# Framing Mobile Information Needs: An Investigation of Hierarchical Query Sequence Structure

Shuguang Han[1], Xing Yi[2], Zhen Yue[2], Zhigeng Geng[2], Alyssa Glass[2]
[1] University of Pittsburgh, 135 N Bellefield Ave., Pittsburgh, PA, USA
[2] Yahoo! Research, 701 First Ave., Sunnyvale, CA, USA
shh69@pitt.edu, {xingyi,zhenyue,zgeng,alyssag}@yahoo-inc.com

## ABSTRACT

When using search engines, people often issue multiple related queries to accomplish a complex search task. A simple query-task structure may not fully capture the complexity of query relations since people may divide a task into multiple subtasks. As a result, this paper applies a three-level hierarchical structure with query, goal and mission - a mission includes several goals, and a goal consists of multiple queries. Particularly, we focus on analyzing query-goal-mission structure for mobile web search because of its increasing popularity and lack of investigation in the literature. This study has three main contributions: (1) we study the query-goal-mission structure for mobile web search, which was not studied before. (2) We identify several differences between mobile and desktop search patterns in terms of goal/mission length, duration and interleaving. (3) We demonstrate that the query-goal-mission structure can be applied to design better user satisfaction metrics. Specifically, goal-based search success rate and mission-based abandonment rate are better aligned with users' long-term engagement than query and session based metrics.

## Keywords

Mobile web search, search goal, search mission, query structure, user engagement

## 1. INTRODUCTION

Modern search engines are experiencing several important changes - the involvement of more complex search tasks and an increasing usage on mobile devices. Complex search tasks often require users to issue multiple related queries while traditional search analysis seldom considers the query relations [10]. This drives researchers to investigate users' search behaviors at the session level [2, 3, 8], assuming that two queries in the same session should share the same goal. A search session is often defined as the 30-minute timeout for user inactivity [16]. However, as many studies pointed out, under this definition, a search session frequently involves queries from multiple unrelated tasks [10, 15]. As a result, recent research has started focusing on the task-level analysis for web search [2, 16]. To the best of our knowledge, existing findings on this topic are mainly drawn from the analysis of desktop search logs. Due to the apparent differences between mobile and desktop search behaviors [19], it is critical to understand how people organize their search tasks in mobile sessions and how it differs from desktop search.

Consider a query sequence – "goggle ski, goggle for men, ski pants, ski pants size, ski pants price, restaurants, restaurant near me". The first two queries belong to the same task since the user was searching for "ski goggle"; the middle three queries also belong to the same task because the user was looking for "ski pants". However, it is unclear whether "ski goggle" and "ski pants" should be treated as the same task. To capture this complex query structure, we adopt a three-level hierarchical structure that organizes a query sequence into query, goal and mission, as proposed by Jones and Klinkner [10]. Following the definitions in [10], a search goal is an atomic information need that includes one or more related queries, and a search mission consists of one or more related search goals. Under these definitions, the above query sequence can be seen as having three search goals: buying ski goggles, buying ski pants, and finding a nearby restaurant. The first two goals are associated with the same mission (buying ski equipment) and the third goal is associated with a different mission.

Therefore, we adopt the query-goal-mission structure for analyzing mobile search sequences. Our study is based on a web-scale search log sampled from Yahoo! search engine, which involves millions of search sessions. To obtain a more comprehensive result, we also examine the desktop query-goal-mission structure for comparison. The first step of conducting the study is to develop an algorithm that automatically constructs such complex structure in a scalable way. We employ the approach adopted in both Jones and Klinkner [10] and Lucchese et al. [16], and improve the algorithm by incorporating several novel query log features. The details of our method is provided in §3.

After obtaining a large-scale, segmented search query sequence, our next task is to understand the mobile query-goal-mission structure: (1) how it differs from desktop query structure, and (2) how to apply the query structure information for supporting real-world mobile searches. To answer the first question, we compare several goal and mission measurements, including goal/mission duration and goal/mission length, between mobile and desktop searches. For the second question, prior studies have shown that the query-task



*CIKM'16 , October 24 - 28, 2016, Indianapolis, IN, USA*



ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983654

structure (a task resembles to a search goal in our study) can be applied in many applications, including recommending better queries [2], improving search results ranking [21] and designing better user satisfaction metrics [15]. In this paper, we study how the three-level query structure is applied for designing better user satisfaction metrics since it is the least-studied application in literature and an important topic for search evaluation [12, 15].

This paper has three key contributions: (1) we develop a novel two-step approach, adapting the methods developed in [10] and [16], to automatically construct the query-goal-mission structure for a search sequence (§3). Our method employs a more comprehensive and larger set of features to achieve improved performance (§4). More importantly, (2) we perform a large-scale query-goal-mission structure analysis with millions of search sessions, and identify important differences between mobile and desktop search patterns (§5). (3) We discover that the query-goal-mission structure can be used to design better user satisfaction metrics - goal is better in determining search success while mission is better for understanding search abandonment (§6). These are not investigated in prior studies.

## 2. RELATED WORK

Studies [10, 15, 16] based on large-scale search log analysis identified that search queries are not independent. To discover the connections among different queries, researchers have been long investigating the ways of segmenting query sequences into meaningful search units. Queries from the same unit are assumed to have tighter relations than those from different units [2, 7, 16]. Search session is one type of unit, and it is often determined by a fixed timeout cutoff [7, 16]. However, under this definition, a session may include multiple unrelated tasks [2, 10]. This work resulted in recent studies [2, 16] focused on more fine-grained analysis at the task level. Although there have been several attempts [10, 14, 16] to identify search tasks within a search session, accurate task extraction remains a challenge due to the complexity of the task definition - a task can involve sub-tasks, and several tasks may share the same mission. Therefore, Jones and Klinkner [10] proposed a three-level hierarchical structure containing query, goal and mission. Following this study, our work is built on top of the three-level query structure, but differs in terms of: 1) we propose several novel features for goal/mission identification, which are shown to be more effective; and 2) we provide a large-scale analysis of such complex structure and explore its potential applications in mobile search.

Rich query structure information can be applied in various applications. First, task and session information were often adopted to enrich users' search contexts and further improve their search experience [2, 21]. Feild and Allan [2] found that the task information can help improve query recommendation performance. White et al. [21] enhanced personalized search by mining users' search behaviors within the same task. Second, task and session information were also applied to design better user satisfaction metrics [11, 15]. For instance, Liao et al. [15] found that sessions and queries are not as precise as tasks in determining user satisfaction. Since the query-task structure cannot capture the complex structure as in a query-goal-mission schema, we expect that the query-goal-mission structure can provide additional value and define better user satisfaction metrics.

The above studies were mostly performed using desktop search logs, while the study of mobile search is rare. With the increasing popularity of mobile devices, dedicated research on mobile search patterns is needed. Indeed, Song et al. [19] found significant differences between mobile and desktop search patterns in query and click patterns. Other researchers [4, 5, 20] discovered that mobile users are frequently involved in complex search tasks, yet a fine-grained analysis of mobile search sequence structure is missing. Thus, we focus our investigation on these mobile query structures.

## 3. FRAMING SEARCH SEQUENCES

The first task of our study is to build a hierarchical query-goal-mission structure for each mobile query sequence. In this paper, a query sequence refers to a mobile search session. Following [19], we use 30-minute user inactivity as the timeout cutoff to determine a session. The search goal and mission are defined the same as [10] - a **search goal** is an atomic information need resulting in one or more queries, and a **search mission** is a related set of information needs, resulting in one or more search goals.

**Problem definition**. We formalize our search sequence structure identification problem as following: suppose that we have an $m$-query mobile session $\mathcal{S} = (q_1, q_2, q_3, ..., q_m)$. We want to partition it into a set of goals ($\mathcal{G}$) and missions ($\mathcal{M}$). To achieve this, we propose a two-step approach. In the first step, we determine if a query pair belongs to the same goal or mission through the Logistic Regression (LR) classifier as did in Jones and Klinkner [10], for which we need manually-labeled ground-truth datasets (§4.1) and classification features (Table 1). The model output is the probability of a query pair belonging to the same goal/mission.

In the second step, we group queries into goals and missions, where we adopt the average-linkage hierarchical clustering algorithm. The goal clustering starts with initializing each query as a goal. Then, we loop through all goal pairs and merge the two most similar goals into one (the similarity is defined as the probability of a query pair belonging to the same goal, as computed in the goal classifier). We repeat this process until the similarities of all goal pairs are smaller than a threshold $\mathcal{T}_G$. For mission clustering, we apply the same approach, except: (1) we compute with the mission similarity (i.e., the probability of a query pair belongs to the same mission); (2) the clustering process stops when reaching the mission pair similarity threshold $\mathcal{T}_M$; and (3) the initialized missions are the identified goal clusters.

**Features**. The logistic regression classifiers employ four groups of features for each query pair $(q_i, q_j)$, in which $i$ and $j$ $(i < j)$ denote the sequential positions of the two queries in the session $\mathcal{S}$. The four feature groups are provided in Table 1. The first two groups are directly adopted from [10, 16], and the mobile query global features are computed based on a large-scale search log with millions of search sessions (§4.1). However, when computing the mobile query pair global features, we must also properly handle data sparsity – we find that only 3% of mobile query pairs occur in two consecutive weeks. To resolve this problem, we propose an additional feature group that calculates query pair features based on the aggregation of query term pairs. A term pair global feature is calculated in the similar manner as the query pair global feature, except the statistics are based on terms rather than queries. The extracted term-pair based features form our third feature group. Furthermore, we consider a fourth

**Table 1: Features for query pair classification. Feature groups with ‡ indicates the new features we proposed.**

| | |
|---|---|
| **Mobile query pair local feature (local)** | |
| jaccard, edit distance and time interval | Same as [10]. |
| **Mobile query pair global feature (queryglobal)** | |
| pmi and pq12 | $\frac{p(q_i \to q_j)}{p(q_i \to *)p(* \to q_j)}$ and $\frac{p(q_i \to q_j)}{max_{q_m} p(q_i \to q_m)}$, [10]. |
| llr and $\chi^2$ | Log-likelihood ratio and $\chi^2$ for the query pair $(q_i, q_j)$ [17]. |
| entropy_x_q1 and entropy_q1_x | $\sum_x p(q_i\|q_x) \log_2 p(q_i\|q_x)$ and $\sum_x p(q_x\|q_j) \log_2 p(q_x\|q_i)$ [10]. |
| **Mobile query term pair global feature (termglobal.mobile) ‡** | |
| t_pmi | $\frac{1}{N} \sum_{t_m \in q_i, t_n \in q_j} \frac{p(q_i \to q_j)}{p(q_i \to *)p(* \to q_j)}$. |
| t_pq12, t_llr, t_$\chi^2$, t_entropy_x_q1, t_entropy_q1_x are computed using the same method as t_pmi. | |
| **Desktop query term pair global feature (termglobal.desktop) ‡** | |
| dt_pmi, dt_pq12, dt_llr, dt_$\chi^2$, dt_entropy_x_q1 and dt_entropy_q1_x are computed using desktop search logs with the above process. | |

feature group that obtains term-based features from desktop search logs, enriching the feature set with additional search queries. The last two groups are our proposed new features (highlighted by ‡ in Table 1).

# 4. SEGMENTATION EXPERIMENTS

## 4.1 Data Collection

Our experiments are performed on a large-scale search log sample from the Yahoo search engine on both mobile and desktop devices during a three-week period (2/23 to 3/15 in 2015), covering millions of search queries and sessions. To train and evaluate the goal and mission classification and clustering models, we need a ground-truth dataset. This paper constructs three ground-truth data sets, i.e., $D_0$, $D_1$ and $D_2$ as shown in Figure 1. Each dataset consists of approximately 1,000 query pairs randomly sampled from the search log. $D_0$ is extracted from mobile search logs in Week 2 and it is used to train our goal classifier and mission classifier. $D_1$ and $D_2$ are two testing data sets for mobile search and desktop search, respectively. Note that we do not collect a desktop training dataset because the model trained on the mobile training dataset can be effectively applied for desktop testing dataset and achieves reasonably good performance.

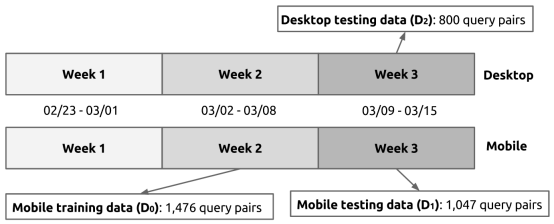


**Figure 1: An illustration of data corpus and sampled training and testing datasets**

For each of the ground-truth datasets, we collect editorial judgments from two independent paid editors. We observed more than 80% agreements for both goal and mission judgments (85.80% for goal and 83.21% for mission in $D_0$, 91.31% for goal and 87.39% for mission in $D_1$, and 91.70% for goal and 82.57% for $D_2$), indicating high precision quality labels for our data sets. The conflicts were resolved by a third editor. The following experiments in this paper are performed on the conflict-resolved datasets.

## 4.2 Query Pair Classification Performance

Our experiment begins with training a goal classifier and a mission classifier using $D_0$. Here, we report the performances of these two classifiers on the testing dataset $D_1$ (we do not provide results for $D_2$ as the findings were the same). The results are provided in Table 2. We use the classification algorithm in [10] as the baseline, for which they use the combination of *local + queryglobal* features (the italic row in Table 2). We adopt F1 score for the same goal/mission to measure the classification performance. We note multiple findings from the data in Table 2.

**Table 2: Search goal and mission classification performances on $D_1$. The row in italics denotes the baseline. Rows with ‡ denote the new features.**

| Feature groups | F1(goal) | F1(mission) |
|---|---|---|
| local | **0.6574** | 0.8038 |
| queryglobal | 0.3384 | 0.7903 |
| termglobal.mobile ‡ | 0.5444 | 0.8165 |
| termglobal.desktop ‡ | 0.5937 | 0.8206 |
| *local + queryglobal* | *0.6485* | *0.8128* |
| local + termglobal.mobile ‡ | 0.6099 | 0.8505 |
| local + termglobal.desktop ‡ | 0.6339 | **0.8622** |

First, the query local feature is the strongest signal for both search goal and search mission classification. In particular, in goal classification, the query local feature achieves the best performance, while adding global features does not result in much improvement. The results show a different trend for mission classification, for which query global features do provide a positive contribution. This reflects the inherent difference between a goal and a mission - a mission defines the connection between two queries at a more abstract level and thus requires external knowledge for understanding the query relation. Such knowledge can be learned from both query pair and term-pair co-occurrence statistics and is consistent across different device types.

Second, compared to the baseline, our proposed term-pair based features (both local + termglobal.mobile and local + termglobal.desktop) generate better performance in mission classification while having no difference for goal classi-

fication. In particular, the term-based query pair features obtained from desktop search logs achieve the best performance. We believe this result is because the desktop search logs contain richer information, and thus produces more robust global features and resolves the data sparseness for tail query pairs. This indicates the possibility of transferring desktop knowledge into mobile search when mobile search information is missing or sparse.

## 4.3 Goal/Mission Clustering Performance

Outputs of the best performing classifiers are then used in the second step for goal and mission clustering. We evaluate the clustering performance using the following procedure. Suppose that we have $m$ human-labeled query pairs from $n$ search sessions. We first run the clustering algorithm on all $n$ sessions. Then, based on the clustering outputs, we evaluate the labeled query pairs on how accurately they are clustered, i.e., whether the same-goal queries are in the same goal cluster, and whether the same-mission queries are in the same mission cluster. We employ three evaluation measurements - overall accuracy, accuracy for the same goal/mission and accuracy for different goals/missions.

The clustering algorithm requires specifying the stopping thresholds $\mathcal{T}_G$ and $\mathcal{T}_M$ (see §3). Here, they are tuned based on the training dataset $D_0$ with a 10-fold cross validation. We set $\mathcal{T}_G = 0.35$ for goal clustering and $\mathcal{T}_M = 0.50$ for mission clustering as they produce the best performances. The clustering performances are presented in Table 3. We find that our algorithm achieves reasonably high accuracy. We further examine the algorithm performance for desktop testing dataset (i.e., $D_2$). In Table 3, the overall accuracy seems to be even better for both goal and mission identification (comparing to $D_1$). This confirms that our classifiers are robust in identifying search goals and search missions.

**Table 3: Clustering performance on $D_1$ and $D_2$**

| | Goal/mission | Acc. | Acc. same | Acc. diff. |
|---|---|---|---|---|
| $D_1$ | Goal | 0.8700 | 0.7117 | 0.9017 |
| | Mission | 0.8424 | 0.8160 | 0.8747 |
| $D_2$ | Goal | 0.9122 | 0.7484 | 0.9816 |
| | Mission | 0.8586 | 0.7724 | 0.9715 |

# 5. CHARACTERIZING MOBILE SEARCH SEQUENCE STRUCTURE

This section provides a large-scale analysis of search sequence structures with millions of mobile and desktop search sessions randomly sampled from one-week search log of the Yahoo search engine (i.e., Week 3 in Figure 1). In particular, we focus on analyzing how mobile search goals/missions differ from desktop search goals/missions. Note that the following statistical significance is reported based on Wilcoxon singed-rank test since data is not normally distributed.

**Goal/Mission/Session Length**. Table 4 reports a list of descriptive statistics about the metric difference between mobile and desktop for the query-goal-mission structure. The difference is computed in the following way. First, we calculate the mean values of each metric in Table 4 for both mobile and desktop. Then, the mobile metric value is subtracted by the desktop metric value for computing the difference. We find that #goals/mission is relatively stable across different devices, meaning that people tend to frame their information needs with a steady hierarchical structure. However, a desktop session contains more queries, goals and missions, which may indicate that people on desktop are more likely to arrange a continuous time slot (i.e., a session) to complete more goals and missions, while people may need multiple discrete time periods on mobile devices to complete the same number of goals and missions.

Regarding the goal and mission lengths, we find that a search goal/mission contains more queries on mobile than on desktop. This finding may be because mobile search goals/missions are more complex than desktop, and thus require more queries to complete; or because mobile search queries are more likely to fail, requiring more frequent query refinement for each goal. In analyzing the first scenario, we find no substantial difference for #goals/mission between mobile and desktop (on the contrary, this number is slightly larger on desktop). For the second hypothesis, goal/mission duration analysis provides more insights; however, further experiments are needed to test this hypothesis in the future.

**Table 4: The differences of goal/mission/session metrics between mobile and desktop.**

| Measurements | $\Delta$ = desktop - mobile | Sig. |
|---|---|---|
| $\Delta$(#queries/goal) | -0.1913 | p<0.001 |
| $\Delta$(#queries/mission) | -0.2172 | p<0.001 |
| $\Delta$(#queries/session) | 0.2461 | p<0.001 |
| $\Delta$(#goals/session) | 0.4082 | p<0.001 |
| $\Delta$(#missions/session) | 0.3251 | p<0.001 |
| $\Delta$(#goals/mission) | 0.0012 | p<0.001 |

**Goal/Mission Duration**. Following [10], we define the goal/mission duration as the time difference between the first query and the last query within the search goal/mission. Same as the above analysis, we also compute the goal/mission duration difference between desktop and mobile. The above analysis indicates that a mobile search goal/mission contains more queries than desktop. Thus, we expect that a goal/mission lasts longer on mobile. We do find that mobile search goals are longer (4.8 seconds longer on mobile than desktop); however, mobile search missions are actually significantly shorter (30.6 seconds longer on desktop than mobile). This indicates that people spent less time on mobile for each query. It is possible that users do not spend as much time exploring search results associated with failed queries, and quickly refine and resubmit another query.

**Goal/Mission Interleaving**. Multitasking is a common behavior on search engines. Our findings support that search goals and missions are sometimes interleaved with other goals and missions. Therefore, we compute the percentages of interleaved goals and missions on both desktop and mobile devices and provide the results in Table 5. We find that the overall interleaving percentage is relatively small since most of the search sequences only have one goal and one mission. Besides, search missions are more likely to be interleaved than search goals since a search goal usually refers to a simple easy-to-achieve sub-task while a search mission may last for a longer time. More interestingly, compared to mobile search, desktop search has more interleaved goals and missions. This may either indicate that people tend to be more focused on mobile or suggest that the current support for mobile multitasking search is still insufficient. We expect that this situation may change as mobile search interfaces continue to evolve. Our goal/mission interleaving analysis

will similarly need to evolve as these supportive interfaces change.

**Table 5: Goal/mission interleaving percentages. A number in bold (italics) indicates a significance comparing goal and mission (mobile and desktop).**

| | mobile | | desktop | |
|---|---|---|---|---|
| measurements | goal | mission | goal | mission |
| Mean | 0.0299 | **0.0319** | *0.0321* | ***0.0481*** |

**Summary**. The above analysis provides several interesting findings for mobile search sequence structures. First, comparing to desktop search, we find that people are likely to search for simpler missions on mobile devices. However, we still observe that people tend to issue more queries on mobile. This suggests that mobile search requires further improvement on effectiveness. Second, we find that mobile search tends to have fewer goal/mission interleaving, which might be related to a lack of effective support for mobile multitasking. Future studies may focus on developing proper techniques/interfaces for search task switching. Third, observing the differences among search goals/missions/sessions, we believe that session-based search analysis in prior studies should be re-examined at a more fine-grained level such as on top of search goals and missions.

# 6. MOBILE SEARCH EVALUATION WITH GOALS AND MISSIONS

To understand the usefulness of the hierarchical query sequence structure, this section explores its application in defining better user satisfaction metrics (e.g., search success and search abandonment). Existing user satisfaction metrics are mostly measured at the query or session level. With the identified goals and missions, we are able to measure user satisfaction at four different levels (query, goal, mission and session) and make comparisons. Here, we focus on metrics for mobile search in particular.

## 6.1 Experiment Setup

To understand the effectiveness of a user satisfaction metric, the canonical method is to manually label a large set of search queries with user satisfaction levels [6, 9]. A better user satisfaction metric should be highly correlated with the labeled data. However, this is hard to scale. Our paper adopts an alternative approach by correlating the defined user satisfaction metrics with users' long-term engagement metric [1]. The underlying assumption is that if users are satisfied with the search in the short term, they are more likely to engage with the search system in the long run. Therefore, a strong correlation with the long-term engagement indicates a good user satisfaction measurement.

A typical metric that measures long term engagement is number of sessions per day per user. This metric is not very sensitive and usually needs to be tracked for a long term. To conduct the experiment, we tracked individual users in a stable, controlled experimentation environment for five weeks in the Yahoo search engine. We then used the first week as the qualifying period to compute the short-term user satisfaction metrics such as query success rate and query abandonment rate for each user. The subsequent four weeks were used to compute the long-term engagement metric - number of sessions per day per user. We then looked at the correlation between the short-term and long-term metrics, which is indicated by the goodness-of-fit (we use $R^2$) in the regression model. Higher $R^2$ indicates a stronger correlation thus better predictive power. Through this study, we can compare the goal/mission level user satisfaction metrics against the query/session level user satisfaction metrics to find out if they can predict long-term engagement more accurately.

## 6.2 Measuring Search Success

A widely used signal for measuring search success is click with dwell time longer than 30 seconds (i.e., long-dwell click). Although this cutoff is derived from desktop search logs, it is widely used in mobile search as well [18]. Thus, we set 30 seconds as the cutoff for long dwell click in our experiments. We consider the following four metrics to measure search success at the query, goal, mission and session levels, respectively. Then, we compare the $R^2$ of each search success metric in predicting the long term engagement metric (Table 6).

- query success rate: #queries with long-dwell clicks divided by total #queries.
- goal success rate: #goals for which the last query has long-dwell clicks divided by total #goals.
- mission success rate: #missions for which the last query has long-dwell clicks divided by total #missions.
- session success rate: #sessions for which the last query has long-dwell clicks divided by total #sessions.

According to Table 6, all of the four search success metrics have significant correlations with the long term engagement metric. Among the four metrics, goal success rate achieves the best performance for the long term engagement. The relatively low performance of query success rate is likely due to the unmeasured presence of query reformulations. Researchers have found that query reformulation can indicate satisfaction more accurately than pure clicks [6]. If a following query is a similar query submitted within a short time interval, even though the current query has a long dwell click, it may still indicate dissatisfaction because the user has not completed the task. Therefore, measuring goal success rate using the last query in a goal has the best correlation with the long term metric. The mission and session success rate also show strong correlations with long-term engagement. However, they do not provide any further benefits compared to the goal success rate.

**Table 6: Correlation analysis of different success and abandonment metrics for predicting long term engagement. $*/\diamond/\dagger$ denotes a significant increase comparing to query/session/mission success rate (or session/query/goal abandonment rate).**

| Level | $R^2$ (success rate) | $R^2$(abandonment rate) |
|---|---|---|
| query | *0.6738* | 0.7404 $\diamond$ |
| session | 0.7338 $*$ | *0.6949* |
| goal | 0.7485 $* \diamond \dagger$ | 0.7627 $* \diamond$ |
| mission | 0.7473 $* \diamond$ | 0.7845 $* \diamond \dagger$ |

## 6.3 Measuring Search Abandonment

Search queries without any result clicks, called search abandonment, is often viewed as an important behavior signal for measuring user dissatisfaction. However, not all abandonment is bad. When users are able to get direct answers from a search result page, they may be satisfied without the need to click on any results [13]. When defining search abandonment for user dissatisfaction, we remove all the queries with good abandonment by examining the direct answer on the search result page. As in our analysis of search success, we define four search abandonment metrics at the query, goal, mission and session level respectively. We then examine the $R^2$ of each search abandonment metric with the long term engagement metric. The search abandonment metrics we investigated:

- query abandonment rate: #queries with no click divided by total #queries.
- goal abandonment rate: #goals for which the last query has no click divided by total #goals.
- mission abandonment rate: #missions for which the last query has no click divided by total #missions.
- session abandonment rate: #sessions for which the last query has no click divided by total #sessions.

The results are shown in Table 6. All four abandonment metrics show significant correlations with long-term engagement. Among the four metrics, we observe that the session abandonment rate has the lowest correlation with long term engagement. This finding might be because the traditional session definition does not measure task boundaries accurately. When segmenting query sequences more precisely with goals or missions, we do find significant improvements. In addition, if a no-click query is the last query in a goal, but not the last query in a mission, it's possible that users may still get some information from the search results which inspires the next search goal in the same mission. When the last query in a mission has no click, it's very likely that the user has given up searching for related information. Thus, the mission abandonment rate is better aligned with users' search satisfaction, and thus their long term engagement with the search engine.

# 7. CONCLUSION AND FUTURE WORK

This paper developed an automatic algorithm to identify query-goal-mission structures from mobile query sequences. The algorithm can effectively extract and incorporate features derived from both mobile and desktop search logs, and produced good performance on multiple datasets. We further applied this algorithm in a web-scale search logs, and found several significant differences in goal/mission length, goal/mission duration and goal/mission interleaving behaviors between mobile and desktop. Compared to desktop search, mobile users tend to handle fewer and simpler search missions within a session, issue more queries within a goal and a mission, and exhibit decreased goal/mission interleaving behaviors. We also examined the utility of the identified hierarchical query sequence structure by applying it to define user satisfaction metrics. We find that the goal-based search success rate and mission-based abandonment rate can predict the long-term user engagement more accurately than the query and session-based metrics.

Defining user satisfaction metrics is only one application of the hierarchical structure. The hierarchical query structure can also be applied to other problems, such as re-ranking relevant search results, recommending and suggesting search queries. We would like to explore these topics in the future. We additionally plan to study this same query structure in a cross-session, rather than within-session, context.

# 8. REFERENCES


[1] A. Drutsa, G. Gusev, and P. Serdyukov. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *WWW*, pages 256–266, 2015.

[2] H. Feild and J. Allan. Task-aware query recommendation. In *SIGIR*, pages 83–92, 2013.

[3] D. Guan, S. Zhang, and H. Yang. Utilizing query change for session search. In *SIGIR*, pages 453–462, 2013.

[4] S. Han, D. He, Z. Yue, and P. Brusilovsky. Supporting cross-device web search with social navigation-based mobile touch interactions. In *UMAP*, pages 143–155, 2015.

[5] S. Han, Z. Yue, and D. He. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. *TOIS*, 33(4):16, 2015.

[6] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *CIKM*, pages 2019–2028, 2013.

[7] D. He, A. Göker, and D. J. Harper. Combining evidence for automatic web session identification. *IP&M*, 38(5):727–742, 2002.

[8] J. Jiang, S. Han, J. Wu, and D. He. Pitt at trec 2011 session track. In *TREC*, 2011.

[9] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM*, pages 57–66, 2015.

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

[11] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *SIGIR*, pages 1053–1062, 2011.

[12] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR*, pages 113–122. ACM, 2014.

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

[14] L. Li, H. Deng, Y. He, A. Dong, Y. Chang, and H. Zha. Behavior driven topic transition for search task identification. In *WWW*, pages 555–565, 2016.

[15] Z. Liao, Y. Song, L.-w. He, and Y. Huang. Evaluating the effectiveness of search task trails. In *WWW*, pages 489–498, 2012.

[16] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Discovering tasks from search engine query logs. *TOIS*, 31(3):14, 2013.

[17] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[18] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *SIGIR*, pages 695–704. ACM, 2015.

[19] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *WWW*, pages 1201–1212, 2013.

[20] Y. Wang, X. Huang, and R. W. White. Characterizing and supporting cross-device search tasks. In *WSDM*, pages 707–716. ACM, 2013.

[21] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang. Enhancing personalized search by mining and modeling task behavior. In *WWW*, pages 1411–1420, 2013.