User Intent, Behaviour, and Perceived Satisfaction in Product Search

Ning Su
DCST, Tsinghua University
Beijing, China
sn-40@163.com

Jiyin He
CWI
Amsterdam, The Netherlands
j.he@cwi.nl

Yiqun Liu*
DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Min Zhang
DCST, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Shaoping Ma
DCST, Tsinghua University
Beijing, China
msp@tsinghua.edu.cn

ABSTRACT

As online shopping becomes increasingly popular, users perform more product search to purchase items. Previous studies have investigated people's online shopping behaviours and ways to predict online purchases. However, from a user perspective, there still lacks an in-depth understanding of why users search, how they interact with, and perceive the product search results. In this paper, we address the following three questions: (1) what are the intents of users underlying their search activities? (2) do users behave differently under different search intents? and (3) how does user perceived satisfaction relate to their search behaviour as well as search intents, and can we predict product search satisfaction with interaction signals?

Based on an online survey and search logs collected from a major commercial product search engine, we show that user intents in product search fall into three categories: *Target Finding* (TF), *Decision Making* (DM) and *Exploration* (EP). Through a log analysis and a user study, we observe different user interaction patterns as well as perceived satisfaction under these three intents. Using a series of user interaction features, we demonstrate that we can effectively predict user satisfaction, especially for TF and DM intents.

ACM Reference Format:
Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining (WSDM 2018).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3159652.3159714

1 INTRODUCTION

Online shopping has overtaken traditional store shopping in popularity. A 2016 survey[1] shows that 54% of the shoppers worldwide buy products online weekly or monthly, and 34% agree that mobile phones have become their primary shopping devices. The search engine provided by a shopping site is one of the main entrances for its users to find specific models, brands or products to buy. In order to effectively design product search engines, it is important to understand how users interact with these systems and how this relates to their perceived system performance. Previous studies on online shopping have studied how users make online purchase decisions [30, 31]; and aimed to enhance product search engines by proposing new ranking models [25, 29]. However, there still lacks an in-depth understanding of product search from the user perspective, i.e. why users search, and how they examine and perceive the product search results.

User satisfaction, a key concept in measuring search success [21], is not yet well understood in the context of product search. Since information search is an important stage of buyer decision process [7, 32], offering result lists that satisfy users' need can increase the loyalty of users and further promote sales. While many shopping sites use purchase as a measure of user satisfaction, satisfaction is not always associated with purchase, as shown in an example search session (Table 1) below. Here user U_1 was satisfied as she found the desired product at rank 2 of the result list. Yet this did not lead to a purchase—e.g. she may decide to make the purchase later in a physical store. Further, depending on their search intents, two users issuing the same query may have very different interaction patterns and perceived satisfaction with the same search results. For instance, different from U_1 who was seeking a specific target product and satisfied after one click, U_2 wanted to investigate different price/model options, and kept exploring results while remained unsatisfied given her goal.

In summary, in the context of product search, user satisfaction depends on a variety of factors including user search intents, search result relevance, personal shopping habits, etc; and the relation between a user's perceived search satisfaction, search intent, and observable search activities can be complex and remains unclear.

Table 1: Example real world search sessions from two users who issued the query "iPhone" to a product search engine.

User	U_1	U_2
Information Need	I'm looking to buy a white iPhone 5	I'd like to investigate the most cost-effective iPhone version to buy
Query	iPhone	iPhone
Rank #1 iPhone 4		click (37 sec)
Rank #2 iPhone 5	click (35 sec)	click (44 sec)
Rank #3 iPhone accessories		click (42 sec)
Satisfaction Feedback	satisfied	not satisfied
Purchase	no	no

*Corresponding author
[1]http://www.pwc.com/totalretail

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3159714

In this paper, we focus on providing insights for the understanding of user search behaviour, and models for predicting their perceived search satisfaction in the context of product search. As mobile phones are becoming the main tools for online shopping, all the data involved is collected from mobile devices. Specifically, we aim to answer the following three research questions:

RQ1. What are the intents underlying user search activities?
RQ2. Do users behave differently under different search intents and if so, how?
RQ3. a. How does user satisfaction relate to their search behaviour and search intents? and b. Can we predict product search satisfaction with interaction signals?

With RQ1 we aim at identifying different types of user intents during product search. Based on the coding of an online survey, we propose to categorise user intents into three categories: *Target Finding* (TF), *Decision Making* (DM) and *Exploration* (EP) (discussed in Section 3). The proposed taxonomy is then verified with the survey as well as a log sample from a popular commercial product search engine. We find that the three categories cover over 95% of the search sessions from the log; and observe a similar distribution of search intents in the survey and in the log data.

By addressing RQ2 (Section 4) we study how users interact with search results under different types of search intents. From the commercial product search log we see users behave differently under different intent categories. Users with *TF* intent exhibit focused search behaviour—they use few specific queries and only browse and click a few top ranked results. In contrast, users with *DM* intent use shorter queries, browse deeper in the result list and click more results. Users with *EP* intent tend to issue many semantically dissimilar queries.

We further investigate the relation between user satisfaction, their search intent and behavioural patterns with RQ3(a) (Section 5). To do so, we carry out an in-lab user study to collect usage data of an experimental mobile product search system with designed search intents and explicit user feedback on search satisfaction. Finally, we address RQ3(b) by exploring the effectiveness of a series of classification models and interaction features in predicting user satisfaction (Section 6). Results show these models and features achieve reasonably good performance, epecially for intent categories *DM* (F1=0.808, AUC=0.760) and *TF* (F1=0.758, AUC=0.651).

Our study has the following contributions:

- We propose a novel product search intent taxonomy that is verified with both an online survey and real world search logs.
- Our analysis over the search logs from both the commercial product search engine and the in-lab study provides insights in the relation between user behavioural patterns, their underlying intents, and their perceived search satisfaction. These insights can, for instance, help search engines to personalise the search results with respect to user intents, optimising their search experience.
- We demonstrate that it is possible to effectively predict user search satisfaction using interaction features, revealing oppotunities for new evalution measures and optimisation methods for product search systems.

2 RELATED WORK

Previous research falls into three categories: taxonomy of search activities, user search satisfaction, and product search.

2.1 Taxonomy of Search Activities

In order to develop search engines satisfying diverse types of information needs, goals or intents, it is important to understand what users are searching for, why they search, and how they search. Studies have attempted to characterise search activities from various perspectives. Broder [4] proposed a taxonomy that categorises Web search into three types: navigational, informational, and transactional. Based on this, Rose and Levinson [34] created a framework for understanding user goals; and some studies proposed algorithms for automatic query type identification [24, 28]. Broder's taxonomy is rather coarse-grained, e.g. the informational category covers a diverse types of search tasks [6]. Considering information seeking from a learning perspective, studies have proposed to categorise search tasks using Anderson and Krathwohl's taxonomy of educational objectives [3], e.g. Jansen et al. [17] observed different user behaviours with respect to different task types.

While the above studies provide an overall view of Web search, the resulting taxonomies are not readily tailored to characterising product search. We focus on providing insights for the understanding of why, and how users search products and propose a novel product search intent taxonomy.

2.2 User Satisfaction

User satisfaction is widely used as a subjective measurement of search success. The concept of satisfaction was first introduced in information retrieval research in the 1970s [35], and is defined as *"the fulfillment of a specified desire or goal"* in recent literature [21]. To predict user search satisfaction, studies have focused on analysing user interactions with search engines. Guo et al. [13] employed interaction features (e.g. click-through based features), query features, and result features to predict query performance; fine-grained features such as cursor position and scrolling speed [10, 16] were considered in later studies. Unlike relevance based measures [1, 9, 11, 15], search satisfaction is based on the overall search experience, e.g. including information gained as well as the effort spent on examining SERPs and landing pages [2]. Therefore, studies have employed the cost-benefit framework that considers both document relevance and the efforts users spend [18, 19].

In the context of mobile search, studies have found interaction patterns different from those on desktops [14, 20], leading to different features for predicting user satisfaction. For example, Li et al. [26] found the good abandonment rate from mobile search is significantly higher than that from PC search. Williams et al. [36] studied different good abandonment scenarios in mobile search. Kiseleva et al. [22] found that the notion of satisfaction varies across different scenarios. Some studies (e.g. [12, 23]) analysed the viewport (the visible portion of a web page) of the mobile devices, and used it to measure user satisfaction in the absence of clicks.

Our work differs from these existing efforts in that we study user search satisfaction with respect to interaction patterns that are specific to product search.

2.3 Product Search

Research of online customer behaviour has gained much attention recently. Using page-to-page clickstream data and the general content of the pages from a given online store, Moe [31] categorized visit patterns as *buying*, *browsing*, *searching*, and *knowledge-building*. Lu et al. [30] proposed a framework for mining and predicting users' movements and purchasing transactions in the context of mobile

commerce. These studies focused on the overall shopping experiences rather than the search processes. Li et al. [25] pointed out that the decision mechanism underlying a purchase process is different from that of locating relevant documents/objects. Many studies therefore focused on enhancing product search with consumption features such as consumer preferences [25] and volume of sales [29]. However, no previous work has aimed to understand users' satisfaction perception during product search. We make a first attempt to classify user intent in product search and predict user satisfaction in different search scenarios.

3 PRODUCT SEARCH INTENT TAXONOMY

In this section, we address RQ1 by proposing a product search taxonomy, followed by an empirical verification of its validity.

3.1 Establishing the Taxonomy

Human interacts with devices in response to certain intents or goals [4]. One would assume the goal of online shopping is to make a purchase. However, sometimes people only search to obtain information from the shopping site; and sometimes people simply do online window shopping to kill time. In an attempt to propose a taxonomy to characterise different types of search activities during online shopping, we designed an online survey to collect users' product search experience. Following basic demographic questions, participants were asked to answer three open-ended questions:

1. Please describe your latest product search experience with as many details as possible (target, motivation, time, place).
2. Please provide your queries. (You can view your search history to find all the queries used in this search.)
3. Please provide the name of the product you purchased (if any).

We spread the survey via a popular social software (WeChat). We received responses from 355 people with 60.56% male and 39.44% female (42% participants were between 18 and 25, 35% between 26 and 30, 20% between 31 and 50, and 3% were either above 50 or below 18). After removing replies with unclear descriptions, we obtained 295 valid replies.

Three Web search professionals reviewed the survey data and coded each response. The coding process takes a grounded theory inspired iterative process, where each iteration consists of two steps: a) conceptualizing the user descriptions, and b) clustering the concepts to form categories of user intents. A random sample of 50 responses was first coded by the researchers individually. They then discussed to resolve conflicts and established an initial coding standard. Based on this, they continued to code the remaining responses and this process repeated. The final coding scheme consists of three concepts that were used to describe user search activities.

Motivation: Why a user searches for a (type of) product(s), which includes two primary types:
- a. With an immediate purchase need (e.g. to solve a problem such as replacing a broken phone, or to buy gifts for others)
- b. Without immediate purchase need (e.g. to kill time, browsing new arrivals)

Target specificity: How sepcific a user's requirement of the target product(s) is—we have identified four levels of specificity from the survey descriptions:
- a. Known product category (e.g. clothes, snacks)
- b. Known product name/type (e.g. wind coat, chips)
- c. Known product brand (e.g. Zara wind coat, Pringles chips)
- d. Undetermined (unknown/multiple product categories)

Table 2: Example search scenarios from the survey.

Cat.	Examples	Query list
TF	"My phone was broken. I wanted to buy the latest iPhone."	"iPhone 7"
	"Recently I started to learn table tennis, so I needed to buy a pair of Double Happiness table tennis rackets. "	"DHS table tennis rackets"
DM	"The weather was getting cooler. I wanted to buy some long sleeves for my son."	"long sleeve T-shirt"; "autumn T-shirt"
	"I would like to buy a refrigerator in the dormitory. After comparing the different brands, I chose the Haier refrigerator."	"refrigerator"; "Haier refrigerator"
EP	"I just felt bored at class and didn't buy anything at last."	"dress"; "shoes"
	"I would like to track Nike's new products."	"Nike"

Search Strategy: How the user searches—we identified the following strategies from the survey replies:
- a. Specific keyword (keyword with detailed specification)
- b. Evaluate (Inspect a product's specification, price, reviews, etc);
- c. Compare (two or more items—typically involves evaluation of individual items);
- d. Browse (seasonal sales, new arrivals, etc);
- e. Direct purchase (directly buy a targeted item);

After coding the detailed user descriptions with the above concepts, the researchers then attempted to form categories of user product search intents. Three top level categories were identified, namely *Target finding* (TF), *Decision making* (DM), and *Exploration* (EP). From TF to EP, the categories represent increasingly more explorative search strategies and less determined search target. The categories are defined as follows:

- **Target Finding (TF):** The user has a specific target in mind (target specificity is at least at level b, typically at level c). Typical search strategy includes direct purchase and specific keywords. Users normally do not need to compare different products in this category (except price comparison from different sellers); the choice of color, model and other details can often be made in the same shop after a click; evaluation may occur as users may want to check if the product is exactly as he/she expected. Search in TF is typically conducted with an immediate purchase need.
- **Decision Making (DM):** The user has an immediate purchase need. He/she has a vague idea of what to buy (i.e. target specificity is at level b) but would typically explore and compare related products of different brands/models in the result list in order to make a purchase decision.
- **Exploration (EP):** Here the user explores the search result without a specific target in mind (i.e. target specificity is at level a or d). Typical search strategy is browsing. The user may or may not have an immediate purchase need, described by the following two sub-groups:
  - a. Casual exploration (e.g. kill time, exploring seasonal sales);
  - b. Purposeful exploration (e.g. search for birthday gifts).

Table 2 shows some examples from the survey.

3.2 Taxonomy Verification

Having established our taxonomy, we now verify its validity using data collected from the survey as well as search logs from a commercial product search engine.

3.2.1 User Survey. Three annotators (different from the researchers who established the taxonomy) were employed to examine the described search scenarios and queries, and categorize these into one of the three types of search intents as defined in our proposed taxonomy. This process examines how well our proposal can be applied in practice. We provide annotators with the definitions of

Table 3: Distribution of the three types of user intent.

	TF	DM	EP
Data from the survey	16.3%	61.8%	21.9%
Data from the search logs	15.7%	60.1%	24.2%

each user intent and asked them to select the category based on coding criteria as described in Section 3.1. They were asked to label a scenario as "others" if they could not determine which category it belongs to, or that it does not belong to any of the three.

Out of the 295 valid replies, 288 (97.6%) cases fall into one of the three categories (excluding "others") with at least two annotators agreed on the labels. The Fleiss' κ [8] is 0.756 among three annotators. This result shows that the proposed taxonomy can be understood and employed by external annotators (Fleiss' κ is 0.679 between the two sets of annotations). Additionally, we see our taxonomy has a high coverage of the search scenarios in the survey and the majority of the cases clearly belong to one of the categories. Table 3 shows the distribution of the three categories (row 1): with *DM* being the most common user intent, accounting for 61.8% of the cases, followed by *EP* (21.9%) and *TF* (16.3%).

3.2.2 Search Logs. We further verified the proposed taxonomy with the search logs of a popular commercial shopping site. We randomly sampled 1000 anonymised users and took all the queries they issued on mobile devices during one day in October 2016.

We created search sessions by setting a time-out threshold of 30 minutes [33] between consecutive queries. This resulted in 1800 search sessions. We then employed 18 assessors to annotate these sessions with our taxonomy. Every three annotators were grouped together to annotate the same 300 sessions. The annotators were shown the query list and the title of the products that users clicked/added to cart/bought on each SERP, which can be used to infer the concepts mentioned. Again, we provided them with the definitions, the coding criteria and some examples of each user intent in the taxonomy as described above (Table 2). The option of category "others" was also given.

In total, 1,713 sessions (95.2%) had at least two annotators agreeing on a user intent category other than "others", i.e. that our proposed taxonomy covers most of the real world product search sessions. Further, the distribution of the three categories among the search sessions in the log is similar to what we observed in the online survey (χ^2=0.753, p=0.686) (See Table 3). In this case, the value of Fleiss' κ among each three annotators varies between 0.492 and 0.639 (mean = 0.545), indicating a moderate agreement.

3.3 Discussion

We now briefly discuss how our taxonomy relates to some of the existing taxonomies. Both the *TF* and *DM* intents would fall into the "transactional" category of Broder's taxonomy [4], as they are both motivated by a purchase intent; while *EP* does not correspond to an existing category. Meanwhile, *DM* can also be "informational" as users explore and collect information to make purchase decisions, while *TF* is similar to "navigational". These overlaps and gaps between categories indicate that Broder's taxonomy of Web search is not directly applicable to product search and a dedicated taxonomy is needed. Moe [31]'s taxonomy, on the other hand, describes overall activities on online shopping sites, and our taxonomy essentially provides a more fine-grained description of the search activities.

Table 4: Session statistics by user intent (* indicates one-way ANOVA with Bonferroni post hoc test, significant at p<0.01)

	TF	DM	EP
Queries per Session (mean)	2.05	2.20	4.20*
Reformulation (%)	40.89	45.19	81.93
Search Depth in Pages (mean)	2.91*	5.11	4.82
Page Turning (%)	45.82	66.08	61.53
Buy (%)	10.04	10.98	8.67
Add to Cart (%)	7.06	12.83	14.22

4 USER BEHAVIOUR AND SEARCH INTENTS

In this section, we address RQ2 by exploiting the annotated search logs (see Section 3.2.2) to provide insights into user product search behaviour in relation to their search intents.

From the logs we extract user interaction data for events logged on two types of pages: the *search results pages (SERPs)* and the landing *product pages*. The product pages provide detailed product-related information, and operational options such as *add to cart* and *buy*. From each log entry we extract the following: query string, page number, result list (a set of product ids), clicked product ids, *add to cart* events and *buy* events.

4.1 Session Level Statistics

We start by examining the session level user interaction patterns in relation to their intents, providing an overview of user behaviour in product search. See Table 4.

In terms of query behaviour, *EP* sessions contain significantly more queries than *TF* and *DM* sessions. Further, 81.93% of the *EP* sessions contain more than one query—indicating that *EP* users tend to reformulate queries. *DM* sessions contain slightly more queries and query reformulations than *TF* sessions, but the difference is not significant.

In terms of search depth (defined as the number of SERPs a user browses per query), users in both *EP* and *DM* sessions browse significantly deeper than in *TF* sessions (4.82, 5.11, and 2.91 pages respectively), and tend to turn pages more often (60% of the *EP* and *DM* sessions compared to 45.8% of the *TF* sessions). It is reasonable since users seek specific products with specific queries in *TF* sessions and there is no need to explore many products as in the *DM* and *EP* sessions.

Finally, we look into purchase behaviours, i.e. the *buy* and *add to cart* events. *DM* and *TF* sessions show a higher purchase rate than *EP* sessions, which is consistent with the result collected from the online survey, based on the answers to the third question. However, *EP* sessions have the highest *add to cart* rate (14.22%), followed by *DM* and *TF*. Given the similar purchase rates, the difference in *add to cart* rates between the *TF* and *DM* sessions suggests that users tend to buy directly when they find the right product in *TF* sessions, while in *DM* sessions, they tend to add it to cart first and make decisions later.

4.2 Event Statistics

We now zoom in to examine user behaviour in terms of individual event types. This includes: click events which indicate how users interact with search results after a query is issued; and query events including query strings and query reformulations.

4.2.1 Click Event. We first measure the click-through rate (CTR) for each session type. From the search result list and the ids of the

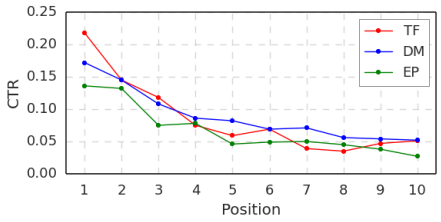


Figure 1: Click-through rate for top 10 positions

Table 5: Click statistics by user intent (* indicates one-way ANOVA with Bonferroni post hoc test, significant at p<0.01)

	TF	DM	EP
Clicks per Session (mean)	3.47*	5.51*	8.27*
Clicks per Query (mean)	1.67	2.51*	1.97
nCS (%)	43.43	29.87	36.10
nRS (%)	39.39	21.82	22.87

clicked products, we obtain the positions of each clicked product. We then calculate the average CTR as the number of clicks at each position divided by the number of queries. In Figure 1, we see the CTR for *TF* sessions is the highest at the first position, followed by a sharp decay. After the 4th position, the CTR for *DM* sessions is higher than *TF* sessions. The relatively higher position bias of the *TF* sessions suggests that users tend to find what they are looking for at top ranks in these sessions. The CTR for *EP* sessions is constantly lower than *DM* sessions suggesting that users tend to click less under the *EP* intent, perhaps because they do not have a particular target in mind.

Table 5 shows click statistics for each session type. The three types of sessions show significant difference in terms of number of clicks per session, with *EP* being the highest and *TF* the lowest, that is, from *TF* to *EP* users explore increasingly more products. Meanwhile, *DM* has a significantly higher number of clicks per query than *TF* and *EP*, suggesting users tend to spend more effort in examining the results of each query.

We also computed nCS and nRS [28], where nCS is defined as the percentage of the queries with less than n clicks (n=2) and nRS is the percentage of queries with clicks only on top n results (n=5). The higher nCS and nRS of the *TF* sessions compared to *DM* and *EP* sessions again suggests a more focused search (examining few results at the top of the list), which is consistent with what we find from the CTR analysis (Figure 1).

4.2.2 Query Strings. As shown in Table 6, *TF* sessions have significantly higher average query length (7.94) than the *DM* (6.66) and *EP* (6.46) sessions, suggesting users tend to issue more specific queries in these sessions. *TF* sessions also have the longest initial queries (7.35). From the survey we observed that users in *TF* sessions often start their search with more specific queries including the brand name and even the specific product features, suggesting that they have specific targets in their mind and therefore are able to describe it clearly. In addition, we see that for all sessions the average query lengths are higher than that of the initial query. This suggests that users tend to expand their queries.

We further investigate query reformulations by calculating the similarity between two consecutive queries, as well as that between the first and the last query, within a session. We use the Jaccard index as a measure of similarity. *EP* sessions have the lowest values

Table 6: Query string statistics by user intent (* indicates one-way ANOVA with Bonferroni post hoc test, significant at p<0.01)

	TF	DM	EP
Query Length (Chinese characters)	7.94*	6.66	6.46
Initial Query Length (Chinese characters)	7.35*	5.76	5.58
Consecutive Queries Similarity	0.42	0.43	0.29*
Head&Tail Queries Similarity	0.38	0.34	0.06*

for both statistics, suggesting that compared to *TF* and *DM* sessions, users tend to issue more diverse queries and drift between different products under this intent.

4.3 Summary

In conclusion, users tend to conduct more focused searches in *TF* sessions compared to those in the *DM* and *EP* sessions. Specifically, they use fewer but more specific queries and only browse and click a few top results. Compared to *TF* sessions, users in *DM* sessions tend to issue shorter queries, browse the result list deeper and click on more results. In *EP* sessions, users issue many more queries that are semantically diverse. They browse deep down the result lists but perform fewer clicks for each query. These findings have implications for both retrieval algorithms and interface designs. For instance, search engines may determine how the results should be displayed to users depending on their intents: e.g. in *TF* sessions accurate results may be important, while in *DM* or *EP* sessions more diverse results may be appropriate and exploratory interface elements may be useful.

5 USER BEHAVIOUR, SEARCH INTENT, AND SATISFACTION

In this section, we address RQ3(a). *How does user satisfaction relate to their search behaviour and search intents?*

5.1 Data Collection

To collect user search interaction data and their corresponding satisfaction feedback, we conducted a user study using an experimental mobile product search system. We did our best to simulate realistic product search scenarios to ensure the data we collected are credible.

Experiment procedure. The experiment consists of two steps.
Step one. Each participant was asked to complete 12 product search tasks (6 *TF* and 6 *DM* tasks). At the start of each task, the participant was shown the task description and an initial search query. After that he/she was guided to the SERP of a popular commercial shopping site. He/she can then browse the result list, click any item of his/her interests, view the details and add products to the shopping cart. We adapted the "Add to Cart" button on the product page to skip log-in. The participant can reformulate their queries if he/she was not satisfied with the current result list. Once clicked on a "Finish" button at the bottom, the participant was requested to select the most desired product he/she might eventually buy, and provide feedback of satisfaction with the search results in a 5-point scale for each query. Invalid queries caused by typing errors were discarded. He/she was then guided to step two.
Step two. After finishing the *TF* and *DM* tasks, each participant was given time to search for anything he/she liked (*EP* task). The "Finish" button appears after 10 minutes so that the participant

Table 7: A list of search tasks for TF and DM user intents

	Initial Query	Task description
TF	JanSport backpack	Your backpack is broken, and you want to buy a JanSport backpack.
	Edifier headset	You want to buy an Edifier headset to listen to music.
	3M head-wearing mask	You want to buy a 3M head-wearing mask to protect against the haze.
	Kyocera peeler	You want to buy a Kyocera peeler.
	Yonex badminton racket	You want to learn badminton, so you plan to buy a pair of Yonex badminton racket.
	Asics running shoes	You hope to lose weight by running, so you want to buy a pair of Asics running shoes.
DM	router	The router fails often recently, so you want to find a good quality router.
	rechargeable lamp	The lights out schedule will start soon, so you want to buy a long-lasting rechargeable lamp.
	tf card 64G	You want to buy a 64G TF card to expand the memory of your mobile phone.
	water cooler	You want to buy a water cooler for your dorm.
	coffee beans	Your lab ran out of coffee, so you want to find some tasty coffee beans.
	television	You want to buy a new TV for your family.

could end the session if he/she wanted to. Upon completing the task, he/she was also asked to select the products with final purchase intention and label the satisfaction scores as in step one.

Apparatus. We used a mobile phone with a 5-inch (diagonal) LCD screen and 1280-by-720-pixel resolution as the search device, similar to most modern mobile phones on the market. In the background, we ran an in-house Android application to collect user interactions and their satisfaction feedback.

Participants. We recruited 20 participants (10 female and 10 male). from our university with moderate product search engine utilization experiences. Each of them was asked to finish two example tasks to get familiar with the experiment process in a training session.

Search tasks. To cover various search intents, we selected 6 *TF* and 6 *DM* tasks from the results of the online survey (see Section 3.2.1) as shown in Table 7. We randomized the task order for each participant. The SERPs were crawled from a popular commercial shopping site with sorting and filtering functions removed as we wanted to focus on how users interact with the default result pages. We leave the investigation on more advanced SERPs for future work. The task instruction for *EP* task is: "Image yourself have ten minutes of free time to shop online before boarding an airplane/bus/train. You can search anything you like, and add products to the shopping cart or make a purchase as your like."

Data collected. We collected a total of 659 valid search queries each associated with satisfaction scores for these 13 tasks (6 *TF* tasks, 6 *DM* tasks and 1 *EP* task).

5.2 Satisfaction and Purchase

Due to the lack of user satisfaction feedback, many shopping sites use purchase as a measure of user satisfaction. However, in Table 4 we see the purchase rate is around 10% for each intent type. Users do not always make a purchase right after search: some may make the purchase later; some may buy the product offline. Hence purchase may not accurately reflect whether users are satisfied with their search experience or the performance of the search engine. It is therefore important to understand the relationship between user satisfaction and their purchase behaviour.

We consider the following two purchase scenarios: *adding to cart*, i.e. the user adds a product to the shopping cart; and *buying*, i.e. the user adds the product to cart and eventually selected to buy. We call the search sessions that contain *adding to cart*/*buying* actions ADD/BUY cases, and the others NOT ADD/BUY cases.

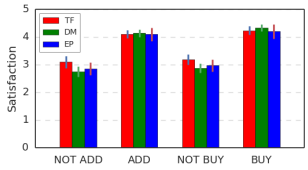
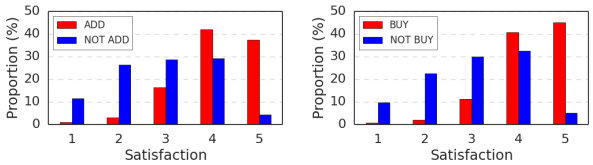


Figure 2: Average satisfaction scores of different intents.



(a) ADD and NOT ADD cases (b) BUY and NOT BUY cases

Figure 3: Distribution of the satisfaction scores over different user intents.

We first compute the average satisfaction scores of the two cases. As shown in Figure 2, ADD/BUY cases have significantly higher scores than NOT ADD/BUY cases (t-test, $p<0.01$). This implies that users are usually more satisfied if desired products are found.

Figure 3 shows the distribution of the satisfaction scores of different cases. We see that most of the ADD and BUY cases have satisfaction scores of 4 or 5, which means users are usually satisfied in these cases. However, there are 15 (4.23%) ADD cases and 8 (2.84%) BUY cases with scores less than 3, indicating users may not be satisfied even if they find the target products, e.g. it may be that the search process took too much effort [18]. For NOT ADD/BUY cases, the distribution is more uniform: 37.8% NOT ADD cases and 32.4% NOT BUY cases have satisfaction scores less than 3; and 33.6% NOT ADD cases and 37.7% NOT BUY cases have scores of 4 or 5. This incidates users may not be dissatisfied even if they did not save or buy anything during the search process.

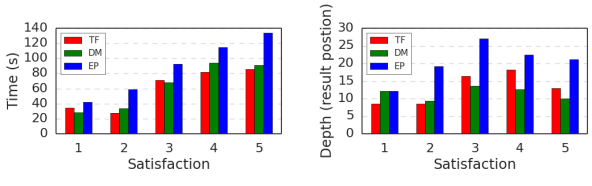
In summary, *adding to cart/buying* behaviour has a high correlation with user satisfaction, i.e. users usually feel satisfied if they find the desired products. However, satisfaction does not necessarily lead to purchase. On the other hand, users may be moderately satisfied even if they do not add products to the cart or make a purchase (although to a much less extent). Therefore, it is not appropriate to use buying activities as a measure of user satisfaction.

5.3 Satisfaction and Interaction Signals

Having observed that purchase behaviour alone is not an accurate indicator of user satisfaction in product search, we now investigate which other user interaction signals reflect the different levels of user satisfaction.

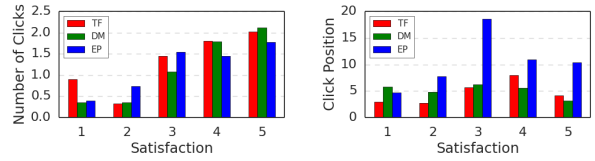
Query Level Statistics. We first look at the relation between user satisfaction and two query-level statistics: query duration (the time duration of a query session); browse depth (the maximum rank of results users browsed on a SERP).

In Figure 4(a), we see that satisfaction is positively correlated with query duration under each type of user intent (Spearman's ρ=0.270 (*TF*); 0.400 (*DM*); 0.336 (*EP*), $p<0.01$), indicating users tend to spend more time when they are satisfied with the result list; and they also tend to browse more products (Figure 4(b)). However, we observe that queries with a satisfaction score of 5 have a smaller browse depth than those with a score of 3 or 4 (t-test, $p<0.01$). It

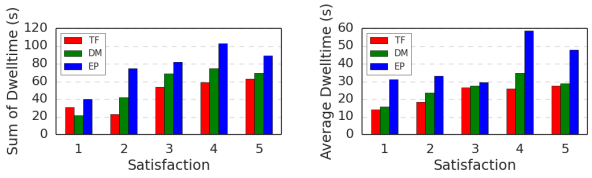


(a) Query duration (b) Browse depth

Figure 4: Query level stats w.r.t satisfaction.



(a) Number of clicks (b) Average click position



(c) Sum of dwell time (d) Average dwell time

Figure 5: Click stats w.r.t satisfaction

suggests that users may feel more satisfied if they find the needed products in the earlier search result ranking.

Further, we find that *TF* sessions have a deeper average browse depth than *DM* sessions, which is inconsistent with previous findings in Section 4.1. This may be an artefact of the task design, e.g. some users may not be familiar with the products in the *TF* tasks and needed to browse a bit more to make sure that they find the right ones. Meanwhile, users tend to browse deeper in *EP* sessions. This may because users are more interested in this part of the experiment as they can search anything they like. However, these inconsistencies are not significant, so the current experiment data is not enough to make conclusions.

Click Statistics. Previous studies have shown that click behaviour is an important signal both for result relevance and search satisfaction for Web search [13]. Our data support this finding in the context of product search. In Figure 5(a) we see that users tend to click more products when they are satisfied with the result list (Spearman's ρ=0.334 (*TF*); 0.509 (*DM*); 0.321 (*EP*), p<0.01). We also compute the average position of the clicked products. In Figure 5(b), we see this statistic is small for queries with low satisfaction scores (1 and 2) in *TF* and *EP* sessions. It suggests that users tend to reformulate the query if they are not satisfied with the first few results. Meanwhile, queries with a score of 5 have a smaller average position than those with a score of 3 or 4 (t-test, p<0.01), which is consistent with Figure 4(b). However, this difference is not significant in *DM* sessions. It is reasonable as users need to click several products to make the comparison before they are satisfied.

Dwell time is also a strong signal for satisfaction [13, 18] in Web search. Therefore we also compute the dwell time for the clicked products. In Figure 5(c), we see users tend to spend a longer time on product pages when they feel satisfied. However, it may be due to the fact that they click more in these cases. Therefore, we also calculate the average dwell time for each query. As we see in Figure 5(d), lower satisfaction (1 and 2) is associated with shorter average dwell time in *TF* and *DM* sessions. In *EP* sessions, this statistic is approximately the same for queries with a satisfaction score less than 4 and is significantly higher for queries with a score of 4 and 5 (t-test, p<0.01), suggesting users examine carefully if they are interested in the product under this intent.

Summary. To summarise, user interaction patterns are closely related to their satisfaction, while their detailed relation may vary under different intents. Next, we explore features derived from the findings of this analysis to predict user satisfaction.

Table 8: Features for satisfaction prediction.

	Feature Description
Q1	Query duration (in second)
Q2	Browse depth
C1	Number of clicks
C2-C4	The average, max and min position of clicks
C5-C6	The sum and average of dwell time on landing pages
C7-C12	Number of clicks with dwell time of 0-5/5-10/10-15/15-30/30-60/>60 seconds

6 SATISFACTION PREDICTION

We now address RQ3(b). *Can we predict product search satisfaction with interaction signals?*. In Figure 3, we see the average satisfaction score is 2.88/3.01 for NOT ADD/BUY cases and 4.12/4.27 for ADD/BUY cases. As discussed in Section 5.2, users usually feel satisfied if they find the desired products, therefore we bin the satisfaction scores, setting scores 4 and 5 as SAT and the remainder as DSAT for satisfaction prediction. This results in 384 SAT cases and 275 DSAT cases.

6.1 Feature Extraction

Based on the findings in Section 5.3 and previous studies [13, 18], we extract two types of features to predict search satisfaction in different intent scenarios: query level features (Q); and click features (C). Table 8 lists all the features. These features can all be easily collected in search logs for online shopping site.

Query Level Features. We define two query level features: query duration (Q1) and browse depth (Q2)—both are highly correlated with the user satisfaction scores (see Figure 4).

Click Features. Based on the findings in user click behaviour, we extract the following click-based features: the number of clicks that a user performed on a result list (C1); the average, maximum, and minimum position of user clicks on the result list (C2-C4). We also compute the sum and average of the dwell time on each product page (C5-C6). We further divide C1 by different dwell time intervals (C7-C12), as different users may have different product search habits. Some may read many details such as specifications, comments, and after-sales service, while others may only care about the price. Therefore we do not divide the clicks into SAT clicks and DSAT clicks with a simple threshold on dwell time [13, 18].

6.2 Prediction Results

We investigate the results of our proposed prediction models in three aspects: the performance of the classifiers; the impact of different feature groups as well as individual features; and the model performance across different user intents.

Following Cheng et al. [5], we use F1 score and AUC (area under the ROC curve) to evaluate the prediction performance. Since the relation between user satisfaction and interaction patterns varies

Table 9: Comparison of classifiers. * indicates significant difference by t-test, p<0.05.

	TF		DM		EP	
	F1	AUC	F1	AUC	F1	AUC
SVM	**0.773**	0.500	0.738	0.508	0.516	0.500
DT	0.675	0.577	0.669	0.637	0.484	0.521
NB	0.394	0.548	0.697	0.703	0.418	0.537
GBDT	0.764	**0.631***	**0.802***	**0.759***	**0.593***	**0.569***

Table 10: Comparison of different feature groups, * indicates significant difference by t-test, p<0.05

	TF		DM		EP	
	F1	AUC	F1	AUC	F1	AUC
Q	0.704	0.544	0.672	0.608	0.590	0.594
C	0.704	0.648	0.781	0.756	0.539	0.568
Q+C	**0.758***	**0.651***	**0.808***	**0.760***	**0.630***	**0.610***

Table 11: Feature importance for different intents

TF		DM		EP	
Feature	Weight	Feature	Weight	Feature	Weight
Q1	0.352	Q1	0.307	Q1	0.354
Q2	0.218	C6	0.211	Q2	0.239
C6	0.136	Q2	0.179	C6	0.110
C5	0.122	C5	0.104	C5	0.072

across different intent types, we perform the predictions for *TF*, *DM* and *EP* tasks separately.

Comparison of Classifiers. We apply Support Vector Machine with RBF kernel (SVM), Decision Tree (DT), Naive Bayes (NB) and Gradient Boosting Decision Tree (GBDT) to predict user satisfaction with 5-fold cross validation. GBDT achieves the best performance, followed by DT, NB, and SVM (Table 9). Therefore, in the rest of the investigation we continue with GBDT.

We experimented with two sampling strategies for cross-validation: (1) "random sampling" (Table 9); and (2) "sampling by user": query sessions from the same user are grouped in either the training or the test set [27]. We found that the two strategies result in very similar prediction performance, meaning the prediction model can deal with previously unseen users. Therefore, we adopt the "sampling by user" strategy to compare different feature groups next.

Comparison of Feature Groups. We first make predictions with each type of features individually (Table 10). Click features achieve better performance than query features, suggesting click behaviour provides more information about user satisfaction. After combining two type of features, we get the best performance for all tasks.

Feature Importance. We now investigate the impact of individual features on predicting user satisfaction with a more detailed feature importance analysis. Table 11 shows the top 4 features using GBDT and the Q+C feature group. We find that *query duration (Q1)* has the highest importance for all user intents. The second important feature is *average dwell time (C6)* for *DM* tasks, but *browse depth (Q2)* for *TF* and *EP* tasks. It is reasonable as users tend to focus more on a product page to compare and evaluate, which is captured by *C6. Sum of dwell time (C5)* is also a useful feature across user intents. Meanwhile, click position and click count features have relatively small weights in all cases, suggesting users may not be sensitive to the positions and the number of clicked results.

Comparison of Intent Groups. From both Table 9 and 10 we see that the prediction performance for *TF* and *DM* tasks is much better than *EP* tasks. We believe this is because people do not have specific purchase targets in *EP* tasks. Their perception of satisfaction may be influenced by many factors, such as the serendipity of the results, and the attractiveness of the contents; and it may vary widely among different users. We leave the research for future work on pursuing more effective satisfaction prediction for the more challenging *EP* search intent.

Summary. We conclude that, by exploiting interaction signals, we can predict product search user satisfaction reasonably well, especially for *TF* and *DM* search intents. In addition, since we collected data with mobile devices, we also experimented with gesture features (e.g. the number and direction of swipes). However, these features has limited impact. We leave the mobile specific investigation for future work.

7 DISCUSSION AND CONCLUSIONS

In this paper we investigated the relation between user intents, search behaviour, and perceived satisfaction in the context of product search. Our findings have several implications for future studies.

To characterise different types of user intents, we proposed and verified a taxonomy that identifies three types of product search intents: *Target Finding* (TF), *Decision Making* (DM) and *Exploration* (EP). This taxonomy bridges a gap between existing taxonomies that describe general Web search and online shopping activities. We found users with different intents behave differently. *TF* users tend to issue few specific queries and browse only top ranked results; *DM* users tend to issue short queries, browse deep, and click more results; and *EP* users issue many diverse queries, browse deep, but do not click often. These findings will help search engines make choices of retrieval algorithms and interface designs depending on the user intents.

Using interaction data and explicit satisfaction feedback collected from a user study, we found that user interaction patterns are closely related to satisfaction, but their detailed relation varies across different intents. Based on these findings, we demonstrate that we are able to predict user satisfaction using interaction signals with reasonable prediction performance, especially for *TF* and *DM* tasks. That is, to effectively predict user satisfaction, one need to consider not only the right interaction features but also the underlying user intent with respect to search and purchase.

One limitation of our study is that the notion of "satisfaction" is subject to users' own interpretation. While it allows users to describe an overall experience, it may also lead to variance in satisfaction due to different interpretations. We leave the investigation of more find-grained types of satisfaction for future work. As mobile phones are becoming the main tools for online shopping, we conduct log analysis and user study on mobile devices. However, user satisfaction may vary from PC to mobile. We leave the analysis on PC, and the comparison between PC and mobile product search for the future investigation.

8 ACKNOWLEDGEMENTS

We thank Dr. Ke Zhou for providing very useful suggestions for this paper. This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61532011, 61672311), National Key Basic Research Program (2015CB358700), and the Netherlands Organisation for Scientific Research (NWO) under project nr. 13675.

REFERENCES

[1] Mikhail Ageev, Dmitry Lagun, and Eugene Agichtein. 2013. Improving search result summaries by using searcher behavior data. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 13–22.

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

[3] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. D. Raths, and M. C. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. 1013–1014 pages.

[4] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.

[5] Justin Cheng, Caroline Lo, and Jure Leskovec. 2017. Predicting Intent Using Activity Logs: How Goal Specificity and Temporal Range Affect User Behavior. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 593–601.

[6] Susan Dumais. 2013. Task-based search: a search engine perspective. In *NSF Workshop on Task-Based Search*. 1.

[7] James F Engel, David T Kollat, and Roger D Blackwell. 1973. Consumer behavior. (1973).

[8] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[9] Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 569–578.

[10] Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2012. Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2050–2054.

[11] Qi Guo, Dmitry Lagun, Denis Savenkov, and Qiaoling Liu. 2012. Improving relevance prediction by addressing biases and sparsity in web search click data. In *Proc. of Int. Conf. on Web Service and Data Mining workshop on Web Search Click Data. New York: ACM*. 71–75.

[12] Qi Guo and Yang Song. 2016. Large-Scale Analysis of Viewing Behavior: Towards Measuring Satisfaction with Mobile Proactive Systems. (2016).

[13] Qi Guo, Ryen W White, Susan T Dumais, Jue Wang, and Blake Anderson. 2010. Predicting query performance using query, result, and user interaction features. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 198–201.

[14] Jeff Huang and Abdigani Diriye. 2012. Web user interaction mining from touch-enabled mobile devices. In *HCIR workshop*. Citeseer.

[15] Jeff Huang, Ryen W White, Georg Buscher, and Kuansan Wang. 2012. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 195–204.

[16] Jeff Huang, Ryen W White, and Susan Dumais. 2011. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1225–1234.

[17] Bernard J. Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management* 45, 6 (2009), 643–663.

[18] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 57–66.

[19] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 607–616.

[20] Maryam Kamvar and Shumeet Baluja. 2006. A large scale study of wireless search behavior: Google mobile search. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 701–709.

[21] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1âĂŤ2 (2009), 1–224.

[22] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 121–130.

[23] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 113–122.

[24] Lee, Uichin, Liu, Zhenyu, and Junghoo. 2005. Automatic identification of user goals in Web search. (2005).

[25] Beibei Li, Anindya Ghose, and Panagiotis G Ipeirotis. 2011. Towards a theory model for product search. In *Proceedings of the 20th international conference on World wide web*. ACM, 327–336.

[26] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 43–50.

[27] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 493–502.

[28] Yiqun Liu, Min Zhang, Liyun Ru, and Shaoping Ma. 2006. Automatic query type identification based on click through information. In *Asia Information Retrieval Symposium*. Springer, 593–600.

[29] Bo Long, Jiang Bian, Anlei Dong, and Yi Chang. 2012. Enhancing product search by best-selling prediction in e-commerce. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2479–2482.

[30] Eric Hsueh-Chan Lu, Wang-Chien Lee, and Vincent Shin-Mu Tseng. 2012. A framework for personal mobile commerce pattern mining and prediction. *IEEE transactions on Knowledge and Data engineering* 24, 5 (2012), 769–782.

[31] Wendy W Moe. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology* 13, 1 (2003), 29–39.

[32] Francesco M Nicosia. 1966. CONSUMER DECISION PROCESSES; MARKETING AND ADVERTISING IMPLICATIONS. (1966).

[33] Jaimie Y Park, Neil O'Hare, Rossano Schifanella, Alejandro Jaimes, and Chin-Wan Chung. 2015. A large-scale study of user image search behavior on the web. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 985–994.

[34] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. ACM, 13–19.

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

[36] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 495–505.