Understanding and Predicting Graded Search Satisfaction

Jiepu Jiang[1], Ahmed Hassan Awadallah[2], Xiaolin Shi[2], and Ryen W. White[2]
[1] School of Computer Science, University of Massachusetts Amherst
[2] Microsoft Research Redmond
jpjiang@cs.umass.edu, {hassanam, xishi, ryenw}@microsoft.com

ABSTRACT
Understanding and estimating satisfaction with search engines is an important aspect of evaluating retrieval performance. Research to date has modeled and predicted search satisfaction on a binary scale, i.e., the searchers are either satisfied or dissatisfied with their search outcome. However, users' search experience is a complex construct and there are different degrees of satisfaction. As such, binary classification of satisfaction may be limiting. To the best of our knowledge, we are the first to study the problem of understanding and predicting graded (multi-level) search satisfaction. We examine sessions mined from search engine logs, where searcher satisfaction was also assessed on multi-point scale by human annotators. Leveraging these search log data, we observe rich and non-monotonous changes in search behavior in sessions with different degrees of satisfaction. The findings suggest that we should predict finer-grained satisfaction levels. To address this issue, we model search satisfaction using features indicating search outcome, search effort, and changes in both outcome and effort during a session. We show that our approach can predict subtle changes in search satisfaction more accurately than state-of-the-art methods, affording greater insight into search satisfaction. The strong performance of our models has implications for search providers seeking to accurately measure satisfaction with their services.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *performance evaluation (efficiency and effectiveness)*. H.1.2 [**Models and Principles**]: User/Machine Systems – *human factors.*

Keywords
Search satisfaction; evaluation; utility; effort; session.

1. INTRODUCTION
The evaluation of search systems can be performed on the basis of result relevance, or alternative measures of the search experience (e.g., satisfaction, usability). When the focus is on relevance, *metrics* such as average precision or nDCG [17] are often computed using human relevance judgments from third-party assessors. Once relevance judgments are gathered, they are reusable and metrics can be computed across systems, allowing direct comparisons of system performance. However, studies indicate there are imperfect correlations between these evaluation metrics and searchers' actual ratings of their search experience (usually considered as the gold standard) [2, 15]. Methods to elicit searcher feedback on the search experience have been proposed in [9, 10]. However, this feedback is difficult to collect (e.g., it can require costly laboratory studies).

An emerging approach for evaluating Web search engines is to predict the quality of the search experience from search interaction data [1, 8, 10, 11, 14]. Search interactions can be studied at scale using retrospective analysis of existing search engine log data or intentional manipulations of the search experience (e.g., in controlled experiments [25]). For example, Hassan et al. used search logs to predict search success [13], searcher satisfaction [11], and whether searchers are struggling or exploring during search episodes [14]. Feild et al. [8] focused on searcher frustration, Ageev et al. [1] on search success, Guo et al. [10] on search engine switching, and Liu et al. [26, 27] and Arguello et al. [3] on search task difficulty. The evaluation approach epitomized by these studies combines aspects of evaluation metrics and asking searchers directly for feedback. It relies on signals from searchers' behavior in natural settings, but requires neither relevance judgments nor explicit feedback. These advantages make it a useful evaluation technique for search providers, who have abundant search logs comprising query and click behavior on which to perform such analysis.

Despite the strong progress, we believe current methods and accompanying studies have the following two important limitations:

First, existing research did not advance our understanding of the searchers' experience during searching. They are mostly black box techniques that receive search interaction data as input and return predictions of outcomes such as satisfaction, frustration, and success. Despite the availability of these predictions, we still do not understand why searchers are satisfied, frustrated, or successful.

Second, in existing studies, the quality of searchers' experience in using the system is estimated on a binary scale. For example, searcher satisfaction is predicted as either satisfied (SAT) or dissatisfied (DSAT) in previous studies [11, 13, 35]. In these studies, datasets of randomly sampled search sessions extracted from search engine query logs had only about 20% of the sessions as DSAT. Important unanswered questions include: Are the 80% SAT search sessions equivalent? What are the differences among SAT sessions? Can we predict subtle differences in satisfaction from features such as observable behavior?

To the best of our knowledge, this paper makes the first attempt to address these two shortcomings in the specific case of satisfaction prediction. In doing so, we make the following contributions:

First, we perform an in-depth analysis characterizing differences in searcher interaction per graded levels of search satisfaction. We show the value of search outcome compared with search effort best explains and most strongly correlates with search satisfaction. In addition, there are rich and non-monotonous differences in search behavior in sessions with different satisfaction levels. Our study extends the current understanding by disclosing the effects of search outcome, effort, and their dynamics over the course of a session on levels of search satisfaction.

Second, we identify effective sets of features related to search outcome, effort, and their changes. Each group of features makes a unique contribution in satisfaction prediction. We divide the sessions into four satisfaction levels. For example, effort features are

WSDM'15, February 2–6, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2684822.2685319

critical in distinguishing between sessions with very high and high satisfaction levels, while features associated with changes in effort and outcomes are most effective in discriminating dissatisfied and moderately satisfied sessions.

Third, our prediction methods are effective and significantly outperform prior methods. Our regression model predicts continuous satisfaction values with moderate correlation to annotators' satisfaction ratings. Additionally, our approach outperforms prior methods in classifying sessions between adjacent satisfaction levels.

The remainder of the article describes our methods and findings.

2. RELATED WORK

In this section, we review related work in the following two areas. First, we summarize previous work on the measurement of satisfaction, both within search settings and beyond. Second, we review approaches to predicting the quality of the search experience and the relevance of search results based on search interaction data.

2.1 Satisfaction

Satisfaction is widely adopted as a subjective measure of search experience. Kelly ([23], p. 117) reviewed a definition: "satisfaction can be understood as the fulfillment of a specified desire or goal." When evaluating search systems, satisfaction can be determined regarding not only the holistic search experience but also some specific aspects [34], e.g., result completeness, result relevance, system response time, caption formatting, and general appearance of the user interface. Although the concept is widely used in retrieval settings, we did not find any literature in information retrieval defining and explaining searcher satisfaction in a principled way.

In contrast, user satisfaction with generic information systems has been studied extensively in the context of management information systems. These studies focused on developing valid instruments (schemes of designing questionnaires) for measuring satisfaction. They usually adopted a factorized approach. For example, Bailey et al. [6] and Ives et al. [16] developed and validated an instrument involving 39 factors. McKinney et al. [31] developed an instrument for customer satisfaction in eCommerce with 34 factors related to information and system quality. These studies provide exhaustive enumerations of the factors that contribute to user satisfaction, but still do not explain satisfaction in a principled way.

We further seek the explanation of satisfaction outside the scope of information and computer science. In economics, there is a close relationship between satisfaction and utility. For example, Mankiw ([29], p. 285) introduced utility as "a person's subjective measure of well-being or satisfaction." Marshall ([30], p. 94) equates both the utility of products and consumer's satisfaction with product purchases to the price that the consumer is willing to pay for the product. This motivates us to explain satisfaction in terms of utility. For example, Su [34] defined utility as a measure of worth of search results versus time, physical effort, and mental effort expended. This suggests that we can consider searcher satisfaction as a compound measure of multiple factors, including search outcome and search effort. In a recent study, Yilmaz et al. [37] also confirmed the impact of searcher effort on document relevance and utility.

Another line of previous work that motivates our explanation of search satisfaction is the study of consumer psychology. Oliver et al. [32] verified that post-purchase consumer satisfaction is a function of pre-purchase consumer expectation and the disconfirmation after purchase (the discrepancy between perceived and expected utility of the product). The disconfirmation factor suggests that we can explain searcher satisfaction by comparing actual search outcomes to expected search outcome. As we will demonstrate in Section 4, with certain assumptions, this explanation is equivalent to the comparison of search outcomes and search effort.

2.2 Predicting Satisfaction and Others

Large-scale search engine logs contain massive amounts of search interaction data that can be useful as implicit searcher feedback. Such implicit feedback data can be employed to predict the quality of the search experience (e.g., searcher satisfaction) as well as other useful information such as result relevance and searcher preferences. In this subsection, we review related studies and approaches.

Hassan et al. [11, 13], Ageev et al. [1], and Wang et al. [35] modeled search success and satisfaction based on the structure of searchers' action sequences. They make the assumption that in successful and unsuccessful, or satisfied and dissatisfied sessions, searchers may perform different types of actions. Hassan et al. [11, 13] represented action sequences using Markov models. Ageev et al. [1] and Wang et al. [35] enhanced these models using conditional random fields and structured learning.

Feature-oriented approaches were also widely used in modeling aspects of the search experience. Researchers focused more on the types of features and adopted general classifiers in prediction. For example, Feild et al. [8] combined search interaction features as well as information from other sensors to predict frustration. Hassan et al. [12, 14] incorporated query reformulation features to further improve satisfaction prediction and identify struggling sessions. Guo et al. [10] predicted search engine switching, and Liu et al. [26, 27] and Arguello et al. [3] predicted search task difficulty.

Our work is different from previous research in that we move beyond binary satisfaction to predict graded search satisfaction. In addition, we model searcher satisfaction based on a combination of search outcome and searcher effort, which is well-grounded in theory, including previous work on consumer psychology.

3. DATA

In order to study graded search satisfaction, we sampled and annotated search tasks from the search logs of consenting users of the Microsoft Bing search engine in April, 2014. These logs were collected through a widely-distributed browser toolbar. These log entries included a unique identifier for each searcher, a timestamp for each page view, a unique browser window identifier, and the URL of the Web page visited. Intranet and secure (https) URL visits were excluded at the source. In order to reduce variability in search behavior caused by geographic and linguistic differences, we only include entries generated in the English speaking United States locale. From these logs we extracted raw search sessions. Every raw session began with a search query and could contain further queries and visits to clicked search results. A session ended if the searcher was idle for more than 30 minutes. These raw sessions were later segmented into search tasks (where a task resolves to a single information need) using Jones et al.'s methodology [21]. Jones et al. [21] showed that many raw search sessions consist of multiple tasks and may result in one or more queries. We use the term *search session* to refer to queries and other search activity belonging to the same search task throughout the paper. To remove navigational and simple informational tasks, we excluded single-query sessions.

The search sessions in our dataset were annotated with session search satisfaction and query result quality by crowd workers. For each session, we showed the annotators: 1) queries in the search session and search engine result page (SERP) dwell time; 2) links to search result pages for these queries; 3) the URLs of any clicked results and associated dwell times. We enforced that annotators visit the SERPs for all queries and visit the clicked result webpages before they can submit their judgments. Since they were third-party annotators, requiring these steps was meant to help annotators maximally restore the original searchers' experience. We paid annotators 10 cents for each of the sessions that they annotated.

Employing third-party crowd workers to annotate logged search sessions is inexpensive and also feasible for search engines, who are unable to contact searchers directly given privacy considerations. One limitation is that this method is less reliable than searchers' own ratings, and thus may introduce some biases. However, in spite of the inaccuracy introduced by the third-party judgment, many previous studies [11, 13, 35] have reached solid conclusions by adopting similar data collection methods. Moreover, collecting searchers' own ratings in a laboratory study could introduce another type of bias. For example, with user studies (e.g., [1]) it is difficult to simulate real search scenarios.

After assessor quality verification, we retained judgments for 476 search sessions. Each session includes three annotators' ratings on 1) search satisfaction of the whole session, and 2) result quality of each individual query in the session. The annotators rated session satisfaction using a 5-point Likert scale ranging from 1 (*very unsatisfied*) to 5 (*very satisfied*). They rated query result quality at three scales: *very useful* (2), *somewhat useful* (1), and *not useful* (0). We use the average score of the three annotators as the gold standard in our study. As will be analyzed in Section 6, individual annotators' ratings strongly correlate with average rating (r = 0.68). Krippendorf's alpha coefficient was 0.62, signifying both reasonable agreement between annotators and the subjectivity of the annotation task.

We assign sessions into four satisfaction levels as follows: *very high* ($4.33 < s \le 5$, 11.1% of the sessions); *high* ($4.0 \le s \le 4.33$, 39.7%); *medium* ($3.33 \le s \le 3.67$, 33.4%); *low* ($s \le 3$, 15.8%). The grouping strategy ensures that there are sufficient sessions at each of the four satisfaction levels to perform statistical analysis. Despite this grouping, most instances are grouped into *high* and *medium* levels; only 27% are in the two extreme groups (*low* and *very high*).

4. GRADED SEARCH SATISFACTION

As reviewed in Section 2.1, we first consider satisfaction as the searcher's subjective measure on the utility of search systems and results. Since utility is usually considered as the value of search results compared to the effort spent on searching, we also examine satisfaction on these two aspects. Considering the notion of utility and satisfaction are not consistently defined, we define our own notion of the constructs to avoid confusion.

Satisfaction. Satisfaction is the searcher's subjective preference on the utility of search systems and their search results. The higher the system utility is for the searcher, the more satisfied the searcher would be. In our analysis and experiments, satisfaction is measured as the average rating of three annotators on session satisfaction.

Search outcome (gain). Search outcome is the value of search results to the searcher, i.e., the quantity of useful information found.

Search effort (cost). Search effort is the cost of acquiring information from the search engine, e.g., issuing queries, examining result snippets on SERPs, reading clicked results, etc.

In the remainder of this section, we present findings showing that: 1) Annotators' satisfaction ratings are strongly correlated with the search outcome compared with searcher effort. This verifies our assumption on the notion of searcher satisfaction. 2) Search outcome and effort have complex and non-monotonous changes within sessions with different satisfaction levels, which suggests novel ways of modeling searcher satisfaction.

4.1 Satisfaction, Search Outcome, and Effort

To verify our assumption that satisfaction is the value of search outcome compared with search effort, we begin by examining the relationship between satisfaction and search outcome and effort. We directly measure satisfaction using annotators' average ratings. We use the following methods to approximate outcome and effort.

We use session cumulated gain (sCG) [18] as a measure of the search outcome in a session, which is the sum of each query's gain as in Equation (1). We do not discount later queries' search gains (i.e., sDCG) as suggested by Järvelin et al. [18], because their reason for discounting is to normalize the value of results by effort. We hope to measure the value of results without considering effort.

In Equation (1), a query q's gain is originally calculated by summing the gains across q's relevant results. Due to the lack of document relevance judgments, we use annotators' query quality ratings as to represent the query's gain. That is, here sCG is calculated as the sum of all queries' quality ratings in the session.

$$outcome = sCG = \sum_{i=1}^{n} gain(q_i) \quad (1)$$

For search effort, we follow the economic model of search interactions [5] and measure search effort in a session by Equation (2), which is proportional to the number of queries issued in a session. In Equation (2), Q is the number of queries. The total search effort is the sum of four components. $c_q \cdot Q$ is the effort of querying; $c_v \cdot V \cdot Q$ is the effort of viewing pages; $c_s \cdot S \cdot Q$ is the effort of inspecting snippets; $c_a \cdot A \cdot Q$ is the effort of assessing results. All of the four parts are proportional to the number of queries. In this research we simply adopt the number of queries as a rough measure of effort.

$$effort = c_q \cdot Q + c_v \cdot V \cdot Q + c_s \cdot S \cdot Q + c_a \cdot A \cdot Q \propto Q \quad (2)$$

Table 1. Correlation of several measures with satisfaction.

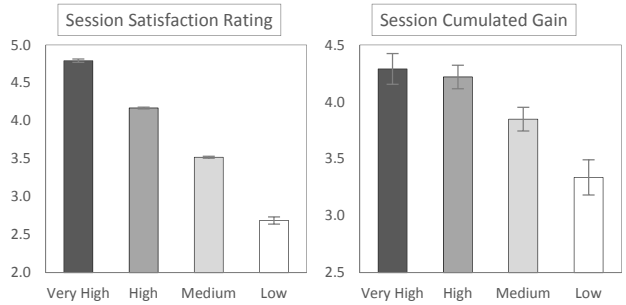
Measure	Correlation w/ Satisfaction	
	Pearson	Kendall
Search Outcome (sCG)	0.27	0.22
Search Effort (# queries)	-0.24	-0.23
Search Outcome / Effort (sCG / #queries)	**0.77**	**0.59**
sDCG (Järvelin et al [18])	0.41	0.29
nsCG	**0.77**	**0.59**
nsDCG (Kanoulas et al. [22])	0.75	0.57

All the correlation values are statistically significant ($p < 0.01$).

Table 1 shows the correlations between satisfaction and search outcome (sCG), effort (# queries), and average search outcome per effort (sCG / # queries). Results show that satisfaction has a weak positive correlation with search outcome (Pearson's r = 0.27) and a weak negative correlation with search effort (r = -0.24). In comparison, we observed a strong correlation (r = 0.77) between satisfaction and average search outcome per effort (sCG / # queries). Such a strong correlation verifies our assumption that it is appropriate to explain satisfaction as a relative measure of outcome versus effort.

Our findings also shed light on the validity of search session evaluation metrics. Järvelin et al.'s sDCG [18] only has a moderate correlation with satisfaction (r = 0.41), while its normalized version (nsDCG) has a strong correlation (r = 0.75), which is adopted as the evaluation metric in the TREC session track 2010. The version of nsDCG that we adopted is the same as that used in Kanoulas et al.'s studies [22] (bq = 4). We found that the normalized version of sCG without discounting (nsCG), as in Equation (3), has the strongest correlation with satisfaction (r = 0.77). When calculating sDCG, nsDCG, and nsCG, we use the sCG defined in Equation (1).

We found that under certain assumptions, nsCG is equivalent to our explanation of satisfaction (sCG / # queries). As in Equation (3), nsCG is the ratio between the actual search outcome (sCG) and the ideal search outcome (sCG_{ideal}). We assume that the ideal search outcome can be achieved via the maximum gain in each query, i.e., $gain_{max}(q)$. If we further assume the maximum gain on each query is a constant, i.e., $gain_{max}(q)$ is the same for different q, nsCG is proportional to sCG/Q, which is exactly our explanation of satisfaction. This explains why nsCG is also strongly correlated with searcher satisfaction ratings in our dataset.



(a) (b)

Figure 1. Average human rating and session cumulated gain (sCG) in sessions of different satisfaction levels.

Note that with a different assumption, Equation (3) also offers the second explanation of satisfaction we reviewed in Section 2.1. Although originally introduced as a normalization factor for *sCG*, it seems reasonable to equate sCG_{ideal} to the searcher's expected outcome of the session, i.e., assuming they expect to obtain maximum search gain for each issued search query. Under such a premise, *nsCG* and Equation (3) is exactly a measure for the discrepancy between actual search outcome and expected search outcome.

$$nsCG = \frac{sCG}{sCG_{ideal}} = \frac{\sum_q gain(q)}{\sum_q gain_{\max}(q)} = \frac{\sum_q gain(q)}{Q \cdot gain_{\max}(q)} \propto \frac{sCG}{Q} \qquad (3)$$

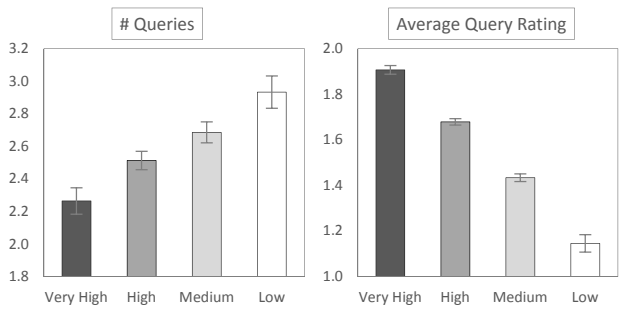
To sum up, findings in this section show that it is appropriate to explain satisfaction as search outcome compared with search effort, or equivalently, the difference between the actual and the expected search outcome. Here we focus on the first explanation given the lack of searcher expectation annotation. In the remainder of the section, we will analyze the differences in search outcome, effort, and related behavior in sessions with different satisfaction levels.

4.2 Tradeoff between Outcome and Effort

Figure 1 shows annotators' average satisfaction ratings and sCG (measuring search outcome) in sessions of four satisfaction levels. The error bars show standard error of each measure. Figure 1(a) shows that less satisfied sessions always have significantly lower ratings from the assessors ($p < 0.01$). However, Figure 1(b) shows that less satisfied sessions do not always have less search outcome. The sCG of sessions with *very high* and *high* satisfaction are 4.29 and 4.22 respectively (no significant difference at $p < 0.05$). But there are significantly less search outcomes in sessions with *medium* and *low* satisfaction (3.85 and 3.34, $p < 0.01$). This shows that low search outcome is indicative of low satisfaction, but high search outcome does not always mean complete satisfaction.

Figure 2 further explains how search outcome is accumulated in a session by showing the number of queries and the average gain of the queries in the session (measured by the assessors' average query quality rating). Figure 2(b) shows that in less satisfied sessions, searchers issue less effective queries (with significantly less gain per query, $p < 0.01$). Here we measure a query's gain using the average of annotators' query quality ratings (ranges from 0 to 2). In order to accomplish the search goal, searchers tend to issue more queries to compensate for the limited useful information gained in a session. As shown in Figure 2(a), searchers issue significantly more search queries in less satisfying sessions ($p < 0.01$). Table 1 also shows that the number of queries has a significant negative correlation with satisfaction.

Our results show adaptive searchers behaviors similar to those reported in a study by Smith and Kantor [33]. They assigned participants to search on systems with different performance and found



(a) (b)

Figure 2. Number of issued queries and average query rating in sessions of different satisfaction levels.

that searchers will adapt their search strategy to ineffective search systems, by issuing more queries. Our results differ in that all searchers used the same system, but similar adaptive behavior is observed when query performance of the system is different. This suggests that it is common for searchers to issue more queries to compensate for the limited search effectiveness, regardless of the cause, e.g., ineffective systems, novice searchers, or tasks.

However, we note that searchers do not always expend additional effort to the extent of achieving the same amount of search outcome as they can accomplish when queries are effective. As shown in Figure 2(b), searchers issue more queries in *high* satisfaction sessions and achieve comparable search outcomes as others do in the *very high* satisfaction sessions. But in the *medium* and *low* satisfaction sessions, searchers do not expend sufficient extra effort (issue a sufficient number of additional queries) to accomplish the similar amount of search outcome as they could in *very high* and *high* satisfaction sessions.

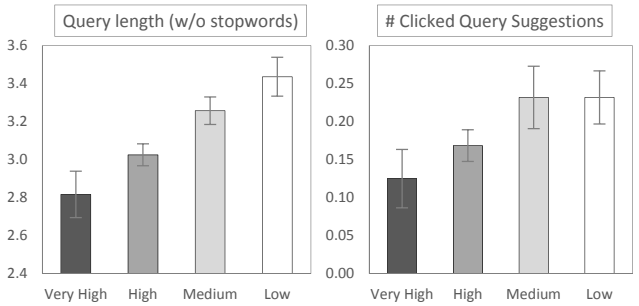
Findings reported in this section suggest we can distinguish sessions with different satisfaction levels from search behavior related to search outcome, query effectiveness, and effort (e.g., the number of search queries issued). In less satisfying sessions, searchers issue fewer effective queries and reformulate more queries to compensate for the loss. In our dataset, the total search outcome does not differ greatly in sessions with *very high* and *high* satisfaction, but is significantly lower in sessions with *medium* and *low* satisfaction.

4.3 Different Types of Search Effort

In the previous section we used the number of queries as an estimate of search effort. In this section, we further examine different types of effort in sessions with different satisfaction levels. We adhere to the effort function in the economic model of search interaction [5] (Equation 2) and consider three types of effort: 1) issuing queries, 2) examining result snippets, and 3) assessing clicked results. We ignore the cost of reading multiple SERPs for a query ($c_V \cdot V \cdot Q$) because SERP pagination is rare in Web search scenarios.

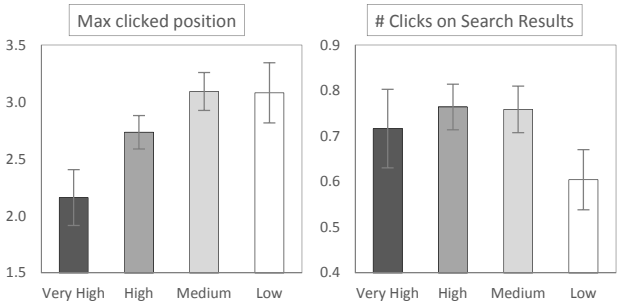
We first examine the cost of issuing queries by considering query length. We assume that longer queries incur higher cost because they require more effort to type into the query box. We also examine the frequencies of using query suggestions, which is less expensive than formulating a query statement directly [4].

Figure 3 shows that in less satisfied sessions, searchers have higher cost of issuing queries and are more likely to switch to alternative query mechanisms such as query suggestions. As shown in Figure 3(a), in sessions with *very high* and *high* satisfaction, searchers issue significantly shorter queries than others do in sessions with *medium* and *low* satisfaction ($p < 0.05$). The trends are the same regardless of whether we remove stop words. Figure 3(b) shows that searchers adopt significantly more query suggestions in sessions with *medium* and *low* satisfaction than others do in ses-



(a) (b)

Figure 3. The cost of formulating queries (in terms of query length) and the frequencies of using query suggestions.



(a) (b)

Figure 4. The number and the average rank of the clicked results in session of different satisfaction levels.

sions with *very high* and *high* satisfaction ($p < 0.05$). In both figures, the differences are not significant between the *very high* and *high* satisfaction sessions, and between the *medium* and *low* satisfaction sessions. Note that due to some data limitations, we lacked sufficient information to normalize the frequencies of using query suggestions by whether the SERP showed query suggestions to the searcher or not. Therefore, the chances of using query suggestions in Figure 3(b) is underestimated.

We further measure the effort involved in examining result snippets using the ranks of the clicked results, i.e., deeper examination of the result lists equates to higher cost. Many user studies [19, 20, 28] have suggested that in general searchers examine search results from top to bottom, with a decreasing likelihood of clicking on results as a function of rank position, and terminate their review of the search results at a certain position. Therefore, the depth of examining result snippets is equal to or deeper than the ranks of the clicked result snippets, because searchers are likely to have examined these result snippets before clicking on a particular result. The effort of assessing results is measured in terms of the average number of clicks per query. In addition to the click depth, there are also costs associated with reading the text on the landing pages. For simplicity, we ignore these reading effects in our study.

Figure 4 illustrates the effort of examining snippets and assessing results in sessions with different satisfaction levels. Figure 4(a) shows that the maximum clicked rank (as the lower bound of the position of examined snippets) in sessions with *very high* satisfaction is significantly fewer than those in other three sessions ($p < 0.01$). However, we did not observe significant difference among sessions with *high*, *medium*, and *low* satisfaction levels. Figure 4(b) shows that searchers clicked significantly fewer results in sessions with *low* satisfaction compared with the other outcomes ($p < 0.05$), but there is no significant difference between sessions with *very high*, *high*, and *medium* satisfaction.

Table 2. Correlation of effort measures with satisfaction.

Search Cost	Correlation with Satisfaction	
	Pearson's r	Kendall's tau-b
Query length (w/ stop words)	-0.10 **	-0.10 **
Query length (w/o stop words)	-0.14 **	-0.11 **
Maximum click position	-0.14 **	-0.12 **
Average click position	-0.16 **	-0.12 **
Minimum click position	-0.16 **	-0.10 *
# clicks / query	0.05	0.04
# total clicks in a session	-0.02	0.00
# queries	-0.24 **	-0.23 **
# adopted query suggestions	-0.16 **	-0.13 **

* and ** denote correlation values that are significant at $p < 0.05$ and $p < 0.01$.

One important question is: how do different types of search effort affect satisfaction? Table 2 provides some insight by showing correlations between different types of effort and searcher satisfaction. The effort of issuing queries (measured by query length) and examining result snippets (measured by maximum click position) negatively affect search satisfaction. However, we did not observe significant correlation between search satisfaction and the effort of assessing the search results (i.e., the number of result clicks).

The findings in this section show that we can further incorporate different types of search effort into satisfaction modeling. In general, the effort of issuing queries and examining result snippets have overall negative correlations with searcher satisfaction. In addition, some of the related search behavior does not change consistently as a function of search satisfaction level. For example, the maximum click rank may only be able to distinguish sessions with *very high* satisfaction from others, but cannot further discriminate among sessions with *high*, *moderate*, and *low* satisfaction. Conversely, the effort of assessing result webpages seems only indicative of *low* satisfaction sessions but does not differ a lot in other sessions. The complexity of search effort in sessions with different degrees of satisfaction further confirms that it is over-simplified to classify session satisfaction into a binary scale.

4.4 Changes of Gain and Cost in a Session

Since previous user studies have identified changes in search behavior within a session over time [19], we sought to understand whether search outcome and effort also change over time. If so, are some changes indicative of searcher satisfaction? In this section, we examine changes in search outcome and effort at the beginning and the end of the session, and their relationship with satisfaction.

Figure 5 shows that both search outcome and assessing effort will increase over time in sessions with *high* or *medium* satisfaction levels, but not in the *very high* or *low* satisfaction sessions. This suggests that we may further predict satisfaction via changes of search outcome and effort. The figures show the comparison of query rating (measuring per-query search outcome) and the number of clicks (measuring the effort of assessing results) between the first query and the last query of the sessions. The changes in both measures are significant in sessions with *high* and *medium* satisfaction levels ($p < 0.01$), but insignificant in the *very high* and *low* satisfaction sessions. For sessions with a *very high* degree of satisfaction, we suspect that this is because the first query is already very successful (the average quality rating is 1.87, with a maximum value of 2) and there is little room for improvement. For sessions with *low* satisfaction, searcher may be unable to reformulate better queries.

From Figure 5, we can observe that query reformulation success and failure may affect search satisfaction. We further tested this hypothesis in sessions with *high*, *medium*, and *low* levels of satisfaction. We excluded the sessions with *very high* satisfaction level since there is little chance of the searcher reformulating better que-

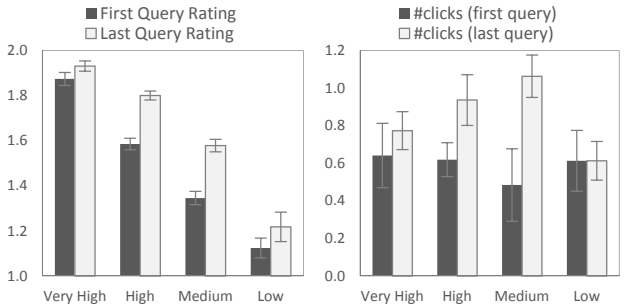


Figure 5. Changes of query search gain and clicking effort.

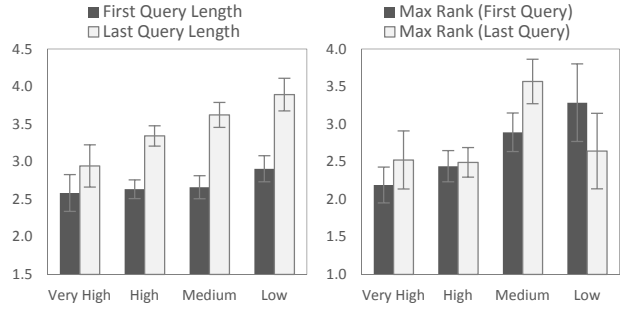


Figure 6. Changes of cost in querying and examining snippets.

Table 3. Correlation of query reformulation with satisfaction in sessions other than *very high* satisfaction.

Search Cost	Correlation with Satisfaction	
	Pearson's r	Kendall's tau-b
Last - First Query Rating	0.20 **	0.13 *
% reform better	0.11	0.09
% reform equal	0.10	0.07
% reform worse	-0.22 **	-0.20 **

* and ** denote correlation values that are significant at $p < 0.05$ and $p < 0.01$.

ries. Table 3 shows the results. We found a positive correlation between the change of query quality rating and satisfaction, and a negative correlation for the likelihood of query reformulation failure.

Figure 6 shows the comparison of query length (cost of formulating queries) and maximum click rank (cost of examining snippets) at the beginning and the end of the session. We can observe overall significant increase of query length in sessions, but the magnitude of increase in the *very high* satisfaction sessions are significantly smaller ($p < 0.05$) than those in other sessions. For the effort of examining snippets (Figure 6(b)), we did not observe significant change in the maximum click rank of results in sessions with *very high* and *high* satisfaction. However, there are significant increases in the maximum click rank in sessions with *medium* satisfaction, but significant decreases in maximum click rank in dissatisfied sessions. As we will show in Section 5.3, the former helps distinguish between *very high* and *high* satisfaction, while the latter can help distinguish *medium* from *low* satisfaction.

4.5 Summary of Findings

To conclude, in this section we first verify that satisfaction can be best explained as the value of the search outcome compared with the degree of search effort. After further examination, we found from our analysis that sessions with different satisfaction levels vary greatly but non-monotonously in search outcome, effort, and changes in these measures. This suggests that we can leverage the search behaviors that are related to search outcome and effort to better model search satisfaction. Conversely, the non-monotonous changes suggest that it is non-trivial and necessary to understand and predict the subtle differences between graded satisfaction levels. In the next section, we will describe features for predicting search outcome, effort, and their changes over a search session.

Table 4. Search outcome features and correlations. Values that are not significant at $p < 0.05$ are omitted ("-").

Features	Correlation w/ SAT				Correlation w/ sCG			
	All	VH/H	H/M	M/L	All	VH/H	H/M	M/L
ClickDwell	0.16	-	-	-	0.17	0.19	0.14	-
QSumClickDwell	0.13	-	-	-	0.10	-	-	-
SSumClickDwell	0.11	-	-	-	0.23	0.26	0.20	0.18
QueryInterval	0.22	0.18	-	0.13	-	-	-	-
SSumQueryInterval	0.18	0.15	-	-	0.18	-	0.15	0.21
SessionDuration	0.14	-	-	-	0.15	0.15	0.14	-
Q#Click T<5	-0.12	-0.13	-0.14	-	-	-	-	-
Q#Click T<15	-0.12	-	-0.15	-	-	-	-	-
Q#Click T≥60	0.16	-	-	0.15	0.14	-	-	0.17
S#Click	-	-0.13	-	-	0.24	0.27	0.22	0.26
S#Click T<10	-0.18	-0.14	-0.15	-	0.13	0.22	0.15	0.16
S#Click T≥185	0.16	-	-	0.18	0.23	0.18	0.19	0.24

5. FEATURES FOR PREDICTION

We now introduce our methods of predicting search satisfaction. In doing so, we adopt a feature-oriented approach. In accordance with our findings from Section 4, we select features related to search outcome, effort, and their changes in a search session. We examine the predictive power of features by their correlations (Pearson's r) with search satisfaction, search outcome (sCG), and the changes in query quality in a session. We suspected that some features may only be indicative of satisfaction within a particular range. Therefore, correlations are examined in all sessions (ALL) and three ranges: 1) sessions with *very high* and *high* satisfaction (VH/H); 2) sessions with *high* and *medium* satisfaction (H/M), and; 3) sessions with *medium* and *low* satisfaction (M/L). The results in the remainder of this section confirm our suspicions about satisfaction ranges.

5.1 Search Outcome Features

Previous studies (e.g., [24]) have found that SERP and click dwell time can indicate the quality of a SERP and clicked result. Thus, we include features related to SERP and click dwell time as search outcome features. Table 4 shows some representative features and their correlations with satisfaction and search outcome (sCG).

Click Dwell Time. We include the average dwell time of a click (ClickDwell) and the sum of click dwell time in a query and in a session (QSumClickDwell and SSumClickDwell). We consider the average, maximum, and minimum values of QSumClickDwell in multiple queries of a session (Table 4 shows the average value). We found that ClickDwell and SSumClickDwell have significant correlations with sCG in nearly all the ranges, but they do not significantly correlate with satisfaction in VH/H, H/M, and M/L.

Query Dwell Time. We measure query dwell time by the difference of the time that two adjacent queries were submitted (QueryInterval). We ignore this measure for the last query of each session. We compared this measure with the server-side query dwell time measure and found it has stronger correlation with satisfaction and sCG. We also consider the average, maximum, and minimum values of QueryInterval and the sum of the value in a session (SSumQueryInterval). Table 4 shows that query dwell time measures have significant correlations with satisfaction in the VH/H range, but no significant correlations in H/M and M/L ranges. The sum of query dwell times in a session also correlates with search outcome (sCG).

SAT and DSAT Clicks. We follow previous studies and divide clicks into satisfactory (SAT) clicks and dissatisfactory (DSAT) clicks by dwell time. SAT clicks are those with dwell time $\geq T_{sat}$ and DSAT clicks $< T_{dsat}$. We consider the number of SAT and DSAT clicks in a query (Q#Click) as well as its total number in a

session (S#Click). Previous studies [11] adopted T_{sat}=30s and T_{dsat}=15s. However, we noticed that there is no universal threshold that can obtain the best correlations with sessions in three ranges simultaneously. Thus, for each of the three ranges, we select T_{sat} and T_{dsat} with the strongest correlations with satisfaction. For example, T_{dsat}=5s has the strongest correlation in the VH/H range, but T_{dsat}=15s has a stronger one in H/M. Table 4 shows that the number of DSAT clicks have significant correlations with satisfaction in VH/H and H/M ranges (e.g. Q#Click T<5, T<15, and S#Click T<10), but the number of SAT clicks has stronger correlations in the M/L range (e.g. Q#Click T≥60 and S#Click T≥185). The total number of clicks in a session (even the DSAT clicks) have a significant strong correlation with the search outcome (sCG).

Table 4 shows that most of our search outcome features do correlate with the actual search outcome (sCG) and satisfaction in general (for all sessions). However, many features do not have consistent significant correlation with satisfaction in all three ranges of satisfaction. This confirms our suspicion expressed at the beginning of the section, and suggests that different measure may contribute differently and predict satisfaction within a certain range.

5.2 Search Effort Features

Search effort features include the effort metrics that we analyzed in Section 4.3 and their variants. The variants include: the number of queries without clicks (#Queries w/o Click), the average, MAX, and MIN rank of clicks in a query (QAvgClickPos), the longest and shortest query length in a session (SMaxQLength), and the total number of unique query terms in a session (#UniqQTerms).

Query Similarity. We also include query similarity in the search effort measures. Hassan et al. [14] discovered that a high similarity between queries may indicate that the searcher is "struggling" in a search session. Struggling inevitably increases searcher effort. Table 5 shows the correlations between satisfaction and some of the query similarity features. QEditDistanceBin is a binary feature with value 1 if the edit distance between two queries ≤ 2. QJaccardSim is the Jaccard similarity between two queries' term sets. Q#CommonCharLeft is the longest subsequence of characters in common between two queries starting from the left side of the query strings. We calculate these measures between each adjacent query pair in a session and adopt the average, MAX, and MIN values in a session as search effort features (Table 5 shows only the average values). Results show that query similarity measures have negative correlations with satisfaction in general. The effectiveness of these features is also different across the three satisfaction ranges, e.g., query similarity cannot distinguish satisfaction in M/L sessions.

The results in Table 5 validate our search effort features in that they do negatively correlate with search satisfaction. In addition, the results also suggest that for each adjacent satisfaction level, we may need to use different sets of effort features in prediction.

5.3 Outcome and Effort Change Features

Many of the features in Section 5.1 and 5.2 are based on query-level measures that indicate a query's search outcome or effort. For these measures, we calculate the difference of value in the last query of a session versus that of the first query. We include these measures as those indicating changes in search outcome and effort during the course of a search session. Table 6 shows some representative features and their correlations with satisfaction and the actual change of search outcome in a session (measured by the difference of query quality ratings between the first and last query).

Table 6 shows that the search outcome change features are usually correlated with the actual changes in search outcome, e.g. the sum of click dwell time in a query (Δ QSumClickDwell) and the number of SAT clicks (Δ Q#Click T≥60). However, these features do not always correlate well with satisfaction. For example, changes in the number of SAT clicks (Δ Q#Click T≥60) can only distinguish between the *medium* and *low* satisfaction sessions (M/L), and changes in the number of DSAT clicks (Δ Q#Click T<50) only weakly correlate with the *high* and *medium* satisfaction sessions (H/M).

To conclude, results in this section show that the discriminative power of features is not consistent in all sessions. For many of the features, it may only be able to indicate changes in search satisfaction within certain ranges. As we will show in the next section, different features have different degrees of effectiveness in predicting search satisfaction of sessions at different satisfaction ranges.

Table 5. Search effort features and correlations. Values that are not significant at $p < 0.05$ are omitted ("-").

Features	Correlation w/ SAT			
	All	VH/H	H/M	M/L
#Queries	-0.24	-	-0.11	-0.18
#Queries w/o Click	-0.25	-	-0.14	-0.18
QAvgClickPos	-	-0.20	-	-
QMaxClickPos	-	-0.19	-	-
QMinClickPos	-0.09	-0.18	-	-
SMaxQLength	-0.21	-0.15	-	-
#UniqQTerms	-0.16	-	-	-0.13
QEditDistanceBin	-0.22	-	-0.16	-
QJaccardSim	-0.14	-0.18	-	-
Q#CommonCharLeft	-0.14	-	-0.11	-

Table 6. Outcome and effort change features and correlations. Values that are not significant at $p < 0.05$ are omitted ("-").

Feature	Correlation w/ SAT				Correlation w/ Q Rating Change			
	All	VH/H	H/M	M/L	All	VH/H	H/M	M/L
Δ QSumClickDwell	-	-	-	-	0.15	0.16	0.16	0.15
Δ Q#Click T≥60	-	-	-	0.15	0.25	0.24	0.26	0.27
Δ Q#Click T<50	-	-	-0.11	-	-	-	-	-
EndClick	0.23	-	-	0.26	0.16	0.19	0.18	0.15
Δ QEditDistanceBin	-0.10	-	-	-0.16	-	-	-	-
Δ QJaccardSim	-	-	-0.11	-	-	-	-	-
Δ Q#CommonWord	-	-	-0.13	-	-	-	-	-
Δ QLength	-0.11	-0.14	-	-	-	-	-	-
Δ QMaxClickPos	-	-	-	0.16	-	-	-	-

5.4 Incorporating Action Transition Features

To capitalize on successful prior methods for satisfaction modeling, we include action transition features into our model. These transitions come from the Markov model approach [11], which takes advantage of the occurrence of action transitions. To leverage this approach in our predictive model, we use the count of each action transition in the Markov model as a part of our feature set.

6. EVALUATION

In this section, we describe the evaluation of our model. We consider the following two scenarios. First, we evaluate how well we can predict the actual session satisfaction through regression analysis (Section 6.1). Second, we examine how well we can distinguish between sessions of two adjacent satisfaction levels (Section 6.2). Results in this section help answer the following questions:

RQ1: How well can we predict searchers' satisfaction in a continuous interval and between adjacent satisfaction levels?

RQ2: How much does each category of features contribute to the prediction of search satisfaction?

RQ3: Where are the difficulties in predicting search satisfaction?

6.1 Prediction in a Continuous Interval

We experimented with different regression models including linear regression, Poisson regression, and boosted tree regression. We found that Poisson regression yielded the best results and report its results in this section. We evaluate regression using Normalized

Table 7. Regression of average session satisfaction ratings.

Features	NRMSE (0-1, the lower the better)	Correlation w/ SAT
Outcome	0.171	0.283
Effort	0.167	0.352
Change	0.165	0.374
All	**0.161**	**0.431**
All + MML	0.162	0.429
Individual Annotator	-	**0.687**

The darker and lighter shadings indicate differences with "All" that are significant at $p < 0.01$ and 0.05. All the correlations are statistically significant at $p < 0.01$.

Root Mean Square Error (NRMSE, smaller value means better prediction) and the Pearson's correlation (r) between the predicted and actual satisfaction values. For each group of features, we generate 10 runs and evaluate each run using 10-fold cross validation. This results in NRMSE values on 100 folds in total. We report the average value of NRMSE and test statistical significance using a Welch t-test. Due to the variability of correlation value in small samples, we report the overall correlation value in 100 folds.

Table 7 shows the results. Among the three groups of features, outcome features are the least predictive and resulted in the largest prediction error and the lowest correlation with the actual values. Whereas, the correlation of all outcome features is still stronger than those of any single outcome features shown in Table 4. The effort and change features have comparable predictive power. Unsurprisingly, using all three groups of features (All) results in better prediction of satisfaction than any of them alone. The predicted satisfaction values also have a moderate correlation with the actual satisfaction value ($r = 0.431$). However, little performance is gained by further combining with action transition features (All + MML).

A moderate correlation with the actual satisfaction values indicates a reasonably effective prediction. However, since we are the first to predict search satisfaction in a continuous interval, there is no accepted baseline method. To further study the effectiveness of our prediction, we examine how well an individual annotator can assess satisfaction (the average value of three annotators). As shown in Table 7, the correlation between an individual annotators' ratings and the average ratings is 0.687. This indicates the subjectivity of the task itself and can be considered as an upper bound for automatic methods. In comparing against individual annotators' assessments (with also an imperfect correlation), our prediction is reasonably effective, yet there is still some room for improvement.

6.2 Predicting Adjacent Satisfaction Levels

6.2.1 Experiment Settings

Given the prediction task in Section 6.1, it is difficult to make comparisons with previous approaches which predicted satisfaction as binary states. In addition, the overall effectiveness of regression in all sessions masks some details of the strengths and weaknesses of features in predicting satisfaction at different ranges. Therefore, we further evaluate how well our approaches can classify sessions with two adjacent satisfaction levels, e.g., between sessions with *very high* and *high* satisfaction levels (Section 6.2.2), between *high* and *medium* levels (Section 6.2.3), and between *medium* and *low* levels (Section 6.2.4). We begin by applying the features separately. Then we combine them (All) and further combine them with the Markov action transition features (All + MML). We tested logistic regression, SVM, and FastRank and report results using logistic regression since it has best performance in our dataset.

This evaluation scenario enables us to compare our methods and those described in previous work. We adopt the Markov model (MML) [13] as a baseline. The set of actions for MML is the same as those used by Hassan [13]. Ageev et al. [1], Hassan [11] and Wang et al. [35] modeled similar set of actions using Conditional Random Fields (CRF), Generative Models and Structured Learning (AcTS). According [11] and [35], the CRF, the generative model and AcTS approaches improved average F_1 score of MML by about 4%, 4% and 8% respectively on their dataset. Here we did not implement CRF, the generative model and AcTS for comparison, but we will compare the relative improvements of our method with the scale of improvements reported in their papers as an indirect comparison with them.

Table 8. Classification of *very high* and *high* sessions.

	F_1			Accuracy
	Average	*Very high*	*High*	
Outcome	0.507	0.138	0.876	0.784
Effort	**0.567**	**0.263**	0.872	0.783
Change	0.474	0.072	0.877	0.782
All	0.558	0.235	**0.880**	**0.794**
All + MML	0.556	0.234	0.879	0.791
MML	0.470	0.076	0.864	0.767
All *high*	0.438	-	0.877	0.781

Darker and lighter shadings indicate sig. diffs with MML (at $p < 0.01$ and 0.05).

In a similar way to the dataset employed in previous studies [1, 11, 13, 35], the satisfaction labels in our dataset are unevenly distributed. Therefore, we evaluate classification mainly by the F_1 scores of the two classes and the macro-average F_1 score. For each approach, we generate 10 runs, each using 10-fold cross validation. We report the average value on the 100 folds and test set, and compute the statistical significance of any observed differences.

6.2.2 Classify Very High and High Satisfaction

Table 8 shows results of classification between sessions with the *very high* and *high* degree of satisfaction. Among the three groups of features, effort features are the most predictive. The differences in F_1 scores between effort features and other two groups are significant at $p < 0.01$. This is consistent with our observations in Section 4.2, i.e. *very high* and *high* satisfaction sessions mainly differ in terms of the search effort instead of the search outcome. We did not observe any improvements from combining the three features. Using effort features alone has the highest F_1 in classifying sessions with *very high* satisfaction levels, but combining features results in significantly lower F_1 on sessions with this grading ($p < 0.05$).

Our approach (All) significantly outperforms the Markov model in both average F_1 score (+18.7%) and accuracy (+3.5%). The magnitude of the gain sis also greater in comparison with that of CRF and AcTS in studies by Wang et al. [35]. In fact, using outcome or effort features alone results in significantly better F_1 than MML. The main reason is the poor effectiveness of MML in classifying sessions with *very high* satisfaction levels. Due to the limited effectiveness of MML in distinguishing VH/H sessions, little is gained from combining our features with those derived from the action transitions (All + MML).

It is worth noting that even our best approach (using effort features alone) also did not perform all that effectively in recognizing the *very high* satisfaction sessions (the F_1 score is only 0.263). This indicates that it is more challenging to classify adjacent satisfaction levels, whose sessions may only have very subtle differences.

6.2.3 Classify High and Medium Satisfaction

Table 9 shows the evaluation results of classification between *high* and *medium* satisfaction levels. All of our approaches significantly outperform MML in both F_1 score and accuracy ($p < 0.01$). This further demonstrates the difficulty of the MML method in recognizing subtle differences of adjacent satisfaction levels.

We observe that the effort and change features are more effective than the outcome features (the difference between outcome and change in average F_1 score is significant at $p < 0.05$). In comparing

Table 9. Classification of the *high* and *medium* sessions. All the methods are significantly better than MML at $p<0.01$.

	F_1			Accuracy
	Average	*High*	*Medium*	
Outcome	0.546	0.661	0.431	0.579
Effort	0.578	0.705	0.451	0.618
Change	0.583	0.681	0.485	0.609
All	0.601	0.714	0.488	0.636
All + MML	**0.612**	**0.724**	**0.500**	**0.647**
MML	0.489	0.599	0.379	0.519
All *high*	0.352	0.704	-	0.543

The darker and lighter shadings indicate differences with "All" that are significant at $p < 0.01$ and 0.05.

Table 10. Classification of the *medium* and *low* sessions.

	F_1			Accuracy
	Average	*Medium*	*Low*	
Outcome	0.407	0.810	0.004	0.680
Effort	0.498	0.806	0.190	0.688
Change	0.644	0.810	**0.478**	0.724
All	0.623	0.813	0.433	0.721
All + MML	**0.649**	**0.821**	0.477	**0.736**
MML	0.577	0.797	0.356	0.699
All *high*	0.405	0.809	-	0.680

The darker and lighter shadings indicate differences with MML that are significant at $p < 0.01$ and 0.05.

the performance of the effort and change features, it appears that effort features can better recognize sessions with *high* level of satisfaction, but change features can better classify sessions in *medium* satisfaction level (with higher F_1 scores in each class). We also observed significant improvements after combining three groups of features. This shows that the three groups of features are effectively modeling different aspects of the differences between the sessions with *high* and *medium* satisfaction levels. There is also a small but insignificant improvement by further combining with action transition features (All + MML).

We achieved better performance in distinguishing between the *high* and *medium* satisfaction sessions, than between *very high* and *high* satisfaction levels. Table 8 shows that the best average F_1 score for *high*/*medium* is 0.612 versus 0.567 for *very high* / *high*. It may be relatively easier to classify the *high* and *medium* satisfaction levels. Due to the limited discriminative power of MML in this task scenario, our approach also achieved remarkable improvements in both average F_1 score (+25%) and accuracy (+24.6%).

6.2.4 Classify Medium and Low Satisfaction

We further evaluate classification between *medium* and *low* satisfaction levels. Sessions with *low* level of satisfaction have dissatisfactory ratings ($s \leq 3$, ranging from 1 to 5), which is similar to the DSAT class in binary satisfaction classification in previous studies. We observed that the Markov model performs effectively in this scenario, with an average F_1 score 0.577 compared with 0.470 and 0.489 in previous two satisfaction ranges (Tables 8 and 9).

Among the three groups of features, change features are the most effective. Outcome and effort features have limited effectiveness in recognizing dissatisfied sessions (F_1 scores < 0.2). Change features also achieved the highest F_1 score in classifying the dissatisfied sessions compared with those that combine all features. We achieved even higher average F_1 score and accuracy from combining with the action transition features (All + MML). One explanation is the strength of the MML model for this particular task.

In comparison with the baseline approach (MML), our best approach (All + MML) still achieved significant improvements in average F_1 (+12.5%) and accuracy (+5.3%). The high performance of the Markov model shows that the strength of our method probably lies in classifying sessions with *medium* and *low* satisfaction. That is, instead of recognizing highly satisfactory sessions, MML is more likely to be a strong classifier for DSAT sessions.

6.2.5 Summary of Results

To sum up, our results answer our three research questions:

First (regarding RQ1 and RQ3), we found limited effectiveness of a well-established previous method (the Markov model) in identifying the subtle differences among *very high*, *high*, and *medium* satisfaction levels. Results indicate that our approach can identify such minor differences of satisfaction more effectively. We also found that the true strength of the Markov model may lie in recognizing DSAT sessions. In this scenario, our approach still achieved significantly better performance. The regression analysis in Section 6.1 also shows that our model can predict satisfaction scores with moderate correlation with the actual scores from judges ($r = 0.43$).

Second (for RQ2), as we suspected, different groups of features contribute diversely in different scenarios. For example, effort features contribute remarkably in distinguishing *very high* and *high* satisfaction sessions. The change features distinguish *medium* and *low* satisfaction sessions. All three feature groups contribute substantially to discriminating *high* and *medium* satisfaction sessions.

7. DISCUSSION AND IMPLICATIONS

We have studied the important issue of graded search satisfaction. Although degrees of satisfaction may be captured during laboratory studies, in many settings, including retrospective analysis of large-scale log data collected by search engines, satisfaction is modeled as a binary variable. We suspected that this may be insufficient to accurately assess and compare the performance of these search systems. Our findings clearly show that satisfaction is sufficiently nuanced that it needs to be modeled at greater granularity than has been done traditionally. Rich and non-monotonous differences in behavior are noted at all satisfaction levels studied, suggesting that there may be subtle distinctions between the levels of search satisfaction and a non-binary assessment may be justified. In addition, we show that by considering both the amount of information gained during a session (the outcome) and the effort involved in obtaining that information, we can accurately model search satisfaction. Leveraging information about outcome and effort let us develop a predictive model to accurately predict *graded* search satisfaction.

Despite the promise of our methods, we also acknowledge some limitations. The satisfaction judgments collected as part of this study were provided by third-party judges. Ideally the judgments of search satisfaction would be sourced from searchers in-situ at search time. This creates significant additional overhead for searchers, especially if we are asking them to provide both quality ratings for the results for each query and an overall multi-level rating for the session. We were concerned that such overhead would dissuade searchers from providing feedback both in our experiment, or more generally in natural settings as part of a broader deployment of satisfaction elicitation methods (limiting the impact of our conclusions). More research is needed into methods to incentivize searchers to provide ratings, to streamline the rating process, or to intelligently sample sessions so that we do not need to probe searchers at every session. In addition, search satisfaction in previous studies as well as in this paper is measured using extremely simple instruments (e.g., simply asking searchers or annotators whether they are satisfied). As we reviewed in Section 2.1, in many other fields researchers have developed complex and factorized instruments for measuring user satisfaction in other systems. This suggests we may also need more robust instruments for measuring searcher satisfaction. We also focused on gain and effort in this study. However, there are other factors that can influence the types of search behavior observed, including the nature of the search task and the tenacity

of the searcher. Further studies are required to understand the role of these and other factors on behavior and search satisfaction.

The strong performance of our models has implications for search providers who aim to accurately measure satisfaction with their services. Typically, satisfaction with online services is evaluated on a binary scale in automatic evaluation approaches. However, modeling *degrees* of satisfaction allows for a more complete understanding of the performance of systems, and the more accurate prioritization of cases where potentially-negative experiences are identified. Moving forward, we must develop and evaluate our predictive models in large-scale settings to understand whether they can yield richer insights about aggregated search experiences than their binary equivalents in practice. In addition, search satisfaction estimates are used in a number of settings beyond system measurement, including evaluating short- and long-term personalization [7], as labels for implicit feedback [9], and in learning about biases in search behavior [36]. The satisfaction judgments in these settings have traditionally been binary, but understanding the impact of graded search satisfaction is also an important future direction.

8. CONCLUSION

Search satisfaction is a complex construct. Despite the complexity, to our knowledge it has hitherto been modeled as a binary variable in the measurement of satisfaction with search systems. In this paper, we performed a detailed study of graded search satisfaction in the context of Web search. We used logged search session data mined from users of the Microsoft Bing search engine and detailed judgments about session satisfaction on a multi-point scale from human annotators. Through our analysis we observe clear differences in search behavior in sessions with different satisfaction levels. Given the presence of these differences, we developed a predictive model that can more accurately estimate graded search satisfaction than existing satisfaction modeling methods by considering search outcomes and searcher effort, both independently and in combination. Our findings are of critical importance to search providers as they attempt to measure system performance based only on implicit signals in log data. Future work involves improving the performance of our model, additional testing against binary equivalents, and experimenting with applying graded satisfaction in different settings, especially surrounding the richer analysis of search engine performance that our graded satisfaction modeling enables.

REFERENCES

[1] M. Ageev et al. 2011. Find it if you can: A game for modeling different types of web search success using interaction data. In *SIGIR'11*: 345–354.

[2] A. Al-Maskari et al. 2007. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR'07*: 773–774.

[3] J. Arguello. 2014. Predicting search task difficulty. In *ECIR'14*: 88–99.

[4] L. Azzopardi et al. 2013. How query cost affects search behavior. In *SIGIR'13*: 23–32.

[5] L. Azzopardi. 2014. Modelling interaction with economic models of search. In *SIGIR'14*: 3–12.

[6] J. E. Bailey and S. W. Pearson. 1983. Development of a tool for measuring and analyzing computer user satisfaction. *Management Science*, 29(5): 530–545.

[7] P. N. Bennett et al. 2012. Modeling the impact of short- and long-term behavior on search personalization. In *SIGIR'12*: 185-194.

[8] H. Feild et al. 2010. Predicting searcher frustration. In *SIGIR'10*: 34–41.

[9] S. Fox et al. 2005. Evaluating implicit measures to improve web search. *ACM TOIS*, 23(2): 147–168.

[10] Q. Guo et al. 2011. Why searchers switch: understanding and predicting engine switching rationales. In *SIGIR'11*: 335-344.

[11] A. Hassan. 2012. A semi-supervised approach to modeling web search satisfaction. In *SIGIR'12*: 275–284.

[12] A. Hassan et al. 2013. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *CIKM'13*: 2019–2028.

[13] A. Hassan et al. 2010. Beyond DCG: User behavior as a predictor of a successful search. In *WSDM'10*: 221–230.

[14] A. Hassan et al. 2014. Struggling or exploring? Disambiguating long search sessions? In *WSDM'14*: 53–62.

[15] S. B. Huffman and M. Hochster. 2007. How well does result relevance predict session satisfaction? In *SIGIR'07*: 567-574.

[16] B. Ives et al. 1983. The measurement of user information satisfaction. *CACM*, 26(10): 785–793.

[17] K. Järvelin and J. Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *SIGIR'00*: 41–48.

[18] K. Järvelin et al. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *ECIR'08*: 4–15.

[19] J. Jiang et al. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *SIGIR'14*: 607–616.

[20] T. Joachims et al. 2005. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR'05*: 154–161.

[21] R. Jones and K. Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM'08*: 699–708.

[22] E. Kanoulas et al. 2011. Evaluating multi-query sessions. In *SIGIR'11*: 1053–1062.

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

[24] Y. Kim et al. 2014. Modeling dwell time to predict click-level satisfaction. In *WSDM'14*: 193–202.

[25] R. Kohavi et al. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*. 18(1): 140–181.

[26] J. Liu et al. 2012. Exploring and predicting search task difficulty. In *CIKM'12*: 1313–1322.

[27] J. Liu and N. J. Belkin. 2010. Personalizing information retrieval for multi-session tasks. In *SIGIR'10*: 26–33.

[28] L. Lorigo et al. 2006. The influence of task and gender on search and evaluation behavior using Google. *IP&M*, 42(4): 1123–1131.

[29] G. Mankiw. 2010. *Principles of Macroeconomics*. South-Western Cengage Learning.

[30] A. Marshall. 2009. *Principles of Economics: Abridged Edition*. Cosimo Classics.

[31] V. McKinney et al. 2002. The measurement of Web-customer satisfaction: an expectation and disconfirmation approach. *Information Systems Research*, 13(3): 296–315.

[32] R. L. Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*. 17(4): 460–470.

[33] C. L. Smith and P. B. Kantor. 2008. User adaptation: Good results from poor systems. In *SIGIR'08*: 147–154.

[34] L. T. Su. 2003. A comprehensive and systematic model of user evaluation of Web search engines. *JASIST*, 54(13): 1175–1192.

[35] H. Wang et al. 2014. Modeling action-level satisfaction for search task satisfaction prediction. In *SIGIR'14*: 123–132.

[36] R. W. White. 2013. Beliefs and biases in web search. In *SIGIR'13*: 3–12.

[37] E. Yilmaz et al. 2014. Relevance and effort: an analysis of document utility. In *CIKM'14*: 91–100.