User Satisfaction Prediction with Mouse Movement Information in Heterogeneous Search Environment

Ye Chen, Yiqun Liu, Min Zhang, and Shaoping Ma, *Member, IEEE*

Abstract—Satisfaction prediction is one of the prime concerns in search performance evaluation. It is a non-trivial task for three major reasons: (1) The definition of satisfaction is subjective and different users may have different opinions in the process of satisfaction judgment. (2) Most existing studies on satisfaction prediction mainly rely on users' click-through or query reformulation behaviors but there are many sessions without such interactions. (3) Most existing works primarily rely on the hypothesis that all results on search result pages (SERPs) are homogeneous, but a variety of heterogeneous search results have been aggregated into SERPs to improve the diversity and quality of search results recently. To shed light on these research questions, we construct an experimental search engine that could collect users' satisfaction feedback as well as mouse click-through/movement data. Inspired by recent studies in predicting search result relevance based on mouse movement patterns (namely, motifs), we propose to estimate search satisfaction with motifs extracted from mouse movement data on SERPs. Besides the existing frequency-based motif selection method, two novel selection strategies (distance-based and distribution-based) are also adopted to extract high-quality motifs for satisfaction prediction. Experimental results show that the proposed strategies outperform existing methods and have promising generalization capability for unseen users and queries in both a homogeneous and heterogeneous search environment.

Index Terms—Search satisfaction, user behavior, mouse movement, federated search, prediction

1 Introduction

Search satisfaction prediction is essential in Web search performance evaluation researches. Although there have been plenty of existing studies [2], [3], [4], [5] on this research topic over the past years, it is still a challenging task for three major reasons: (1) The definition of satisfaction is rather subjective and different users may have different opinions in satisfaction. Therefore, satisfaction feedback from different users for the same result ranking list may be very different (see Section 5.4). (2) There usually lacks enough explicit feedback information to infer users' opinions in satisfaction for practical search engines. Different from relevance prediction researches in which result clicks can be regarded as strong signals of user preference, the feedback information of satisfaction is related with a number of different interaction behaviors. Many existing approaches on satisfaction prediction rely on users' click-through or query reformulation behaviors [3], [6]. However, for many search sessions neither mouse clicks nor query reformulations are available [7], [8] and these solutions are therefore not applicable. (3) Most previous works on search satisfaction rely on the hypothesis that all results on search engine result pages (SERPs) share a similar presentation style (one hyperlink with a short snippet). However, as more and more heterogeneous vertical results (videos, images, knowledge graphs and so on) are aggregated into modern SERPs to improve the diversity and quality of search results, the differences between users' satisfaction perception process in the homogeneous and heterogeneous search environment remain uninvestigated. We therefore try to explore the following three research questions in this work:

- RQ1: Do users have different perceptions of satisfaction and how can we design experiments to study the effect of user variability? (subjectivity in satisfaction judgment)
- RQ2: Besides click-through behaviors, what other interaction information can be used to suggest user satisfaction? (lack of explicit feedback information)
- RQ3: How user satisfaction are affected by vertical results and how can we predict user satisfaction in heterogeneous search environment? (effect of heterogeneous search results)

For the first problem, the definition of satisfaction itself is subjective and different users may have different opinions in satisfaction judgement process. We use the definition proposed by Kelly et al. [2] throughout the paper to ensure the consistency of satisfaction judgment criteria. The definition in [2] states that "satisfaction can be understood as the fulfillment of a specified desire of goal". In our work, we define satisfaction as "*the fulfillment of the search goal*" because we require users to finish search tasks. For satisfaction judgment collection, some researchers design systems to collect users'

- *The authors are with the Department of Computer Science and Technology, Tsinghua Unviersity, Beijing 100084, China. E-mail: chenye617@gmail.com, {yiqunliu, z-m}@tsinghua.edu.cn, msp@mail.tsinghua.edu.cn.*

Manuscript received 4 Oct. 2015; revised 24 Feb. 2017; accepted 5 Aug. 2017. Date of publication 0 . 0000; date of current version 0 . 0000.
(Corresponding author: Yiqun Liu.)
Recommended for acceptance by F. Silvestri.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2739151

(a) Example of a satisfied(SAT) search session

(b) Example of a dissatisfied(DSAT) search sessioin

Fig. 1. Examples of users' mouse movement trails on SERPs.

explicit feedback as the ground truth for satisfaction [3], [4]. However, the quality of data cannot always be ensured because collecting feedback information explicitly usually affects users' search processes. Other researchers choose not to interrupt users' search process. Instead, they employ external assessors to review the original searchers' behavior logs and make judgments according to their own experiences [9]. According to recent studies on query intent labelling and relevance annotations [7], [10], external assessments may be very different from users' self-annotations. In our work, we manipulate the SERPs in our experiments to investigate how users' perception of satisfaction differs across different search result pages. We try to quantitatively measure the effect of user variability in satisfaction prediction.

For the second problem, although click-through and query reformulation behaviors are not always available for all search sessions, there are other interactions that can be collected in most cases. Among these interaction behaviors, mouse movement has recently been paid much attention to. It can be adopted as a proxy of eye fixation behavior [11], [12] and can be easily collected at large scale as well. Existing studies indicate that mouse movement behaviors can provide insights into result examination [12] and result relevance estimation [13], [14], [15], [16]. Guo et al. [4] are among the first to predict search satisfaction (namely search success in their work) with fine-grained mouse interactions (e.g., hovers, scrolls, etc.) in addition to clicks. However, mouse movement data contains much richer interaction information between users and search engine result pages than these behavior signals. Recent studies [17] already show that automatically discovered mouse movement subsequences (namely motifs) can be utilized to infer result relevance. Therefore, we try to extract the rich information stored in mouse movement logs and investigate whether satisfaction prediction can benefit from such information.

For the third problem, the appearances of the vertical results can be quite different from the non-vertical results [18], [19] and may provide information in a completely different way. Previous works showed that a user's examination and clicking behavior can be quite different [20], [21] in a heterogeneous search environment. Because vertical results may provide richer information than the traditional non-vertical results, the sense of fulfilling information needs during the search process may also be different. Therefore, we try to study how vertical results affect user satisfaction and investigate whether there exists any difference between satisfaction prediction in the homogeneous and heterogeneous search environment.

To shed light on these research questions, we construct an experimental search engine system which can collect users' click-through and mouse movement information simultaneously. The explicit feedback of users on search satisfaction are collected as well. Fig. 1 shows two examples of users' mouse movement process on SERPs with the constructed experimental search engine (see Section 3), where Fig. 1a shows an example of SAT (self-reported satisfactory) case and Fig. 1b shows a DSAT (self-reported dissatisfactory) case. Mouse movement trail is shown in circles and the numbers in them correspond to the sequence of mouse movement positions. The red circles in both figures are movement patterns (namely motifs, which means frequently appearing subsequences in mouse movement data) extracted and selected by the algorithms described in Section 4. In Fig. 1a, the user appears to examine the first result (which is a key resource to the corresponding query) carefully and just take a quick look at other results before ending the search session. This sequence means that he/she succeeds in finding necessary information with relatively little effort. In contrast, most results on the SERP in Fig. 1b seem not to meet the user's information need. We can see from the mouse trail that the user examines almost all results on the SERP carefully during the session, which means he/she may take much effort without obtaining much useful information. Therefore, mouse movement information can help us infer that the user in search session shown in Fig. 1a is likely to be satisfied while the one in Fig. 1b is not.

The examples in Fig. 1 indicate that mouse movement data records rich information in the sequence of examining, reading relevant/irrelevant results and so on. Our work focuses on extracting these movement patterns from the sequence of cursors on SERPs to help predict search satisfaction. To avoid too much subjectivity in satisfaction judgment, we introduce manipulated SERPs to control annotation qualities.

The major difference between our work and existing studies in search satisfaction prediction lies in that we adopt rich interaction patterns (or motifs) in mouse movement data and we try to predict satisfaction in both a homogeneous and heterogeneous search environment. Although previous studies such as [4] already introduce mouse behavior features in addition to result clicks, motifs are not among their investigated features. According to the cases in Fig. 1, motifs may contain important feedback information and should not be ignored. Our work also differs from the motif extraction method proposed by Lagun et al. [17] in that they focused on the problem of relevance estimation instead of search satisfaction prediction. We further propose two specific strategies (distance-based and distribution-based) in the motif extraction process to efficiently select effective patterns. Compared with the frequency-based strategy proposed in [17], they are more suitable for the task of satisfaction prediction by achieving better prediction performance with fewer motifs.

Our contributions in this paper include:

- To the our best knowledge, this is the first attempt to predict search satisfaction with mouse movement patterns (or motifs) in both a homogeneous and heterogeneous search environment.
- We propose to use distance-based and distribution-based strategies in the selection of motifs, which outperforms existing frequency-based strategy and other traditional feature selection methods (e.g., lasso regression) in choosing the most effective motifs to separate SAT sessions from DSAT ones.
- With an experimental search system, we adopt manipulated SERPs to study how search satisfaction judgment criteria differs across different users. We investigate the effect of user variability on satisfaction prediction quantitatively.

The rest of this paper is organized as follows: Related studies are discussed in Section 2. The experimental system and corresponding data collection process are presented in Section 3. Motif extraction method and corresponding selection strategies are proposed in Section 4. Experimental results in satisfaction prediction are introduced and discussed in Section 5. Finally come the conclusions and future work directions.

2 Related Work

Three lines of researches are related to this work. The first line of work focuses on user satisfaction understanding and prediction. Some researchers tried to collect users' explicit feedback to be the ground truth of satisfaction while others invited external assessors to make satisfaction judgments according to the original users' search logs. However, users' satisfaction judgments tend to be subjective and the consistency of data cannot always be ensured while external assessments may be quite different from users' annotations. In our work, we try to investigate the effect of user variability on satisfaction prediction with manipulated SERPs. The second line focuses on search performance evaluation with interaction information. Both coarse-grained and fine-grained features were adopted in searh performance prediction in the recent years. We extend this line by testing the effectiveness of mouse movement patterns extracted directly from SERPs. The third line focuses on federated search. We are inspired by these researches and try to investigate whether vertical results will make any difference and try to predict user satisfaction in a heterogeneous search environment.

2.1 Search Satisfaction Study

The concept of satisfaction was first introduced in IR researches in 1970s according to Su et al. [22]. A recent definition by Kelly et al. states that "satisfaction can be understood as the fulfillment of a specified desire or goal" [2]. Various models involving user behaviors [23] and SERP layouts [24] have been set up to quantify user satisfaction in recent years. However, search satisfaction itself is a subjective construct and is difficult to measure. Some existing studies tried to collect users' explicit feedback as the ground truth of satisfaction. For example, Guo et al.'s work [4] on predicting Web search success and Feild et al.'s work [3] on predicting searcher frustration were both based on searchers' self-reported judgements. Differently, other researchers employed external assessors to restore the users' search experience and make annotations according to their own experience. For example, Guo et al.'s work [25] on predicting query performance and Huffman et al.'s work [26] on predicting result relevance were based on this kind of annotations. Recent research [10] showed that annotations on result relevances from external assessors may not be a good estimator of users' self-judgements. Recently, a benefit-cost framework was proposed [9] to analyze the satisfaction judgement process. In this framework, both the benefit factors (result utility) and the search effort users spend on examining SERPs and browsing landing pages are taken into consideration. In this work, we study the subjectivity in satisfaction perception across different users. We try to investigate the effect of user variability on satisfaction prediction.

2.2 Mouse Interaction Features

A number of different interaction behaviors have been adopted in the prediction of search performance over the

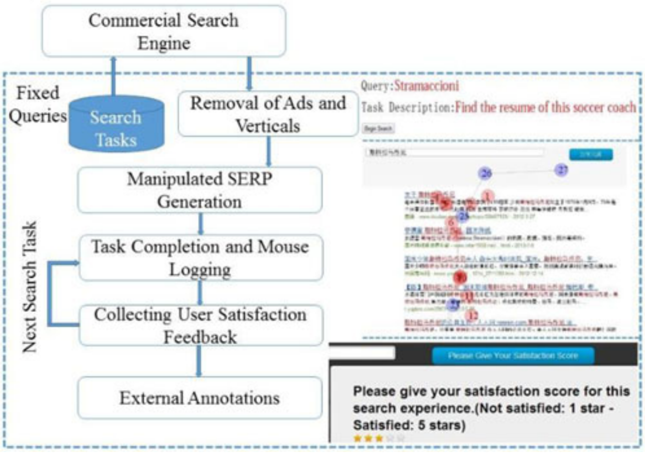


Fig. 2. Data collection procedure.

past years, including both coarse-grained features (e.g., SERP components, click-through based features in [25]) and fine-grained ones (e.g., cursor position, mouse hover and scrolling speed in [4]). The benefit-cost framework was also used to predict users' graded search satisfaction [5], [9].

Mouse movement information like scroll and hover have proven to be valuable signals in inferring user behavior and preferences [11], [14], [16], [27], [28], [29], user attention [30], [31], [32], search intent [33], search examination [12] and predicting result relevance [7], [34]. More recently, viewport informational is also adopted to analyze user behavior patterns [35], [36]. However, none of these studies tried to extract mouse movement patterns and adopt them to predict search satisfaction.

With the advancement of technology, more detailed and scalable mouse information can be collected. Arapakis et al. extracted mouse gestures to measure within-content engagement [37]. Navalpakkam et al. [38] used mouse tracking to predict user experience on the web. Lagun et al. [17] introduced the concept of frequent cursor subsequences (namely motifs) in the estimation of result relevance. Different from their work, we focus on how to extract and select effective mouse movement patterns from SERPs to help predict satisfaction at a search task level instead of result level in both a homogeneous and heterogeneous search environment. We also propose different motif selection strategies to improve the prediction performance.

2.3 Federated Search Study

As more and more heterogeneous search results are aggregated into search result pages to promote users' search experiences, there are a number of existing works focused on this kind of federated search, among which most works focused on predicting whether a vertical result is relevant to a query (vertical selection). Diaz et al. [39] first carried out a system to collect news dynamically and aggregated them into web search results. Arguello et al. [40], [41] demonstrated the effectiveness of query logs when selecting relevant verticals. Zhou et al. [21] further presented an approach that considers both reward and risk within the task of vertical selection.

Because the display form of a vertical result may be different from that of a non-vertical result, users examination behavior may change when SERPs become more heterogeneous. Some existing studies tried to analyze users new behavior patterns on heterogeneous SERPs. Wang et al. [20] found that different verticals may create examination biases on users search behavior. They suggested that images and videos will attract a users attention more than other search results. Liu et al. [19] showed three behavior effect in federated search, namely, the vertical attraction effect, the examination cut-off effect and the examination spill-over effect. Chen et al. [18] further studied the effect of vertical results with different presentation styles, positions and qualities on user satisfaction. Navalpakkam et al. [31] also showed that a knowledge graph will influence a users attention distribution on SERPs.

Traditional search result evaluation metrics may also become inappropriate when dealing with federated search pages. Various diversity aware IR metrics have been proposed [42], [43], [44], which may be adjusted to evaluate heterogeneous result pages. Zhou et al. [45] introduced the concept of vertical orientation and instantiated a suite of metrics for evaluating aggregated search pages. Markov et al. [46] further proposed two vertical-aware metrics based on user click models for federated search.

Inspired by these existing works on the differences between vertical results and non-vertical results, we incorporate vertical results with different presentation styles into SERPs. We predict satisfaction on such pages and try to demonstrate the effectiveness of our proposed prediction framework in both homogeneous and heterogeneous search environment.

3 Data Collection

3.1 Experiment Procedure

To collect user behavior data during search process and corresponding satisfaction annotation data, we implemented a lab-based search engine system as shown in Fig. 2. During the experimental procedure, satisfaction feedback as well as a variety of mouse movement information, including mouse coordinates, clicks, hovers and scrolls are logged by injected Javascript on SERPs.

As shown in Fig. 2, the process of this study is as follows. First, we prepared a set of search tasks and their corresponding queries (one query for each task). To make sure that the same SERP for a certain task is shown to all the participants in the experiment, we crawled and stored in advance the corresponding SERPs of all search tasks. The results are shown on the same screen whose resolution is 1920*1080 for all participants.

Each participant was asked to perform two "warm-up" practice tasks to be familiar with the study flow, followed by the 30 tasks that we used in our analysis. Before each task, the participant was shown the search query and corresponding explanations to avoid ambiguity. After that, he/she would be guided to a pre-designed search result page where the query is not allowed to change. The participants were asked to examine the results provided by our system and end the search session either if the search goal was completed or he/she was disappointed with the results. Each time they end a search session, they were required to label a 5-point satisfaction score to the session where 5 means the most satisfactory and 1 means the least. As mentioned

TABLE 1
Examples of Search Queries in Different Search Tasks

Task Type	ID	Query	Task Description
Organic Search	1	what is a sound card	find a brief introduction about sound card
	2	"A Little Thing Called Love"	find a online movie resource of "A Little Thing Called Love"
	3	Meizu official website	find the official website of Meizu
	4	Stramaccioni	find a biographical sketch of Stramaccioni
	5	Beijing International Conference Center	find a brief introduction of Beijing International Conference Center
Vertical Search	1	interview of Lee Sedol	find the interview of Lee Sedol after his match against AlphaGo
	2	Arrow	find online watch resources of Arrow Season 4
	3	vehicle mounted refrigerator	find the brand ranking of vehicle mounted refrigerator
	4	price of the laser freckle	find the price of the laser freckle
	5	prophat	find the equipment list of Dota hero "prophat"

before, the judgment criteria of satisfaction is defined as "the fulfillment of the search goal". Then they would be guided to continue to the next search task.

During the search process of each task, the users' mouse movements/click-through behaviors were logged by the injected JavaScript code on SERPs. We implemented our own version of mouse movement recorder but researchers may also rely on open source solutions such as EMU toolbar for the Firefox browser [33]. We tried our best to simulate a practical Web search environment for our participants. They were allowed to click any result link on the SERP and visit the landing page without time limits during the search process.

3.2 Search Tasks and SERP Generation

We generated two sets of search tasks to collect search satisfaction feedback in both a homogeneous and heterogeneous search environment, namely the *organic search tasks* and *vertical search tasks*.

3.2.1 Search Tasks in Heterogeneous Search

For the *organic search tasks*, we first selected 30 search tasks from NTCIR IMine task [47], among which there are 10 navigational tasks and 20 informational ones. All these queries were collected from a commercial search engine and were neither long-tailed nor popular ones to avoid unnecessary biases. Different from the IMine task, we also provided detailed task explanations to the participants to avoid any possible ambiguity. An example set of the search queries are shown in Table 1. The search results were collected from a popular commercial search engine and only top 10 organic results were retained. We excluded the vertical results and advertisements to study user satisfaction in the homogeneous search environment. We fixed the query and results for the consistency of result sets across users. Such task design is similar with previous researches on web user study [31].

Considering the fact that users may have different criteria or even be distracted during the satisfaction annotation process, we manipulate the SERPs to study the variability across different users. We invite three professional assessors from a commercial search engine to label the relevance scores for all query-result pairs. The KAPPA coefficient of the their annotation is 0.70, which can be characterized as a substantial agreement according to Cohen [48]. Two different types of SERPS are designed for each query based on the relevance annotations. For each query, the results on two SERPs are the same but in different ranking orders. On the first page, the results were ranked in the order of relevance and on the second one they were ranked in the reverse order of relevance. We call these two pages ordered-page and reversed-page, which should entail different levels of satisfaction. The pages are used to verify the subjectivity of user satisfaction and to study the effect of user variability on satisfaction prediction.

For the data collection process, we had 60 (30 queries * 2 different SERPs) SERP conditions in total. Each participant needs to complete 30 tasks with our search engine system, which contain 15 SERPs from each kind of conditions (ordered-pags and reversed-page). We adopted a Graeco-Latin square design and randomized sequence order to ensure that each task condition had the same opportunity to be shown to users. It is reasonable to believe that searchers tend to be more satisfied with ordered-pages and less satisfied with reversed-pages. Therefore, we can study the subjectivity in users' satisfaction judgement based on their satisfaction annotation on these manipulated SERPs.

3.2.2 Search Tasks in Heterogeneous Search

For the *vertical search tasks*, we adopt SERPs which are exactly the same as those in real-life scenario. We sampled a large number of search queries based on the search logs from a major commercial search engine and use such queries to organize our search tasks. Considering that the results crawled from the search engine are generally good and users will tend to be satisfied in most cases. We sampled some "difficult" search tasks manually in order to generate enough negative examples for a comparatively balanced dataset. Some examples of the search queries are shown in Table 1. Top 10 results from the commercial search engine are retained for each search task and there are 7.4 vertical results on each SERP in average. Each participant is required to finish all these 30 tasks and the sequence order of the tasks are randomized.

3.3 Participants

We recruited 40 and 30 participants for the data collection in organic search tasks and vertical search tasks, respectively. All participants are first-year undergraduate students and have a variety of self-reported search engine utilization experiences. Their majors vary from biology, life science,

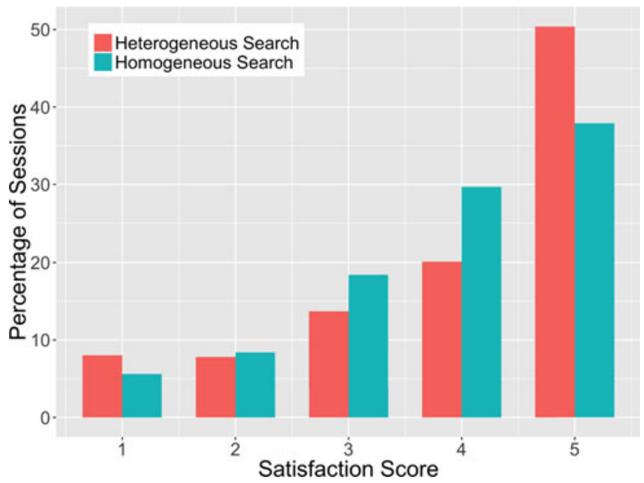


Fig. 3. Distribution of user satisfaction in homogeneous/heterogeneous search.

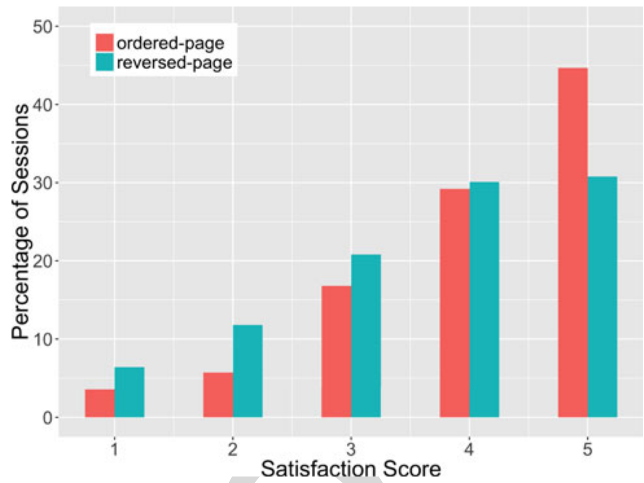


Fig. 4. Distribution of satisfaction scores on manipulated SERPs.

arts, economics to social science. We didn't invite computer science or electrical engineering students because they may be too familiar with the use of search engines and cannot represent ordinary search engine users. Each participant was paid 10 US dollars for completing the 30 search tasks.

3.4 Satisfaction Distribution

With the data collected in the experiment process, we show the distribution of satisfaction scores from users in both homogeneous and heterogeneous search environment in Fig. 3. From this figure we can see that users tend to give a high satisfaction score for the search tasks in both homogeneous and heterogeneous search environment, which shows that the commercial search engine generally provides promising results for these non-long-tailed queries. The percentage of sessions labelled 5 in heterogeneous search (50.4 percent) is higher than that in homogeneous search (37.9 percent), which may indicates that vertical results can help improve SERP quality.

We use the two kinds of pre-defined SERPs (ordered-page and reversed-page) in organic search tasks to verify the subjectivity of satisfaction annotations from users. The distribution of the satisfaction scores on the manipulated SERPs are shown in Fig. 4. Results show that users tend to feel more satisfied with ordered-pages and less satisfied with reversed-pages, which is in line with our expectations. It indicates that users' satisfaction scores will be affected by the relevance of search results but the impact is not as large as we have imagined. More detailed study on the effect of user variability on satisfaction prediction will be shown in Section 5.4.

4 Motif Extraction and Selection

The motif-based satisfaction prediction framework can be described as Algorithm 1. We first extract large amount of motif candidates from the training set and then adopt specific selection strategies to pick out the ones with high quality. Then we train a satisfaction classifier with the selected motifs and the training dataset. For a new testing data without satisfaction annotation, we only need calculate features based on the selected motifs and mouse movement information in the testing data and input them into the classifier, the output will be the prediction result of satisfaction. The algorithm shows that, once the motifs are selected, we only need to calculate some features for a new coming data, which makes our method a fast and scalable way for satisfaction prediction.

Algorithm 1. Motif-Based Satisfaction Prediction

Input:
training user sessions. $TrainD$
$TrainD$'s satisfaction annotation. $TrainSAT$
testing user sessions. $TestD$
Output:
$TestD$'s satisfaction annotation. $TestSAT$
1: Generate motif candidates MC from $TrainD$
2: Pick out motifs M of high quality from MC for prediction with specified selection strategy
3: Generate feature sets $TrainF$ based on $TrainD$ and M
4: Train a classifier C with $TrainF$ and $TrainSAT$
5: Generate feature sets $TestF$ based on $Test$ and M
6: Predict $TestSAT$ with C and $TestF$

In this section, we first give a brief introduction of the motif extraction method, which is similar with the method in [17]. In Section 4.2, we make a detailed description of the novel motif selection strategies and we show some examples of the predictive motifs in Section 4.3.

4.1 Motif Candidate Extraction

The concept of motif is first introduced by Lagun et al. [17] and defined as frequent subsequences in mouse cursor movement data. They proposed to automatically extract motifs from web search examination data and used it for document relevance prediction and search result ranking. Although the method can be adopted to all kinds of Web pages, they focused on extracting motifs from landing pages so that users' implicit preference feedback could be inferred. Different from their work, we try to extract motifs from mouse cursor movement logs on SERPs because we believe that whether users are satisfied can be predicted by their interaction behaviors on SERPs. We first introduce the definition of motif in our work and explain the extraction process from cursor movement data to motifs.

Definition. *A motif is a frequently-appeared sequence of mouse positions, which can be represented by* $T = \{(x_i, y_i)\}_{i=1}^{N}$, *where* (x_i, y_i) *is the coordinates of the cursor at time* t_i.

To extract motifs from cursor data, we first use a sliding window to perform data pre-processing and generate candidates from raw data, which means we shift a given length of window in the mouse log and every shift will generate a motif candidate. In the generation of motifs, we also use Dynamic Time Warping (DTW) algorithm [49] for distance measurement as in [17]. DTW algorithm calculates the smallest possible distance between two time series by aligning one time series with another [50]. Different from Lagun et al.'s work, we try both euclidean and Manhattan distances in calculation. euclidean distance which is not selected by [17] is also used in our method because we believe that motif extraction on SERPs and ordinary Web pages are different. The size and number of components on SERPs are generally fixed and the direct distances between points are mostly comparable across different search sessions.

During the process of clustering similar motifs, we adopted a similar early abandonment and lower bounding strategy as in [17] and a number of time series mining studies such as [51]. The difference is that we just remove the candidate motifs which have overlapping subsequences instead of using a range parameter R to distinguish good motifs from candidates. By this means, we are able to get more candidate motifs and adopt specific strategies to select out motifs with high quality for satisfaction predicting.

4.2 Motif Selection Strategies

A major difference between our motif extraction method and the one in [17] is that we use a number of selection strategies to find the most predictive motifs from candidates. Different from the frequency-based strategy in [17] which selects motifs with the most appearances in training set, we make use of the data distribution information to locate the motifs which can separate SAT sessions from DSAT ones. We believe that frequently-appeared motifs may not always be predictive ones because they may appear in both SAT and DSAT sessions. Therefore, a better selection strategy should use both frequency information and the differences between different kinds of sessions.

We first define $SAT\_DATA/DSAT\_DATA$ as the search sessions which are labelled as satisfactory/unsatisfactory ones annotated by users/assessors. M_SAT and M_DSAT are then defined as the sets of motifs extracted from SAT_DATA and $DSAT\_DATA$. When we select proper motifs with high predictive power from M_SAT and M_DSAT, they could be adopted to generate features for each search session. If we get a series of predictive motifs $C_1, C_2, \ldots, C_N$, we can obtain N distance features for a certain search session S: $Dist(C_1, S), Dist(C_2, S) \ldots\ Dist(C_N, S)$, which will then be used as the N features in the prediction method.

One should note that although the motif selection strategies adopted in our method is different from that in [17], the efficiency of online satisfaction prediction process is similar with the existing method if the same number (N) of motifs are selected. This is because in the prediction process, both methods require the calculation of similarity between predictive motifs and motifs from search sessions. The computation complexity is therefore mostly unchanged if both adopt the same number of motifs.

4.2.1 Distance-Based Selection

This strategy is based on a *Difference Hypothesis*: predictive motifs in M_SAT should be quite different from the ones in M_DSAT and vice versa. This hypothesis probably holds because it is reasonable to assume that users have different mouse movement patterns when they are satisfied / unsatisfied with the search results. The examples in Fig. 1 also agrees with this assumption.

To select the motifs that are significantly different, we use the average distance between motifs in different sets to measure the difference. For example, for a motif candidate C_SAT_i in M_SAT, we have

$$S_{dist}(C\_SAT_i) = \frac{\sum_{C_j \in M\_DSAT} DTW(C\_SAT_i, C_j)}{|M\_DSAT|}. \quad (1)$$

$DTW(C\_SAT_i, C_j)$ represents the DTW distance of two candidate motifs, C_SAT_i and C_j. Intuitively, this equation represents the average DTW distance between C_SAT_i and all motifs in M_DSAT. Similarly, for motifs in M_DSAT, we have

$$S_{dist}(C\_DSAT_i) = \frac{\sum_{C_j \in M\_SAT} DTW(C\_DSAT_i, C_j)}{|M\_SAT|}. \quad (2)$$

With Equations (1) and (2), we can select motifs with large difference from the motifs in the other kind of sessions, which have large chances to be predictive ones.

4.2.2 Distribution-Based Selection

This strategy is based on a *Covering Hypothesis*: predictive motifs in M_SAT/M_DSAT should cover sufficient sessions in $SAT\_DATA/DSAT\_DATA$. We introduce this hypothesis because when a certain motif can only cover a small number of sessions, it is not reasonable to select it even if it is quite different from the motifs in the other set. We want to focus on the general behavior patterns in satisfied / unsatisfied sessions. Therefore, it is necessary to use the distribution information to filter possible noises and retain the ones with large coverage.

We define the distance of a motif C and a session S first to determine whether a motif covers a specific session

$$Dist(C, S) = min\{DTW(C_i, C) | C_i \in S\}. \quad (3)$$

As shown in (3), we use a sliding window to capture several motif candidates (C_i) from session S and calculate the distance between C and these motifs. The smallest distance is defined as the distance between C and S. We then define the coverage rate of a motif C on a dataset D

$$CR(C, D) = \frac{\left| \left\{ \frac{Dist(C, S_i)}{\frac{1}{|D|}\sum_{S_j \in D} Dist(C, S_j)} < r | S_i \in D \right\} \right|}{|D|}. \quad (4)$$

In (4), r is the parameter to ensure we can select enough motifs, which we set as $\frac{1}{30}$ in our experiment. Intuitively, if the distance between a motif candidate C and a seesion S_i

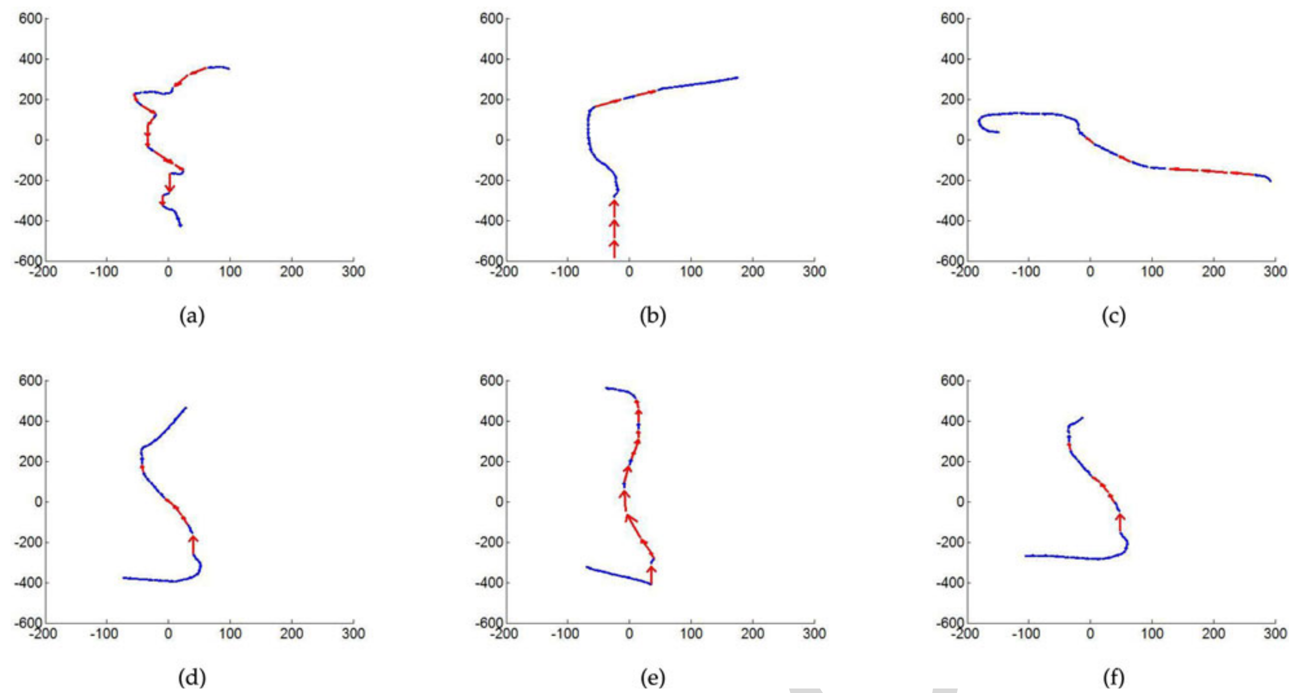


Fig. 5. Predictive motifs discovered from SAT_DATA (a-c) and $DSAT\_DATA$ (d-f).

divided by the average distance between C and all sessions in dataset D is smaller than the threshold r, we consider the motif candidate C covers session S_i. With the concept of coverage rate, we can define the score for each motif based on distribution difference as follows:

$$S_{distri}(C\_SAT_i) = \frac{CR(C\_SAT_i, SAT\_DATA)}{CR(C\_SAT_i, DSAT\_DATA)} \tag{5}$$

$$S_{distri}(C\_DSAT_i) = \frac{CR(C\_DSAT_i, DSAT\_DATA)}{CR(C\_DSAT_i, SAT\_DATA)}. \tag{6}$$

As shown in Equations (5) and (6), if a motif from M_SAT/M_DSAT has a large coverage rate on $SAT\_DATA/DSAT\_DATA$ and a small coverage rate on $DSAT\_DATA/SAT\_DATA$, it will get a higher score and is considered to be predictive. We select motifs with high scores since they tend to have a large distribution difference.

4.3 Example of Predictive Motifs

The proposed distance-based and distribution-based strategies can help discover predictive motifs from mouse movement data and a few examples are shown in Fig. 5. Figs. 5a, 5b, and 5c show 3 of the 10 most predictive motifs extracted from SAT_DATA while Figs. 5d, 5e, and 5f show 3 of the 10 most predictive motifs extracted from $DSAT\_DATA$. We tried to extract motifs from the datasets collected in both a homogeneous search and heterogeneous search. Although vertical results are quite different from organic results in presentation styles, the extracted motifs from different search environments appear to be similar in general. It seems that in both aggregated and non-aggregated search, predictive motifs have similar characteristics as shown in Fig. 5. The motifs are selected based on distribution-based strategy while distance-based strategy produce similar results according to our experiments. The movement directions are annotated by arrows and the coordinate axis is in pixels.

We can see that the motif in Fig. 5a shows a process that user examines the top results carefully and then take a quick look at the lower-ranked results and Fig. 1a can be regarded a practical example. Fig. 5b probably shows the process of re-visiting a previous checked result while Fig. 5c mainly indicates the behavior of using the mouse as a reading aid or the action of moving mouse to click. In contrast, the three motifs show in Figs. 5d, 5e, and 5f are similar and all reflect the process of moving the mouse from bottom to the top after carefully examining a result at a lower position. This is reasonable since we can infer that a searcher may not be satisfied if he has to re-examine a number of results after examining a lower-ranked one. These motifs extracted automatically from mouse data will play an important role in satisfaction predicting. The distance calculated based on Equation (3) will be the features of the classification learning algorithm, as will be discussed in the next section.

5 Experimental Results

5.1 Experiment Setups

In this section, we demonstrate the value of our method by predicting users' satisfaction annotations in both homogeneous and heterogeneous search environment. After the motif extraction and selection process described in Section 4, the motifs from the data sets collected in Section 3 are adopted to generate features in the prediction process.

We compare the performance of the proposed model in predicting user satisfaction scores in both homogeneous and heterogeneous search environment. We compare the effectiveness of different parameter settings and motif selection strategies based on data collected with homogeneous search tasks in Sections 5.2 and 5.3. With the ordered-pages and reversed-pages designed in homogeneous search, we

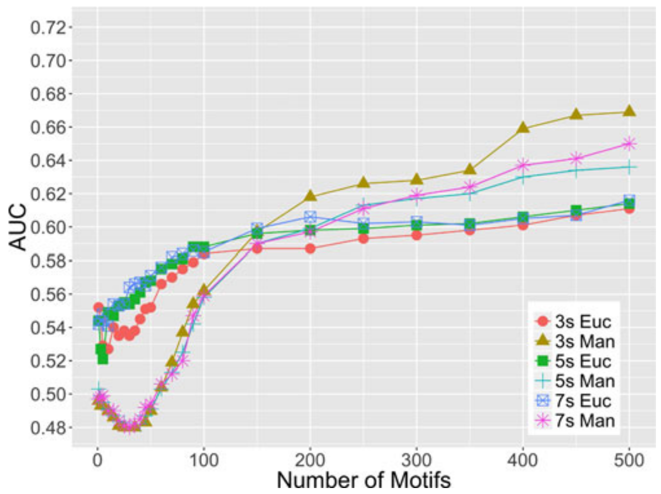


Fig. 6. AUC of satisfaction prediction with different parameter settings.

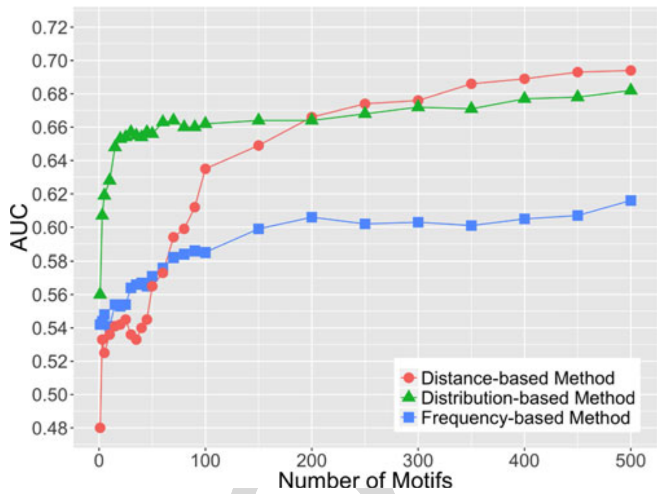


Fig. 7. AUC of satisfaction prediction with different motif selection strategies.

study the effect of user variability in Section 5.4. With two state-of-the-art methods, we demonstrate the predictive power of motifs in both homogeneous and heterogeneous search for unseen users/queries in Sections 5.5 and 5.6.

For satisfaction prediction, we exclude sessions with a satisfaction score of 3 because we consider users do not have a satisfaction preference in such sessions. We consider sessions with a score of 4 or 5 are regarded as SAT cases and those with a score of 1 or 2 as DSAT ones. Based on our dataset, there are 807 SAT sessions and 167 DSAT sessions in homogeneous search, 589 SAT sessions and 132 DSAT sessions in heterogeneous search, which is imbalanced. We use all the DSAT sessions and downsample the SAT ones to make the satisfaction prediction a balanced learning task (The training sets are balanced while the testing sets still remain imbalanced). The learning algorithm in the prediction process is logistic regression,[1] which is widely used in prediction tasks [4]. We use Area Under roc Curve (AUC) to be the evaluation metric because it is less sensitive to the ratio of positive and negative daata samples and is more reliable in imbalanced learning [52]. All results reported in the following sections are the average AUC of five-fold cross validation (The motifs are recomputed for each training and testing set).

5.2 Comparison of Parameter Settings

There are two parameters in the motif extracting algorithm we discussed in Section 4.1, namely the length of sliding window and the distance measurement method for two basic points. Fig. 6 shows the prediction results with different sliding windows and distance measurement methods. The motif selection method used in Fig. 6 is the frequency-based method, which is the one used in [17]. We compare the effectiveness of three different length of sliding windows (3s, 5s and 7s) and two distance measurement methods (Manhattan(**Man**) and euclidean(**Euc**)) in Fig. 6.

From the figure we can see that all models perform better when the number of used motifs increases. With the same distance measurement method, the model's prediction performance does not differ much with the three tested length (3s, 5s and 7s) of sliding windows. A model with euclidean distance can achieve comparatively better predicting results with fewer motifs, which is quite important because the calculation of motifs is quite time-consuming. It will be of great value if we can predict satisfaction with comparatively fewer motifs in both academic and industrial applications. A sliding window of 7s is slightly better than others at the early stage. As a results, we set the length of sliding window to be 7s and use euclidean distance measure in the next sections.

5.3 Comparison of Motif Selection Strategies

To compare the different strategies for motif selection, we use the method used in [17] as a baseline, which selects motifs based on frequency in training set. Experimental results with different motif selection strategies described in Section 4.2 are shown in Fig. 7.

Results in Fig. 7 show that the proposed distance and distribution-based motif selection strategy outperform the baseline frequency-based strategy. Moreover, the distribution-based method can achieve a good performance with quite a small number of motifs. We consider the distribution-based method the best one because we want to predict satisfaction with a small number of motifs so that the motif extraction process can be efficient. Therefore, we adopt the distribution-based selection strategy in the prediction models in the next sections.

We also try the lasso-based feature extraction method to further demonstrate the effectiveness of our proposed motif selection strategy. The results are shown Fig. 8. Different penalty coefficients (from C=0.001 to C=100) of lasso regression are adopted and we can observe that lasso-based method does outperform the frequency-based selection strategy. However, the distribution-based method still outperforms the lasso regression with all tested penalty coefficients, which further demonstrates the effectiveness of our proposed method.

We note that the AUC performance on users' satisfaction annotations is only around 0.65, which may be because that users' self-annotations may be quite subjective and are difficult to be predicted. Such findings further validate the necessity of investigating the subjectivity of users' satisfaction perception.

1. The LR model used in this paper is the one implemented in scikit-learn (http://scikit-learn.org) with all default parameter settings.

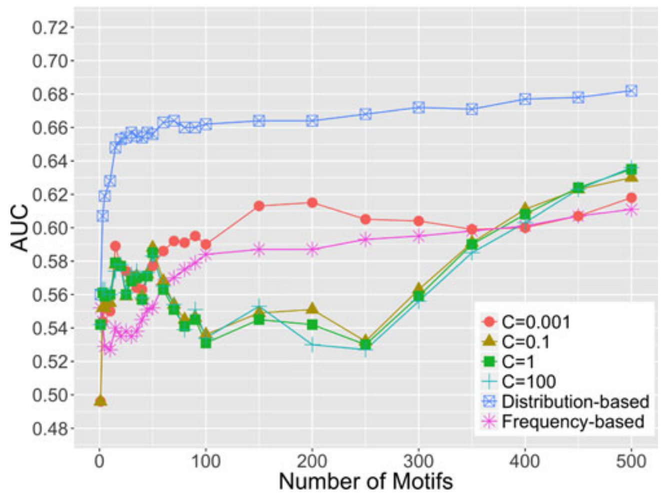


Fig. 8. AUC of satisfaction prediction with Lasso regression.

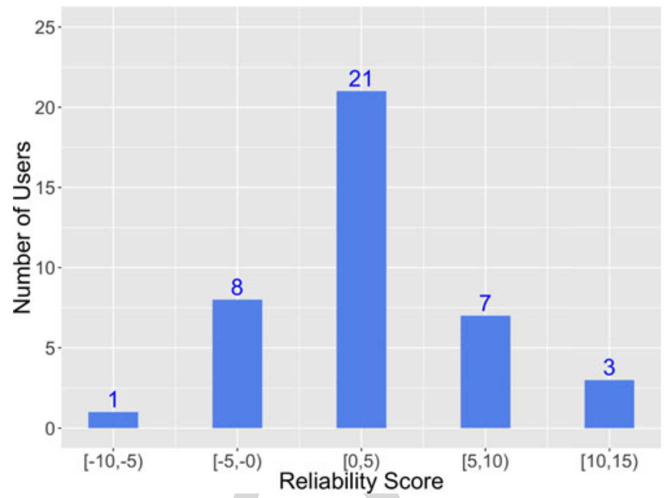


Fig. 9. Distribution of user variability scores.

5.4 User Variability Study

Considering the fact that different users may have different opinions in satisfaction judgement, satisfaction annotations collected from users may be subjective. To study the effect of user variability on satisfaction prediction, we test the performance model on the following three different datasets based on the the manipulated SERPs described in Section 3.2:

The Original Dataset. All search sessions with satisfaction scores of 1, 2, 4 or 5 collected in homogeneous search are included in this dataset.

The Controlled Dataset. As described in Section 3.2, we use the manipulated SERPs and assume users should perceive different levels of satisfaction on different SERPs. For each participant, we define x_1 to represent the number of ordered-pages which he/she gave a satisfaction score of 1 and y_1 to represent the number of reversed-pages which he gave a satisfaction score of 1. Similarly, we get $x_i, y_i (i = 2, 3, 4, 5)$. With these variables, we can define a combination of x_i, y_i to measure user variability quantitatively

$$\mathbf{S}(participant) = f(x_1, x_2 \dots x_5, y_1, y_2 \dots y_5). \tag{7}$$

In general, we assume that users should to be more satisfied with ordered-pages and less satisfied with reversed-pages. Based on this assumption, we define a score for users as following:

$$\mathbf{S}(participant) = x_5 + y_1 + y_2 - x_1 - x_2 - y_5. \tag{8}$$

With the definitions of x_i and y_i, we can see that larger x_5, y_1 and y_2 indicate that the user labelled more ordered-pages with high satisfaction scores and more reversed-pages with low satisfaction scores. Meanwhile, larger y_5, x_1 and x_2 indicate that the user labelled more reversed-pages with high satisfaction scores and more ordered-pages with low satisfaction scores. Therefore, it is reasonable to think that the higher the defined score is, the more the users' satisfaction judgment criteria is consistent with our assumption. We do not include x_3/y_3 because users do not have clear satisfaction preference in such search sessions. Meanwhile, we do not include x_4/y_4 because x_5/y_5 can denote the number of the most SAT sessions and is already larger than $x_1 + x_2/y_1 + y_2$ (see Fig. 3). The combined score will be mostly determined by the number of SAT sessions if x_4/y_4 are also included. The distribution of the reliability scores of the 40 participants are shown in Fig. 9. We can see that the score varies across users and the range is from -10 to 15, which demonstrates the variability in users' satisfaction judgement. To investigate the effect of user various, we exclude the search sessions collected from some users to reduce user variability. We remove the sessions collected from users with the five lowest scores (which are below -2) and the remaining 827 search sessions are regarded as the reliable dataset.

The Manipulated Dataset. We define the sessions collected with ordered-pages as SAT cases and those collected with reversed-pages as DSAT ones. It should be noticed that this dataset has nothing to do with users' original satisfaction feedback. The only information we use is the mouse movement information collected during users' search process.

Performance of our predict model on these three datasets are shown in Fig. 9. We can see that the proposed model gain comparatively better results on the controlled dataset and manipulated dataset. The model performance is the best on the controlled dataset, which is probably because user variability is reduced. Such results indicate that users' annotations on search satisfaction are rather subjective and we can improve prediction performance to some extent if user variability is reduced.

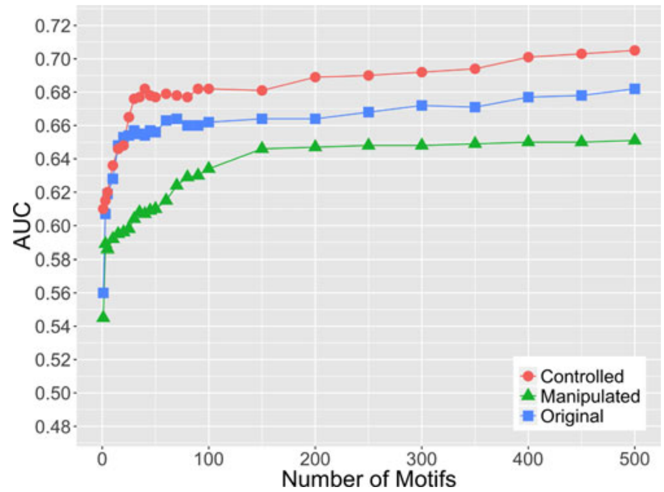


Fig. 10. AUC of satisfaction prediction on datasets with different user variability. Q2

TABLE 2
AUC of Search Satisfaction Prediction across Different Users and Queries in Homogeneous and Heterogeneous Search
(* Indicates Statistical Significance at $p < 0.05$ Level, ** Indicates Statistical Significance at $p < 0.01$ Level)

	Sampling strategy	Guo et al. [4]	Jiang et al. [9]	motif	motif + Guo et al. [4]	motif + Jiang et al. [9]
Homogeneous Search	random sample	0.654	0.596	0.666	0.671 (+2.6%)	0.682 (+14.4%**)
	sample by user	0.630	0.542	0.664	0.658 (+4.4%)	0.663 (+22.3%*)
	sample by query	0.624	0.546	0.669	0.674 (+8.0%*)	0.673 (+23.3%**)
Heterogeneous Search	random sample	0.892	0.877	0.865	0.932 (+4.5%*)	0.930 (+6.0%**)
	sample by user	0.890	0.877	0.856	0.936 (+5.2%**)	0.931 (+6.2%**)
	sample by query	0.923	0.871	0.831	0.925 (+0.2%)	0.931 (+6.9%**)

5.5 Prediction across Users and Queries

According to Section 4, the motif selection strategy relies on data distributions on training sets to locate the most predictive motifs. Therefore, it is important to investigate the generalization power of the proposed prediction model across different users and queries. According to previous studies on predicting examination sequence with mouse movement information [53], different users may have rather different mouse movement patterns and this may lead to poor generalization power of proposed prediction models.

To verify the prediction performance of the proposed models while dealing with new users and queries, we adopt three different training strategies. *Random sampling*: the segmentation of training and testing data in cross validation is completely random. *Sampling by user*: in the segmentation of training and testing data in cross validation, sessions from a same user can only be grouped into either the training set or the testing set. *Sampling by query*: in the segmentation of training and testing data in cross validation, sessions for a same query can only be grouped into either the training set or the testing set. With the latter two strategies, we can ensure that data from the same user/query cannot be adopted for both training and testing.

We implement the satisfaction prediction method proposed in [4] (with both coarse-grained features such as number of clicks and fine-grained features such as scroll speed) and adopted it as a baseline method. We choose this method because it is also based on mouse behavior data (although without motifs) and is one of the most closely related studies. The predictive model in [9] is also used as a baseline method because in this work the features are extracted in a benefit-cost framework and can estimate graded search satisfaction more accurately than most existing works in the homogeneous search environment. We combine the features calculated based on motifs (as shown in Equation (3)) and the features in baseline methods to make combined classification methods. Note that in our experiment, there is only one query in a search task. So any feature that is related with multi-queries is not included in the implementation. The baseline methods and our proposed method are both tested with the three different training strategies and the prediction results are shown in Table 2. The numbers in parentheses show the improvement of the prediction method with combined features over the corresponding baseline method. We also conduct bivariate statistical test for the significance of the performance improvement according to [54].

TABLE 3
Feature Coefficients of LR Model for Satisfaction Prediction

	motif + Guo et al. [4]		motif + Jiang et al. [9]	
rank	feature	coefficient	feature	coefficient
1	max_y_coordinate	-0.844	exist_of_click (bool)	0.851
2	DSAT_ratio	0.702	min_clicked_rank	0.606
3	SAT_ratio	0.428	sesson_dwell_time	-0.463
4	avg_scroll_speed	0.417	max_clicked_rank	-0.418
5	session_dwell_time	-0.413	# DSAT_click	0.407
6	max_scroll_speed	0.325	sum_click_dwell	-0.307
7	avg_click_dwell	0.304	motif #1	0.295
8	motif #1	0.294	motif #2	0.280
9	motif #2	0.288	avg_click_dwell	0.257
10	motif #3	0.274	motif #3	-0.253

Results in Table 2 reveal a number of interesting findings: 1) The prediction performance of the proposed method with motif features is effective with different training strategies. It means that the method can be adopted to deal with previously-unseen queries and users, which is important for practical Web search applications. 2) The motif-based method performs better and can achieve a significant improvement of around 5 percent over [4] and around 20 percent over [9] in most cases, which may indicate that the proposed method makes use of more details in users' interaction process and can be used for improving state-of-the-art technologies.

To get deep insight in the predictive power of selected motifs, features with the top ten logistic regression coefficients are shown in Table 3. The selected models are those trained with random sampling, while the feature rankings are similar in other cases. The detailed feature descriptions can be found in [4] and [9]. Results in Table 3 show that while the traditional user behavior based features have the highest regression weight, the selected motifs are also comparatively important.

The results in this section show that the motif features can achieve a promising performance in predicting satisfaction in the homogeneous search environment and are extremely useful for the satisfaction prediction of previous-unseen users/queries.

5.6 Prediction in Heterogeneous Search

As mentioned in previous sections, the existence of vertical results will affect users' search behaviors. Fig. 11 shows the heatmap of users' mouse movement behavior on SERPs with verticals placed at different positions. The search task in the four subfigures of Fig. 11 are the same and the query is "pictures of wine cabinet". We can see that users' mouse

(a) Without vertical (b) Vertical inserted at rank 1 (c) Vertical inserted at rank 3 (d) Vertical inserted at rank 5

Fig. 11. Heatmap of user mouse movement behavior on SERPS with verticals inserted at different positions.

movements are attracted around the vertical results, which indicates that movement patterns may be different in heterogeneous search environment and varifies the necessity of investigating the effectiveness of motifs in heterogeneous search.

Based on the vertical search tasks, we show the prediction performance in heterogeneous search environment in Table 2. The results reveal similar findings with those in homogeneous search environment: The motif-based features can help improve the performance of state-of-the-art methods under all data sampling strategies. Most improvements are significant based on the bivariate test. Such findings verifies the effectiveness of the motif-based method in heterogeneous search and demonstrate that the proposed model will be effective in real-life settings (in which verticals are usually included and predicting previous-unseen users' opinions is important).

We achieved prediction results at different levels in different search scenarios (around 0.7 in homogeneous search and higher than 0.85 in heterogeneous search), which is because the datasets used are based on different SERP settings and generated by different participants. The tasks for homogeneous search are sampled from NTCIR IMine, which are composed of torso queries. Search behaviors as well as satisfaction judgement may be quite different across users. For the heterogeneous search tasks, we incorporated some difficult search tasks, which may make users struggle. Therefore, we may get more sufficient information to help predict satisfaction. Regardless of the variability in these two datasets, the baselines we used are both state-of-the-art and are reported to have good performance in similar tasks. The performance may be affected by the constructing and sampling of datasets, which makes the absolute values not comparable to some extent. It is important to note that the proposed motif-based method can improve the performance of the baseline methods in different search scenario, which demonstrates the effectiveness of our method.

It will be interesting if we try to extract motifs from specific areas of SERPs, e.g., the vertical result area, because users' mouse behavior will probably be affected by vertical results. However, in this work we extract motifs from the entire result page due to the limited size of dataset. This interesting research topic can be left for future work.

6 Conclusions and Future Work

Search satisfaction prediction is a non-trivial task in search performance research. The definition of satisfaction is subjective, which makes the consistency of feedback from users can't be ensured. External assessors are employed to annotate the satisfaction scores but such annotations may be different from those of users. In this work, we study the subjectivity in users' satisfaction perception. We study the satisfaction judgment criteria across different users and demonstrate that we can improve the prediction performance by reducing user variability.

We further propose a motif based learning framework to predict users' search satisfaction annotations. We introduce specific methods for extracting high quality motifs directly from SERPs and demonstrate that our proposed distance-based and distribution-based strategies outperforms existing solutions. The proposed method is shown to be more effective than state-of-the-art satisfaction prediction methods in predicting previously-unseen users' opinions, which makes it applicable for practical Web search environment. We also carry out a study with aggregated search result pages to investigate the effect of vertical results on user satisfaction. We demonstrate that the findings in the homogeneous search environment are also applicable in heterogeneous search and verify the effectiveness of our proposed motif-based method in the heterogeneous search environment.

However, there are some potential limitations besides all these contributions made in this paper. We removed the advertisements in our experiment setup and we only use torso queries to organise our search tasks. Meanwhile, we only collect data from undergraduate students for convenience. Such experiment setup will help to reduce potential distractions and make the collected data more consistent. However, such specific experimental settings may be very different from the real-life search environment and therefore will cause potential biases. A large-scaled and real-life search enviornoment based study should be carried out in the future to verify the effectiveness of proposed method. Meanwhile, the model we discussed in this paper adopts a batch training approach, which may need to be further revised to be more adaptable for industrial use. Other interesting directions for future work include further improving the efficiency of mining motifs and try to incorporate other effective features into satisfaction predicting models.

Acknowledgments

This work is supported by the Tsinghua University Initiative Scientific Research Program (2014Z21032), National Key Basic Research Program (2015CB358700), and Natural Science Foundation (61622208, 61532011) of China. This paper is an extension of [1]. Compared with the previous conference version, it further verifies the effectiveness of the proposed motif selection strategy by comparing with lasso based feature extraction methods. It also introduces detailed discussion about the performance of the proposed method in heterogeneous search environment.

References

[1] Y. Liu, et al., "Different users, different opinions: Predicting search satisfaction with mouse movement information," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2015, pp. 493–502.

[2] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Found. Trends Inf. Retrieval*, vol. 3, no. 1/2, pp. 1–224, 2009.

[3] H. A. Feild, J. Allan, and R. Jones, "Predicting searcher frustration," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 34–41.

[4] Q. Guo, D. Lagun, and E. Agichtein, "Predicting Web search success with fine-grained interaction data," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 2050–2054.

[5] J. Jiang, D. He, and J. Allan, "Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 607–616.

[6] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein, "Find it if you can: A game for modeling different types of Web search success using interaction data," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 345–354.

[7] J. Huang, R. W. White, and S. Dumais, "No clicks, no problem: Using cursor movements to understand and improve search," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 1225–1234.

[8] J. Li, S. Huffman, and A. Tokuda, "Good abandonment in mobile and pc internet search," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2009, pp. 43–50.

[9] J. Jiang, A. H. Awadallah, X. Shi, and R. W. White, "Understanding and predicting graded search satisfaction," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 57–66.

[10] S. Verberne, et al., "Reliability and validity of query intent assessments," *J. Assoc. Inf. Sci. Technol.*, vol. 64, no. 11, pp. 2224–2237, 2013.

[11] Q. Guo and E. Agichtein, "Towards predicting Web searcher gaze position from mouse movements," in *Proc. CHI Extended Abstracts Human Factors Comput. Syst.*, 2010, pp. 3601–3606.

[12] Y. Liu, C. Wang, K. Zhou, J. Nie, M. Zhang, and S. Ma, "From skimming to reading: A two-stage examination model for Web search," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 849–858.

[13] M. Ageev, D. Lagun, and E. Agichtein, "Improving search result summaries by using searcher behavior data," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 13–22.

[14] Q. Guo and E. Agichtein, "Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 569–578.

[15] Q. Guo, D. Lagun, D. Savenkov, and Q. Liu, "Improving relevance prediction by addressing biases and sparsity in Web search click data," in *Proc. Int. Conf. Web Search Data Mining*, 2012, pp. 71–75.

[16] J. Huang, R. W. White, G. Buscher, and K. Wang, "Improving searcher models using mouse cursor activity," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 195–204.

[17] D. Lagun, M. Ageev, Q. Guo, and E. Agichtein, "Discovering common motifs in cursor movement data for improving Web search," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2014, pp. 183–192.

[18] Y. Chen, Y. Liu, K. Zhou, M. Wang, M. Zhang, and S. Ma, "Does vertical bring more satisfaction? predicting search satisfaction in a heterogeneous environment," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1581–1590.

[19] Z. Liu, Y. Liu, K. Zhou, M. Zhang, and S. Ma, "Influence of vertical result in Web search examination," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2015, pp. 193–202.

[20] C. Wang, et al., "Incorporating vertical results into search click models," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 503–512.

[21] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose, "Evaluating reward and risk for vertical selection," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 2631–2634.

[22] L. T. Su, "Evaluation measures for interactive information retrieval," *Inf. Process. Manage.*, vol. 28, no. 4, pp. 503–516, 1992.

[23] A. Hassan, R. Jones, and K. L. Klinkner, "Beyond DCG: User behavior as a predictor of a successful search," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 221–230.

[24] A. Chuklin and M. de Rijke, "Incorporating clicks, attention and satisfaction into a search engine result page evaluation model," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 175–184.

[25] Q. Guo, R. W. White, S. T. Dumais, J. Wang, and B. Anderson, "Predicting query performance using query, result, and user interaction features," in *Proc. Adaptivity Personalization Fusion Heterogeneous Inf.*, 2010, pp. 198–201.

[26] S. B. Huffman and M. Hochster, "How well does result relevance predict session satisfaction?" in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 567–574.

[27] K. Rodden, X. Fu, A. Aula, and I. Spiro, "Eye-mouse coordination patterns on Web search results pages," in *Proc. CHI Extended Abstracts Human Factors Comput. Syst.*, 2008, pp. 2997–3002.

[28] F. Mueller and A. Lockerd, "Cheese: Tracking mouse movement activity on Websites, a tool for user modeling," in *Proc. CHI Extended Abstracts Human Factors Comput. Syst.*, 2001, pp. 279–280.

[29] I. Arapakis and L. A. Leiva, "Predicting user engagement with direct displays using mouse cursor information," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 599–608.

[30] F. Diaz, R. White, G. Buscher, and D. Liebling, "Robust models of mouse movement on dynamic Web search results pages," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2013, pp. 1451–1460.

[31] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola, "Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 953–964.

[32] M. D. Smucker, X. S. Guo, and A. Toulis, "Mouse movement during relevance judging: Implications for determining user attention," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 979–982.

[33] Q. Guo and E. Agichtein, "Exploring mouse movements for inferring query intent," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 707–708.

[34] R. W. White and G. Buscher, "Text selections as implicit relevance feedback," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 1151–1152.

[35] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam, "Towards better measurement of attention and satisfaction in mobile search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 113–122.

[36] D. Lagun and M. Lalmas, "Understanding user attention and engagement in online news reading," in *Proc. 9th ACM Int. Conf. Web Search Data Mining*, 2016, pp. 113–122.

[37] I. Arapakis, M. Lalmas, and G. Valkanas, "Understanding within-content engagement through pattern analysis of mouse gestures," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 1439–1448.

[38] V. Navalpakkam and E. Churchill, "Mouse tracking: Measuring and predicting users' experience of web-based content," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 2963–2972.

[39] F. Diaz, "Integration of news content into Web results," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2009, pp. 182–191.

[40] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo, "Sources of evidence for vertical selection," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2009, pp. 315–322.

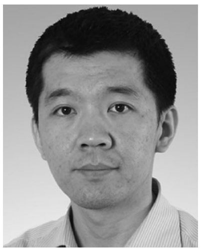
[41] J. Arguello, F. Diaz, and J.-F. Paiement, "Vertical selection in the presence of unlabeled verticals," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 691–698.

[42] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2009, pp. 621–630.

[43] C. L. Clarke, et al., "Novelty and diversity in information retrieval evaluation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 659–666.

[44] T. Sakai and R. Song, "Evaluating diversified search results using per-intent graded relevance," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 1043–1052.
[45] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose, "Evaluating aggregated search pages," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 115–124.
[46] I. Markov, E. Kharitonov, V. Nikulin, P. Serdyukov, M. de Rijke, and F. Crestani, "Vertical-aware click model-based effectiveness metrics," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 1867–1870.
[47] Y. Liu, et al., "Overview of the NTCIR-11 IMine task," in *Proc. NTCIR Conf. Eval. Inf. Access Technol.*, 2014, pp. 8–23.
[48] J. Cohen, "Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, 1968, Art. no. 213.
[49] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-26, no. 1, pp. 43–49, Feb. 1978.
[50] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, 2005.
[51] T. Rakthanmanon, et al., "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 262–270.
[52] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[53] J. Huang, R. White, and G. Buscher, "User see, user point: Gaze and cursor alignment in Web search," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 1341–1350.
[54] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

Ye Chen is currently working toward the master's degree in the Department of Computer Science, Tsinghua University, Beijing, China. His research interests mainly include information retrieval and user behavior analysis.



Yiqun Liu is currently working as an associate professor in the Department of Computer Science, Tsinghua University, Beijing, China. Since 2010, he has regularly served a a program committee member of SIGIR, ACL, and WSDM and he also worked as coordinator for the NTCIR Intent and IMine tasks. His research interests mainly include information retrieval, Web search, and user behavior analysis.

Min Zhang is an associate professor in the Department of Computer Science and Technology, Tsinghua University, Beijing, China. She served on the editorial board of the *ACM Transactions of Information Systems*, and as a PC or senior PC member of SIGIR, WWW, IJCAI, CIKM, WSDM etc. She has also jointly coordinated NTCIR Intent and IMine tasks since 2010. Her major research interests include Web search and recommendation, and user behavior analysis.

Shaoping Ma is currently working as a professor in the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests mainly include intelligent information processing, information retrieval, and Web data mining. His recent research projects span over Web page quality estimation, Web spam page and illegal resource identification, search performance evaluation, on-line advertising performance evaluation, and search engine query recommendation. He is a member of the IEEE.