Searcher Struggle Detection via the Reversal Theory

Jiyun Luo
Pinterest Inc
Yan Yang
University of Nevada
Valerie Nayak
Carnegie Mellon University
Grace Hui Yang (grace.yang@georgetown.edu)
Georgetown University

Research Article

Keywords: Web Search, searcher struggle detection, reversal theory, information retrieva

Posted Date: July 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1762506/v1

Searcher Struggle Detection via the Reversal Theory

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

[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, D.C., USA.

*Corresponding author(s). E-mail(s): grace.yang@georgetown.edu;
Contributing authors: jluo@pinterest.com;
yy490@georgetown.edu; vjn@andrew.cmu.edu;

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, recent studies have suggested 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. The reversal theory (RT) 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. Evaluations on week-long Web search logs confirm that the proposed method can statistically significantly improve state-of-the-art struggle detection methods.

Keywords: Web Search, searcher struggle detection, reversal theory, information retrieval

1 Introduction

Searcher struggle is the event that a searcher "makes strenuous efforts in the face of difficulties"[1] during a search process. At a struggling moment, the user experiences a negative emotion of feeling frustrated, upset, or annoyed by the search activity. Detecting searcher struggle is an important task for Web search engines because it can alert the search engines to timely adjust their algorithms. Most existing struggle detection methods are classifiers using features that measure user efforts [1, 2]. This is because struggling behaviors are often demonstrated through excessive repetitions of user actions. For instance, when searching for "curly hair dye", a user issued over twenty queries and still could not find a satisfactory result. It is straightforward to assume that a large number of user efforts suggest user struggles.

However, this conjecture does not always align with real-world observations. One observation is that non-struggling search tasks may also possess many user-issued queries and clicks. For instance, a user searched for "guy proposing ideas," and the log shows that he browsed over 900 search results and bookmarked 27; the user reported he enjoyed the search throughout the session. Large amount of efforts seem to map to both struggling and non-struggling sessions. Recent lab studies have raised similar concerns. Edwards and Kelly [3] pointed out that although an increase in user effort might help predict searcher struggle, such an increase can also signify engagement or exploration. An engaged user, much like a frustrated one, tends to make many queries on the same topic and click on many presented links. Without an effective mechanism, the distinguishing power of effort-based features diminishes, making it difficult to separate a negative experience such as struggle from positive experiences such as engagement and thus challenging to improve Web user experiences.

This paper refers to the psychology literature and proposes a new feature modulation method for struggle detection. Our work is motivated by the Reversal Theory (RT) [4–8]. The reversal theory is a "mode-based" psychology theory that challenges some fundamental assumptions in the fields of motivation and personality. Rather than assuming static personality traits, revercal theory focuses on the "complexity, changeability, and dynamics of human motivation and personality" [7]. Its key insight is that people change in the flow of everyday life, and their personalities and motivations can reverse [6]. The theory suggests some counter-intuitive ideas, such as whether a user is conforming to rules or rebelling has little impact on his struggling behaviors, but whether

[1] merriam-webster.com

she works on a serious or playful task would matter. We conducted statistical hypothesis tests on these arguments and confirmed them on commercial search logs.

Moreover, reversal theory states that human motivations can be organized into a few dimensions and each dimension consists of two opposing states. Within one dimension, a person can reverse between the pair of states but can be only active at one of them at any given moment, which indeed suggests a bi-modal distribution instead of a single-modal distribution. We leverage this insight in this paper and propose a novel and effective feature modulation method for struggle detection. We propose to (1) select features responsible for the related reversal theory dimension, (2) moderate the selected features by reducing the bias between the two distributions in the bi-modal arousal model, and (3) send the modulated features to classifiers to identify if a search session has struggling moments.

Our method can be used in combination with any feature-based struggle detection method. We evaluate our approach on one-week-long search logs collected from both mobile and PC platforms. The experimental results show that our method is highly effective; it can significantly improve a few top-performing methods by ∼5% accuracy and ∼9% positive precision.

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, Section 8 concludes the paper.

2 Related Work

In this section, we review related work both from information retrieval (especially search struggle research) and from reversal theory.

2.1 Search 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 struggles and their search behaviors.

2.1.1 Lab studies

The lab studies on search struggles 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. [9] 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.

Xu et al.'s lab study [10] 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 more complicated bi-modal arousal model and our focus is on feature transformation.

A highly relevant lab study to ours is Edwards and Kelly's work [3]. They observed that although user efforts increase might help predict searcher struggles, such increases can also indicate engagement, the opposite of 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.

2.1.2 Log-based studies

Studies on search struggles that are based on search logs are 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. [11] worked on detecting "struggling" and "exploring" (including "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 lent 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 re-use existing features.

J. Li et al. [1] 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.

Table 1 Motivation Dimensions and States.

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

Feild et al. [12] 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 [11, 13], our work is a novel method to modulate these features for more effective struggle detection.

Other well-studied, negative search experiences besides struggles include irrelevancy and dissatisfaction [12]. 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 Reversal Theory

Reversal theory (RT) is a psychology theory that studies personality dynamics and motivations. It recognizes that people "are essentially changeable and move between different motivational styles" [5]. This theory "sheds light on the paradoxes of risk-taking, addiction, rebelliousness, and other areas of motivation, emotion, and personality" [5].

The key ideas in the reversal theory 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.[2] Table 1 lists the two opposing states in each of the four RT 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.
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 "challenges some of the basic concepts of mainstream psychology, such as the trait concept of personality" [5]. For example, RT "explains how anxiety can be reverted almost instantaneously into excitement and vice versa" [14]. It also explains that a person "may now experience his job as an obligation and at a later time experience the very same job as a kind of game." In another example, RT improves the *model of arousal* [15] in traditional motivational theories from a single-modal model into a bi-modal model, which measures the complex relationship between one's happiness level[3] and effort level.[4] This new bi-modal model of arousal is the basis for our work to perform feature modulation.

The first two RT dimensions, "means-ends" and "rules," describe how users perform tasks. They are relevant to our discussion and examined in this paper. The last two dimensions, "transactions" and "relationships," describe interpersonal interactions instead of user and task; thus are less relevant and not discussed in this paper.

2.3 Struggle vs. Irrelevancy vs. Dissatisfaction

Struggle [9, 11, 16], irrelevancy, and dissatisfaction [12] 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* or frustrated. Studies about results irrelevancy, user dissatisfaction are not within the scope of this paper.

Search success and search satisfaction Concepts related to struggles, search satisfaction, and search success has been extensively studied. Search success has been interpreted as content relevance [17], fulfillment of information need in [18–20] and the searcher experience of pleasure in [2, 21]. Fox et al. [22] 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

[2] Some books call these states "meta-motivational states", "motivational styles," or motives. For simplicity, we call them motivational states or states in this paper.
[3] Also called hedonic level in some books.
[4] Also known as arousal level.

page or a search session well. Through a lab study, Huffman and Hochster [21] 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. [23] concluded that user satisfaction is related to the relevance of examined webpages and the effort needed to locate the relevant content in these webpages. Jiang et al. [2] 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. [20] could predict search session success with high accuracy. Hassan et al. [19] 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.

Mobile search There is also prior work about detecting relevance in mobile search. Guo et al. [24] conducted lab studies and provided a predictive model for detecting URL relevance in mobile search. They revealed that "inactivity" indicated "reading" in mobile web search, but not in search using PCs. They also found that swiping was similar to scrolling on PCs in that both were signals that suggest content irrelevance. Han et al. [25] found that mobile touch interaction signals on SERP were more effective than landing web page signals for predicting content relevance. Lagun et al. [26] proved that scrolling past search cards and spending more time on contents below search cards are clear signals of non-relevance. Huang and Diriye [27] pointed out that changing viewport coordinates are more accurate than user touch coordinates for predicting content relevance in mobile search. Kim et al. [28] 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 Web Search Struggle Detection

A Web search **struggle** is an event that a user feels frustrated by the search results at some point during a search session. We formulate the task of Web search struggle detection as a binary classification problem, with the two classes "struggling" and "non-struggling". The time unit to study whether a struggle happen is chosen to be a single search session, as it is 1) a natural block for a search task and 2) produces more stable responses than every single user action. We define:

- A **struggling session** as a session that contains a struggling moment, where a user feels frustrated by the search results at some point during the session.

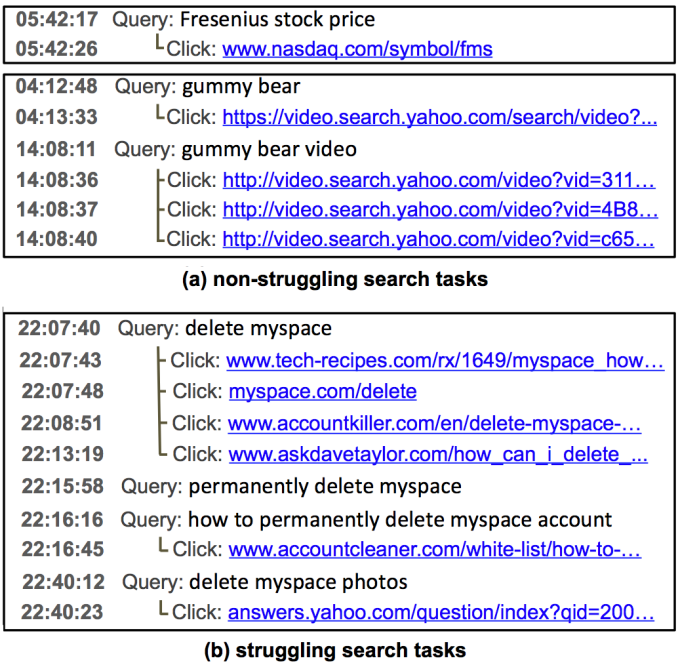

Fig. 1 Examples for non-struggling sessions vs. struggling sessions.

- A **non-struggling session** is a session that contains no struggling moment throughout the session. In this work, we obtain search sessions by segmenting them from query logs (See details in Section 6).

Figure 1 shows examples for both non-struggling and struggling search sessions.

Moreover, we denote a struggle predictor Y and an input search session s, described by a vector of features $X(s)$. The probability of s being struggling is $P(Y = 1|X(s), \Theta)$, with Θ being the model parameter. The classifier predicts "struggling" when $P(Y = 1|X(s), \Theta) > .5$ and "non-struggling" otherwise. In Section 7, we experiment on multiple classifiers to show the effect of the proposed feature modulation method. The class labels are obtained by third-party manual annotation (Section 6).

3.2 Features

The input feature vector $X(s)$ is automatically extracted from a query log. Our features include features proposed by prior works and new features presented in this paper. Most features are indicators of certain user efforts, which measure the quantity and diversity of user actions in a session.Table 2 lists these effort-based features and groups them into seven feature groups.

Efforts to Query. The first group of features measures user efforts spent on writing queries. We derive most of them from Edwards and Kelly's work [3].

Table 2 Effort-Based Features. * marks new features. ** marks feature groups selected for feature modulation.

Efforts to Query**	**Efforts to Scroll**
Number of queries in a session [3]	Screen size*
Number of unique queries in a session [3]	Total and avg. number of scrolling down actions*
Avg. number of terms per query [3]	**Efforts to Re-formulate Queries**
Avg. number of characters per query [3]	Avg. cosine similarity between every query and the first query [11]
Number of manually-typed queries*	Avg. cosine similarity of every query pair in a session [11]
Percentage of manually-typed queries [3]	Avg. edit distance per adjacent query pair [11]
Percentage of suggested queries (that are automatically corrected, suggested, or completed by the search engine) [3]	Number of query generations (when removing one or more terms from its previous query)[11]
The longest query's position in a session*	Number of query specifications (one or more terms are added into its previous query [11]
Efforts to Click**	Difference between the first query length and the avg. query length*
Total and avg. number of clicks in [11, 12]	Standard deviation of query lengths in a session*
Total and avg. number of Satisfactory (SAT) clicks [11, 12]	Avg. number of terms appear in the previous query [11]
Percentage of queries without clicks [17, 21, 22]	Avg. number of terms added to the previous query [11]
Maximum and avg. number of adjacent queries without clicks*	Avg. number of terms deleted from the previous query [11]
Total and avg. number of images clicked in a session*	Avg. number of terms that substitute terms in the previous query [11]
Total and avg. number of ads clicked in a session*	**Efforts to Diversify****
Total and avg. number of bookmarks clicked in a session*	Percentage of unique URLs among all clicked URLs [11]
Number of events (clicks, bookmarks, and queries) in a session*	Percentage of the unique domain (DNS) names among all clicked URLs [11]
Number of clicks at the first two queries*	Total number of unique clicks*
Number of clicks at the third and fourth queries*	Total number of unique topics [11]
Number of clicks at the fifth and sixth queries*	Entropy of topic distribution in a session [11]
Whether the session ends with a click*	**Efforts to Issue Rare Queries & Rare Clicks**
Efforts to Read**	Log(1 + avg. query frequency in popularity data) [11, 13, 21]
Total dwell time of all clicks [11, 12]	Log(1 + a query's avg. SAT clicks in popularity data) [11, 13, 21]
Avg. number of image impressions per SERP*	Log(1 + a query's avg. clicks in popularity data) [11, 13, 21]
Total number of zoom-in on result images*	A query's avg. click entropy in the popularity data [11, 13, 21]
Log(1 + avg. dwell time per click in a session)*	Log(1 + a query's avg. number of fast-back clicks (whose dwell time is less than 15s) in the popularity data) [11, 13, 21]
Log(1 + avg. dwell time per click exclude clicks for the last query)*	Log(1 + a clicked URL's avg. click frequency in the popularity data)*
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)*	

Usually, many queries in a session indicate a user has spent a lot of effort typing queries. But, redundant queries may be "copy-and-pasted", which requires little effort. Likewise, using system-suggested queries also requires little user effort. We, therefore, measure their ratios among all queries in the session. The new feature we propose is the longest query's position in a session, inspired by Aula et al.'s work [9], where they pointed out that the longest query often appears at the end of a successful session.

Efforts to Click. The second group of features measures user efforts spent on clicking the search result URLs. Existing features include total numbers of clicks and satisfactory (SAT) clicks, widely used to infer relevance, satisfaction [17, 21, 22], as well as struggles [11, 12]. In addition to them, we propose new features to measure clicks for different types of search results, including images, ads, and Web pages. We also propose new features indicating abandonment and little effort – for instance, the average number of adjacent queries that receive no clicks. Moreover, we also introduce bookmarked results and clicks on different phases of a session.

Efforts to Read. The third feature group measures user efforts spent on reading and examining the content of search results [11, 12]. Except for the standard dwell time feature, we propose to count the number of *zoom-in* actions on image results and specific types of *dwell time* on different kinds of returning items and search results sections.

Efforts to Scroll. The fourth feature group measures user efforts spent on reaching out to results beyond the current sight. These features are all new. Here we propose to count the numbers of scrolling-downs and screen-resizing. When a user scrolls down or resizes her screen, a new search request is sent to the back-end engine to get and fresher search results for the same query. We obtain the numbers of scrolling-downs and resizing by counting the number of pagination requests from the user.

Efforts to Reformulate Queries. The fifth feature group measures user efforts spent on editing queries and re-articulating the information needs. [11]. We add new features to measure the variance of the query lengths after editing.

Efforts to Diversify. The sixth feature group measures user efforts in diversifying the search results' content and the examination process. Features in this group indicate how exploratory the user and the search are. We mainly use features proposed by [11] for click diversity and topical diversity. A new feature added is the total number of unique clicks.

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. We obtain the Web population's click data from a commercial search engine from 11/15/2020 to 11/21/2020 and use that as the basis to derive which queries and clicks are rare.

Besides these seven feature groups, we also recruit features that are not effort-related. For instance, we use the taxonomy topic of the search task as a categorical feature. However, our method mainly acts on the effort-based features shown in Table 2.

4 The Reversal Theory

RT is a psychology theory that studies *personality dynamics* [6]. It states that in the flow of daily life, a person regularly reverses, like a teeter-totter, between opposing motivational states. This section introduces the basic concepts in RT and how they relate to Web search struggle detection.

4.1 Opposing Motivational States

RT groups human motivations into four dimensions (also known as "domains"). They are "means-ends", "rules", "transactions", and "relationships" [14]. The first two dimensions describe how a user performs tasks and are relevant to this paper. The last two dimensions describe interpersonal interactions and are less relevant to this paper. We can find the dimensions and states in Table 1.

The first dimension, "means-ends", is about achievements, goals, and enjoyments of process. It has two opposing motivational states that "reflect people's motivational styles, the meaning they attach to a given situation at a given time, and the emotion they experience" [6]. The two states are *telic*, at which one is motivated by achievement, task completion, and fulfilling of goals; and *paratelic*, at which one is playful and seeks excitement and fun. When one is at the telic state, she is serious about the task at hand and focuses on achieving the task goal, whereas when one is paratelic, the activity she does is not for the sake of the task's goal but the task's own sake. E.g., people run because they enjoy running themselves, not because they want to win a medal [6].

In the context of web search, telic and paratelic states can parallel to goal-oriented and non-goal-oriented search tasks. A user at the telic state would focus on finishing the search task, for instance, to look for job vacancies or medical help. A user at the paratelic state would seek to enjoy the search process itself, for example, browsing for fun videos on YouTube.

The second dimension, "rules", is how routines, expectations, and constraints could direct a person's activity. The two opposing states are *conformist*, at which a person tends to operate within rules and expectations; and *negativistic*, at which a person wishes to push against regulations and explore new possibilities. An example for a conformist is "I am eating because this is what I am supposed to do at this moment." [6]. And when one thinks "I am eating because I am not supposed to eat at this moment", she is at the negativistic state. Interestingly, the same behavior can be motivated by opposite reasons.

In the context of web search, conformist and negativistic are parallel to non-exploratory and exploratory search behaviors. When a user is in a conformist state, she obeys rules and meets others' expectations. For instance, the user

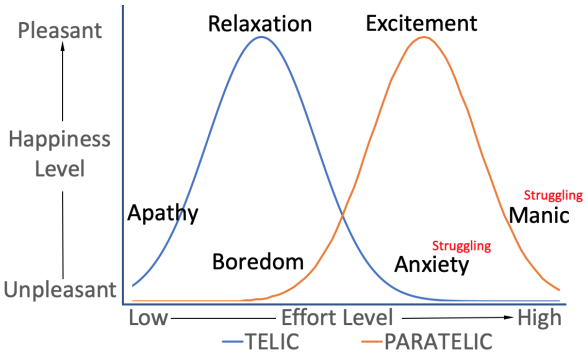

Fig. 2 Arousal Model for Means-ends; Both Anxiety and Manic indicate struggling. (adapted from [29]).

would use the query suggested by the search engine instead of creating her own. On the contrary, when a user is in a negativistic state, she can explore novel ideas. For instance, she would issue rare queries, read from different URLs, and prefer more novel and diverse search results.

The reversal theory's key insight is that people "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 trait, a person teeter-totters in her motivational states and the states are completely opposite to each other.

4.2 RT's Bi-Modal Arousal Model

Struggle detection can be thought of to detect one's (un)happiness level when one's search behaviors change. In psychology, the *model of arousal* [15] tells the relationship between a person's (un)happiness level and arousal level. Arousal is a central concept in motivational theories, which indicates the intensity of activities and feelings a person experiences. In our context, it corresponds to the intensity of search actions, i.e., the level of effort. Therefore, the arousal model can help establish a relationship between a user's happiness and effort levels.

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 [15]. 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 the reversal theory, the arousal model is a bi-modality model. RT assumes that there are two optimums present in the arousal model.

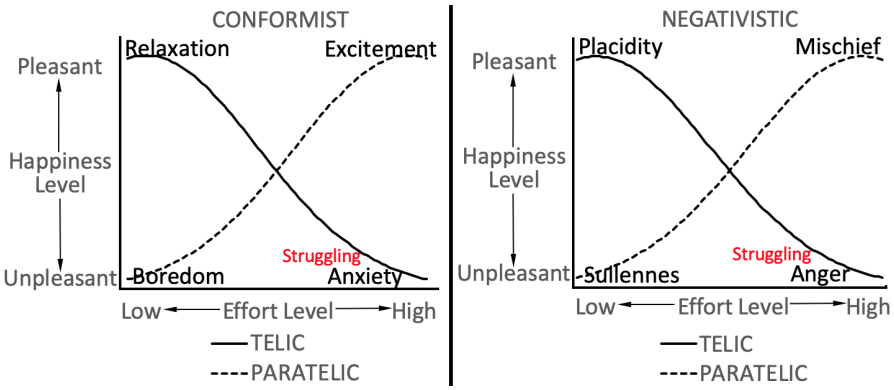

Fig. 3 Interplay of "means-ends" and "rules" (adapted from [5]).

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 2 [29] 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 are present.

In the context of Web search, this bi-modal arousal model could be the cause of inconsistent prediction of user struggles because the same level of user efforts can indeed map to two different happiness levels, depending on at which state the user is at the moment. For instance, the same effort level could mean "struggling/anxiety" for a user at the telic state and "excitement" for a user at the paratelic state. To resolve this inconsistency, we propose to shift the two (state) curves horizontally closer to each other until they overlap and then separate the left and right end, given that the struggling instances lie at the right end.

4.3 "Rules" Dimension is Irrelevant

As we mentioned before, the first two RT dimensions are seemingly relevant to Web search because they care about users and tasks. However, contrary to our intuition, RT suggests that the "rules" dimension has little impact on struggling, and only the "means-ends" dimension matters. It [5]'s interplay of the first two dimensions (Figure 3). The two sub-figures in Figure 3 depicts RT's arousal model when the second dimension state is "conformist" and "negativistic," respectively. We notice that in both sub-figures, struggles happen

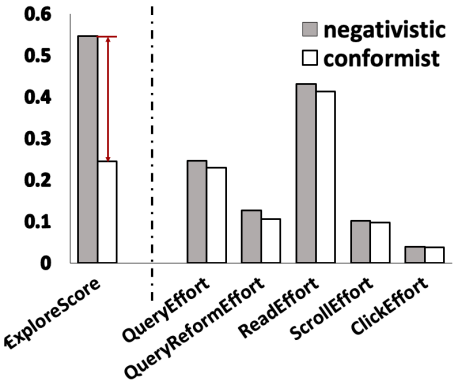

Fig. 4 Differences in Features at Two "Rules" States. ↕ marks significant differences at p<.05.

at 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 RT dimension and one for the second, on a whole week's query log collected from a commercial search engine (more details about the data in Section 7). 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:

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

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.

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.

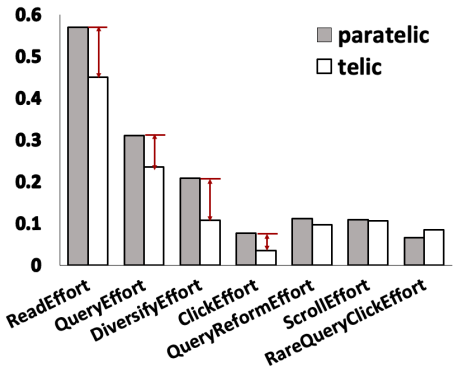

Fig. 5 Feature value gaps along "Means-ends". The red ↕ indicates the difference is statistically significant at p<.05.

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 [30] test across all feature groups and 5 ANOVA [31] 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 $p > .05$ 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 RT.

Further, we plot the mean feature values for the conformist and negativistic states in Figure 4. 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 RT suggests that when the first two 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 RT that the first "means-ends" dimension is influential to a user's struggle. We, therefore, use "means-to-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 the 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 RT, a search session would be at any one moment only at either state, not both. Second, based on what RT'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

RT's bi-modal model of arousal (Figure 2) 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 6 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 [32]. In the mixture model, a data point can have a soft mapping onto both Gaussians. However, based on RT, at any one moment, a user can only be at one of the opposing states, not both. We choose to follow what RT 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. RT considers telic states are associated with "serious" tasks, and paratelic states are associated with "playful" tasks [6]. Other research also pointed out that search topic shows the impact on searcher behaviors [1]. We determine a session's search topic using a taxonomy. Without losing generality, an internal taxonomy, constructed by graph-based algorithms [33, 34], at Pinterest, is used. Note that the proposed method is general and should be compatible with most other Web taxonomies.

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. It measures the similarity between the URL link's text with the category's name using tf-idf and word2vec cosine similarity. The taxonomy category with the highest similarity score to the URL text becomes the label to the URL. We use Gradient Boosted Decision Trees (GBDT) to combine the similarity scores (with a learning rate of 0.1, minimum split loss 0.5, and maximum tree depth 8). Third, We chose the most frequent URL label in the session as the search topic for the session.

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".

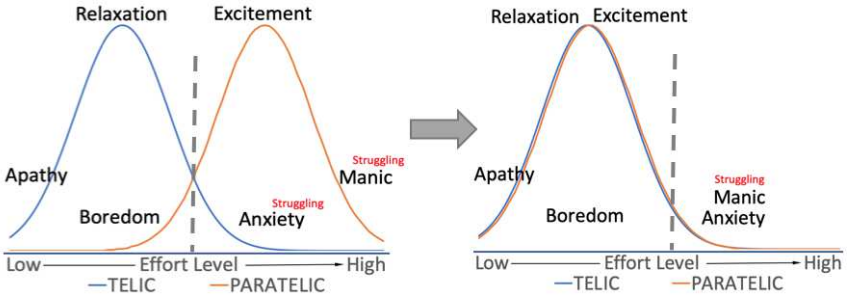

Fig. 6 Modulation to separate struggles from non-struggles.

5.2 Select "Means-Ends" Features

RT 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
$$\frac{value - min\_value}{max\_value - min\_value}. \tag{2}$$
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 RT 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.

Figure 5 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 RT states. We leverage this information to remove the bias between the two Gaussians.

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 Figure 6, 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 un-modulated features with a wide range of classifiers for struggle detection.

5.4 Detect 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":

$$\mathcal{I} = \begin{cases} 1 \text{ if } P(Y=1 \mid \mathbf{X'(s)}, \Theta) > 0.5 \text{if } P(Y=1 \mid X'(s), \Theta) > 0.5 \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

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

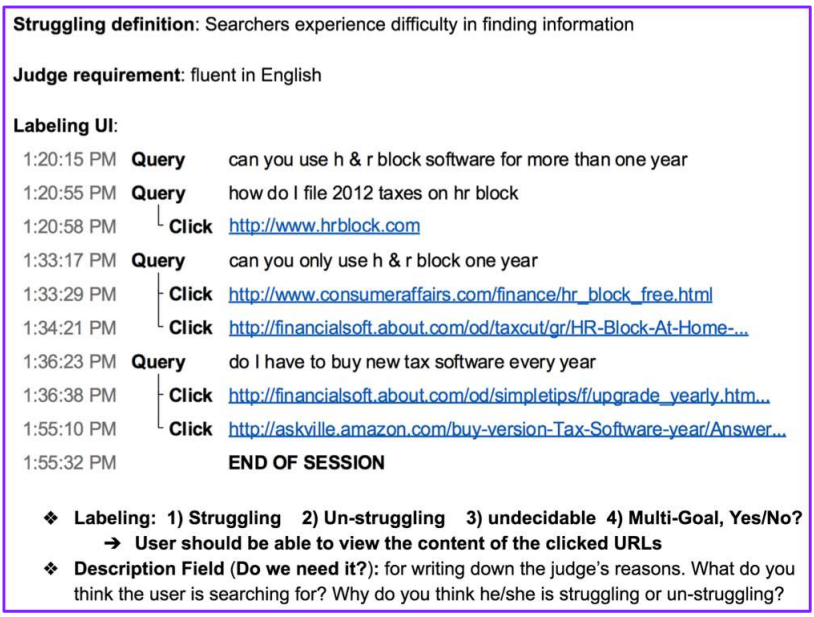

Fig. 7 Annotation Labeling User Interface.

6 Experimental Setup

We conduct 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 a commercial search engine in the period of 11/22/2020 ∼ 11/28/2020. We created two datasets, consisting of search logs using a PC browser and a mobile App. The user activities on the two platforms are slightly different due to different platform interfaces.

We take the following steps to prepare our data. First, we segment the search log into topically coherent segments [35], each corresponding to a session. We segment the sessions following [36]. 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% for 10-fold cross-validation.

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) Struggle, 2) Non-struggle, 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 [37]. 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 Figure 7.)

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

> *We have a set of user search sessions from year 2020. 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 share some examples of struggles and non-struggle sessions and our reasoning with the assessors. For example,

> *A person 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. I think she found what she's looking for, hence there is no struggle here.*
>
> *A person 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 also makes me think they're struggling. 4) Then there are things like: "What is the search topic? Is someone just having fun and trying to find the story behind the movie? Why continue re-querying on that 'true story' angle and still not focus on any article?" This to me feels like struggle, but I'd be hard-pressed to explain why 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 achieve an inter-coder agreement of 73.3%.

Third, in the end, we obtained 2,157 labeled sessions in total, including 601 struggling and 1,556 non-struggling sessions. We then processed these sessions to produce a feature vector for each of them. Table 3 reports the dataset statistics.

Table 3 Dataset statistics.

	Duration	#Sessions	#Struggle	#NonStruggle	#Query/Session
Mobile		1,123	299	824	5.39
PC	11/22 ∼ 28,	1,034	302	732	4.62
Total	2020	2,157	601	1,556	5.02

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.

6.2 Baseline Classifiers

We experimented with several classifiers for struggle detection. They include common baselines as well as best-performing struggle detection methods.

- **ZeroRule** is a naive baseline that classifies instances based on the majority label in the ground truth.
- **SVM** is the support vector machine classifier [38], which is a top linear classifier. We use a radial kernel with a kernel coefficient of 0.016 and a cost of 2.0.
- **LM** is a logistic regression classifier [39], which is also a leading linear classifier.
- **MART** is a Multiple Additive Regression Trees (MART) classifier [40], which is a top-performing non-linear classifier widely used in Web search. We set MART's n.tree to be 8000 and shrinkage 0.005.
- **Transformer** [41] is a binary classifier built on top of a multi-head self-attentive deep neural network. We use a batch size of 64, a learning rate of 0.00005, and a dropout rate of 0.1. All classifiers use features in Table 2 and the categorical feature, search topic.
- We also re-implement a state-of-the-art struggle detection system proposed by **Hassan et al.** [11] since it shares the most features with us. We used their features presented in [11] and experiment on our dataset. This model performs similarly compared to their reported results. All classifiers are trained and validated with 10-fold cross-validation.

6.3 Runs under Comparison

For each baseline, we experiment with three different settings. (1) The original setting. (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".

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

	accu.	impr.	pos. p	impr.	pos. r	impr.
ZeroRule	0.7337	–	–	–	0.0000	–
LM	0.8413		0.7239		0.6561	
LM+FMNS	0.8621	2.5%↑	0.7342	1.4%↑	0.6393	2.5%↓
LM+FM	0.8910	5.9%↑ †	0.7513	3.7%↑ †	0.6116	6.8%↓
SVM	0.8565		0.7956		0.7414	
SVM+FMNS	0.8675	1.3%↑	0.7940	0.2%↓	0.7180	2.9%↓
SVM+FM	0.8928	4.2%↑ †	0.8511	7.0%↑ †		
Hassan et al.[11]	0.8507		0.7729		0.6439	
Hassan et al.+FMNS	0.8626	1.4%↑	0.7962	3.0%↑	0.6447	1.3%↑
Hassan et al.+FM	0.8786	3.3%↑ †	0.8419	8.9%↑ †		
MART	0.8740		0.7968		0.7305	
MART+FMNS	0.8835	1.1%↑	0.8042	0.9%↑	0.7144	2.2%↓
MART+FM	0.9055	3.6%↑ †	0.8666	8.8%↑ †	0.7754	
Transformer	0.8811		0.8036		0.7457	
Transformer+FMNS	0.8902	1.0%↑	0.8062	0.3%↑	0.7414	0.6%↓
Transformer+FM	0.9207	4.5%↑ †	0.8725	8.6%↑ †	0.7629	2.3%↑

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). Among them, *accuracy* and *positive precision* are the main metrics. Positive precision is important for web search [42]. Because precise assistance is preferred over generic assistance by human users; hence precisely predicting user struggles is very important.

7 Experimental Results

7.1 Main Results – Search Struggle Detection Effectiveness

Tables 4 and 5 report the effectiveness of struggle detection by the runs compared to the mobile and PC datasets. We also highlight the percentage improvement of a run from its original run. In addition, we report one-tailed t-test statistical results between the "+FM" runs and the initial runs.

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 PC 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

Table 5 PC: performance of struggle detection (Up and down arrows indicate absolute performance increase and decrease. † 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".

	accu.	impr.	pos. p	impr.	pos. r	impr.
ZeroRule	0.7079	–	–	–	0.0000	–
LM	0.8384		0.7624		0.8316	
LM+FMNS	0.8425	0.5%↑	0.7742	1.5%↑	0.8260	0.7%↓
LM+FM	0.8676	3.5%↑ †	0.8141	6.8%↑ †	0.8395	0.9%↑
SVM	0.8617		0.7843		0.8810	
SVM+FMNS	0.8757	1.6%↑	0.8008	2.1%↑	0.8810	0.0%
SVM+FM	0.9100	5.6%↑ †	0.8625	10.0%↑ †	0.8935	1.4%↑
Hassan et al.[11]	0.8418		0.7646		0.8374	
Hassan et al.+FMNS	0.8604	2.2%↑	0.7927	3.7%↑	0.8416	0.5%↑
Hassan et al.+FM	0.8981	6.7%↑ †	0.8499	11.2%↑ †	0.8557	2.2%↑
MART	0.8775		0.8223		0.8605	
MART+FMNS	0.8952	2.0%↑	0.8416	2.3%↑	0.8683	0.9%↑
MART+FM	0.9293	5.9%↑ †	0.8874	7.9%↑ †	0.9024	4.9%↑ †
Transformer	0.8813		0.8266		0.8649	
Transformer+FMNS	0.9005	2.2%↑	0.8494	2.8%↑	0.8737	1.0%↑
Transformer+FM	0.9372	6.3%↑ †	0.8988	8.7%↑ †	0.9090	5.1%↑ †

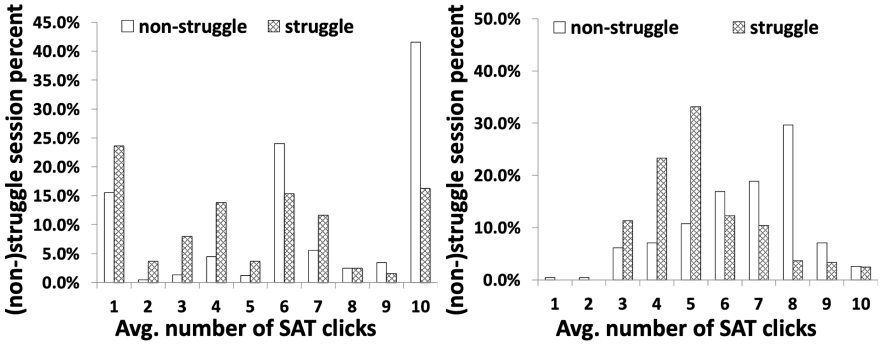

Fig. 8 distribution of struggling and non-struggling sessions over avg. number of Satisfactory (SAT) clicks.

what RT 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 RT considers irrelevant. The weak performance from the "+FMNS" runs again supports this insight from RT, besides our hypothesis test in Section 4.3.

7.2 Impact of Feature Modulation

We further investigate the effect on individual features before and after feature modulation. In this investigation, we study the magnitude of the features for both struggling and non-struggling sessions and their distributions. We pick several features from feature groups that show significant differences between paratelic and telic states in Figure 5 to demonstrate the effect of our feature

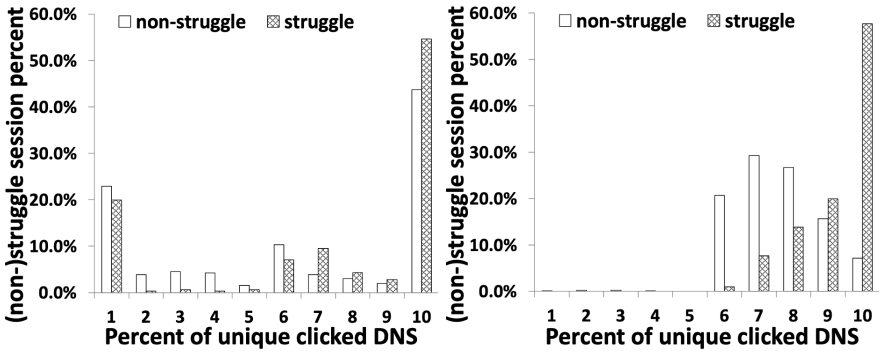

Fig. 9 distribution of struggling and non-struggling sessions over the percentage of unique clicked DNS domains.

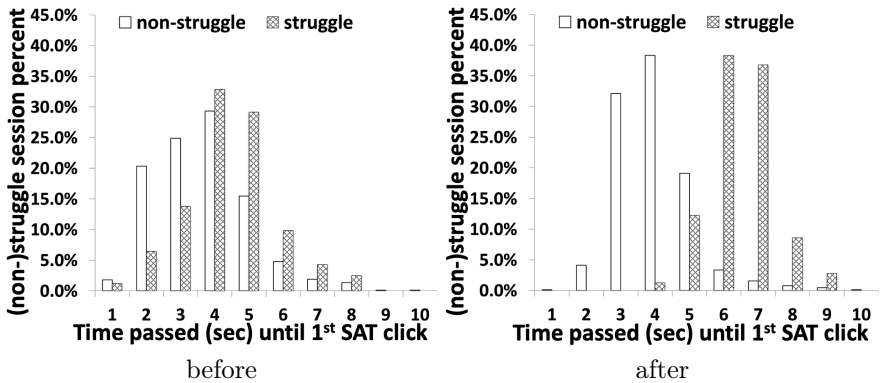

Fig. 10 distribution of struggling and non-struggling sessions over time passed until the 1st SAT click.

modulation. These features are average number of SAT clicks in the "Click Effort" group; percentage of unique clicked DNS domains in the "Diversify Effort" group; time passed until the first SAT click in the "Read Effort" group; number of unique queries, average number of characters per query, and number of manually typed queries in the "Query Effort" group'.

The figures are generated by first dividing the magnitudes of avg. SAT clicks per session, percentage of unique clicked domains, time passed until 1st SAT click, number of unique queries, number of average query characters, and number of manually typed queries into ten bins evenly and then plotting the ratio of struggling and non-struggling sessions in each bin.

Figure 8 shows that before feature modulation, non-struggle and struggle sessions' average number of SAT clicks distributions are mixed and present no obvious patterns. After feature modulation, average number of SAT clicks in non-struggle sessions and struggle sessions forms two different "bell-shape" curves, one peaks at five, and the other peaks at eight. The new distribution

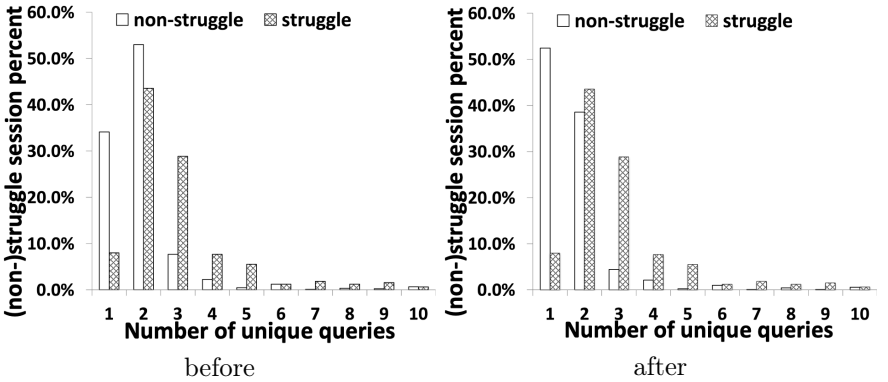

Fig. 11 distribution of struggling and non-struggling sessions over the number of unique queries.

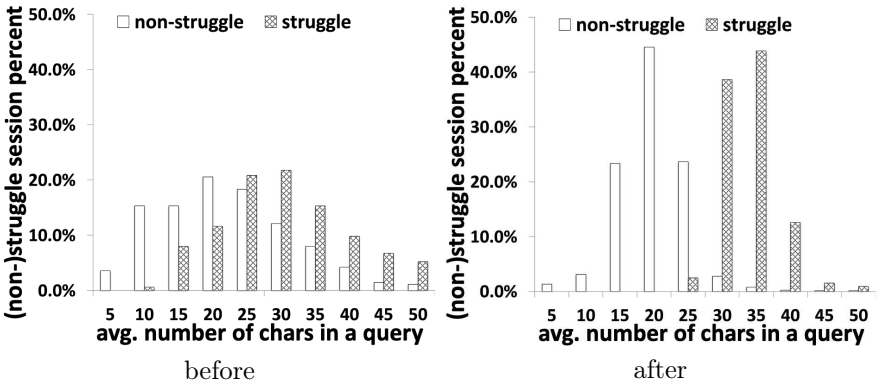

Fig. 12 distribution of struggling and non-struggling sessions over average number of characters in a query.

suggests that users tend to SAT click more documents in the non-struggle sessions than in the struggle sessions.

Similarly, Figure 9 shows that after modulation, the value curve of "percentage of unique clicked DNS" in non-struggled sessions peaks at seven; while in struggled sessions, it peaks at 10. This suggests that users more frequently browse results from different websites in the struggle sessions than in the non-struggle sessions.

In Figure 10, before modulation, the curves of non-struggle and struggle sessions are close to each other. After modulation, the curve of non-struggle sessions shifts left and centers at value four, while the curve of struggle sessions shifts right and centered at value 6. After modulation, it is obvious that users need more time to find the first SAT click document in struggle sessions than in non-struggle sessions.

In Figure 11, before modulation, both curves peak at value two. After modulation, the non-struggle sessions' curve peaks at one, while the struggle

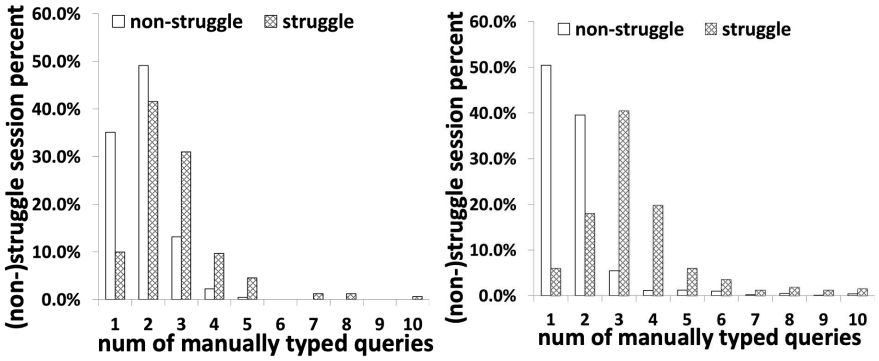

Fig. 13 distribution of struggling and non-struggling sessions over number of manually typed queries.

sessions' curve peaks at two, which suggests that users tend to issue more unique queries in struggle sessions than in non-struggle sessions.

In Figure 12, after modulation, the distributions of average number of characters per query in struggle and non-struggle sessions are further separated. Users generate longer queries in struggled sessions than in on-struggle sessions.

In Figure 13, after modulation, we observe that users usually type one or two queries manually in non-struggle sessions, while the number increases to two to four in struggle sessions.

To summarize, as we can see, before feature modulation, the distributions of struggling and non-struggling sessions do not present any exciting patterns. After feature modulation, however, the distributions of struggling and non-struggling sessions are centered at different bins. It suggests our method helps these features better separate the binary classes and become more valuable features in this classification task.

8 Conclusion & Future Work

This paper takes a unique solution path, uses insights from established psychology theories, and derives working solutions from there. Inspired by the Reversal Theory in psychology, we propose a novel feature modulation method to work as a component in Web search struggle detection. Our method is highly-effective and straightforward. It can be easily re-implemented and run on other datasets.

Our work is an initial application of the reversal theory, which we think will show great potential in other sub-fields of information retrieval. For instance, users' psychological states can be reversed when proper triggers present. What triggers a reversal to happen can be inherent tendency, situations, or just natural biological rhythm. To reverse from telic to paratelic state, situational triggers can be removal of threat, entertainment, or humor. To reverse from

paratelic to telic, situational triggers can be unavoidable tasks, sudden threat, and need for strategic decisions [6].

Interestingly enough, Web searcher struggle itself can be a trigger to reversal. It implies that across multiple sessions, if a struggling session is detected, we may take advantage of this knowledge to infer states for the later sessions. We leave these interesting directions as future work.

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

Statements and Declarations

- Funding: The first, second, and last authors were supported by the United States National Science Foundation grant number IIS-1453721 while they worked on this project at Georgetown University.
- Competing interests:
  - Organizations: Pinterest Inc, Georgetown University, Carnegie Mellon University, and University of Nevada, Reno.
  - Individuals: Jamie Callan, Tat-Seng Chua, Charlie Clarke, Sean MacAvaney, Razieh Rahimi (Negin), Eugene Yang.
- Compliance with Ethical Standards: The human annotators were all signed informed consent. The user study was conducted internally in a commercial search engine company.
- Consent for publication: All authors agree for the manuscript submission and publication.
- Availability of data and materials: The dataset belongs to a commercial search engine, containing user click and browsing behavior data, which unfortunately cannot be made available to the public. However, the methodology described in the paper is general enough to be easily adapted to other datasets.
- Code availability: The code can be made available to the public upon publication.
- Authors' contributions:
  - Jiyun Luo: Methodology, Formal analysis and investigation, Writing - original draft preparation, and Resources;
  - Yan Yang: Formal analysis and investigation, and Writing - original draft preparation;
  - Valerie Nayak: Conceptualization, Formal analysis and investigation, and Writing - review and editing;
  - Grace Hui Yang: Conceptualization, Methodology, Writing - original draft preparation, Writing - review and editing, Funding acquisition, and Supervision.

References

[1] 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. https://doi.org/10.1145/1571941.1571951

[2] Jiang, J., Hassan Awadallah, A., Shi, X., White, R.W.: 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. http://doi.acm.org/10.1145/2684822.2685319

[3] 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. https://doi.org/10.1145/3077136.3080818

[4] Gerrig, R.J., Zimbardo, P.G.: Psychology and Life, p. 704. Pearson Higher Education AU, Australia (2015)

[5] Apter, M.J.: Motivational Styles in Everyday Life: A Guide to Reversal Theory, p. 373. American Psychological Association, USA (2001)

[6] Apter, M.J.: Personality Dynamics: Key Concepts in Reversal Theory, p. 98. Apter International, Manassas, VA (2005)

[7] Apter, M.J.: Reversal theory and personality: A review. Journal of Research in Personality **18**(3), 265–288 (1984). https://doi.org/10.1016/0092-6566(84)90013-8

[8] Four domains in Reversal Theory. http://tuckertalk.net/blog2/welcome/reversal-theory/. Accessed: 2019-08-03

[9] Aula, A., Khan, R.M., 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. https://doi.org/10.1145/1753326.1753333

[10] 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. https://doi.org/10.1145/3331184.3331353

[11] Hassan, A., White, R.W., Dumais, S.T., 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. http://doi.acm.org/10.1145/2556195.2556221

[12] Feild, H.A., 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. https://doi.org/10.1145/1835449.1835458

[13] Odijk, D., White, R.W., Hassan Awadallah, A., Dumais, S.T.: 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. https://doi.org/10.1145/2806416.2806488

[14] Apter, M.J.: Reversal theory: A new approach to motivation, emotion and personality. Anuario de psicología/The UB Journal of psychology **42**, 17–29 (1989)

[15] Hebb, D.O.: Drives and the c. n. s. (conceptual nervous system). Psychological Review **62**(4), 243–254 (1955). https://doi.org/1939-1471(Electronic);0033-295X(Print)

[16] Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., Shneiderman, B.: Determining causes and severity of end-user frustration. International journal of human-computer interaction **17**(3), 333–356 (2004)

[17] Kim, Y., Hassan, A., White, R.W., 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. http://doi.acm.org.proxy.library.georgetown.edu/10.1145/2556195.2556220

[18] Hassan, A., Jones, R., Klinkner, K.L.: Beyond dcg: User behavior as a predictor of a successful search. In: WSDM '10

[19] Hassan, A., Shi, X., Craswell, N., Ramsey, B.: Beyond clicks: Query

reformulation as a predictor of search satisfaction. In: CIKM '13

[20] Wang, H., Song, Y., Chang, M.-W., He, X., Hassan, A., White, R.W.: Modeling action-level satisfaction for search task satisfaction prediction. In: SIGIR '14

[21] Huffman, S.B., 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. https://doi.org/10.1145/1277741.1277839

[22] Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve web search. ACM Trans. Inf. Syst. **23**(2), 147–168 (2005). https://doi.org/10.1145/1059981.1059982

[23] Verma, M., Yilmaz, E., Craswell, N.: On obtaining effort based judgements for information retrieval. In: WSDM '16

[24] 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

[25] Han, S., Yue, Z., He, D.: Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. ACM Trans. Inf. Syst. **33**(4), 16–11634 (2015)

[26] Lagun, D., Hsieh, C.-H., Webster, D., Navalpakkam, V.: Towards better measurement of attention and satisfaction in mobile search. In: SIGIR '14

[27] Huang, J., Diriye, A.: Web user interaction mining from touch-enabled mobile devices. In: HCIR Workshop '12

[28] Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., Yoon, H.-J.: Pagination versus scrolling in mobile web search. In: CIKM '16

[29] Apter, M.J.: The Experience of Motivation: the Theory of Psychological Reversals, p. 378. Academic Press London ; New York, New York (1982)

[30] Weinfurt, K.P.: Multivariate analysis of variance., 245–276 (1995)

[31] Stahle, L., Wold, S.: Analysis of variance (anova). Chemometrics and Intelligent Laboratory Systems **6**(4), 259–272 (1989). https://doi.org/10.1016/0169-7439(89)80095-4

[32] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical

Society: Series B (Methodological) **39**(1), 1–22 (1977) https://arxiv.org/abs/https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[33] 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. https://doi.org/10.1145/3366423.3380271

[34] Gonçalves, R.S., Horridge, M., Li, R., Liu, Y., Musen, M.A., Nyulas, C.I., 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. (eds.) The Semantic Web – ISWC 2019, pp. 418–435. Springer, Cham (2019)

[35] Jones, R., Klinkner, K.L.: 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. https://doi.org/10.1145/1458082.1458176

[36] 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. https://doi.org/10.1145/2983323.2983654

[37] Yang, H., Mityagin, A., Svore, K.M., 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. https://doi.org/10.1145/1835449.1835526

[38] Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1023/A:1022627411411

[39] Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics **28**(2), 337–407 (2000). https://doi.org/10.1214/aos/1016218223

[40] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189–1232 (2001). https://doi.

org/10.1214/aos/1013203451

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., 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)

[42] 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. https://doi.org/10.1145/2600428.2609523