How Much is Too Much? Whole Session vs. First Query Behaviors in Task Type Prediction

Matthew Mitsui
Department of Computer Science
Rutgers University
New Brunswick, New Jersey, USA
mmitsui@cs.rutgers.edu

Jiqun Liu, Chirag Shah
School of Communication & Information
Rutgers University
New Brunswick, New Jersey, USA
{jl2033,chirags}@rutgers.edu

ABSTRACT

One of the emerging and important problems in Interactive Information Retrieval research is predicting search tasks. Given a searcher's behavior during a search session, can the searcher's task be predicted? Which aspects of the task can be predicted, and how quickly and how accurately can the prediction be made? Much past literature has examined relationships between browsing behavior and task type at a statistical level, and recent work is moving towards prediction. While one may think whole session measures are useful for prediction, recent findings on common measures have suggested the contrary. Can less of the session still be useful? We examine the opposite end: the first query. Using multiple data sets for comparison, our results suggest that first query measures can be at least as good as – and sometimes better than – whole session measures for certain task type predictions.

CCS CONCEPTS

• **Information systems → Task models**; **Retrieval tasks and goals**; *Personalization*;

KEYWORDS

interactive information retrieval; task type; prediction; statistical significance; user behavior

ACM Reference Format:
Matthew Mitsui and Jiqun Liu, Chirag Shah. 2018. How Much is Too Much? Whole Session vs. First Query Behaviors in Task Type Prediction. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3209978.3210105

1 INTRODUCTION

For decades, Interactive Information Retrieval (IIR) researchers have recognized that people ultimately search in order to accomplish a task. Searchers interact with a search system through a series of actions – such as entering query terms and browsing results – to accomplish a larger goal. IIR researchers have taken to discovering how they can assist searchers in accomplishing their goals. This has been done by examining relationships between searchers' behaviors and the tasks they are trying to accomplish. Interest in task has also manifested in efforts like the TREC Tasks Track.[1] It is a standard by which researchers can tackle various aspects of the task prediction problem, such as predicting the task (or sub-tasks) of a person's search as early as possible.

Intuitively, a searcher's behavior across an entire session is the best indicator the searcher's task. The whole session offers the most complete information of a searcher's activity. Much past work has in fact related whole session behavior to aspects of a user's task [9, 13]. This is not the most realistic type of data to use for prediction and associated applications (e.g., real-time recommendations). Ideally, a task prediction should be made as soon as possible; predictions in the middle of a session can be used in real-time applications, such as result re-ranking and recommendation. It has been shown that information about task mid-session can be useful in ranking [11]. Yet whole session prediction serves as a reasonable baseline to start with task type prediction. Can we predict the task type of an unseen session given the information from previous sessions?

Recent work has suggested the contrary with common whole-session measures [15]. However, some other recent statistical work has drawn relationships between the first query and a user's task (e.g., [2]). A searcher's first query provides a fraction of the information given by the searcher's whole session. On the other hand, the first query is a more realistic moment for real-time recommendation and is one of the earliest such possible moments. Is the first query useful? We ask the following research question:

- To what extent do whole session or first query features outperform each other in predicting a search session's task type?

In this work, we find a somewhat surprising answer. We show on multiple datasets that first query measures perform at least as well as common whole session measures for task type prediction. Sometimes, first query measures even perform better.

2 TASK TYPE, FIRST QUERY, AND WHOLE SESSION BEHAVIOR

Several models characterizing a user's "task" have arisen in information seeking literature. In all of these, the user is a central actor. The user brings some search task when they begin to search for information – e.g., on a search engine. Search tasks are influenced by the work task that drives them or are associated with a problematic situation [3]. They are also driven by intentions and can be well-defined or ill-defined [8]. Tasks have varying levels of complexity [4].

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210105

[1] http://www.cs.ucl.ac.uk/tasks-track-2017/

Some facets of a task are cognitive, such as the perceived difficulty of a task. Some influence cognition [7], such as the interdependence between subgoals of a task. However, it is still believed that important aspects of the task can be captured and characterized through externally observable browsing behavior, such as typing, querying, and scanning through results. Some research in this thread has taken the form of eye tracking research [6, 13], yet since the people searching on computers do not commonly have eye trackers, a large portion of task-related research still analyzes log data. This data can be noninvasively captured while a searcher is browsing, e.g., from a Chrome browser, and includes implicit feedback such as clicking and dwelling on pages or explicit feedback such as bookmarking pages.

Extensive literature has compared behaviors over completed search sessions to their task types. The presumption is that accomplishing more complex tasks may require more complex actions that are manifested throughout the duration of the session. For instance, more complex tasks may generally take longer or require more queries, and indeed, behaviors such as query length, session time, and number of URLs per query can distinguish task type [9, 13]. Such work is based on finding statistically significant differences between behaviors on a controlled set of distinct task types. Similarly, work has examined the relationships between browsing behavior and task difficulty [1, 12] as well as users' topic knowledge [14]. Little work has been done in seeing how useful such features would be in classifying the task type of unseen sessions, but recent work has suggested that the utility of such features may be limited in classification [15].

Contrastingly, some work has compared first query behavior for task type, using statistics. [2], for instance, examined the statistical relationships between first query behaviors - such as the query length, number of clicks, and maximum scroll depth - and various types of exploratory and lookup tasks, and they found that such tasks can be distinguished by subsets of these behaviors. Can first query features be used in the predictive setting? If so, do they offer more success than the whole session features? [9] has suggested that for some task types, there is a significant difference between whole session behaviors and first query behaviors, so perhaps there is at least a predictive difference between the two. We seek to address this open question in the following sections.

3 METHOD AND DATASET

For task type prediction, we chose to predict the goal and product of the task, in the sense defined in [10]. The goal is a binary classification of "specific" or "amorphous", analogous to the well-defined or ill-defined goals in [8]. The product is either "intellectual" – producing new findings, or "factual" – locating facts or data. These task types are largely dependent on a task specification that can be controlled by an experimenter. Moreover, pairing these products and goals yields 4 possible task types: "known-fact search" ("factual"/"specific") , "known-subject search" ("factual"/"amorphous"), "interpretive search" ("intellectual"/"specific"), and "exploratory search" ("intellectual"/"amorphous").

We ran experiments across 2 datasets. Multiple experiments with similar statistically significant results – e.g. experiments showing a relationship between smoking and cancer – can provide a replication, or at least strong evidence, for a phenomenon. Likewise, we hoped to demonstrate the strength of the relationship between whole session and first query with multiple datasets. Our first dataset is the TREC 2014 Session Track [5], which has been used in the evaluation of retrieval over the course of a search session. This data is comprised of search logs of users conducting searches over sessions, rather than performing ad-hoc retrieval. The data comprises 1,257 sessions, with 260 unique users conducting searches across 60 different topics (15 unique prompts per task type). We used 1,021 of these sessions; the others do not include a current or final query and cannot be applied for our work [5]. The searchers in the TREC data were recruited through Mechanical Turk. In contrast, our second data is from a traditional lab study, composed of users conducting 2 sessions in a controlled environment in a university setting. This is a common setting in IIR research to explore relationships between task and behavior, as with [6, 7, 13]. The participants were undergraduate journalism student and were given journalism-based task prompts across 4 task types following the same faceted classification: copy editing (factual/specific), story pitch (factual/amorphous), relationships (intellectual/amorphous), interview preparation (intellectual/amorphous). There are 80 sessions total: 22 copy editing, 18 story pitch, 18 relationships, 22 interview preparation. 40 participants conducted 2 sessions each. The task prompts can be found in [16].

We predict the task type (goal and product) using either first query features or whole session features listed below. We conduct traditional machine learning classification experiments, with the task product, goal, or type of a session as the classification label. Hence, we had 1,021 total data points and 80 data points for the TREC and journalism data, respectively. The TREC data contained 513 amorphous tasks, 508 specific tasks, 529 factual tasks, and 492 intellectual tasks. The journalism data contained 58 amorphous tasks, 22 specific tasks, 40 factual tasks, and 40 intellectual tasks. We compared several machine learning classifiers against two naive baselines. Each of these are explained in Table 1. For both datasets, we used 80% training data and 20% test data.

While a rich set of features can be extracted from a controlled laboratory study, we limited ourselves to first query and whole session features that could be extracted from both the TREC 2014 Session Track data and journalism data, to come as close to creating a genuine replication as possible. Since we are examining the effects of first query and whole session features in task type prediction, we use the same features in our 2 experiments on our 2 datasets. The features are as follows:

First query features

- Query length, total dwell time on SERPs and content pages, % time on SERPs. These are directly drawn or derived from [2], and associated with differences between task types.
- # pages visited. This is a specific case of the # pages visited over a session [13].

Whole session features

- # pages, # queries, and completion time. These distinguish task types in [13].
- Dwell time on content pages per SERP, % time on SERPs, dwell time on SERP per query, dwell time on content pages per query, total dwell time on SERP pages, total dwell time

Table 1: Accuracy on the TREC data set for several algorithms: AdaBoost (ADA), Decision Tree (DCT), Naive Bayes (GNB), k-nearest neighbors (KNN), multilayer perceptron (MLP), and Support Vector Machine (SVM). Baselines are a most frequent (MFQ) and stratified random (STR) baseline. Best performers in each column are boldfaced. Significant values indicate 1) whether the predictor is significantly better than its baseline (*=p<.05,**=p<.01) and 2) whether the whole session best predictor is significantly better than the first query best predictor, or vice versa. ($\dagger = p < .05, \dagger\dagger = p < .01$)

	Prediction Type & Feature Set					
	Goal		Product		Type	
Classifier	First Query	Session	First Query	Session	First Query	Session
ADA	**0.532****††	**0.504**	0.549	0.516	0.288	0.245
DCT	0.529	0.493	0.531	0.525	0.270	0.255
GNB	0.520	0.498	0.558	0.505	0.267	0.261
KNN	0.522	0.501	0.560	**0.544****	**0.295****††	0.255
MLP	0.507	0.479	**0.562****††	0.541	0.281	0.262
SVM	0.526	0.495	0.558	0.539	0.287	**0.266****
MFQ	0.474	0.476	0.515	0.514	0.248	0.250
STR	0.494	0.495	0.498	0.492	0.251	0.251

Table 2: Accuracy on the journalism data. Keys are identical to Table 1. ($* = p < .05, ** = p < .01, \dagger = p < .05, \dagger\dagger = p < .01$)

	Prediction Type & Feature Set					
	Goal		Product		Type	
Classifier	First Query	Session	First Query	Session	First Query	Session
ADA	0.641	0.679	0.595	0.536	0.442	0.493
DCT	0.582	0.651	0.594	**0.556****	0.443	0.440
GNB	0.654	0.698	0.628	0.513	0.474	0.455
KNN	0.623	0.715	0.631	0.514	0.438	**0.500**
MLP	0.672	**0.738***	**0.656****††	0.474	0.480	0.463
SVM	0.685	0.709	0.638	0.444	**0.501**	0.498
MFQ	**0.715**	0.708	0.423	0.428	**0.501**	**0.500**
STR	0.615	0.595	0.510	0.490	0.373	0.363

on content pages. These are directly drawn from or derived from [9].

- Average dwell time on content pages, pages per query, average query length, range of query lengths.

4 RESULTS

Tables 1-2 show accuracy scores for different classifiers, using our feature sets to predict task product, goal, and type. In our subsequent analyses, we located the best classifier for a given task type and feature set. For instance, for predicting task product on TREC data using first query features, the multilayer perceptron classifier was best and was significantly better than the stratified baseline (**). We then compared the best classifiers on first query vs. whole session features to see which performed better. For instance, a multilayer perceptron on first query features performed significantly better than k-nearest neighbors with whole session features (††). Our results show the following 3 simple conclusions:

(1) **Prediction with first query features can be more accurate than with whole session features.** – This happens in all cases in both datasets excepting only the task goal in the journalism data. Here, whole session features obtained 0.738 accuracy, versus 0.715 accuracy for first query features.

(2) **Additionally, first query features can be significantly more accurate.** - This happens for task product in both datasets ($p < .01$), providing strong evidence that first query features are generally better for predicting whether a product is factual or intellectual. This only happens once with goal and once with task type in the TREC dataset ($p < .01$).

(3) **If #1 or #2 above do not hold, whole session features are still not significantly more accurate than first query features.** - Our result for task goal in the journalism data (0.738 whole session vs. 0.715), violates #1. Similarly for first query features on task type (0.5 whole session vs 0.501). In both cases, the best whole session predictor is not significantly better than the best first query predictor, despite differences.

These conclusions are applicable across task type, product, and goal on both data sets. For task product, first-vs-whole significant differences were duplicated across data sets. One possible reason is because the first query features help determine whether a person is locating facts or data. For instance, a person locating facts or data may issue long, specific queries or only dwell briefly on a few content pages, scanning for facts. Another consideration is that information about individual query lengths or dwell times is lost by averaging across a session - a point we will briefly address in the conclusion. For task goal and type, first-vs-whole significance values were not duplicated across the data sets. However, we never once draw opposite conclusions – i.e., first is significantly better than whole in one dataset significantly worse in the other. Extra datasets are required to draw a stronger conclusion, though we believe our 3 findings above show that the first query measures we listed are *better than whole session features* in predicting task product and *at least as good as whole session features* in predicting goal and type. This has implicates that information the first query – one of the earliest possible times help a user – perhaps useful for task-based recommendation and re-ranking, and first query information is more useful than the whole session when distinguishing fact-finding tasks from intellectual ones.

5 CONCLUSION AND FUTURE WORK

Past work has drawn various relationships between whole session behavior, a searcher's first query, and the type of their search task. Here, we expanded on such work by showing the relative utility of different features in predicting task type on unseen sessions. In particular, we explored the relative utility of averages over a session, totals over a session, and special cases of such features in the first query segment, which have been popular in past work. We found that the first query segment is consistently at least as useful as the whole session, and sometimes better.

The next natural step is to explore the range in between the first query and the whole session. Perhaps the relationship between the first query and whole session is not linear. Perhaps the two queries are better than one and three are better than two, and some threshold point is reached in the middle until noisy data produces features that are difficult to use. Noise can pollute averages and totals, hence polluting whole session feature calculations. The next natural step is to use more complex features. Our first query and whole segment features, while drawn from literature, are simple measures, averages, or totals. Since information is lost in averages and totals, the next extension would be pairwise features, such as the frequencies and types of transitions between subsequent pages and queries. These would make a 1st order Markov model. Larger orders of Markov models can be used, but with increasing complexity of the predictor, more data is required. [6] used a 1st order Markov chain on a small set of "eye tracking states" to distinguish task types (again in the statistically significant sense). Perhaps Markov models can be used to model transitions between "query segment types" or "page types" and use them for prediction.

[9] showed that the whole session may be significantly different from the first query. Perhaps a weighted model could trade off the importance of first query features and whole session features, with their weights being a function of time. The importance of different features may fluctuate throughout a session.

This work is not an exact replication. To be a replication, users from both data sets need to be drawn from similar populations and exposed to identical tasks, with the tasks randomly but evenly assigned among users, such as with a Latin square design. Such replications are difficult or rare in information retrieval studies, though we designed this experiment to illustrate a point rather than to create a clean replication. We showed similar results on distinct data sets of homogeneous search behavior over similar task types. On datasets on different search topics, capturing similar behaviors over similar task types, we showed that the first query can be at least as good as – and sometimes better than – the whole session.

6 ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) grant IIS-1717488.

REFERENCES

[1] Jaime Arguello. 2014. Predicting Search Task Difficulty. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416 (ECIR 2014)*. Springer-Verlag New York, Inc., New York, NY, USA, 88–99. https://doi.org/10.1007/978-3-319-06028-6_8

[2] Kumaripaba Athukorala, Dorota GÅĆowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651. https://doi.org/10.1002/asi.23617

[3] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *Journal of the Association for Information Science and Technology* 56, 10 (2005).

[4] D. J. Campbell. 1988. Task Complexity: A Review and Analysis. *Academy of Management Review* 13, 1 (1988).

[5] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. [n. d.]. Overview of the TREC 2014 Session Track. ([n. d.]).

[6] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2015. User Activity Patterns During Information Search. *ACM Trans. Inf. Syst.* 33, 1, Article 1 (March 2015), 39 pages. https://doi.org/10.1145/2699656

[7] Ashlee Edwards and Diane Kelly. 2017. Engaged or Frustrated?: Disambiguating Emotional State in Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

[8] Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc.

[9] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval (SIGIR '14)*. ACM.

[10] Yuelin Li and Nicholas J Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management* 44, 6 (2008), 1822–1837.

[11] Chang Liu, Nicholas J. Belkin, and Michael J. Cole. 2012. Personalization of Search Results Using Interaction Behaviors in Search Sessions. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM.

[12] Chang Liu, Jingjing Liu, and Nicholas J. Belkin. 2014. Predicting Search Task Difficulty at Different Search Stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 569–578. https://doi.org/10.1145/2661829.2661939

[13] Jingjing Liu, Michael J. Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J. Belkin, Jun Zhang, and Xiangmin Zhang. 2010. Search Behaviors in Different Task Types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. ACM.

[14] Jingjing Liu, Chang Liu, and Nicholas J. Belkin. 2016. Predicting information searchers' topic knowledge at different search stages. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2652–2666. https://doi.org/10.1002/asi.23606 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23606

[15] Matthew Mitsui, Jiqun Liu, and Chirag Shah. 2018. The Paradox of Personalization: Does Task Prediction Require Individualized Models?. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 277–280. https://doi.org/10.1145/3176349.3176887

[16] Matthew Mitsui, Chirag Shah, and Nicholas J. Belkin. 2016. Extracting Information Seeking Intentions for Web Search Sessions. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 841–844. https://doi.org/10.1145/2911451.2914746