# Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information

Yiqun Liu†, Ye Chen†, Jinhui Tang‡, Jiashen Sun⋆, Min Zhang†, Shaoping Ma†, Xuan Zhu⋆
†Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science & Technology, Tsinghua University, Beijing, China
‡School of Computer Science & Engineering, Nanjing University of Science and Technology
⋆Samsung R&D Institute China - Beijing
yiqunliu@tsinghua.edu.cn

## ABSTRACT

Satisfaction prediction is one of the prime concerns in search performance evaluation. It is a non-trivial task for two major reasons: (1) The definition of satisfaction is rather subjective and different users may have different opinions in satisfaction judgement. (2) Most existing studies on satisfaction prediction mainly rely on users' click-through or query reformulation behaviors but there are many sessions without such kind of interactions. 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. Different from existing studies, we compare for the first time search users' and external assessors' opinions on satisfaction. We find that search users pay more attention to the utility of results while external assessors emphasize on the efforts spent in search sessions. Inspired by recent studies in predicting result relevance based on mouse movement patterns (namely motifs), we propose to estimate the utilities of search results and the efforts in search sessions with motifs extracted from mouse movement data on search result pages (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 on over 1,000 user sessions show that the proposed strategies outperform existing methods and also have promising generalization capability for different users and queries.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## Keywords

Search satisfaction; User behavior; Mouse movement

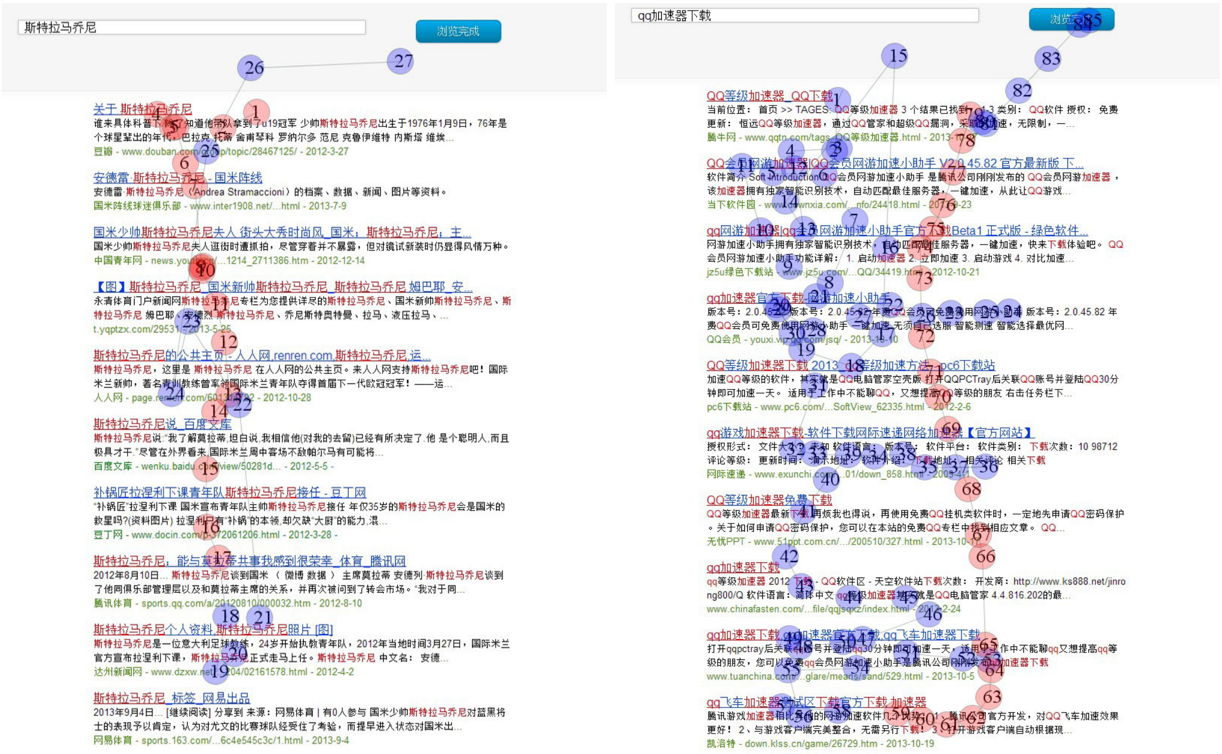
*SIGIR'15,* August 09 - 13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767721.

## 1. INTRODUCTION

Search satisfaction prediction is essential in Web search performance evaluation researches. Although there are plenty of existing studies [18, 4, 8, 17, 16] on this research topic, it is still a challenging task for two 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 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 [23, 4]. However, for many search sessions neither mouse clicks nor query reformulations are available [13, 20] and these solutions are therefore not applicable.

For the first problem (subjectivity in satisfaction judgment), some researchers design systems to collect users' explicit feedback as the ground truth for satisfaction [8, 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 [16]. According to recent studies on query intent labelling and relevance annotations [13, 28], external assessments may be very different from users' self-annotations. However, the question of whether there exist such differences in search satisfaction evaluation remains uninvestigated (*RQ1*).

For the second problem (lack of explicit feedback information), 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 [6, 22] and can easily be collected at large scale as well. Existing studies indicate that mouse movement behaviors can provide insights into result examination [22] and result relevance estimation [1, 7, 9, 12]. Guo et al. [8] 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



斯特拉马乔尼

关于 斯特拉马乔尼
谁来具体科普下 除了知道他带队拿到了u19冠军 少帅斯特拉马乔尼出生于1976年1月9日，76年是个球星辈出的年代，巴拉克 托蒂 舍甫琴科 罗纳尔多 范尼 克鲁伊维特 内斯塔 维埃...
豆瓣 - www.douban.com/group/topic/28467125/ - 2012-3-27

安德雷·斯特拉马乔尼 - 国米阵线
安德雷·斯特拉马乔尼（Andrea Stramaccioni）的档案、数据、新闻、图片等资料。
国米阵线球迷俱乐部 - www.inter1908.net/...html - 2013-7-9

国米少帅斯特拉马乔尼夫人 街头大秀时尚风_国米；斯特拉马乔尼；主...
国米少帅斯特拉马乔尼夫人逛街时遭抓拍，尽管穿着并不暴露，但对镜试新装时仍显得风情万种。
中国青年网 - news.youth.cn/...1214_2711386.htm - 2012-12-14

【图】斯特拉马乔尼_国米新帅斯特拉马乔尼_斯特拉马乔尼 姆巴耶_安...
永清体育门户新闻网斯特拉马乔尼专栏为您提供详尽的斯特拉马乔尼、国米新帅斯特拉马乔尼、斯特拉马乔尼 姆巴耶、安德烈 斯特拉马乔尼、乔尼斯特奥特曼、拉马、液压拉马、...
t.yqptzx.com/29531 - 2013-5-25

斯特拉马乔尼的公共主页 - 人人网,renren.com,斯特拉马乔尼,运...
斯特拉马乔尼，这里是 斯特拉马乔尼 在人人网的公共主页。来人人网支持斯特拉马乔尼吧！国际米兰新帅，著名青训教练曾率领国际米兰青年队夺得首届下一代欧冠冠军！——运...
人人网 - page.renren.com/6013[illegible]792 - 2012-10-28

斯特拉马乔尼说_百度文库
斯特拉马乔尼说:"我了解莫拉蒂,坦白说,我相信他(对我的去留)已经有所决定了.他 是个聪明人,而且极具才干."尽管在外界看来,国际米兰周中客场不敌帕尔马有可能将...
百度文库 - wenku.baidu.com/view/50281d... - 2012-5-5 -

补锅匠拉涅利下课青年队斯特拉马乔尼接任 - 豆丁网
"补锅匠"拉涅利下课 国米宣布青年队主帅斯特拉马乔尼接任 年仅35岁的斯特拉马乔尼会是国米的救星吗?(资料图片) 拉涅利只有"补锅"的本领,却欠缺"大厨"的能力,混...
豆丁网 - www.docin.com/p-372061206.html - 2012-3-28 -

斯特拉马乔尼：能与莫拉蒂共事我感到很荣幸_体育_腾讯网
2012年8月10日... 斯特拉马乔尼谈到国米 （ 微博 数据 ） 主席莫拉蒂 安德列·斯特拉马乔尼谈到了他同俱乐部管理层以及和莫拉蒂主席的关系，并再次被问到了转会市场。"我对于同...
腾讯体育 - sports.qq.com/a/20120810/000032.htm - 2012-8-10

斯特拉马乔尼个人资料,斯特拉马乔尼照片 [图]
斯特拉马乔尼是一位意大利足球教练，24岁开始执教青年队，2012年当地时间3月27日，国际米兰官方宣布拉涅利下课，斯特拉马乔尼正式走马上任。斯特拉马乔尼 中文名： 安德...
达州新闻网 - www.dzxw.net/[illegible]204/02161578.html - 2012-4-2

斯特拉马乔尼_标签_网易出品
2013年9月4日... [继续阅读] 分享到 来源：网易体育 | 有0人参与 国米少帅斯特拉马乔尼对蓝黑将士的表现予以肯定，认为对尤文的比赛球队经受住了考验，而提早进入状态对国米出...
网易体育 - sports.163.com/...6c4e545c3c/1.html - 2013-9-4

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

qq加速器下载

QQ等级加速器_QQ下载
当前位置： 首页 >> TAGES: QQ等级加速器 3 个结果已找到 1-3 类别： QQ软件 授权： 免费
更新： 恒远QQ等级加速器，通过QQ管家和超级QQ漏洞，采取[illegible]加速，无限制，一...
腾牛网 - www.qqtn.com/tags_QQ等级加速器.html - 2013-1[illegible]

QQ会员网游加速器|QQ会员网游加速小助手 V2.0.45.82 官方最新版 下...
软件简介 Soft Introduction QQ会员网游加速小助手 是腾讯公司刚刚发布的 QQ会员网游加速器，该加速器拥有独家智能识别技术，自动匹配最佳服务器，一键加速，从此让QQ游戏...
当下软件园 - www.downxia.com/...nfo/24418.html - 20[illegible]-23

qq网游加速器|qq会员网游加速小助手官方下载Beta1 正式版 - 绿色软件...
网游加速小助手拥有独家智能识别技术，自动匹配最佳服务器，一键加速，快来下载体验吧。 QQ会员网游加速小助手功能详解： 1. 启动加速器 2. 立即加速 3. 启动游戏 4. 对比加速...
jz5u绿色下载站 - www.jz5u.com/...QQ/34419.html - 2012-10-21

qq加速器官方下载-网游加速小助手
版本号：2.0.45.82 版本号：2.0.45.82 年费QQ会员可免费使用网游小助手 版本号：2.0.45.82 年费QQ会员可免费使用网游小助手 一键加速 无须自己选服 智能测速 智能选择最优网...
QQ会员 - youxi.vip.qq.com/jsq/ - 2013-10-10

QQ等级加速器下载 2013_qq等级加速方法 - pc6下载站
加速QQ等级的软件，其实就是QQ电脑管家空壳版 打开QQPCTray后关联QQ账号并登陆QQ30分钟即可加速一天。 适用于工作中不能聊QQ，又想提高QQ等级的朋友 右击任务栏下...
pc6下载站 - www.pc6.com/...SoftView_62335.html - 2012-2-6

qq游戏加速器下载-软件下载网际速递网络加速器【官方网站】
授权形式： 文件大小： 未知 软件语言： 版本号： 软件平台： 软件类别： 下载次数： 10 98712
评论等级： 更新时间： 演示地址： 软件介绍： 下载地址： 相关评论 相关下载
网际速递 - www.exunchi.com/...01/down_858.html - 2009-4-1

QQ等级加速器免费下载
QQ等级加速器最新下载 再烦我也得说，再使用免费QQ挂机类软件时，一定地先申请QQ密码保护。关于如何申请QQ密码保护，您可以在本站的免费QQ专栏中找到相应文章。 QQ...
无忧PPT - www.51ppt.com.cn/.../200510/327.html - 2013-10-17

qq加速器下载
qq等级加速器 2012 下载 - QQ软件区 - 天空软件站下载次数： 开发商：http://www.ks888.net/jinrong800/Q 软件语言： 简体中文 qq等级加速器其实就是QQ电脑管家 4.4.816.202的最...
www.chinafasten.com/...file/qqjsqxz/index.html - 2012-2-24

qq加速器下载,qq加速器官方下载,qq飞车加速器下载
打开qqpctray后关联qq账号并登陆qq30分钟即可加速一天，适用于工作中不能聊qq又想提高qq等级的朋友，您可以免费qq会员网游加速小助手是腾讯公司刚刚发布qq加速器下载
www.tuanchina.com/...glare/means/sand/529.html - 2013-10-5

qq飞车加速器测试区下载官方下载 加速器
腾讯游戏加速器相比其他的网游加速软件几个优势：1、腾讯公司官方开发，对QQ飞车加速效果更好！ 2、与游戏客户端完美整合，无需另行下载！ 3、打开游戏客户端自动根据现...
凯洛特 - down.klss.cn/game/26729.htm - 2013-10-19

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

**Figure 1: Examples of Users' Mouse Movement Trails on SERPs**

to clicks. However, mouse movement data contains much richer interaction information between users and search engine result pages (SERPs) than these behavior signals. Recent studies [19] already show that automatically discovered mouse movement subsequences (namely motifs) can be utilized to infer result relevance. Therefore, the question of whether satisfaction prediction can benefit from the rich interaction information stored in mouse movement logs needs to be investigated (*RQ2*).

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 and external assessors' opinions are collected as well. Figure 1 shows two examples of users' mouse movement process on SERPs with the constructed experimental search engine (see Section 3), where Figure 1(a) shows an example of SAT (self-reported satisfactory) case and Figure 1(b) 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 appeared subsequences in mouse movement data) extracted and selected by the algorithms described in Section 5. In Figure 1(a), 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 Figure 1(b) 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 information. Therefore, mouse movement information can help us infer that the user in search session shown in Figure 1(a) is likely to be satisfied while the one in Figure 1(b) is not.

The examples in Figure 1 indicate that mouse movement data records richer 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 adopt and compare two different sources of satisfaction labels from both search users and external assessors and introduce manipulated SERPs to control annotation qualities. Following recent efforts in understanding users' judgments of satisfaction [17, 16], we compare the different roles of result utility (benefit) and user effort (cost) in different satisfaction labels as well.

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. Although previous studies such as [8] already introduce mouse behavior features in addition to result clicks, motifs are not among their investigated features. According to the cases in Figure 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. [19] in that they focused on the problem of relevance estimation instead of search satisfaction prediction. We also 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 [19], they are more suitable for the task of satisfaction prediction by achieving better prediction performance with fewer motifs.

Our contributions in this paper are three-fold: (1) To our best knowledge, this is the first attempt to predict search satisfaction with mouse movement patterns (or motifs) on SERPs. (2) We propose to use distance-based and distribution-based strategies in the selection of motifs, which outperforms existing frequency-based strategy in choosing the most effective motifs to separate SAT sessions from DSAT ones. (3) With an experimental search system, we compare satisfaction labels collected from both search users and external assessors and observe for the first time that users and assessors have different criteria in search satisfaction judgments.

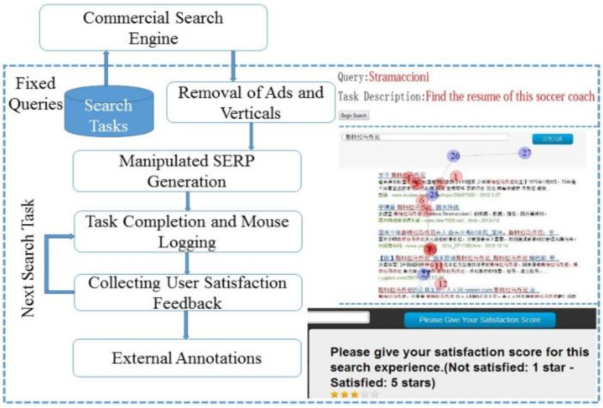
The rest of this paper is organized as follows: Related studies are discussed in Section 2. The experimental system and corresponding data collection process are presented in Section 3. The differences in satisfaction judgments from users and external assessors are investigated in Section 4. Motif extraction method and corresponding selection strategies are proposed in Section 5. Experimental results in satisfaction prediction are introduced and discussed in Section 6. Finally come the conclusions and future work directions.

# 2. RELATED WORK

Two lines of researches are related to the work we describe in this article: (1) User satisfaction understanding and prediction. (2) Search performance evaluation with mouse movement information.

The concept of satisfaction was first introduced in IR researches in 1970s according to Su et al. [27]. A recent definition states that "satisfaction can be understood as the fulfillment of a specified desire or goal" [18]. However, search satisfaction itself is a subjective construct and is difficult to measure. Some existing studies tried to collect satisfaction feedback from users directly. For example, Guo et al.'s work [8] on predicting Web search success and Feild et al.'s work [4] on predicting searcher frustration were both based on searchers' self-reported explicit judgements. Differently, other researchers employed external assessors to restore the users' search experience and make annotations according to their own opinions. For example, Guo et al.'s work [10] on predicting query performance and Huffman et al.'s work [14] on predicting result relevance were based on this kind of annotations. Recent research [28] showed that annotations on result relevances from external assessors may not be a good estimator of users' own judgements. However, the relationship between searchers' and external assessors' opinions in search satisfaction remains uninvestigated.

A number of different interaction behaviors have been taken into consideration in the prediction of search user satisfactions including both coarse-grained features (e.g. click-through based features in [10]) and fine-grained ones (e.g. cursor position and scrolling speed in [8]). Recently, a number of studies (e.g. [17, 16]) chose to use the cost-benefit framework to analyze users satisfaction judgment in search process. In this framework, both document relevance (or attractiveness, or utility) and the efforts users spend on examining SERPs and browsing landing pages are considered. In this work, we also follow the same framework and try to predict result utilities and user efforts with mouse movement features. We also want to investigate the differences in search satisfaction annotations collected from user side and external assessor side.


Commercial Search Engine
Fixed Queries
Search Tasks
Removal of Ads and Verticals
Manipulated SERP Generation
Task Completion and Mouse Logging
Collecting User Satisfaction Feedback
External Annotations
Next Search Task
Begin Search
26
27
1
25
6
90
11
23
12
Please Give Your Satisfaction Score

Query:Stramaccioni

Task Description:Find the resume of this soccer coach

Please give your satisfaction score for this search experience.(Not satisfied: 1 star - Satisfied: 5 stars)

★★★☆☆

**Figure 2: Data Collection Procedure**

In the second line of related studies, mouse movement information like scroll and hover have proven to be valuable signals in inferring user behavior and preferences [6, 7, 12, 25], search intent [5], search examination [22] and predicting result relevance [13]. 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 [2]. Lagun et al. [19] 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 search result utility, searcher effort, and satisfaction at a search task level instead of result level. We also propose different motif selection strategies to improve the prediction performance.

# 3. DATA COLLECTION PROCESS

## 3.1 Experiment Procedure

To collect user behavior data during search process and corresponding satisfaction annotation data, we construct an experimental search engine. With the system, users' interaction process while completing search tasks were recorded, including click-through, mouse movement and satisfaction annotations. 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.

We recruited 40 participants (among which 16 are female and 24 are male) for the data collecting process. All participants are first-year undergraduate students from our university with a variety of self-reported search engine utilization experiences. Their majors include life science, economic and 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.

The procedure of the experiment is shown in Figure 2. In the experiment, each participant was asked to complete 30 search tasks within about 1 hour. 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 participant was 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. Then they would be guided to continue to the next search task.

## 3.2 Search Tasks and Quality Control

To predict search satisfaction at a search task level, we selected 30 search tasks from NTCIR IMine task [21], among which there are 10 navigational tasks and 20 informational tasks. All these queries were collected from a commercial search engine and were neither long-tailed nor hot ones. Different from the IMine task, we provided detailed task explanations to the participants to avoid unnecessary ambiguity.

For each search task, we fix the query and results to ensure the consistency of our data. The search results were collected from a popular commercial search engine and only top 10 organic results are retained. Vertical results and advertisements were not included because they may affect users' search behaviors [29]. The investigation on SERPs involving verticals and advertisement will be left to future research.

Considering the fact that users may have different criteria or even be distracted during the satisfaction annotation process, we manipulate the SERPs to make a quality control of the data collection process. 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 [3]. We then design two different types of SERPS 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 determine if an annotator is reliable by observing how these pages are annotated.

For the data collection process, we had 90 (30 queries * 3 different SERPs) search conditions in total. Each participant needs to complete 30 queries using our search engine system, which contain 10 SERPs from each kind of conditions. 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 determine whether a participant is reliable based on his/her satisfaction annotation on these SERPs.

## 3.3 External Annotations

To compare the satisfaction annotations from actual users and external assessors, we recruited three assessors to annotate the satisfaction scores of the collected search sessions. The assessors are different from the ones in the query-result relevance annotation process and had worked in the commercial search engine company for at least one year. They can be regarded as professionals who have deep understanding in search engine's opinions of search satisfaction.

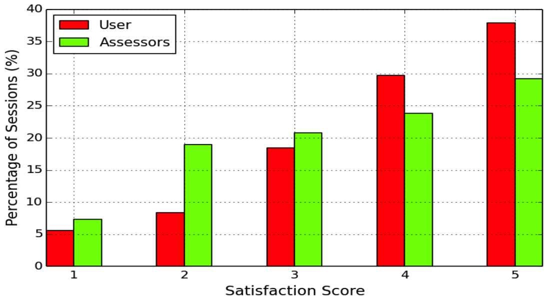


**Figure 3: Distribution of Satisfaction Annotations from Users and External Assessors**

In the satisfaction annotation process, we follow the settings in existing studies [16] and extract key information from users' search behavior logs. The assessors were shown a list of behavior items of users including search dwell time, mouse click action and click dwell time, mouse hover action and hover time, and the action of cursor movement from one search result to another. We used such detailed information to help our assessors maximally restore the original searchers' search experience and make as reasonable annotations as possible. The assessors were also asked to give a 5-point satisfaction score so that the satisfaction scores from two resources would be comparable. The KAPPA coefficient of assessors' annotations is 0.41, which can be characterized as a moderate agreement.

All the collected data is available for download through the first author's Web site[1].

# 4. USERS V.S. EXTERNAL ASSESSORS

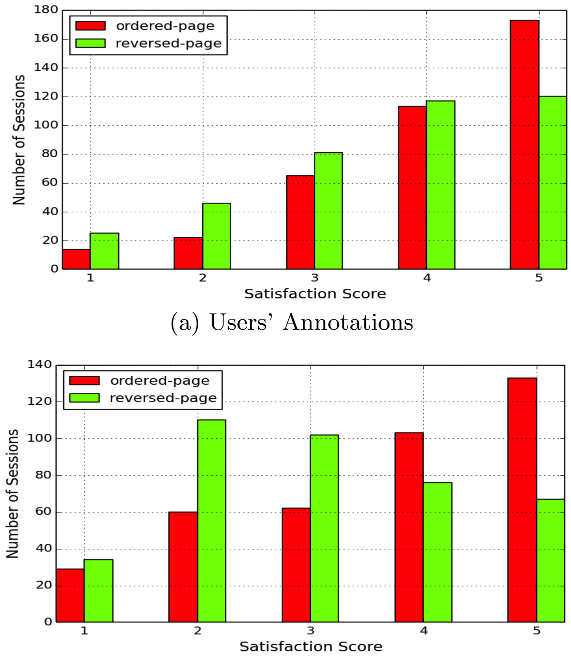
With the data collected in the experiment process, we want to compare the satisfaction annotations from actual users and external assessors. Figure 3 shows the distribution of satisfaction scores from users and assessors.

From the figure we can see that both users and assessors tend to give a high satisfaction score for the search tasks, which shows that the commercial search engine generally provides promising results for these non-long-tailed queries. We can also see from the figure that assessors annotated much more sessions with scores of 1 or 2 (26.3% v.s. 14.0%), which indicates that assessors may be stricter than users and tend to regard search session as DSAT ones.

To verify the reliability of satisfaction annotations from users and assessors, we use the two kinds of pre-defined SERPs (ordered-page and reversed-page) in the experiment process. Figure 4 shows the distribution of the satisfaction scores from two different resources on the pre-defined SERPs.

Results in Figure 4(a) 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. Figure 4(b) shows a similar distribution tendency in assessors' annotations. We notice that the percentages of reversed-pages with scores of 4 or 5 in assessors' annotations are much lower than those in users' annotations. This observation is consistent with our finding from Figure 3 that assessors may be stricter than users. It may also indicate that assessors examine the results more carefully and

[1] http://www.thuir.cn/group/~yqliu



(a) Users' Annotations

(b) Assessors' Annotations

**Figure 4: Distribution of Satisfaction Scores on Three Types of Manipulated SERPs**

their annotations are more reliable because reverse-pages are not supposed to provide satisfactory results.

Inspired by existing researches on the understanding of search satisfactions [17, 28], we also analyze the different roles of utilities and efforts in the satisfaction annotation process. This cost-benefit framework may help us better understand the differences in satisfaction annotations from users and assessors.

As for search result utility, we choose to measure it with session cumulated gain (sCG) and normalized discounted cumulative gain (nDCG@N, N=3, 5, 10) [15] based on the relevance annotations from professional assessors. Three different metrics were adopted to measure search effort, including search dwell time, maximum clicked rank and number of clicks. These effort-based metrics are widely adopted in search satisfaction related studies [4, 8].

In Tables 1 and 2, we show the Pearson correlation coefficients between satisfaction annotations and the utility/efforts metrics including sCG (U1), NDCG@3 (U2), NDCG@5 (U3), NDCG@10 (U4), search duration (E1), maximum clicked rank (E2) and number of clicks (E3). We can see that all measurements of result utility have a positive correlation with both user satisfaction and assessors' annotations. Meanwhile all measurements of user effort have a negative correlation with those annotations. All correlations are statistically significant except that between assessors' annotation and sCG. We also find that NDCG@3, search dwell time and number of clicks may have a comparatively large effect on satisfaction judgement. Such correlation between these metrics and satisfaction will be used for predicting satisfaction in later sections.

**Table 1: Correlations between Satisfaction and Utilities (*indicates statistical significance at p<0.01)**

| | sCG (U1) | NDCG@3 (U2) | NDCG@5 (U3) | NDCG@10 (U4) |
|---|---|---|---|---|
| User | 0.22* | **0.24*** | 0.25* | 0.27* |
| Assessors | 0.03 | **0.22*** | 0.20* | 0.14* |

**Table 2: Correlations between Satisfaction and Effort (All results are statistically significant, p < 0.01)**

| | Session duration (E1) | Maximum clicked rank(E2) | Number of clicks (E3) |
|---|---|---|---|
| User | -0.14 | -0.10 | **-0.15** |
| Assessors | **-0.59** | -0.35 | -0.39 |

From the statistics shown in Tables 1 and 2, we see an interesting phenomenon that users and assessors may have different criteria in the annotation of search satisfaction. We can see that the correlations between users' annotations and utility-based metrics (U1-U4) are higher than those between assessors and U1-U4. Meanwhile, the correlations between assessors' annotations and effort-based metrics (E1-E3) are much higher. It means that users tend to pay more attention to the utility of search results while assessors emphasize on the effect of search effort. Such result is reasonable since users may be satisfied as long as their search need is met in not a very long time (The average search dwell time of all users in our research is 37 seconds). Meanwhile, assessors tend to require search engines to help users locate necessary information within short time periods. They may not be able to judge how much information is enough for search users, but if the search task can be completed promptly, they are usually sure that it is a successful session and user should be satisfied.

We also expand the work in [28] to measure the consistency of the satisfaction scores from users and assessors. We calculate the correlation coefficient between the satisfaction annotations from assessors and users. Inspired by the above findings, we also divide users' satisfaction scores by three different measurements of effort since assessors tend to care more about search effort. Results are shown in Table 3. We can see that there is a fair agreement between the two resources of satisfaction annotations, which is comparable with the findings in [28]. However, if we divide users' satisfaction annotations by search efforts, we achieve a moderate agreement, which further validates the assumption that assessors pay more attention to search efforts.

These results show that utility and effort play a quite different role in users' and assessors' satisfaction judgement. Original users pay more attention to search utilities and external assessors care more about search effort. It indicates that the annotations from professional assessors may be more reliable (considering the results shown in Figure 4) but not always consistent with users' opinions. Although the explicit feedback process during search sessions may interrupt users and information collected may be noisy, we cannot simply replace the annotations with professional assessors' because they emphasize on different factors. It also means that a better satisfaction feedback strategy which collects users' opinions implicitly (e.g. the one proposed in our work which predicts satisfaction based on motifs) is necessary for the evaluation of search performances.

# 5. MOTIF EXTRACTION AND SELECTION

## 5.1 Motif Candidate Extraction

The concept of motif is first introduced by Lagun et al. [19] and defined as frequent subsequences in mouse cursor movement data. They proposed to automatically extract

**Table 3: Correlation between Satisfaction Annotations from Users and External Assessors (All values are statistically significant, p < 0.01)**

| | User | User\E1 | User\E2 | User\E3 |
|---|---|---|---|---|
| Assessors | 0.27 | **0.49** | 0.33 | **0.54** |

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. In the generation of motifs, we also use DTW(Dynamic Time Warping)[26] for distance measurement as in [19] but try both Euclidean and Manhattan distances in calculation. Euclidean distance which is not selected by [19] 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 [19] and a number of time series mining studies such as [24]. 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.

## 5.2 Motif Selection Strategies

A major difference between our motif extraction method and the one in [19] is that we use a number of selection strategies to find the most predictive motifs from candidates. Different from the frequency-based strategy in [19] 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 firstly 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 $C_1, C_2, ..., C_N$, we can obtain $N$ distance features for a certain search session $S$: $Dist(C_1, S), Dist(C_2, S)...Dist(C_N, S)$, 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 [19], 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.

### 5.2.1 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 Figure 1 also validates 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.

### 5.2.2 Distribution-based Selection

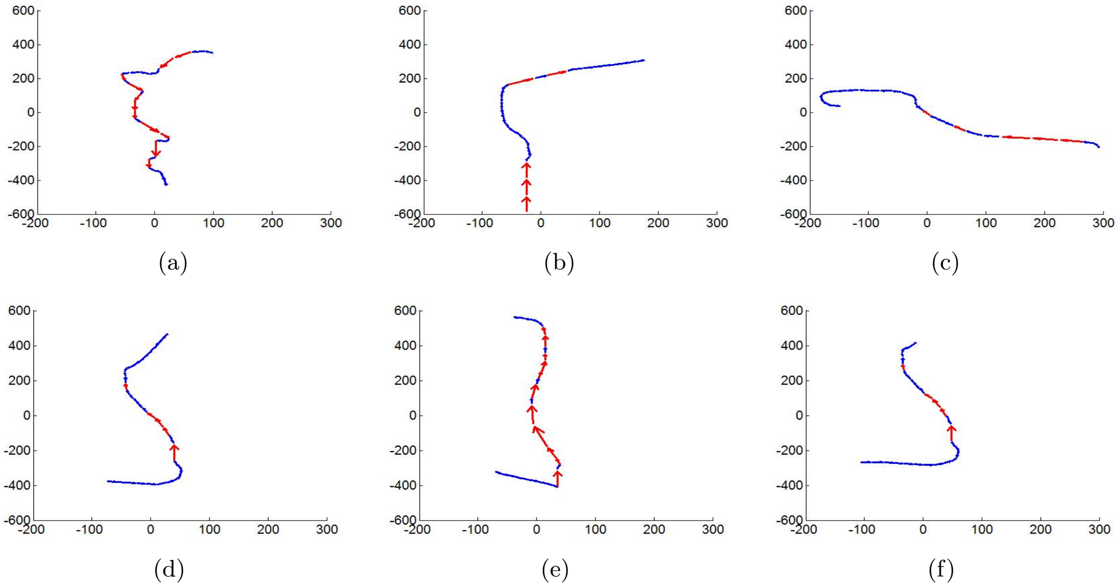
This strategy is based on a **Covering Hypothesis**: predictive motifs in $M\_SAT/M\_DSAT$ should cover sufficient sessions in $SAT\_DATA/DSAT\_DATA$. We introduce this hypothesis because when a certain motif can only cover a small number of sessions, it is not reasonable to select it even if it is quite 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 $D$:

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



(a)

(b)

(c)

(d)

(e)

(f)

**Figure 5: Predictive motifs discovered from $SAT\_DATA$ (a-c) and $DSAT\_DATA$ (d-f)**

In (4), $r$ is the parameter to ensure we can select enough motifs, which we set as $\frac{1}{30}$ in our experiment. With the concept of coverage rate, we can define the score for each motif based on distribution difference as follows:

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

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

Similar to the method in 5.2.1, we select motifs with high scores since they tend to have a large distribution difference.

## 5.3 Example of Predictive Motifs

The distance-based and distribution-based strategies can help discover predictive motifs from mouse movement data and a few examples are shown in Figure 5. Figure 5(a), 5(b) and 5(c) show 3 of the 10 most predictive motifs extracted from $SAT\_DATA$ while Figure 5(d), 5(e) and 5(f) show 3 of the 10 most predictive motifs extracted from $DSAT\_DATA$. 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 Figure 5(a) shows a process that user examines the top results carefully and then take a quick look at the lower-ranked results and Figure 1(a) can be regarded a practical example. Figure 5(b) probably shows the process of re-visiting a previous checked result while Figure 5(c) 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 Figure 5(d), 5(e) and 5(f) 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, as will be discussed in the next section.

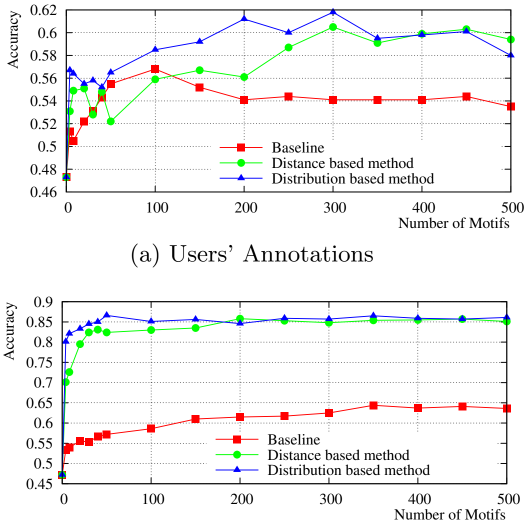
# 6. EXPERIMENTAL RESULTS

## 6.1 Experiment Setups

In this section, we demonstrate the value of our method by predicting the satisfaction annotation results of users and assessors. After the motif extraction and selection process described in Section 5, the motifs from the data sets collected in Section 3 are adopted as features in the prediction process. There are two parameters in the motif extracting algorithm we discussed in section 6.1, namely the length of sliding window and the distance measurement method for two basic points. Based on the result of quite a number of experiments, we find that a sliding window of 3 seconds can make the model perform best. We also found that the Euclidean distance measurement can help the model achieve a promising accuracy with just a small number of motifs. It is extremely important considering the fact that the procedure of extracting motifs may be time-consuming. Such findings are different from that in [19], where a sliding window of 5 seconds and the Manhattan distance measurement is used. This difference may come from the fact that we focus on discovering motifs from SERPs while Lagun et al. mainly try to discover motifs from landing pages. A time period of 3 seconds may be enough for a mouse action on SERPs because there are mainly ten blue links without many other components. The Euclidean distance measurement may better reflect the distance of two points in a two-dimensional space on SERPs because the sizes of components on SERPs are similar with each other.

We compare the performance of the proposed model in predicting satisfaction scores from users and external assessors. We exclude sessions with a satisfaction score of 3 since we consider that users or assessors do not have a satisfaction tendency in such sessions. Thus we use 951 search sessions from users and 923 search sessions with assessors' annotations for satisfaction predicting, where sessions with a score of 4 or 5 are regarded as SAT cases and those with a score of 1 or 2 are regarded as DSAT ones.

The learning algorithm in the prediction process is logistic regression, which is widely used in satisfaction prediction



(a) Users' Annotations

(b) Assessors' Annotations

**Figure 6: Prediction Performance with Different Motif Selection Strategies**

tasks [8]. All results are based on 5-fold cross-validation on the data collections unless specified otherwise.

## 6.2 Comparison of Motif Selection Strategies

To compare the different strategies for selecting motifs, we use the method used in [19] as a baseline, which selects motifs based on frequency in training set. Experimental results on the prediction of both users' and assessors' annotations are shown in Figures 6.

Results on both datasets shown in Figure 6 indicates that the proposed selection strategies based on distance and distribution outperform the selecting strategy based on frequency. The performance of the proposed prediction model performs better as the number of motifs increases. Between the two selecting methods we proposed in section 6.2, we consider the one based on distribution better since it can reach a high accuracy with comparatively a small number of motifs on both two datasets. We want to predict satisfaction with a small number of motifs so that the motif extraction process can be efficient. Therefore, prediction models used in the following sections all adopt the distribution-based selection strategy.

Another important finding from Figure 6 is that our predict model can reach an accuracy of more than 0.85 on assessors' annotations while that on users' satisfaction annotations is only around 0.60. This indicates that users' self-annotations may be quite subjective and may be more difficult to be predicted, which validates the necessity of investigating results based on these two different satisfaction annotations as we did in Section 4.

## 6.3 Comparison of Quality Control Strategies

Considering the fact that different users may have different opinions in satisfaction judgement, satisfaction annotations collected from users may be subjective and sometime even unreliable. To obtain reliable satisfaction annotations from users, we introduce a number of quality control strategies with the manipulated SERPs described in Section 3.3. Specifically, we test the performance of our prediction model on the following three different datasets:

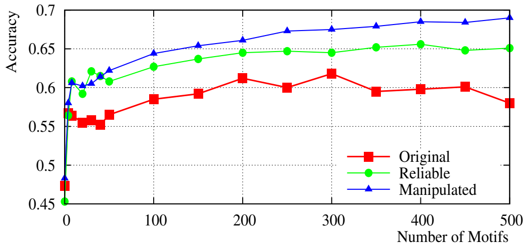


**Figure 7: Comparison of Prediction Accuracies with Different Quality Control Methods**

**The original dataset**. This dataset is the one we use in Figure 7(a), which includes all search sessions with satisfaction scores of 1, 2, 4 or 5 from users.

**The reliable dataset**. We use the manipulated SERPs as a information source for quality control and remove data from unreliable users. 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 $x_i, y_i (i = 2, 3, 4, 5)$. We can then use a combination of $x_i, y_i$ to calculate a score for each participant to measure his/her reliability.

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

We think that users tend to be more satisfied with ordered-pages and less satisfied with reversed-pages. If a searcher gives high satisfaction scores for ordered-pages and low scores for reverse pages, we consider him/her as a reliable source of annotation information. Based on this assumption, we define the reliability score as:

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

It is reasonable to think that the lower the score is, the participant tends to be less reliable. We remove annotations coming from participants with the five lowest reliability scores and the remaining 827 search sessions are regarded as the reliable dataset.

**The manipulated dataset**. We define the sessions with ordered-pages as satisfied cases and those with reversed-pages as unsatisfied ones. It should be noticed that this dataset is objective since it 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 Figure 8. We can see that the proposed model performs best on the manipulated dataset and also gain promising results on the reliable dataset. Prediction performance on the original dataset is the worst because the annotations from unreliable users are not removed and the subjectivity of annotations are also expected. We also notice that the accuracies on manipulated and reliable datasets are still lower than that on the external assessors' annotations in Figure 6(b). It indicates that users' annotations on search satisfaction are rather subjective. Although we can reduce possible noises with certain quality control strategies, it is still difficult for prediction models to provide high-quality predictions.

## 6.4 Predicting Utility and Effort

We try to estimate utility and effort with motifs in this section since some metrics like NDCG@3, search dwell time

**Table 4: NRMSE of Proposed Method for Predicting Utility and Effort**

| Number of Motifs | NDCG@3 | Dwell time | Number of Clicks |
|---|---|---|---|
| 5 | 0.287 | 0.134 | 0.158 |
| 10 | **0.287** | **0.133** | 0.151 |
| 50 | 0.288 | 0.137 | 0.127 |
| 100 | 0.299 | 0.142 | **0.126** |
| 300 | 0.383 | 0.181 | 0.147 |
| 500 | 0.583 | 0.259 | 0.223 |

and number of clicks have a remarkable correlation with satisfaction annotations according to Section 4. If we could predict these factors in user sessions, it also helps us better understand why users are satisfied or unsatisfied with certain result lists.

We use Normalized Root Mean Square Error (NRMSE, smaller value means better estimation) to evaluate the model performance as in most regression-based methods. The experimental results are shown in Table 4.

From the results shown in Table 4 we see that motifs can moderately estimate NDCG@3, session dwell time and number of clicks since the NRMSE value is considerably small. We also notice that motifs can probably better estimate search effort since the NRMSE value is much smaller for session dwell time and number of clicks. Another important finding is that we gain best performance when the number of motifs used is around 50, which indicates that the estimation model may overfit if too much motifs are incorporated.

## 6.5 Prediction across Users and Queries

According to Section 5, 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 predict model across different users and queries. According to previous studies on predicting examination sequence with mouse movement information [11], 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 or queries, we adopted 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 also implement a satisfaction prediction method proposed in [8] (with both coarse-grained features such as number of clicks and fine-grained features such as scroll speed) and adopted it as the baseline method. We choose the method because it is also based on mouse behavior data (although without motifs) and is one of the most closely related studies. The baseline method and our proposed method are both tested with the three different training strategies and results are shown in Table 5.

**Table 5: Comparison of Accuracy with the baseline method in [8] across different users and queries(* indicates statistical significance at p < 0.1 level)**

| Annotation & Sampling strategy | Features in [8] | Motif Features | All Features |
|---|---|---|---|
| User annotation & random sample | 0.570 | 0.580<br>(+1.75%) | 0.598<br>(+4.91%) |
| User annotation & sample by user | 0.523 | 0.578<br>(+10.5%*) | 0.552<br>(+5.54%*) |
| User annotation & sample by query | 0.602 | 0.631<br>(+4.82%) | 0.638<br>(+5.98%) |
| Assessor annotation & random sample | 0.920 | 0.861<br>(-6.41%*) | 0.922<br>(+0.217%) |
| Assessor annotation & sample by user | 0.921 | 0.859<br>(-6.73%) | 0.930<br>(+0.977%) |
| Assessor annotation & sample by query | 0.924 | 0.886<br>(-4.11%*) | 0.938<br>(+1.52%) |

Results in Table 5 reveal a number of interesting findings: 1) The prediction performance of the proposed method with motif features doesn't change much 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) Compared with the "fine-grained" interaction features (e.g. scroll speed, y-axis maximum coordinate, etc) proposed in [8], the proposed motif-based method performs better in predicting users' annotations but worse in predicting assessors' annotations. When we investigate the differences between two sets of features, we found that the features related with y-axis maximum coordinates in [8] are quite predictive. It probably accords with our findings in Section 4 that assessors emphasize on search effort in satisfaction annotations because the maximum coordinates in y-axis is a strong signal for the efforts of users. As for the prediction of users' annotations, the motif-based method studies better, especially in predicting the opinions of unseen users (with a significant improvement of 10.5%). This means that the proposed method makes use of more details in users' interaction process and is probably more suitable for practical applications (in which predicting previous-unseen users' opinion is important). 3) With all the features proposed in baseline method and our proposed method, we gain best prediction performance for both users' and assessors' annotations. It shows that the motif features can be used to improve state-of-the-art technologies and they can be extremely useful for the satisfaction prediction of previous-unseen users.

# 7. CONCLUSIONS AND FUTURE WORK

Search satisfaction prediction is a non-trivial task in search performance researches. 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 collect data from both users and assessors to make a deep analysis. We find that there is a moderate agreement between satisfaction annotations from those two resources. Users pay more attention to the utility of results while external assessors also emphasize on the search effort.

We further propose a motif based learning framework to predict result utility, search effort as well as satisfaction annotations for both users and assessors. 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.

For future work, we would like to further improve the efficiency of mining motifs and try to incorporate other features into satisfaction predicting models. Besides, we also want to research how motifs can be used in a heterogeneous Web search environment and also other web search applications.

## 8. ACKNOWLEDGMENTS

This work is supported by Tsinghua University Initiative Scientific Research Program(2014Z21032), National Key Basic Research Program (2015CB358700) and Natural Science Foundation (61472206, 61073071) of China. Part of the work has been done at the Tsinghua-NUS NExT Search Centre, which is supported by the Singapore National Search Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

## 9. REFERENCES

[1] M. Ageev, D. Lagun, and E. Agichtein. Improving search result summaries by using searcher behavior data. In *SIGIR'13*, pages 13–22. ACM, 2013.

[2] I. Arapakis, M. Lalmas, and G. Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *CIKM'14*, pages 1439–1448. ACM, 2014.

[3] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[4] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR'10*, pages 34–41. ACM, 2010.

[5] Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. In *SIGIR'08*, pages 707–708. ACM, 2008.

[6] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI'10*, pages 3601–3606. ACM, 2010.

[7] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW'12*, pages 569–578. ACM, 2012.

[8] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. In *CIKM'12*, pages 2050–2054. ACM, 2012.

[9] Q. Guo, D. Lagun, D. Savenkov, and Q. Liu. Improving relevance prediction by addressing biases and sparsity in web search click data. In *WSCD'12*, pages 71–75, 2012.

[10] Q. Guo, R. W. White, S. T. Dumais, J. Wang, and B. Anderson. Predicting query performance using query, result, and user interaction features. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 198–201, 2010.

[11] J. Huang, R. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. In *SIGCHI'12*, pages 1341–1350. ACM, 2012.

[12] J. Huang, R. W. White, G. Buscher, and K. Wang. Improving searcher models using mouse cursor activity. In *SIGIR'12*, pages 195–204. ACM, 2012.

[13] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *SIGCHI'11*, pages 1225–1234. ACM, 2011.

[14] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *SIGIR'07*, pages 567–574. ACM, 2007.

[15] K. Järvelin, S. L. Price, L. M. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Advances in Information Retrieval*, pages 4–15. Springer, 2008.

[16] J. Jiang, A. H. Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM'15*, 2015.

[17] J. Jiang, D. He, and J. Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. 2014.

[18] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009.

[19] D. Lagun, M. Ageev, Q. Guo, and E. Agichtein. Discovering common motifs in cursor movement data for improving web search. In *WSDM'14*, 2014.

[20] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR'09*, pages 43–50. ACM, 2009.

[21] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. Kato, H. Ohshima, and K. Zhou. Overview of the ntcir-11 imine task. In *NTCIR*, volume 14, 2014.

[22] 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 *CIKM'14*, 2014.

[23] 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 *SIGIR'11*, 2011.

[24] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *SIGKDD'12*, 2012.

[25] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI'08*, pages 2997–3002. ACM, 2008.

[26] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. 1978.

[27] L. T. Su. Evaluation measures for interactive information retrieval. *Information Processing & Management*, 28(4):503–516, 1992.

[28] S. Verberne, M. Heijden, M. Hinne, M. Sappelli, S. Koldijk, E. Hoenkamp, and W. Kraaij. Reliability and validity of query intent assessments. *JASIST*, 64(11):2224–2237, 2013.

[29] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In *SIGIR'13*, pages 503–512. ACM, 2013.