designed to complement any feature-based struggle detection method, adding a layer of nuance by aligning with the dynamic motivations captured in Reversal Theory. It serves as an augmentation that can refine and enhance the predictive power of existing struggle detection models by providing a more sophisticated interpretation of user behavior.

We assessed our methodology using Pinterest search logs, encompassing data from both mobile and desktop platforms. The evaluation compared the performance of several leading struggle detection methods, with and without the integration of our proposed feature modulation technique. The results of these experiments were highly favorable, demonstrating that our method substantially enhances these top-performing methods-yielding an approximate increase of 5% in accuracy and 9% in precision for the prediction of struggling search sessions.

The remainder of this paper is organized as follows. Section 2 describes the related work. Section 3 defines the research problem and categorizes features that are based on user efforts. Section 4 details the Reversal Theory and how it can be used for Web search. Section 5 presents our proposed method to modulate the features. Sections 6 and 7 describe our experiment setups and experimental results. Lastly, Sect. 8 concludes the paper.

2 Related work

In this section, we review related information retrieval work from searcher struggle detection field.

2.1 Web searcher struggle detection

Studies on searcher struggles can be grouped into (1) laboratory studies and (2) query log studies. Both types of studies look for meaningful relationships between searcher struggle and their search behaviors.

2.1.1 Lab studies

The lab studies on searcher struggle monitor a user's entire search process in a laboratory setting and collect explicit user feedback via questionnaires. They ask a user if they are experiencing a struggling moment during a search session and study interesting behavior patterns when the struggle happens.

For instance, Aula et al. [12] found that when encountering a struggle, a user tends to (a) formulate question-like queries, (b) use advanced search operators, (c) spend more time examining search results, (d) be more likely to write the most extended query in the middle of a search session if the search eventually fails and (e) at the end of the session if the search succeeds. They also suggested that task difficulty may lead to user struggles.

These lab studies can go to great lengths to investigate searcher struggles; however, constrained by monetary costs, they usually only perform with small groups of users and the limited number of search tasks. To address the challenges of scalability in lab studies, Xu et al. proposed a method to generate struggling search tasks by leveraging crowd sourcing and identifying paraphrased sentences [13]. They published 80 struggling search tasks using this approach, although it is still a relatively small dataset when compared to the data used in log-based studies.

A highly relevant lab study to ours is Edwards and Kelly's work [7]. They also observed that although the increase of user efforts might help predict searcher struggles, such increase can also indicate engagement, the opposite of struggles. The prior work mentioned above give us ideas on how to develop features that can measure user effort during search. However, these previous studies were unable to explain the discrepancy we observed between the user's effort level and their struggling experience.

Xu et al.'s lab study [14] suggested that searcher struggles are related to the user's mood. When users are irritated or excited, they tend to issue more queries than in neutral moods. This aligns with the classic single-modality arousal model in psychology. However, the work did not distinguish between negative emotions and positive emotions, leading to more queries. What is different is that our work uses a bi-modal arousal model and our focus is on feature transformation.

2.1.2 Log-based studies

Search log based studies on searcher struggles are quite popular. They record a user's search process in search logs and analyze the historical data to understand the searcher's behaviors and how they relate to struggles. Usually, a struggling event is labeled afterward by third-party annotators. Log-based studies can be large-scale and support automatic detection of searcher struggles. Most methods derive helpful features from the logs and use regressors or classifiers to detect the struggles.

For instance, Hassan et al. [15] worked on detecting *struggling* and *exploring* (including *being both exploring and struggling*) search sessions. Their effort-based features included the number of unique queries, term additions, removals and substitutions, clicks, and dwell time. They reported accuracy of 81.67% for detecting *struggling* sessions. They also acknowledged that a user behaves similarly when exploring and struggling; the search logs for both types of sessions are "similar in terms of the number of queries and the session duration." This is similar to the insight we learned from the Reversal Theory that the same user behavior can happen at different states. However, their focus was on finding new features, such as query transitions and result clicks, that can help distinguish the subtle difference between exploring and struggling sessions; while ours is on new ways to interpret and re-use existing features.

Li et al. [5] studied good abandonment, which is relevant to the absence of searcher struggles. Good abandonment happens when a user abandons her search before clicking any results as the content on the SERP has met the information need. When good abandonment happens, a user's effort is minimal, and struggle is absent. They also reported the important role of search topics in determining good abandonment, which is investigated in our paper as significant features in the "means-ends" motivational dimension.

Feild et al. [16] compared features derived from query logs and physical sensors. They found that using log-generated features is reliable and more effective than using sensor-generated features in detecting searcher struggles.

Our work belongs to the log-based studies. Although we use many prior features in [15, 17], their work primarily concentrates on engineering features to enhance searcher struggle detection performance. Despite acknowledging the discrepancy between high user effort and user struggles, they did not attempt to comprehend and model this phenomenon. In contrast, our work employs a bi-modal arousal model to capture the psychological factors contributing to this inconsistency, ultimately proposing a novel method to modulate these features for more effective struggle detection.

Other well-studied, negative search experiences besides struggles include irrelevancy and dissatisfaction [16]. Note that these concepts are related to struggles but not interchangeable. For instance, dissatisfaction occurs after a search task when a user has not found satisfactory information from the search results. On the other hand, struggles can occur anytime during a session, as soon as the results are frustrating. Even if a user is satisfied at the end of a session, she may still experience struggles during it. Our paper only studies struggles.

2.2 Struggles vs. irrelevancy vs. dissatisfaction

User struggles [12, 15, 18], search results irrelevancy, and user dissatisfaction [16] are all negative search experiences for a user. Although they are all related to negative search experiences that search engines want to detect and avoid, they are different. Our research specifically focuses on determining if a user is *struggling*. Studies about results irrelevancy, user dissatisfaction are not within the scope of this paper.

Search success and user satisfaction

Concepts that are oppositely related to user struggles, such as user satisfaction and search success, have also been extensively studied. Search success has been interpreted as content relevance [19, 20], fulfillment of information need in [21–23] and the searcher experience of pleasure in [6, 24]. Fox et al. [25] built predictive models using a search log gathered from daily search activities of 146 Microsoft employees and revealed that combining click-through, dwell time, and session termination could predict user satisfaction about a SERP page or a search session well. Through a lab study, Huffman and Hochster [24] found that session satisfaction was related to how relevant the first three results of the first query were, whether the information needed was navigational, and how active the user was in the session. By analyzing annotated data using crowdsourcing, Verma et al. [26] concluded that user satisfaction is related to the relevance of examined web pages and the effort needed to locate the relevant content in these web pages. Jiang et al. [6] found that user satisfaction changed within a session by analyzing a commercial search engine log and suggested that predicting satisfaction should be done at different grades. By utilizing click, query, and query transition features, Wang et al. [23] could predict search session success with high accuracy. Hassan et al. [22] studied search success at the query level. They pointed out that query-based signals can predict search success more accurately than click-based signals. We hypothesize that signals that are good indicators of search satisfaction and search success should also influence predicting struggles, which motivates us to include those signals in our framework.

2.3 Searcher struggle detection in mobile search

Most of the previously mentioned work has primarily focused on desktop platforms. However, it is vital to consider both desktop and mobile platforms since they are both significant for Pinterest. These two platforms exhibit certain distinct user behaviors attributed to hardware and UI disparities. Our work aims to tackle the challenges of struggle detection on both platforms, which is why we also incorporate prior research on detecting struggles in mobile search. Guo et al. [27] conducted lab studies and provided a predictive model for detecting URL relevance in mobile search. They revealed that users' inactivity indicated they are reading in mobile search, but not in desktop search. They also found that swiping was similar to scrolling on desktops in that both were signals that suggest content irrelevance. Han et al. [28] found that mobile touch interaction signals on SERP were more effective than landing web page signals for predicting content relevance. Lagun et al. [29] show that scrolling past search cards and spending more time on contents below search cards are clear signals of non-relevance. Huang and Diriye [30] pointed out that changing viewport coordinates are more accurate than user touch coordinates for predicting content relevance in mobile search. Kim et al. [31] provided vertical scrolling and horizontal pagination functions to web searchers in a study. They found that searchers found relevant content faster by using pagination than scrolling due to the time taken for the scroll itself.

3 Problem formulation

3.1 The task of searcher struggle detection

Searcher struggle refers to a challenging event within a search session where the individual conducting the search (the user) experiences frustration stemming from difficulties encountered during the search process. We define the task of detecting searcher struggle as a binary classification issue, where the two categories are defined as *struggling* and *non-struggling*, which corresponds to any struggling event/moment is present or absent in a search session. This approach aligns with the framework used in numerous previous studies. [15–17]

In this paper, we select a single search session as the unit for investigating the occurrence of searcher struggles. This decision is based on two key considerations: (1) a search session naturally encapsulates the entire search task, and (2) it yields more consistent responses than analyzing each user action individually. Instead of employing the conventional approach of isolating search sessions based on 30-min user inactivity, we segment search sessions from query logs based on topical coherency, utilizing the algorithm proposed by [32] (more details are provided in Sect. 6). Prior research [33] has demonstrated that studying topic-based search sessions provides better insights into user behavioral patterns compared to studying time-based search sessions.

In our study, we define a searcher struggle predictor Y for any given search session s that the searcher is in, which is characterized by a feature vector $X(s)$. The probability that session s includes a struggling moment is denoted by $P(Y = 1|X(s), \Theta)$, where Θ represents the parameters of the model. A search session is classified as *struggling* if $P(Y = 1|X(s), \Theta) > c$ exceeds a certain threshold c, and as *non-struggling* if the probability is below this threshold. Our experimental findings indicate that setting c to 0.5 yields the most accurate predictions, optimizing the F1 score, which balances precision and recall for positive instances of struggle. Further details of our experiments with multiple classifiers and the impact of our feature modulation approach are discussed in Sect. 7, with class labels provided by independent manual annotation as described in Sect. 6).

3.2 Features

The input feature vector $X(s)$ used in this study is derived automatically from the query logs. It comprises both previously proposed features from prior research and new features introduced in this study. The majority of these features serve as indicators of user effort, quantifying the extent and variety of user actions within a session, including the time spent reviewing and evaluating search results. These effort-based features are categorized into seven distinct groups, as detailed in Table 1.

Efforts to Query. The first feature group in our study quantifies the effort users expend in formulating queries. This set of features is largely based on the findings of Edwards and Kelly [7], which suggest that a high number of queries within a session could indicate substantial user effort in query composition. However, the mere quantity of queries may not always reflect effort accurately, as users might copy-paste the redundant queries, which is less labor-intensive. To account for the varying levels of effort, we analyze the proportion of queries that are either copied, pasted, or system-generated against the total number of queries in a session. Additionally, we consider the count of manually typed queries as a direct metric of user labor in querying. We also introduce a novel feature: the sequence position of the longest query within a session. This is informed by research from Aula et al. [12], which observed that the most extensive query typically marks the conclusion of a successful search session. Our inclusion of this feature aims to refine the understanding of user effort by considering both the quantity and the strategic placement of queries during a search session.

Efforts to Click. The second group of features assesses the user's effort in interacting with search results through clicks. Traditional metrics in this category include the total number of clicks and clicks that satisfy user needs (SAT clicks), which serve as proxies for gauging relevance, satisfaction [19, 24, 25], and even struggle [15, 16]. Building on this foundation, we introduce new features specific to the multi-modality search environment seen on modern platforms such as Pinterest. These features track user clicks across various search result types, such as images, advertisements, and general web pages, acknowledging the diverse content interactions on such platforms. Additionally, we identify features that signal low effort or abandonment, such as the average number of consecutive queries without a follow-up click. This aspect of user behavior, termed *good abandonment*, is particularly relevant on visual search platforms [5, 34] where users may find what they're looking for without needing to click further, as studied by Chilton and Li. We further examine the bookmarking of results and the timing of clicks within a session. The underlying rationale is that users typically show less struggle

Table 1 Effort-Based Features

Efforts to Query
Number of unique and total queries in a session [7]
Avg. number of terms per query [7]
Avg. number of characters per query [7]
Number of manually-typed queries*
Percentage of manually-typed queries [7]
Percentage of suggested queries (that are automatically corrected, suggested, or completed by the search engine) [7]
The longest query's position in a session*
Efforts to Click
Total and avg. number of clicks in [15, 16]
Total and avg. number of Satisfactory (SAT) clicks [15, 16]
Percentage of queries without clicks [19, 24, 25]
Maximum and avg. number of adjacent queries without clicks*
Total and avg. number of images clicked in a session*
Total and avg. number of ads clicked in a session*
Total and avg. number of bookmarks clicked in a session*
Number of events (clicks, bookmarks, and queries) in a session*
Number of clicks at the first two queries*
Number of clicks at the third and fourth queries*
Number of clicks at the fifth and sixth queries*
Whether the session ends with a click*
Efforts to Read
Total dwell time of all clicks [15, 16]
Avg. number of image impressions per SERP*
Total number of zoom-in on result images*†
Log (1 + avg. dwell time per click in a session)*
Log (1 + avg. dwell time per click exclude clicks for the last query)*
Log (1 + time passed until the first SAT click)*
Log (1 + avg. time spent on each SERP in a session)*
Log (1 + avg. time spent on each SERP exclude the last query)*
Efforts to Scroll
Screen size*
Total and avg. number of scrolling down actions*
Efforts to Re-formulate Queries
Avg. cosine similarity between every query and the first query [15]
Avg. cosine similarity of every query pair in a session [15]
Avg. edit distance per adjacent query pair [15]
Number of query generations (when removing one or more terms from its previous query) [15]
Number of query specifications (one or more terms are added into its previous query [15]
Difference between the first query length and the avg. query length*
Standard deviation of query lengths in a session*
Avg. number of terms appear in the previous query [15]
Avg. number of terms added to the previous query [15]
Avg. number of terms deleted from the previous query [15]
Avg. number of terms that substitute terms in the previous query [15]
Efforts to Diversify
Percentage of unique URLs among all clicked URLs [15]
Percentage of the unique domain (DNS) names among all clicked URLs [15]
Total number of unique clicks*

Table 1 (continued)

Total number of unique topics [15]
Entropy of topic distribution in a session [15]
Efforts to Issue Rare Queries & Rare Clicks
Log (1 + avg. query frequency in popularity data) [15, 17, 24]
Log (1 + a query's avg. SAT clicks in popularity data) [15, 17, 24]
Log (1 + a query's avg. clicks in popularity data) [15, 17, 24]
A query's avg. click entropy in the popularity data [15, 17, 24]
Log (1 + a query's avg. number of fast-back clicks (whose dwell time is less than 15 s) in the popularity data) [15, 17, 24]
Log(1 + a clicked URL's avg. click frequency in the popularity data)*

*Marks the new features that we have added, which differentiate them from the other features used in previous related works. † marks the features that we only use on the mobile platform

when they find and interact with content early in their search session as opposed to later stages, which often indicates a smoother search experience.

Efforts to Read. The third feature group quantifies the effort users invest in reading and examining the contents of search results [15, 16]. Beyond the conventional dwell time metric, we propose to encompass the tallying of zoom-in actions on image results, which offers a direct indication of user effort in reading and examining visual content. Moreover, we distinguish between types of dwell time across various returned items and sections of the search results page (SERP). This differentiation enables a more granular understanding of user behavior, capturing both the overarching browsing activity and the focused attention given to specific search result items. Such detailed analysis can reveal the extent of reading efforts, which is vital for distinguishing between mere skimming of the SERP and in-depth examination of individual results.

Efforts to Scroll. The fourth feature group is focused on capturing the user's effort in navigating to search results that are not immediately visible on the screen. This set of features is entirely novel and is designed to measure how often a user scrolls down or resizes the screen to view additional content. Such actions typically trigger a new search request to the backend engine to retrieve more, and often fresher, search results for the current query. We quantify these efforts by tracking the number of pagination requests made by the user, which serves as a proxy for the number of scroll-downs and screen-resizings.

Efforts to Reformulate Queries. The fifth feature group assesses the amount of work a user puts into refining their search queries to better articulate their information needs. This involves measuring changes in query formulation, such as the variance in query length after edits, which can be indicative of the complexity or evolution of the user's search intent. Frequent query reformulations point to possible ambiguity in the initial information need or suggest that the user's information need is developing and becoming more complex over the course of the search session.

Efforts to Diversify. The sixth feature group evaluates the extent to which users seek diversity in their search results and the effort with which they inspect these results. This category includes metrics for click diversity and topical diversity, drawing on features primarily identified by Hassan et al. [15] to determine the breadth of exploration in user search behavior. A novel addition to this group is the measurement of the total number of unique clicks, which further underscores the user's exploratory efforts. By analyzing these features, we can infer how much users are branching out to consider a wide range of information, rather than focusing narrowly on a single thread of search results.

Efforts to Issue Rare Query & Rare Clicks. The seventh group of features measures user efforts spent on critical thinking and being novel and unique. They include rare queries and rare clicks that a user would create in a session compared to the large Web population who have the exact or similar information need. The idea is that issuing popular queries, like most others, requires fewer efforts, while giving a rare query requires more thinking efforts. Likewise, clicking on unpopular URLs is also an indicator of critical thinking [35]. We obtain the Web population's click data from Pinterest from 11/15/2020 to 11/21/2020 and use that as the basis to derive which queries and clicks are rare.

Efforts to Issue Rare Query & Rare Clicks. The seventh group of features quantifies the effort users put into devising unconventional queries and selecting less common search results, reflecting a higher level of critical thinking and originality in their search behavior. This feature set captures the uniqueness of a user's queries and clicks by comparing them with the actions of a broader web population with similar information needs. The rationale is that common queries and clicks require less cognitive effort, whereas rare ones suggest a deeper level of mental effort and individual thought

Table 2 Motivation Dimensions and States

Means-ends	
Telic	Paratelic
• *Serious. Focus on future goals and achievement. Tend to avoid arousal, risk & anxiety.*	• *Playful, passion and fun. Focus on current moment. Seek excitement and entertainment.*
Rules	
Conformist	Negativistic
• *Conforming. Value rules and tradition. Tend to operate within rules and expectations.*	• *Rebellious. Value innovation and changes. Like to explore new possibilities.*
Transaction	
Mastery	Sympathy
• *One wants to be in control, whether this be over people, tasks, ideas, machinery or anything else that one can interact with.*	• *Wanting to develop close and nurturing relationships, to be tender and sensitive.*
Relationships	
Autic	Alloic
• *Doing things for self rather than for others.*	• *Genuinely concerned with others, and putting them first.*

process [35]. By analyzing data on popular user behavior on Pinterest from November 15 to 21, 2020, we can identify which user actions are considered rare within the context of the larger web community.

In addition to the seven groups of effort-based features, our analysis includes non-effort-related characteristics to provide a more rounded understanding of search behavior. An example of such a feature is the taxonomy topic of the search task, which is used as a categorical variable. These additional features offer context that can significantly influence search behavior and the interpretation of effort. However, the primary focus and application of our method lie in the modulation of effort-based features, as detailed in Table 1. This comprehensive approach allows us to address various facets of search behaviors to enhance the accuracy of struggle detection.

4 Reversal theory and how it can affect search

4.1 Introduction to reversal theory

Reversal Theory [36] is a prominent psychological theory that was primarily developed by British psychologist Dr. Michael J. Apter, in collaboration with psychiatrist Dr. Ken Smith, since its establishment in the mid-1970s. The theory has garnered significant recognition and has been extensively researched, leading to the publication of numerous empirical papers that either test or utilize its concepts. Additionally, Reversal Theory has spawned over twenty books, the establishment of its own journal, the creation of several standardized questionnaires, and the adoption of various training techniques in multiple countries [37]. It studies personality dynamics and motivations. It recognizes that people "are essentially changeable and move between different motivational styles" [9, 38]. This theory "sheds light on the paradoxes of risk-taking, addiction, rebelliousness, and other areas of motivation, emotion, and personality" [9].

Reversal Theory is built around several key ideas that distinguish it from other psychological theories and are the following.

1. In everyday life, people's motivations can be organized along a few dimensions. They include "means-ends," "rules," "transactions," and "relationships."
2. Each dimension consists of a pair of opposing states.[1] Table 2 lists the two opposing states in each of the four Reversal Theory dimensions.
3. A person can only be at one of the two states at any given moment.
4. A person can reverse between the pair of motivational states.

[1] Some books call these states "meta-motivational states," "motivational styles," or motives. For simplicity, we call them motivational states or states in this paper.

5. Although each person has their *dominating* states, i.e., they have a preference to stay more often in a state when in a non-dominant state, people follow the current state to the same extent as they are at the dominating state.

Reversal Theory provides an alternative to traditional personality and motivation models by suggesting that human behaviors and emotions are fluid rather than fixed [9]. This concept is a departure from the idea of stable personality traits, positing instead that people can experience rapid transitions between emotional states-like anxiety and excitement-depending on their current motivations and perceptions. For example, Reversal Theory posits that work can be seen as either an obligation or a source of enjoyment, based on the individual's current state [39]. This dynamic view allows for a more nuanced understanding of job satisfaction and employee motivation.

The theory also revises the model of arousal by proposing a bi-modal approach as opposed to the traditional uni-modal one [40]. It acknowledges that happiness (hedonic level) and effort (arousal level) are not always directly correlated. Sometimes, high effort may be associated with pleasure rather than stress, depending on the individual's state and context.

In the context of user struggle detection in Web search, the Reversal Theory's bi-modal arousal model offers a framework for understanding why higher effort does not necessarily indicate frustration or struggle. By applying this model, we adjust feature assessment to better align with the true nature of user experiences, enabling more accurate identification of struggling and non-struggling sessions based on user behavior data.

In this paper, we focus on the first two dimensions of Reversal Theory—"means-ends" and "rules"—which elucidate the methods and principles guiding user task performance. These dimensions are pertinent to our analysis and are therefore examined in detail. The latter two dimensions, "transactions" and "relationships," deal primarily with interpersonal interactions and are beyond the scope of this discussion, as they do not directly pertain to the interaction between users and tasks. These dimensions are not addressed in this paper.

4.2 Opposing motivational states

Reversal Theory groups human motivations into four dimensions (also known as domains). They are "means-ends," "rules," "transactions," and "relationships" [39]. We can find the dimensions and states in Table 2. The first two dimensions describe how a user performs tasks and will be the focus of this paper.

The "means-ends" dimension of Reversal Theory addresses the interplay between goal attainment and the pleasure derived from participating in a process. It encapsulates two diametrically opposed motivational states that represent the underlying motivations and emotions individuals associate with activities at any given moment. These are the *telic* state, where focus is placed on accomplishing objectives and completing tasks, and the *paratelic* state, where the pursuit of enjoyment and fun prevails. In the telic state, a person's actions are goal-oriented, with a serious dedication to task completion. In contrast, in the paratelic state, the activity itself, independent of any end goal, is the source of satisfaction; for instance, someone may run simply for the joy of running rather than the competitive goal of winning a race.

In the context of web search, the concepts of *telic* and *paratelic* states correspond to goal-oriented and non-goal-oriented search behaviors, respectively. A user in a telic state is driven by definitive objectives, such as finding job openings or seeking medical advice. Conversely, a user in a paratelic state engages in the search activity for pleasure, exemplified by leisurely perusing entertaining videos on YouTube. This distinction emphasizes the varying intentions and experiences that users bring to their search activities.

The "rules" dimension explores the influence of routines, expectations, and constraints on individual behavior. It features two contrasting states: *conformist*, where a person's actions align with established rules and expectations, and *negativistic*, where an individual is inclined to challenge conventions and explore new possibilities. For instance, a conformist attitude is evident in the thought, "I am eating because it is the appropriate thing to do at this moment." In contrast, a negativistic perspective might be, "I am eating precisely because I am not supposed to eat at this time." This intriguing aspect of human behavior highlights how identical actions can stem from opposite motivations.

Within the scope of web search, the states of *conformist* and *negativistic* reflect non-exploratory and exploratory search behaviors, respectively. A user exhibiting conformist behavior adheres to established guidelines and fulfills external expectations, such as using search engine suggestions rather than crafting unique queries. Conversely, a user in a negativistic state actively seeks out new ideas and experiences, evident in the use of uncommon queries, sourcing information from a variety of URLs, and displaying a preference for novel and diverse search results.

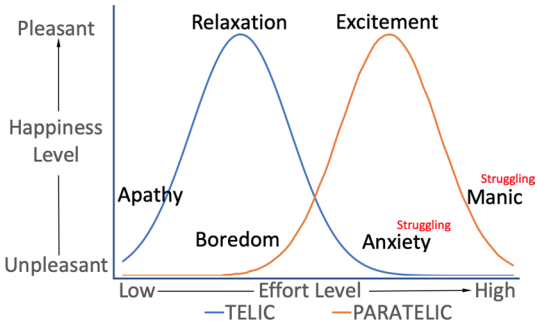

Fig. 3 Arousal Model for Means-ends; Both Anxiety and Manic indicate struggling. (adapted from [41])

Additionally, the Reversal Theory posits that individuals "reverse" back-and-force between the two opposite states in the same motivational dimension. Instead of being at a static state like having an enduring personality trait, a person teeter-totters in her motivational states and the states are completely opposite to each other.

4.3 Reversal theory's bi-modal arousal model

Searcher struggle detection aims to discern the searcher's (un)happiness level as they interact with search engines. The *model of arousal* [40], a psychological concept, tells the interplay between an individual's (un)happiness level and their degree of arousal, which reflects the intensity of activities and feelings a person experiences. In the context of web search, arousal corresponds to the intensity and effort invested in search activities. Thus, the arousal model provides a valuable framework for understanding the correlation between a user's effort and their emotional state during the search process.

The traditional model of arousal in psychology is a single-modality model. It suggests that as the arousal level increases, a single optimal arousal level exists to reach the happiest moment [40]. For instance, there is an optimal usage level of air-conditioning to feel the most comfortable; too much or too little would both reduce a person's happiness. It suggests an inverted U shape or a Gaussian distribution. However, this model cannot capture extreme happiness caused by intense arousal, e.g., riding a roller-coaster. It can neither capture that people experience a high level of happiness with low arousal, e.g., being calm and happy after completing a significant project.

On the contrary, in Reversal Theory, the arousal model is a bi-modality model. Reversal Theory assumes that there are two optimums present in the arousal model. Each of them is for one of the two opposite states within a motivational dimension. The model takes the shape of two inverted U-curves or two Gaussian distributions crossing.

Figure 3 [41] illustrates this bi-modal arousal model for the means-ends dimension. Here the x-axis is effort, and the y-axis is happiness. A low happiness level indicates negative feelings. Among the negative emotions there are apathy, boredom, anxiety, and manic. Both "anxiety" and "manic" happen when efforts are substantial, and happiness is low. In this paper, we consider both of them are struggling and do not distinguish them further. On the graph, the two curves each represent one of the two states, telic or paratelic. We can see the two states peak at different effort levels – the telic curve peaks early when a moderate amount of effort happens; while the paratelic curve peaks late after a significant amount of effort is present.

In the context of Web search, this bi-modal arousal model could be the cause of inconsistent predictions regarding user struggles based on user effort levels. The reason for this inconsistency is that the same level of user effort can correspond to two different levels of happiness, depending on the user's current state. For example, the same effort level may indicate "struggling/anxiety" for a user in the telic state and "excitement" for a user in the paratelic state. To address this inconsistency, we propose shifting the two state curves horizontally closer to each other until they overlap. This ensures that the struggling instances always fall on the right end of the curve, thereby effectively aligning user struggles with higher levels of user effort.

4.4 "Rules" dimension is irrelevant

As we mentioned before, the first two Reversal Theory dimensions are seemingly relevant to Web search because they care about users and tasks. However, contrary to our intuition, the Reversal Theory suggests that the "rules" dimension has little impact on struggling, and only the "means-ends" dimension matters. It [9]'s interplay of the first two dimensions (Fig. 4). The two sub-figures in Fig. 4 depicts the Reversal Theory's arousal model when the second dimension state is conformist and negativistic, respectively. We notice that in both sub-figures, struggles happen at

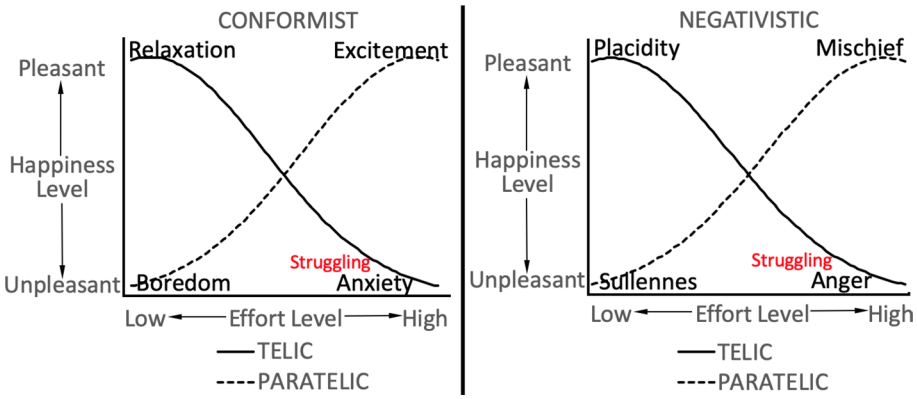

Fig. 4 Interplay of "means-ends" and "rules" (adapted from [9])

the same effort level, which suggests that whether the user is conformist or negativistic has little impact on determining struggles.

To confirm this, we conducted two MANOVA hypothesis tests, one for the first Reversal Theory dimension (described in this section) and another for the second (described in Sect. 5), on a training dataset of a whole week's Pinterest query log (collected from 11/08/2020 to 11/14/2020, one week before the time window that we used for the test dataset.) We segment the sessions following [42] into topically coherent segments [32] in the same way as we did for the testing dataset (more details refer to Sect. 6.1). For the "rules" dimension, we make the following hypotheses:

H_0: The "rules" dimension is irrelevant to a user's happy level. In other words, there is no statistically significant difference in the average effort level from users at the conformist state and users at the negativistic state.

H_1: The "rules" dimension is relevant to a user's happy level. The average effort spent by users in the conformist state differs from that spent in the negativistic state.

We carry out the hypothesis test in the following steps. First, we sort all search sessions in the query log-based on an *ExploreScore*. We define the *ExploreScore*; it is the average score of features in the "efforts to diversity" and "efforts to issue rare queries and clicks" feature groups:

$$ExploreScore = \frac{1}{|F_{diverse}|} \sum_{i \in F_{diverse}} f_i + \frac{1}{|F_{rare}|} \sum_{j \in F_{rare}} f_j \tag{1}$$

where f_i is a feature in the feature group $F_{diverse}$ and f_j is a feature in group F_{rare}. All features are normalized into [0, 1] before taking the average. A bigger *ExploreScore* suggests a more *negativistic* state, where a user puts more effort in diversifying the search process and being against conventions. A smaller *ExploreScore* suggests a more *conformist* state, where the user puts less effort in doing so. This aligns well with the conclusions from Facebook and Twitter's research [43, 44], where they observe that individuals who value rules and tradition (in a *conformist* state) tend to access less diverse content compared to individuals who value innovation and change (in a *negativistic* state). They are less likely to issue unfamiliar queries or click on unfamiliar URLs.

Second, we establish the *conformist* and *negativistic* states from the query log data. To do so, we select the top 15% (we empirically choose 15% to relax a bit from a rigorous top 10%) sessions with the highest *ExploreScore* to represent the negativistic state and the last 15% sessions to illustrate the conformist state.

Third, we conduct a statistical significance test between the two states for all feature groups except the two groups used to calculate *ExploreScore*. For each remaining feature group, we obtain the state averages for features in the group at the two states. Then we conduct a MANOVA [45] test across all feature groups and 5 ANOVA [46] tests for each of them. The detailed results are: *MANOVA* [F(5, 330) = 1.1352, p = 0.3414], *QueryEffort* [F(1, 334) = 0.5753, p = 0.4487], *QueryReformEffort* [F(1, 334)=1.5423, p = 0.2151], *ReadEffort* [F(1, 334) = 1.2738, p = 0.2599], *ScrollEffort* [F(1, 334) = 1.6459, p = 0.2004], and *ClickEffort* [F(1, 334) = 0.5863, p = 0.4444]. The significance tests produce $p > 0.05$ and fail to reject the null hypothesis. In other words, the "rules" dimension is irrelevant to a user's happiness level, which implies it is irrelevant to struggle detection and confirms what is suggested by the Reversal Theory.

Further, we plot the mean feature values for the conformist and negativistic states in Fig. 5. We can see that, except for the feature groups used to generate *ExploreScore*, none of the other feature groups show a statistically significant difference between the two states. Again, this confirms what Reversal Theory suggests that when the first two

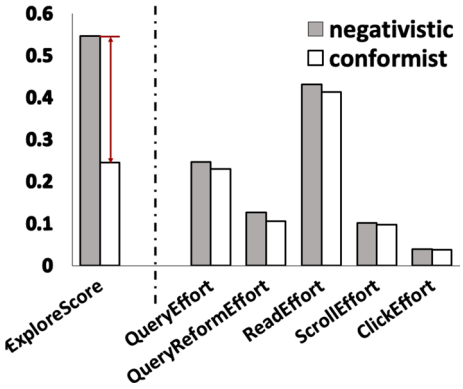

Fig. 5 Differences in Features at Two "Rules" States. ↕ marks significant differences at $p<.05$

dimensions interplay, the "rules" dimension has little impact on user efforts and struggle detection. We, therefore, do not handle features along this dimension.

A similar MANOVA hypothesis test runs for the "means-ends" dimension. That result is statistically significant and confirms what is suggested by the Reversal Theory that the first "means-ends" dimension is influential to a user's struggle. We, therefore, use "means-ends" as the primary dimension for our research.

5 Our approach

This paper presents a novel feature modulation method for search struggle detection based on Reversal Theory's bi-modal arousal model. First, we establish the two "means-ends" motivational states, telic and paratelic, for every search session. Based on Reversal Theory, a search session would be at any one moment only at either state, not both. Second, based on what Reversal Theory's interplay figure suggests and our hypothesis tests confirm, we select highly related features to the "means-ends" dimension. Third, we modulate these features by shifting their values for those in the paratelic state towards those in the telic state until their arousal model's peaks overlap. Fourth, we use the modulated features to fit a classifier and predict whether a session has struggles.

5.1 Put sessions into "means-ends" states

Reversal Theory's bi-modal model of arousal (Fig. 3) tells us that without knowing which motivational state the user is in, it is challenging to separate struggling from excitement or boredom from relaxation. We are thus motivated to (1) detect which state the user (and the session) is at, and then (2) move the two curves closer to each other for a selected group of features so that the struggles would be separable from the rest. Figure 7 illustrates our idea.

Our first step is to put every session into either a telic or paratelic state. The bi-modal arousal model is a two-component Gaussian mixture model, whose means and variances can be found by the Expectation-Maximization (EM) algorithm [47]. In the mixture model, a data point can have a soft mapping onto both Gaussians. However, based on Reversal Theory, at any one moment, a user can only be at one of the opposing states, not both. We choose to follow what Reversal Theory suggests in this work and only associate a search session with one of the two states. We, therefore, propose to take a less common approach to identify the states for each session.

We propose to assign the sessions into states based on the session's topic. Reversal Theory considers telic states are associated with more serious tasks, and paratelic states are associated with more playful tasks [10]. Other research also pointed out that search topic shows the impact on searcher behaviors [5]. We determine a session's search topic using a taxonomy used internally at Pinterest, which is constructed by graph-based algorithms [48, 49] and contains 24 topics. While Pinterest's taxonomy might differ from those used on other platforms such as Google or Reddit, modern search applications often feature rich content that covers a wide range of topics and shares some common elements. Our proposed method has the potential to be adapted to other search platforms, provided their topic taxonomies include topics that can be identified as serious or playful.

To determine a session's search topic, first, we extract every clicked URL in the session. Second, we assign each clicked URL to a taxonomy category using an in-house Gradient Boosted Decision Trees (GBDT) classifier[2] (with a learning rate of 0.1, minimum split loss 0.5, and maximum tree depth 8). The classifier calculates a URL's category score using various features, such as 1) tf-idf feature (here URL's text description is the document, and the category name is the search term here); 2) the embedding cosine similarity feature between the URL link's image embedding and the category name's FastText[3] embedding. The taxonomy category with the highest similarity score to the URL text becomes the label to the URL. Third, we chose the most frequent URL label in the session as the search topic for the session.

We annotate each topic in the Pinterest taxonomy as serious or playful by asking three annotators to label it and taking the majority vote. Then we assign those with a search topic relating to serious, significant events, such as financial, health, and career decisions, to a telic state. For instance, "Health," "Job," and "Finance." To a paratelic state, we assign those with a search topic relating to fun, relaxing events, such as entertainment and hobby. For instance, "Entertainment," "Art," and "Beauty."

5.2 Select "means-ends" features

Reversal Theory suggests that we should modulate the features along the "means-ends" dimension only. To identify the "means-ends" feature groups, we propose to identify feature groups that are significant to distinguish the two "means-ends" states. Other feature groups would remain the same without modulation.

Our goal is to select effort features that are significant to distinguish the two means-ends states. We take the following steps to accomplish it.

1. First, we normalize all effort-based features within a feature group into the range [0, 1] using

$$\frac{value - minValue}{maxValue - minValue}. \tag{2}$$

2. Second, we calculate two state average scores for each feature group by taking the group average for sessions at the telic and paratelic states.
3. Third, we conducted a MANOVA test to compare the state average score for all feature groups in the two states. The significance test result [$F(7, 292) = 34.3121$, $p < 0.0001$] proves that these feature groups are statistically significantly affected by the two states, which agrees with what Reversal Theory suggests.
4. Fourth, we then conducted one ANOVA test for each feature group to select the significant features. We find that four out of seven features groups, *ReadEffort* [$F(1, 298) = 39.4581$, $p < 0.0001$], *QueryEffort* [$F(1,298) = 95.3286$, $p < 0.0001$], *DiversifyEffort* [$F(1, 298) = 30.0176$, $p < 0.0001$], and *ClickEffort* [$F(1,298) = 54.4846$, $p < 0.0001$], are statistically significantly different in paratelic and telic sessions.

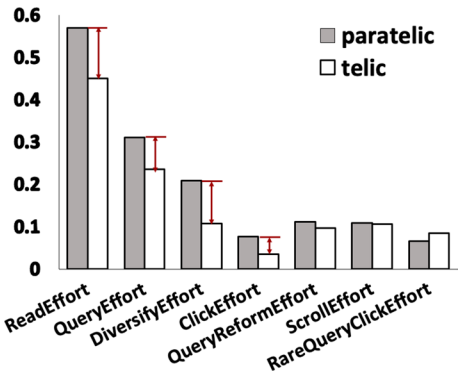

Fig. 6 Feature value gaps along "Means-ends." The red ↕ indicates the difference is statistically significant at p<.05

[2] https://medium.com/pinterest-engineering/pin2interest-a-scalable-system-for-content-classification-41a586675ee7.

[3] https://github.com/facebookresearch/fastText.

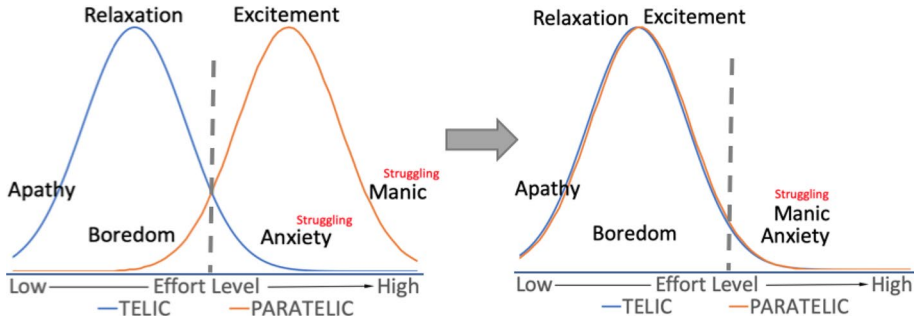

Fig. 7 Modulation to separate struggles from non-struggles

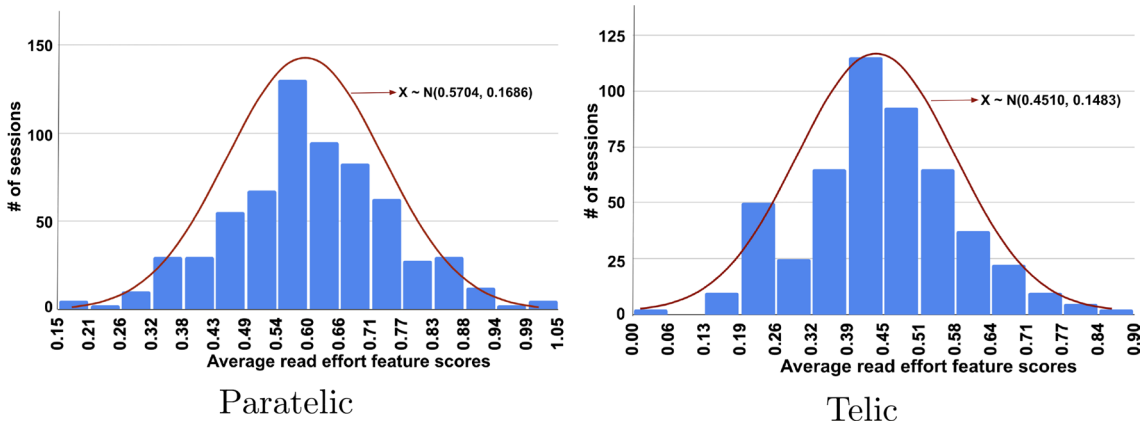

Fig. 8 read effort feature value distribution vs. corresponding Gaussion distribution PDFs

Figure 6 plots the mean feature values from each selected feature group. As we can see, the four feature chosen groups show a large gap between the telic and paratelic sessions. We determine these feature groups as "means-ends" features and modulate them.

5.3 Modulate the features

Although we did not use the EM algorithm to find the means and variances of the two Gaussian distributions, our state assignment method based on the search topic still roughly forms Gaussian distributions, as what Reversal Theory states. We also show this alignment in Figs. 8, 9, 10, and 11, where the feature value distributions fit the corresponding Gaussian distribution probability density functions. We leverage this information to remove the bias between the two Gaussians.

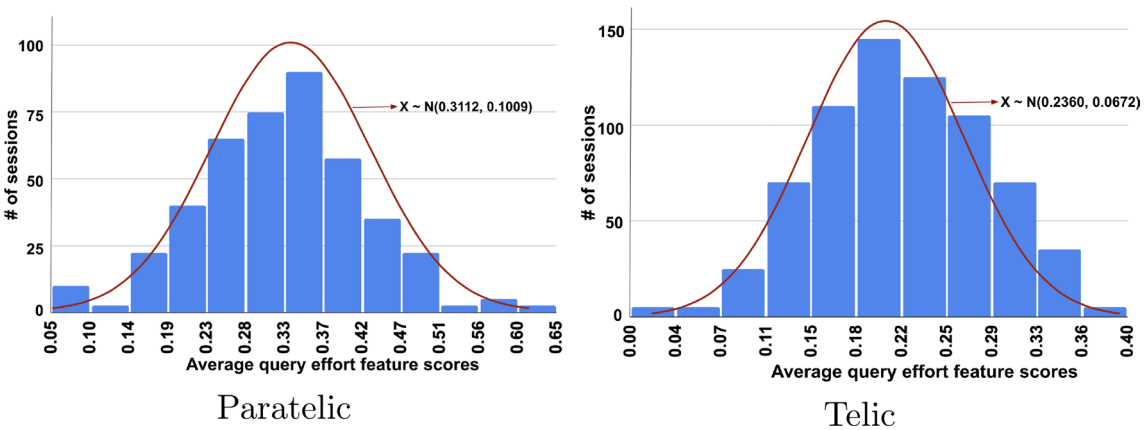

Fig. 9 query effort feature value distribution vs. corresponding Gaussion distribution PDFs

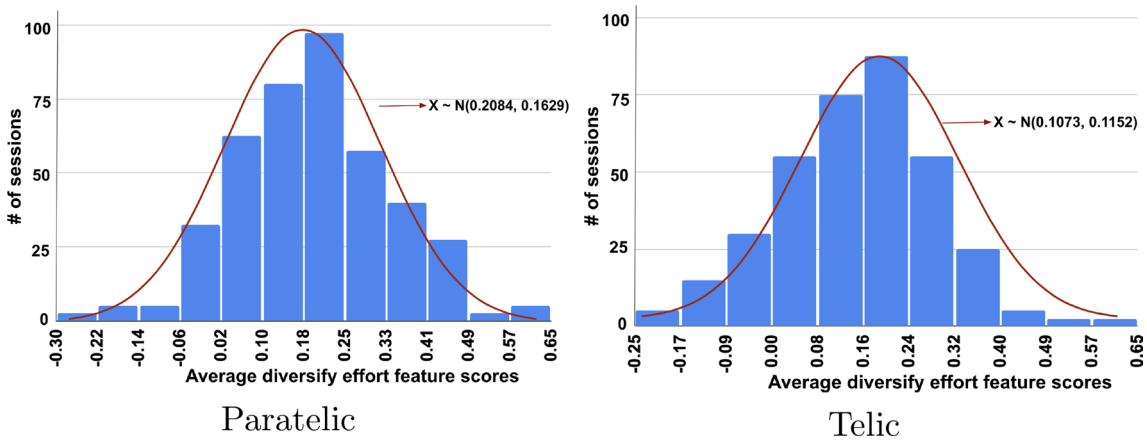

Fig. 10 diversify effort feature value distribution vs. corresponding Gaussion distribution PDFs

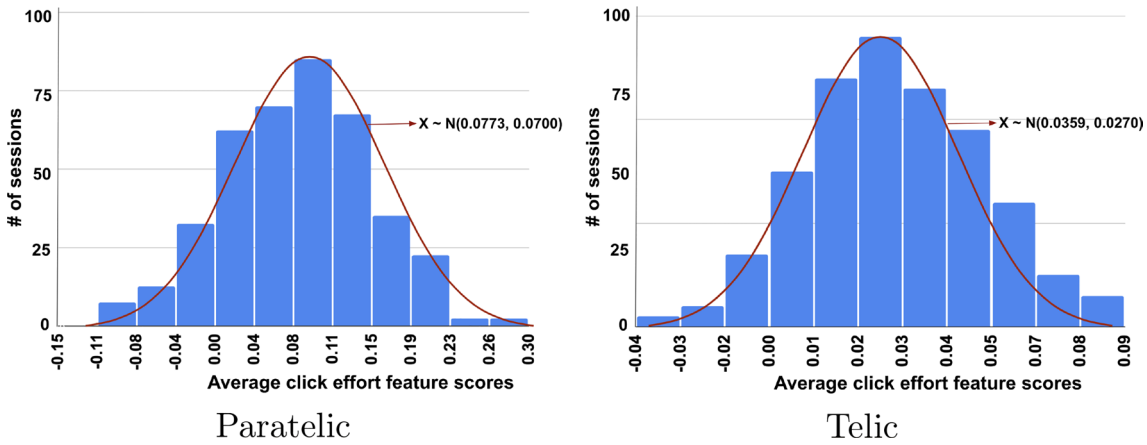

Fig. 11 click effort feature value distribution vs. corresponding Gaussion distribution PDFs

Given a feature X_i in one of the feature groups being selected earlier, {*QueryEffort*, *ClickEffort*, *ReadEffort*, *DiversifyEffort*}, we use X^i_{telic} and $X^i_{paratelic}$ to represent two different Gaussian distributions, each for X_i's feature values in the telic state and paratelic state, respectively:

$$X^i_{telic} \sim \mathcal{N}(\mu_{i_{telic}}, \sigma^2_{i_{telic}}).$$

and

$$X^i_{paratelic} \sim \mathcal{N}(\mu_{i_{paratelic}}, \sigma^2_{i_{paratelic}}).$$

where $\mu_{i_{telic}}$ and $\sigma_{i_{telic}}$ are the mean and standard deviation of the i^{th} feature in all telic sessions; and $\mu_{i_{paratelic}}$ and $\sigma_{i_{paratelic}}$ are the mean and standard deviation of the i^{th} feature in all paratelic sessions. We obtain the states as described in Section 5.1 and calculate the means and variances directly from them.

Next, we propose to reduce the bias between the two distributions by a Bayesian scaling method, shifting the paratelic towards the telic state for the selected "means-ends" features. This transformation is done by Eq. 3:

$$X'_{paratelic} = \frac{\sigma_{telic}}{\sigma_{paratelic}} X_{paratelic} + \mu_{telic} - \frac{\sigma_{telic}}{\sigma_{paratelic}} \mu_{paratelic} \tag{3}$$

where $X_{paratelic}$ is the original feature value in the paratelic state, and $X'_{paratelic}$ is the new feature value after modulation.

As illustrated in Fig. 7, the effort levels previously identified as both "anxiety/struggling" and "excitement" would now be separable after feature modulation. We can now combine these modulated "means-ends" features and other unmodulated features with a wide range of classifiers for struggle detection.

5.4 Detect searcher struggles

To predict struggling sessions and non-struggling sessions, we use our modulated features and formulate the problem as a binary classification problem. Let $X'(s)$ be the modulated effort feature vector of Session s and Y be the random variable of search struggles. The classification would output 1 for *Struggle* and 0 for *Non-struggling*:

$$\mathcal{I} = \begin{cases} 1 & \text{if } P(Y = 1|X'(s), \Theta) > c \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathcal{I}$ is the indicator function, Θ is the classifier's model parameter, $X'(s)$ is the modulated feature vector extracted from a search session, and c is the cutoff value of the prediction score. To compare the modulating effect, we also use the original feature vector $X(s)$ to conduct the prediction and compare the outcomes.

6 Experimental setup

We have conducted experiments to evaluate our method. The design of the experiments focuses on showing the before-and-after effect of using feature modulation in struggle detection. In this section, we describe how we set up the experiments.

6.1 Dataset preparation

We collected a week's search log data from the Pinterest search engine during the period of 11/22/2020 to 11/28/2020, which we used exclusively to create two test datasets: one from desktop browsers and the other from mobile apps. The user activities on the two platforms are slightly different due to different platform interfaces. For training, we utilized a separate week's Pinterest query log, collected from 11/08/2020 to 11/14/2020, one week prior to the testing period. This ensured a clear separation between the training and test datasets, with no overlap in the data used during development and evaluation.

We take the following steps to prepare our data. First, we segment the search log into topically coherent segments [32], each corresponding to a session. We segment the sessions following [42]. It uses logistic regression to classify two neighboring queries as they belong to the same search session or otherwise. Then, the consecutive query pairs are added into the same segment if they show high regression scores. The classification features include query edits, click similarity, and time-related features. We achieved a segmentation accuracy of 99.8% in 10-fold cross-validation on the experiment dataset used in [42].

Second, we recruit human assessors to annotate whether a session is struggling or non-struggling. The assessors were instructed to label a session into (1) Struggling, (2) Non-struggling, or (3) Uncertain. Each session was judged by two assessors independently. If there was a disagreement between the two assessors, a third assessor joined in resolving the dispute [50]. Every assessor carefully examined the query logs, with information about queries, user clicks, documents read by users, and timestamps of every user activity (refer to Fig. 12.)

We hired assessors through a third-party company contracted by Pinterest. This third-party company specializes in data annotation services and recruited assessors from various countries outside the United States. All assessors were fluent in English and were selected based on their qualifications and experience. The compensation was set according to local market standards, ensuring fairness relative to the economic conditions of the assessors' countries. The third-party company provided training to ensure effective task performance and continuously monitored the quality and consistency of the annotations through regular evaluations.

The assessors also went through a training session before they started the actual annotation. At the beginning of the annotation process, the annotation procedure was described as the following:

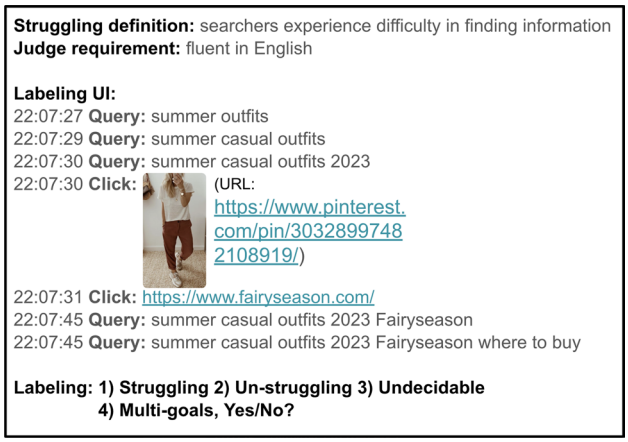

Fig. 12 Annotation User Interface

We have a set of user search sessions. Each session consists of a few time-stamped queries followed by a few clicks or maybe a re-query. Our goal is to just look at the search activities in a search session and levy a judgment on whether or not the searcher was struggling to find information.

Then we shared some examples of struggling and non-struggling sessions and our reasoning with the assessors. For example,

A searcher is not struggling when (1) she uses search engine as a bookmark, for example a user searched "home depot" and clicked www.homedepot.com (2) she is doing research on a topic, e.g., "how many chromosomes are present in interphase of meiosis?" (3) she is just looking up information, such as stock ticker prices (4) she is just checking the same thing over and over to check Facebook or email, or monitor sports results, or see if there are new Craigslist listings. (5) sometimes the initial query shows up twice with minor spelling correction. Then she clicked on a URL that seems to answer the story. Then there's no other action. she might have found what she's looking for, hence the searcher is non-struggling.
A searcher is struggling when (1) they're not finding what they want in that initial query. We see this a lot on ambiguous queries and people's names. (2) A person is probably struggling when they try multiple variations of a query or click into different URLs and then re-query. (3) Clicking on an ad and then re-querying may also suggest they're struggling. (4) Then there are cases like: "What is the search topic? Is someone just having fun and trying to find the story behind the movie? Why do they continue re-querying on that 'true story' angle and still not focus on any article?" "This to me feels like struggling, but I'd be hard-pressed to explain why it is beyond 'This is my gut feeling'." Sometimes I can't decide and go with the "Uncertain" decision. A session of two identical queries with no click tells me nothing (unless the relevant results are just the top few images which requires no clicks at all). Also I can't do anything with session topics that I'm very unfamiliar with.

Eventually, the annotations achieved an inter-coder agreement of 73.3%.

Third, in the end, for the training data, collected from the week prior (11/08/2020 to 11/14/2020), we obtained 2,025 labeled sessions, including 565 struggling and 1,460 non-struggling sessions. Similarly, for the test data (11/22/2020 to 11/28/2020), we obtained 2,157 labeled sessions, including 601 struggling and 1,556 non-struggling sessions. We then processed all of these sessions-both training and test data-into feature vectors for further analysis and model development. Table 3 reports the dataset statistics. We acknowledge that while the test data was kept entirely separate from the training and development phases, there is overlap between the training data and the data used during method development and parameter tuning. This overlap does not affect the validity of the test results but may limit the demonstrated generalizability of the method to entirely new training datasets. Future work will seek to further validate the method on independent training datasets to confirm its robustness across different data sources.

Fourth, we asked the assessors to mark out sessions that contained multiple search tasks. This step served as a sanity check for the effectiveness of our automatic session segmentation.

Table 3 Dataset statistics

Dataset	Duration	#Sessions	#Struggling	#NonStruggling	#Query/Session
Training mobile		1,045	275	770	5.31
Training desktop	11/08 ~ 14, 2020	980	290	690	4.55
Training total		2,025	565	1,460	4.94
Test mobile		1,123	299	824	5.39
Test desktop	11/22 ~ 28, 2020	1,034	302	732	4.62
Test total		2,157	601	1,556	5.02

6.2 Baseline classifiers

We experimented with several baseline classifiers for the task of searcher struggle detection. These include widely-used classifiers (such as SVM, Logistic Regression, MART, and Transformers) as well as a state-of-the-art best-performing searcher struggle detection method Hassan et al. [15].

- **ZeroRule** is a naive baseline that classifies instances based on the majority label in the ground truth. We include this baseline classifier because it is perhaps the simplest possible method and almost equivalent to random guessing. This allows us to set a basic performance benchmark, as we expect an algorithm specifically designed for detecting searcher struggles to perform better than this naive baseline.
- **SVM** is the support vector machine classifier [51], which is one of the top-performing linear classifiers prior to the era of deep learning. We select SVM because the number of features used in our work is not overwhelmingly large, and they do not necessitate the use of deep neural networks. These features can be effectively handled by pre-deep learning models such as SVM. Here we use the *svm()* model provided by the R library,[4] where we use a radial kernel with a kernel coefficient of 0.016 and a cost of 2.0.
- **LM** is a logistic regression classifier [52]. Logistic regression is another top-performing linear classifier prior to the era of deep learning. For a similar reason to why we use SVM, our features do not necessitate the use of deep neural networks and can be effectively handled by pre-deep learning models such as logistic regression. We employ LM to compare our approach on this leading non-neural network classifier. The particular logistic regression model we use is *glm()*, provided by the R library.[5] We set the module parameter "family" as "binomial."
- **MART** is the Multiple Additive Regression Trees (MART) classifier [53], a top-performing non-linear, non-neural network classifier. Prior to the era of deep learning, MART was widely used in applications related to web search, such as learning to rank, which often take a feature-based approach that is similar to our setting. We set MART's n.tree to be 8000 and shrinkage 0.005.
- **Transformer** [54] is a deep neural network classifier that leverages the state-of-the-art multi-head self-attention transformer architecture. We use this classifier because it has demonstrated superior performance in many classification tasks, thanks to its ability to capture complex patterns and dependencies in data. To leverage Transformer, we treat numerical features as dense features, and use one-hot encoding for the categorical features. All features are then concatenated together to form the input embedding for Transformer. We use a batch size of 64, a learning rate of 0.00005, and a dropout rate of 0.1. Deep neural networks require a large number of training samples. To compensate for our limited amount of human-labeled data, we first use MART as a teacher model to generate additional training data for the Transformer. We then fine-tune the model using the human-labeled data.
- We also re-implemented a state-of-the-art searcher struggle detection method that was proposed by Hassan et al. [15] It is perhaps the most similar work to our work and shares the most features with us.

SVM, **LM**, **MART**, and **Transformer** all use features in Table 1 and the categorical feature, search topic. Hassan et al. [15] uses features presented in [15] and experiment on our dataset. This model performs similarly compared to their reported results.

In this study, we employed a 10-fold cross-validation technique to evaluate the performance of all baselines and our own method. This approach involves dividing the dataset into ten equal parts or 'folds.' During each iteration, one fold

[4] https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf.

[5] https://www.statmethods.net/advstats/glm.html.

is used as the test set, while the remaining nine folds are used to train the model. This process is repeated ten times, ensuring that each fold serves as the test set exactly once and as part of the training data nine times. By using 10-fold cross-validation, we maintain a clear separation between training and test data for each iteration, which helps to prevent overfitting and provides a robust estimate of the model's performance on unseen data. The results presented in our tables and plots are the averages obtained across all ten folds, offering a comprehensive overview of the methods' effectiveness across different subsets of the data.

6.3 Runs under comparison

For each baseline, we experiment with three different settings. (1) The original setting described in Sect. 6.2 and without feature modulation. (2) The baseline classifiers running with a variation of the proposed feature modulation method. We skip the "means-ends" features step in the variation and directly use Eq. 3 to modulate all features. These runs have suffix "+FMNS," which stands for feature modulation no selection. (3) The baseline classifiers with only the "means-ends" features are modulated. These runs have the suffix "+FM."

6.4 Evaluation metrics

We evaluate the struggle detection systems using multiple metrics to understand their effectiveness from different perspectives. The metrics include *accuracy*, *positive precision* and *positive recall* (they are precision and recall for the struggling class), and *negative precision* and *negative recall* (they are precision and recall for the non-struggling class). They are defined as follows.

$$accuracy = \frac{\text{number of correct instances}}{\text{total number of instances}} \tag{5}$$

$$positive\_precision = \frac{\text{number of correctly returned struggling instances}}{\text{total number of instances being classified as struggling}} \tag{6}$$

$$positive\_recall = \frac{\text{number of correctly returned struggling instances}}{\text{total number of struggling instances in ground truth}} \tag{7}$$

$$negative\_precision = \frac{\text{number of correctly returned non-struggling instances}}{\text{total number of instances being classified as non-struggling}} \tag{8}$$

$$negative\_recall = \frac{\text{number of correctly returned non-struggling instances}}{\text{total number of non-struggling instances in ground truth}} \tag{9}$$

Among these metrics, *accuracy* and *positive precision* are chosen as the main metrics. Accuracy is important as it measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. This provides a general sense of the model's performance across all classes. Positive precision is crucial for our task because precise assistance is preferred over generic assistance by human users; therefore, accurately predicting user struggles is very important [55].

Table 4 Mobile: performance of struggle detection (Up and down arrows indicate absolute performance increase and decrease

	accu	impr	pos. p	impr	pos. r	impr	neg. p	impr	neg. r	impr
ZeroRule	0.7337	–	–	–	0.0000	–	0.7337	–	1.0000	–
LM	0.8413		0.7239		0.6561		0.8788		0.9088	
LM+FMNS	0.8621	2.5%$^{\uparrow}$	0.7342	1.4%$^{\uparrow}$	0.6393	2.5%$^{\downarrow}$	0.8946	1.8%$^{\uparrow}$	0.9297	2.3%$^{\uparrow}$
LM+FM	0.8910	5.9%$^{\uparrow\dagger}$	0.7513	3.7%$^{\uparrow\dagger}$	0.6116	6.8%$^{\downarrow}$	0.9157	4.2%$^{\uparrow\dagger}$	0.9542	5.0%$^{\uparrow\dagger}$
SVM	0.8565		0.7956		0.7414		0.8823		0.9105	
SVM+FMNS	0.8675	1.3%$^{\uparrow}$	0.7940	0.2%$^{\downarrow}$	0.7180	2.9%$^{\downarrow}$	0.8929	1.2%$^{\uparrow}$	0.9260	1.7%$^{\uparrow}$
SVM+FM	0.8928	4.2%$^{\uparrow\dagger}$	0.8511	7.0%$^{\uparrow\dagger}$	0.7278	1.8%$^{\downarrow}$	0.9052	2.6%$^{\uparrow}$	0.9533	4.7%$^{\uparrow\dagger}$
Hassan et al. [15]	0.8507		0.7729		0.6439		0.8737		0.9287	
Hassan et al.+FMNS	0.8626	1.4%$^{\uparrow}$	0.7962	3.0%$^{\uparrow}$	0.6447	1.3%$^{\uparrow}$	0.8807	0.8%$^{\uparrow}$	0.9408	1.3%$^{\uparrow}$
Hassan et al.+FM	0.8786	3.3%$^{\uparrow\dagger}$	0.8419	8.9%$^{\uparrow\dagger}$	0.6923	7.5%$^{\uparrow\dagger}$	0.8894	1.8%$^{\uparrow}$	0.9501	2.3%$^{\uparrow}$
MART	0.8740		0.7968		0.7305		0.9002		0.9288	
MART+FMNS	0.8835	1.1%$^{\uparrow}$	0.8042	0.9%$^{\uparrow}$	0.7144	2.2%$^{\downarrow}$	0.9065	0.7%$^{\uparrow}$	0.9409	1.3%$^{\uparrow}$
MART+FM	0.9055	3.6%$^{\uparrow\dagger}$	0.8666	8.8%$^{\uparrow\dagger}$	0.7754	6.1%$^{\uparrow\dagger}$	0.9182	2.0%$^{\uparrow}$	0.9548	2.8%$^{\uparrow}$
Transformer	0.8811		0.8036		0.7457		0.9073		0.9318	
Transformer+FMNS	0.8902	1.0%$^{\uparrow}$	0.8062	0.3%$^{\uparrow}$	0.7414	0.6%$^{\downarrow}$	0.9155	0.9%$^{\uparrow}$	0.9402	0.9%$^{\uparrow}$
Transformer+FM	0.9207	4.5%$^{\uparrow\dagger}$	0.8725	8.6%$^{\uparrow\dagger}$	0.7629	2.3%$^{\uparrow}$	0.9327	2.8%$^{\uparrow}$	0.9672	3.8%$^{\uparrow\dagger}$

†Shows statistically significant improvement from "feature modulation (X+FM)" runs over the original runs (one-tailed t-test, p=.01)). "X+FMNS" refers to "Feature Modulation with No feature Selection"

Table 5 Desktop: performance of searcher struggle detection (Up and down arrows indicate absolute performance increase and decrease

	accu	impr	pos. p	impr	pos. r	impr	neg. p	impr	neg. r	impr
ZeroRule	0.7079	–	–	–	0.0000	–	0.7079	–	1.0000	–
LM	0.8384		0.7624		0.8316		0.8917		0.8425	
LM+FMNS	0.8425	0.5%$^{\uparrow}$	0.7742	1.5%$^{\uparrow}$	0.8260	0.7%$^{\downarrow}$	0.8890	0.3%$^{\downarrow}$	0.8526	1.2%$^{\uparrow}$
LM+FM	0.8676	3.5%$^{\uparrow\dagger}$	0.8141	6.8%$^{\uparrow\dagger}$	0.8395	0.9%$^{\uparrow}$	0.9015	1.1%$^{\uparrow}$	0.8846	5.0%$^{\uparrow\dagger}$
SVM	0.8617		0.7843		0.8810		0.9201		0.8497	
SVM+FMNS	0.8757	1.6%$^{\uparrow}$	0.8008	2.1%$^{\uparrow}$	0.8810	0.0%	0.9265	0.7%$^{\uparrow}$	0.8726	2.7%$^{\uparrow}$
SVM+FM	0.9100	5.6%$^{\uparrow\dagger}$	0.8625	10.0%$^{\uparrow\dagger}$	0.8935	1.4%$^{\uparrow}$	0.9385	2.0%$^{\uparrow}$	0.9194	8.2%$^{\uparrow\dagger}$
Hassan et al. [15]	0.8418		0.7646		0.8374		0.8959		0.8374	
Hassan et al.+FMNS	0.8604	2.2%$^{\uparrow}$	0.7927	3.7%$^{\uparrow}$	0.8416	0.5%$^{\uparrow}$	0.9040	0.9%$^{\uparrow}$	0.8714	3.2%$^{\uparrow}$
Hassan et al.+FM	0.8981	6.7%$^{\uparrow\dagger}$	0.8499	11.2%$^{\uparrow\dagger}$	0.8557	2.2%$^{\uparrow}$	0.9237	3.1%$^{\uparrow\dagger}$	0.9204	9.0%$^{\uparrow\dagger}$
MART	0.8775		0.8223		0.8605		0.9134		0.8878	
MART+FMNS	0.8952	2.0%$^{\uparrow}$	0.8416	2.3%$^{\uparrow}$	0.8683	0.9%$^{\uparrow}$	0.9262	1.4%$^{\uparrow}$	0.9100	2.5%$^{\uparrow}$
MART+FM	0.9293	5.9%$^{\uparrow\dagger}$	0.8874	7.9%$^{\uparrow\dagger}$	0.9024	4.9%$^{\uparrow\dagger}$	0.9508	4.1%$^{\uparrow\dagger}$	0.9428	6.2%$^{\uparrow\dagger}$
Transformer	0.8813		0.8266		0.8649		0.9166		0.8911	
Transformer+FMNS	0.9005	2.2%$^{\uparrow}$	0.8494	2.8%$^{\uparrow}$	0.8737	1.0%$^{\uparrow}$	0.9298	1.4%$^{\uparrow}$	0.9152	2.7%$^{\uparrow}$
Transformer+FM	0.9372	6.3%$^{\uparrow\dagger}$	0.8988	8.7%$^{\uparrow\dagger}$	0.9090	5.1%$^{\uparrow\dagger}$	0.9560	4.3%$^{\uparrow\dagger}$	0.9508	6.7%$^{\uparrow\dagger}$

†Shows statistically significant improvement from "feature modulation (X+FM)" runs over the original runs (one-tailed t-test, p=.01)). "X+FMNS" refers to "Feature Modulation with No feature Selection"

7 Experimental results

7.1 Main results—searcher struggle detection effectiveness

Tables 4 and 5 show the effectiveness of the experimental runs for searcher struggle detection on the mobile and desktop datasets, respectively. These tables also highlight the percentage improvement of each run compared to its original run. Additionally, they report the results of a one-tailed t-test comparing the "+FM" runs with the initial runs.

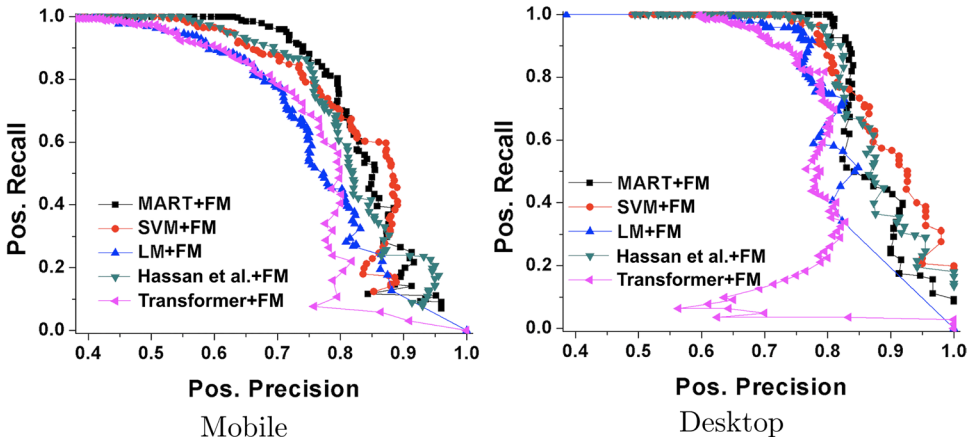

Fig. 13 Mobile and desktop positive precision-and-recall curves

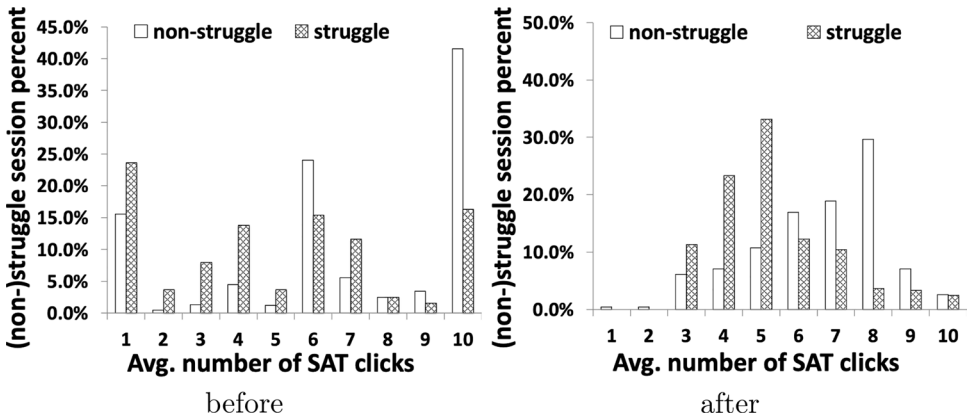

Fig. 14 Distribution of struggling and non-struggling sessions over avg. number of Satisfactory (SAT) clicks

The results show that the proposed feature modulation method is highly effective. The "+FM" runs statistically significantly improve the performance of all classifiers on all metrics. On average, our approach boosts a baseline method's accuracy by ~5% and positive precision by ~9%. Combined with our method, these classifiers have become highly effective. Transformer+FM achieves the best performance among all models and settings, with a high 0.937 accuracy and 0.899 positive precision for the desktop dataset. We observe similar trends on the mobile dataset. The "+FMNS" runs gain slightly better performance than the original baselines and worse than the "+FM" runs. It confirms what Reversal Theory suggests that only the first dimension, "means-ends," impacts the arousal model, thus effective on our struggle detection task. Other features, some of which are more related to the "rules" dimension, which Reversal Theory considers irrelevant. The weak performance from the "+FMNS" runs again supports this insight from Reversal Theory, besides our hypothesis test in Sect. 4.4.

Our experimental results also show that while large-scale machine learning models have the capability to automatically learn feature representations, feature selection and modification still play a crucial role, particularly when we have limited high-quality labels. In many cases, feature selection and modification can reduce the dimensionality of the input space, decrease input data noise, lower the memory requirements, training time, and inference time, and significantly mitigate the risk of overfitting.

7.2 Impact of probability cutoff *c*

In this section, we evaluate the cutoff parameter *c*'s (See Eq. 4) impact on the classifiers' performance. We plot out the positive labels' precision-and-recall curves for the best performed modules in each classifier group, which includes LM+FM, SVM+FM, Hassan et al.+FM, MART+FM, and Transformer+FM. Figure 13 shows the results. We observe that setting $c = 0.5$ leads to the best F1 scores of each classifier. Note that, the numbers we report in Tables 4 and 5 all use this cutoff value.

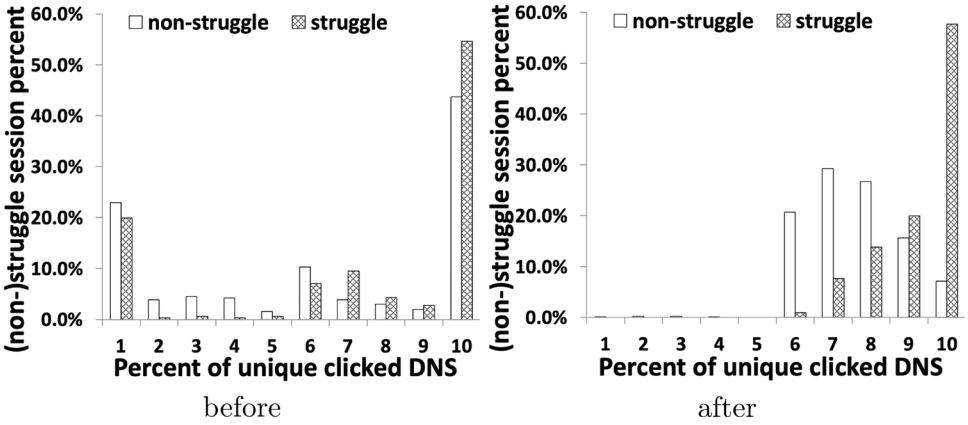

Fig. 15 Distribution of struggling and non-struggling sessions over the percentage of unique clicked DNS domains

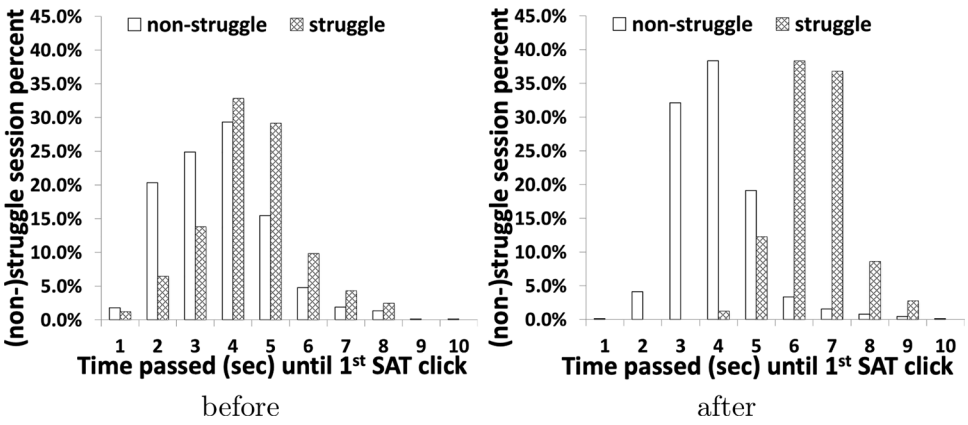

Fig. 16 Distribution of struggling and non-struggling sessions over time passed until the 1st SAT click

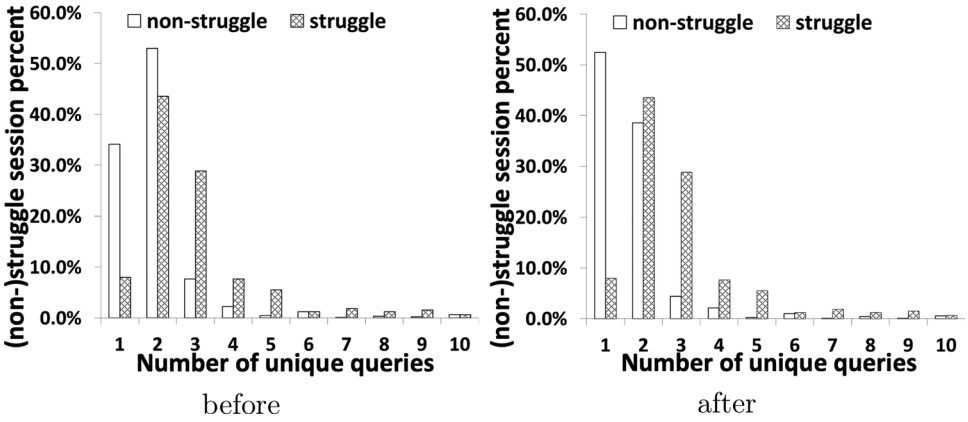

Fig. 17 Distribution of struggling and non-struggling sessions over the number of unique queries

7.3 Impact of feature modulation

To demonstrate the effect of our feature modulation, we investigate its impact on individual features in this section. We select several features from groups that show significant differences between paratelic and telic states, as illustrated in Fig. 6. These selected features are: the average number of SAT clicks in the 'Click Effort' group; the percentage of unique clicked DNS domains in the 'Diversify Effort' group; the time elapsed until the first SAT click in the 'Read Effort' group; and the number of unique queries, the average number of characters per query, and the number of manually typed queries in the 'Query Effort' group.

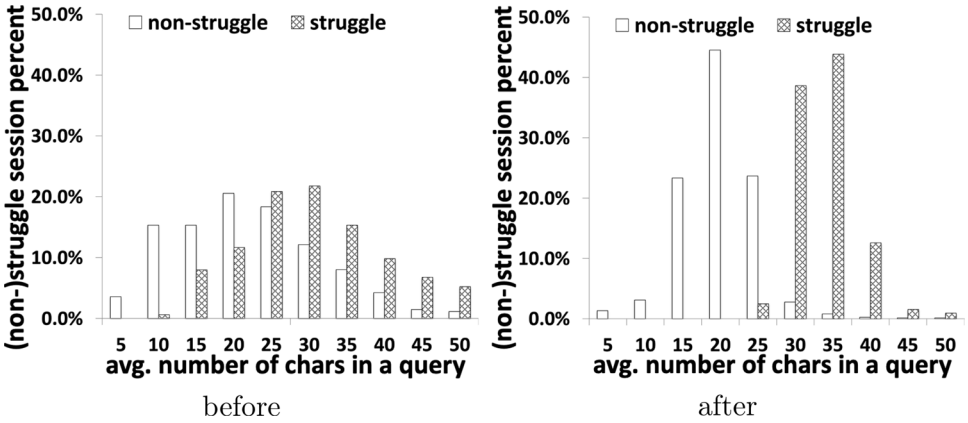

Fig. 18 Distribution of struggling and non-struggling sessions over average number of characters in a query

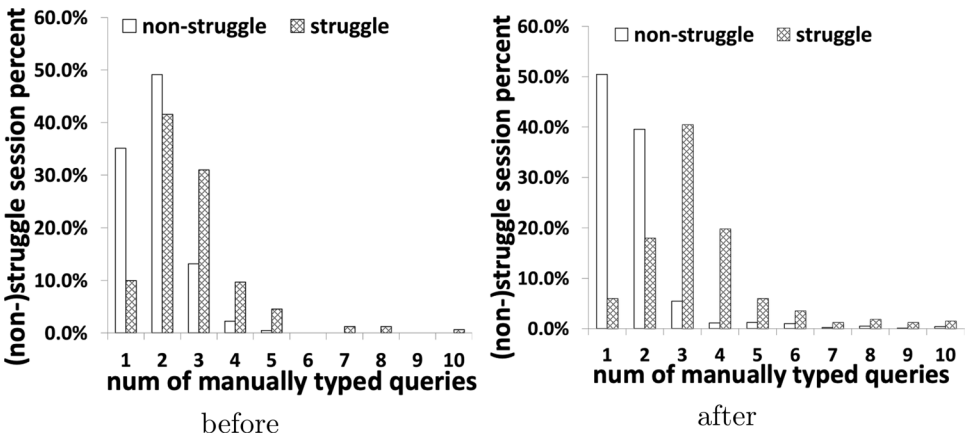

Fig. 19 Distribution of struggling and non-struggling sessions over number of manually typed queries

Figures 14, 15, 16, 17, 18, and 19 demonstrate these comparisons. They aim to show the magnitudes and distributions of these features for both struggling and non-struggling sessions. These figures are generated by first dividing the magnitudes of these features into ten evenly spaced bins, and then plotting the ratio of struggling to non-struggling sessions in each bin.

The figures show that before feature modulation, the feature distributions of non-struggling and struggling sessions are mixed and present no obvious patterns. However, after feature modulation, the features in non-struggling and struggling sessions form two distinct bell-shaped curves with different peaks. For instance, in Fig. 14, after feature modulation, the average number of SAT clicks forms two distinct bell-shaped curves: one peaking at five for non-struggling sessions and the other peaking at eight for struggling sessions. This separation demonstrates that the distributions of these features in struggling and non-struggling sessions are more distinct after modulation.

To summarize, before feature modulation, the distributions of struggling and non-struggling sessions may not present any obvious pattern. However, after feature modulation, the peak feature values of struggling and non-struggling sessions are further separated. This suggests that our method helps these features better distinguish between the binary classes, making them more valuable in this classification task.

8 Conclusion and future work

This paper charts a unique solution path by harnessing insights from established psychological theories to craft practical solutions. Drawing on the principles of Reversal Theory, we introduce a novel feature modulation method to enhance searcher struggle detection during web search. Our method modulates the commonly-used effort-based features according to Reversal Theory's bi-modal arousal model. It begins by isolating features that correspond to the dynamic nature of different user motivations. These features are then adjusted to address any inherent bias between the two different motivational states that users might experience. By refining these features, our method

can better correlate the user's level of effort with their actual experience of struggles. The goal of the modulation is that after adjustments, the features can provide a more accurate representation of the user's experience and better aligned with their struggling experience, thereby improving the detection results. These modulated features are then fed into classification models to detect the presence of searcher struggle within a search session. Evaluations on week-long Pinterest search logs confirm that the proposed method can statistically significantly improve searcher struggle detection methods.

Moreover, our method improves the state-of-the-art understanding of user experience during a search session by refining the assumptions made by most existing methods. Searcher struggle is important feedback to web search engines. Most existing web search struggle detection methods rely on effort-based features to identify the struggling moments. Their underlying assumption is that the more effort a user spends, the more struggling the user may be. However, studies have shown that this simple association might be incorrect. Reversal Theory points out that instead of having a static personality trait, people constantly switch between opposite psychological states, complicating the relationship between the efforts they spend and the level of frustration they feel. This may explain several reasons why the existing assumption may have limitations. First, some methods mix the various reasons for user struggles, not distinguishing between different types of motivational states and contexts that could lead to a user experiencing difficulty. This blending of diverse causes can obscure the specific factors contributing to the struggle, making it harder to address them effectively. Second, they may not fully account for the fact that at any given moment, a user can only have one of two opposing motives, not both.

There are also limitations to our method that should be considered. The accuracy of our approach is highly dependent on correctly identifying motivational states. Incorrect classification of motivational states can occur when there is an overlap in user behavior between telic and paratelic states, or when there are insufficient distinguishing features. If a user's motivational state is misinterpreted, the resulting feature modulation may not reflect their actual experience, leading to inaccurate detection of struggles. This misalignment can reduce the overall effectiveness of the model. Moreover, inaccurate or incomplete logging of user interactions, as well as external factors like technical issues or distractions, can introduce noise into the data. This noise can obscure meaningful patterns and lead to incorrect inferences about user struggles, emphasizing the need for robust data preprocessing and noise reduction techniques. In addition, our method assumes that Reversal Theory applies uniformly across all users and contexts. However, individual differences and cultural factors could influence motivational states in ways that limit the generalizability of the approach. Addressing these limitations will be key to refining our method and broadening its applicability across diverse user populations.

Future work. Our research represents an initial application of Reversal Theory, which we believe holds significant promise for broader applications in information retrieval. By incorporating this theory, we can enhance the personalization and effectiveness of various digital systems. Here are several potential applications:

- Personalized Search Results: Extending our approach to personalize search results based on the dynamic motivational states of users. By understanding and adapting to these states, search engines can deliver more relevant results that align with the user's current needs and motivations, improving overall satisfaction and efficiency.
- Enhanced Recommendation Systems: Utilizing Reversal Theory to enhance user experience in recommendation systems. By tailoring suggestions to the user's shifting motivations, these systems can provide more engaging and satisfying content, increasing user engagement and retention.
- Adaptive User Interfaces: Improving user interface design by creating adaptive interfaces that respond to changes in user motivation. This would increase usability and engagement by providing a more intuitive and responsive user experience.
- Motivational State Detection: Developing methods to detect and analyze shifts in motivational states in real-time. This could lead to the creation of dynamic systems that adapt their responses based on the current psychological state of the user, enhancing the relevance and effectiveness of interactions. By understanding and responding to the dynamic nature of user motivations, digital platforms can create more engaging and satisfying experiences, ultimately leading to improved user retention and loyalty.

Of particular interest is the observation that Reversal Theory suggests users' psychological states may pivot in response to various catalysts, such as inherent tendencies, situational factors, or the body's innate biological rhythms. For instance, a shift from a goal-directed (telic) to a playful (paratelic) state may be triggered by stress alleviation, entertainment, or humor. Conversely, a transition from a playful (paratelic) to a goal-directed (telic) state might occur

due to unavoidable tasks, sudden threats, or the need for strategic decision-making. These shifts are influenced by a range of external and internal factors, underscoring the dynamic nature of user motivation and behavior.

Interestingly, the phenomenon of web searcher struggle may act as a catalyst for motivational reversal. When users experience difficulty or frustration during a search session, it might trigger a change in their motivational state. For example, struggling with a complex search could shift a user from a playful to a goal-directed state as they become more focused on finding a solution. This suggests that detecting struggle within a session could inform our understanding of user states in subsequent sessions.

Recognizing these state transitions could provide valuable insights into user behavior, enabling more personalized and adaptive information retrieval systems. Leveraging this insight represents an exciting avenue for future research, with the potential to significantly enhance user experience and the effectiveness of search technologies. We look forward to exploring this further.

Acknowledgements This research was supported by the United States National Science Foundation Grants IIS-1453721 and IIS-2336768. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor.

Author contributions J. L.: Methodology, Formal analysis and investigation, Writing - original draft preparation, and Resources; Y. Y.: Formal analysis and investigation, and Writing - original draft preparation; V. N.: Conceptualization, Formal analysis and investigation, and Writing - review and editing; G. H. Y.: Conceptualization, Methodology, Writing - original draft preparation, Writing - review and editing, Funding acquisition, and Supervision.

Data availability The data used in this study come from a commercial company's search session logs and reflect the struggling experiences of its users. To protect user privacy and confidentiality, these data are not publicly available. Access to the data may be granted upon reasonable request, subject to the approval of the company and compliance with ethical guidelines and data protection regulations.

Declarations

Ethics approval and consent to participate The data were provided by Pinterest, with administrative permissions granted by the company for research purposes. All experimental protocols were also reviewed and approved by Georgetown University's Institutional Review Board (IRB) under the study protocol STUDY00007635, categorized under (2)(ii) for Tests, surveys, interviews, or observation (low risk). The research was conducted in accordance with these guidelines to ensure participant privacy and data confidentiality. Ethical Guidelines/Accordance: All analyses were performed following relevant ethical guidelines and regulations. No personally identifiable information was collected or used, ensuring compliance with data protection standards and participant confidentiality.

Consent for publication The data were provided by Pinterest and involved anonymized user search logs. Direct informed consent for assessors from individual participants was obtained via a third-party data annotation company as the data were anonymized and provided under the company's data usage policies. However, all necessary administrative permissions and ethical considerations were addressed in accordance with our institutions' data-sharing agreements and legal frameworks. Since the data do not involve participants under the age of 16, the requirement for parental or legal guardian consent does not apply.

Competing interests The authors declare no competing interests.

References

1. Deng Y, Liao L, Zheng Z, Yang GH, Chua T-S. Towards human-centered proactive conversational agents. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24. Association for Computing Machinery, New York, NY, USA 2024.
2. Liao L, Yang GH, Shah C. Proactive conversational agents. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. WSDM '23, pp. 1244–1247. Association for Computing Machinery, New York, NY, USA 2023; https://doi.org/10.1145/3539597.3572724.

3. Liao L, Yang GH, Shah C. Proactive conversational agents in the post-chatgpt world. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23, pp. 3452–3455. Association for Computing Machinery, New York, NY, USA 2023; https://doi.org/10.1145/3539618.3594250.
4. Tang Z, Kulkarni H, Yang GH. High-quality dialogue diversification by intermittent short extension ensembles. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1861–1872. Association for Computational Linguistics, Online 2021; https://doi.org/10.18653/v1/2021.findings-acl.163.
5. Li J, Huffman S, Tokuda A. Good abandonment in mobile and pc internet search. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '09, pp. 43–50. Association for Computing Machinery, New York, NY, USA 2009; https://doi.org/10.1145/1571941.1571951.
6. Jiang J, Hassan Awadallah A, Shi X, White RW. Understanding and predicting graded search satisfaction. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining. WSDM '15, pp. 57–66. ACM, New York, NY, USA 2015; https://doi.org/10.1145/2684822.2685319.
7. Edwards A, Kelly D. Engaged or frustrated? disambiguating emotional state in search. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17, pp. 125–134. Association for Computing Machinery, New York, NY, USA 2017; https://doi.org/10.1145/3077136.3080818.
8. Gerrig RJ, Zimbardo PG. Psychology and life. Pearson Higher Education AU: Australia; 2015. p. 704.
9. Apter MJ. Motivational styles in everyday life: a guide to reversal theory. USA: American Psychological Association; 2001. p. 373.
10. Apter MJ. Personality dynamics: key concepts in reversal theory. Apter International: Manassas; 2005. p. 98.
11. Apter MJ. Reversal theory and personality: a review. J Res Pers. 1984;18(3):265–88. https://doi.org/10.1016/0092-6566(84)90013-8.
12. Aula A, Khan RM, Guan Z. How does search behavior change as search becomes more difficult? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '10, pp. 35–44. Association for Computing Machinery, New York, NY, USA 2010; https://doi.org/10.1145/1753326.1753333.
13. Xu L, Zhou X, Gadiraju U. Taskgenie: Crowd-powered task generation for struggling search. In: Web Information Systems Engineering—WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20-24, 2020, Proceedings, Part II, pp. 3–20. Springer, Berlin, Heidelberg 2020; https://doi.org/10.1007/978-3-030-62008-0_1.
14. Xu L, Zhou X, Gadiraju U. Revealing the role of user moods in struggling search tasks. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'19, pp. 1249–1252. Association for Computing Machinery, New York, NY, USA 2019; https://doi.org/10.1145/3331184.3331353.
15. Hassan A, White RW, Dumais ST, Wang Y-M. Struggling or exploring?: Disambiguating long search sessions. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. WSDM '14, pp. 53–62. ACM, New York, NY, USA 2014; https://doi.org/10.1145/2556195.2556221.
16. Feild HA, Allan J, Jones R. Predicting searcher frustration. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '10, pp. 34–41. Association for Computing Machinery, New York, NY, USA 2010; https://doi.org/10.1145/1835449.1835458.
17. Odijk D, White RW, Hassan Awadallah A, Dumais ST. Struggling and success in web search. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15, pp. 1551–1560. Association for Computing Machinery, New York, NY, USA 2015; https://doi.org/10.1145/2806416.2806488.
18. Ceaparu I, Lazar J, Bessiere K, Robinson J, Shneiderman B. Determining causes and severity of end-user frustration. Int J Hum-Comput Interaction. 2004;17(3):333–56.
19. Kim Y, Hassan A, White RW, Zitouni I. Modeling dwell time to predict click-level satisfaction. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. WSDM '14, pp. 193–202. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2556195.2556220.
20. Borisov A, Wardenaar M, Markov I, de Rijke M. A click sequence model for web search. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18, pp. 45–54. Association for Computing Machinery, New York, NY, USA 2018. https://doi.org/10.1145/3209978.3210004.
21. Hassan A, Jones R, Klinkner KL. Beyond dcg: User behavior as a predictor of a successful search. In: WSDM '10.
22. Hassan A, Shi X, Craswell N, Ramsey B. Beyond clicks: Query reformulation as a predictor of search satisfaction. In: CIKM '13.
23. Wang H, Song Y, Chang M-W, He X, Hassan A, White RW. Modeling action-level satisfaction for search task satisfaction prediction. In: SIGIR '14.
24. Huffman SB, Hochster M. 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. SIGIR '07, pp. 567–574. Association for Computing Machinery, New York, NY, USA 2007; https://doi.org/10.1145/1277741.1277839.
25. Fox S, Karnawat K, Mydland M, Dumais S, White T. Evaluating implicit measures to improve web search. ACM Trans Inf Syst. 2005;23(2):147–68. https://doi.org/10.1145/1059981.1059982.
26. Verma M, Yilmaz E, Craswell N. On obtaining effort based judgements for information retrieval. In: WSDM '16.
27. Guo Q, Jin H, Lagun D, Yuan S, Agichtein E. Mining touch interaction data on mobile devices to predict web search result relevance. In: SIGIR '13.
28. Han S, Yue Z, He D. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. ACM Trans Inf Syst. 2015;33(4):16–11634.
29. Lagun D, Hsieh C-H, Webster D, Navalpakkam V. Towards better measurement of attention and satisfaction in mobile search. In: SIGIR '14.
30. Huang J, Diriye A. Web user interaction mining from touch-enabled mobile devices. In: HCIR Workshop '12.
31. Kim J, Thomas P, Sankaranarayana R, Gedeon T, Yoon H-J. Pagination versus scrolling in mobile web search. In: CIKM '16.
32. Jones R, Klinkner KL. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08, pp. 699–708. Association for Computing Machinery, New York, NY, USA 2008; https://doi.org/10.1145/1458082.1458176.

33. Tamine L, Melgarejo JL, Pinel-Sauvagnat K. What can task teach us about query reformulations? In: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I, pp. 636–650. Springer, Berlin, Heidelberg 2020; https://doi.org/10.1007/978-3-030-45439-5_42.
34. Chilton LB, Teevan J. Addressing people's information needs directly in a web search result page. In: WWW '11.
35. Jackson S. How a critical thinker uses the web. 2021; https://api.semanticscholar.org/CorpusID:221493500.
36. Apter MJ. Developing reversal theory: some suggestions for future research. J Motivation Emotion Personality. 2013;1(1):1–8.
37. Apter MJ, Fontana D, Murgatroyd S. Reversal theory: applications and development. Psychology Press, 2014.
38. Tang Z, Yang GH. A re-classification of information seeking tasks and their computational solutions. ACM Trans Inf Syst. 2022. https://doi.org/10.1145/3497875.
39. Apter MJ. Reversal theory: a new approach to motivation, emotion and personality. Anuario de psicología/ UB J Psychol. 1989;42:17–29.
40. Hebb DO. Drives and the c. n. s. (conceptual nervous system). Psychol Rev. 1955;62(4):243–54.
41. Apter MJ. The Experience of Motivation: the Theory of Psychological Reversals. New York: Academic Press London; 1982. p. 378.
42. Han S, Yi X, Yue Z, Geng Z, Glass A. Framing mobile information needs: An investigation of hierarchical query sequence structure. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16, pp. 2131–2136. Association for Computing Machinery, New York, NY, USA 2016; https://doi.org/10.1145/2983323.2983654.
43. Bakshy E, Messing S, Adamic LA. Exposure to ideologically diverse news and opinion on facebook. Science. 2015;348(6239):1130–2. https://doi.org/10.1126/science.aaa1160.
44. Himelboim I, McCreery S, Smith M. Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on twitter. J Comp-Med Commun. 2013;18(2):40–60. https://doi.org/10.1111/jcc4.12001.
45. Weinfurt KP. Multivariate analysis of variance. 1995;245–276
46. Stahle L, Wold S. Analysis of variance (ANOVA). Chemom Intell Lab Syst. 1989;6(4):259–72. https://doi.org/10.1016/0169-7439(89)80095-4.
47. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc: Ser B (Methodol). 1977;39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
48. Manzoor E, Li R, Shrouty D, Leskovec J. Expanding taxonomies with implicit edge semantics. In: Proceedings of The Web Conference 2020. WWW '20, pp. 2044–2054. Association for Computing Machinery, New York, NY, USA 2020; https://doi.org/10.1145/3366423.3380271.
49. Gonçalves RS, Horridge M, Li R, Liu Y, Musen MA, Nyulas CI, Obamos E, Shrouty D, Temple D. Use of owl and semantic web technologies at pinterest. In: Ghidini C, Hartig O, Maleshkova M, Svátek V, Cruz I, Hogan A, Song J, Lefrançois M, Gandon F, editors. The Semantic Web—ISWC 2019. Cham: Springer; 2019. p. 418–35.
50. Yang H, Mityagin A, Svore KM, Markov S. Collecting high quality overlapping labels at low cost. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '10, pp. 459–466. Association for Computing Machinery, New York, NY, USA 2010; https://doi.org/10.1145/1835449.1835526.
51. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1023/A:1022627411411.
52. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Stat. 2000;28(2):337–407. https://doi.org/10.1214/aos/1016218223.
53. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232. https://doi.org/10.1214/aos/1013203451.
54. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser u, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 6000–6010. Curran Associates Inc., Red Hook, NY, USA 2017.
55. Savenkov D, Agichtein E. To hint or not: Exploring the effectiveness of search hints for complex informational tasks. In: Proceedings of the 37th International ACM SIGIR Conference on Research And Development in Information Retrieval. SIGIR '14, pp. 1115–1118. Association for Computing Machinery, New York, NY, USA 2014; https://doi.org/10.1145/2600428.2609523.