Predicting Information Seeking Intentions from Search Behaviors

Matthew Mitsui
Department of Computer Science
Rutgers University
mmitsui@cs.rutgers.edu

Jiqun Liu, Nicholas J. Belkin, Chirag Shah
School of Communication & Information
Rutgers University
{jl2033,belkin,chirags}@rutgers.edu

ABSTRACT

It has been shown that people attempt to accomplish a variety of intentions during the course of an information seeking session, and there is reason to believe that these different information seeking intentions can benefit from system support tailored to each such intention. We address the problem of predicting the presence of such intentions during an information seeking session, through analysis of observable user search behaviors. We present results of a study of 40 participants, each working on two different journalism tasks, which investigated how their search behaviors could indicate their intentions. Using 725 query-segments captured from this study, we demonstrate that information seeking intentions can be predicted with a simple classification model using a linear combination of search behavior features that can be logged with a browser plug-in.

CCS CONCEPTS

•Information systems →Query intent; Task models;

KEYWORDS

Search intentions; Information seeking intentions; Motivating task; Information seeking episode; Search session analysis

1 INTRODUCTION

The general problem with which this paper is concerned is the identification and understanding of the different "things" an information seeker is attempting to accomplish during the course of an information seeking session, as in interaction with a Web search engine. Previous research (e.g. [9, 10]) has shown that people *intend* to accomplish different things at different times during the course of an information seeking episode. These may include, *inter alia*, learning about a topic, learning about the contents of a database, formulating a query for a specific information object, recognizing objects of interest, evaluating information objects, comparing information objects, saving objects, and so on. We, in common with others [7, 9], call these *information seeking intentions*. As [10] has shown, different search intentions can benefit from different support techniques. However, current search engines are primarily designed to support only one search intention and its associated technique, searching for information objects, by specifying their characteristics through query formulation, system response, and query reformulation.

We propose that, to better support user's interactions with information, information retrieval (IR) researchers need to go beyond the single query-single response pattern, and to develop interaction (and other) techniques that are tailored to the specific search intention at hand. To accomplish this goal, the first, minimal, step is to enable IR systems to "understand" users' information seeking intentions; that is, the goals behind their visible search behaviors, such as query formulation, browsing, and clicking. To do this, researchers must be able to identify, or predict, the user's intention based on the observable behavioral features. This last is the problem which the research reported here addresses: identifying information seeking intentions through observation of information seeking behaviors during the course of an information seeking session. The results of this study can be exploited in characterizing user's intentions, as well as applications in personalization of IR systems and recommendation system design.

2 METHOD AND DATASET

We conducted a study in which we recruited 40 participants from undergraduate journalism courses in a US university, who were required to have taken at least one course in news writing. Participants were brought to a lab and assigned two different journalism work tasks selected from a total of four such tasks, each with an associated information seeking task. These tasks were constructed according to a modification of the faceted task classification scheme of [4], in which the facet values were varied as indicated in Table 1. After completion of each search task, participants reviewed a video of their search behavior and were asked to annotate each query segment (all that occurred between one query and the next (or task completion), including the queries) with their information seeking intentions for that query segment. Participants were shown a list of 20 search intentions, organized into nine types, and were instructed to mark which intention(s) (or "other") they had during that query segment. Our dataset is similar to that of [7]. We used identical tasks and the same intentions, as follows: identify something to get started (**IS**), identify something more to search (**IM**), learn domain knowledge (**LK**), learn database content (**LD**), find a known item (**FK**), find specific information (**FS**), find items sharing a named characteristic (**FC**), find items without predefined criteria (**FP**), keep record of a link (**KR**), access a specific item (**AS**), access items with common characteristics (**AC**), access a web site/home page or similar (**AP**), evaluate correctness of an item (**EC**), evaluate usefulness of an item (**EU**), pick best item(s) from all the useful ones (**EB**), evaluate specificity of an item (**ES**), evaluate duplication

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan.

DOI: http://dx.doi.org/10.1145/3077136.3080737

Table 1: Task characteristics.

Task	Product	Level	Goal	Named	$\|Q\|$
CPE	Factual	Segment	Specific	True	214
STP	Factual	Segment	Amorphous	False	117
REL	Intellectual	Document	Amorphous	True	160
INT	Intellectual	Document	Amorphous	False	234

of an item (**ED**), obtain specific information (**OS**), obtain part of the item (**OP**), obtain a whole item(s) (**OW**). All browsing behavior was collected through a browser plugin in Mozilla Firefox. We classified the same intentions described in [7] and use the same naming convention throughout. Please refer to that paper for a full description.

We treated the problem of identifying intentions for a query segment as a binary classification problem, in which a positive label indicated that an intention was present in a query segment and a negative intention indicated that the intention was absent. The rightmost column of Table 1 shows the number of queries that were issued by all participants, per task. The tasks were: Copy Editing (CPE), Story Pitch (STP), Relationships (REL), and Interview Preparation (INT). We treated each query segment as a data point in the classifier, totaling 725 data points/queries. For each query segment they conducted, participants self-reported whether the intention was present. We found that the correlation between intentions within the same query segment was moderate (with "moderate" as a Pearson correlation of 0.4) for 3 pairs of intentions and less than moderate for the rest. We show this in a truncated Table 2 but do not show the full correlation matrix here for brevity. We therefore simplified our classification problem by training one classifier per intention, totaling 20 classification problems.

As this is a novel classification problem, there are no established machine learning baselines for it. We therefore started with the following simple baselines:

- Stratified (random) sampling of positive/negative labels proportional to their distribution in training data (STR)
- Assigning the most frequent label in the training data (MFQ)

These baselines are based on the intuition that intentions have a prior probability. If we can show that a learning algorithm incorporating browser-based features can improve over this baseline, we can show that there is a relationship between intentions and behavior that is not simply random. For our classifier, we used simple logistic regression on features collected through the Firefox browser as the participant conducted the task. We used feature values accumulated from the start of a query segment (i.e., when the user first issued the query) up until the next query (or for the last query, the end of the segment). Browser-level features have been useful for distinguishing task type [5] and task difficulty [6]. We hence calculated the following browser-based features:

- Bookmark features (BK) – Whether a bookmark was saved in the current segment, and the number of bookmarks saved in the segment
- Content page features (CP) – Mean dwell times on content pages- i.e., non-search engine result pages (SERPs). Dwell time types included total dwell time, total dwell time until a page was saved, total open time, total open time until a page was saved, and first dwell time [1]. Content pages were divided into pages that were saved, unsaved, and never saved, and total pages.
- Bookmark and content page features (CP+BK) – Combining BK and CP
- SERP features (SP) – Dwell times (as in CP) for search engine result pages (SERPs)
- Query features (QU) – The query length (number of words) and query reformulation type (labeled by an expert assessor by the scheme of [8])
- SERP and query features (SP+QU) – Combining SP and QU features
- All features (ALL) – Combining all features

Table 2: Truncated table of Pearson correlations between intentions *i*,*j* within a query segment. Showing the top 5 and bottom 3 in absolute value.

Intent i	Intent j	|Pearson|
FC	AC	0.476
ES	EU	0.414401
LK	KL	0.404794
KL	EU	0.387237
AS	ES	0.373349
LK	FS	0.018916
LK	OS	0.012831
FS	AP	0.004167

We treated each of the 725 query segments as independent data points for classification, even though all users conducted multiple query segments during their search tasks. We do not do explicit personalization, though some may occur by training on a few of a participant's query segments and testing on that participant's other queries. Explicit personalization is a source of future work. For classification, we randomly selected 66% of the data as training and the remainder as test data for evaluation. This is a rather modest amount of data to use as training. We will show that we already significantly improve over the baseline, and we believe a higher amount of training data would give us more improvements.

3 RESULTS

For evaluation, we computed several evaluation metrics. To discuss them, we will use the notions of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

- **Accuracy:**

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- P_{Pres} **(precision for intention being present):**

$$P_{Pres} = \frac{TP}{TP + FP} \tag{2}$$

- P_{Abs} **(precision for intention being absent):**

$$P_{Abs} = \frac{TN}{TN + FN} \tag{3}$$

Our evaluation metrics were influenced by the large variance in the percentage of positive labels in our data. Not every intention is present in every query, and some intentions are much more frequent than others. As shown in our results, many of the labels are largely imbalanced, ranging from 7.5% positive to 57.1% positive. Hence while accuracy gives us an overall measure of performance

improvement, it is an insufficient measure. In the case of rare positive labels, we must see how well positive labels are assigned.

Our results (Tables 3 and 4) show the best performance among our subsets of features and compare them to each baseline. We compared the significant differences between regression classifier performance and performance with the better of the two baselines using the Kolmogorov-Smirnov significance test. All such differences were significant at $p < .01$ with a few exceptions: for accuracy, OP, OW, LD (not significant) and FW ($p < .05$); and for P_{Pres}, LD, OW (not significant), FW, and OP ($p < .05$).

As a first finding, we make gains in accuracy across all intentions except OP, in which accuracy is equal. While the some of the significant gains in accuracy are modest (within 1%−3%), substantial gains are made in IS, KR, and LK of 13%, 5%, and 5%, respectively. Accuracy is a way to gauge overall performance improvement in a classification problem, and we see that logistic regression is a more accurate classifier overall than our baselines.

While these results are promising, the classifier may make various trade-offs, particularly where very small accuracy gains are made. It could trade off correct negative labels for correct positive labels, or vice versa. However, we first see that with P_{Abs} gains are made across all intentions in correctly classifying true negatives, increasing the number of true negatives. Even in places where accuracy gains are not significant, there are non-significant gains in P_{Abs}. So better negative predictions are made. Also, gains in P_{Pres} are rather large – and also significant – for all intentions. When interpreting this extreme result, we considered the ratio of positive and negative labels (i.e., an intention being present or absent). This is a crucial point. The ratio determined whether the MFQ baseline would label all predictions as 0 (giving a P_{Pres} score of 0 by definition). But STR – the better baseline overall – was more liberal about applying a positive label. Specifically, the P_{Pres} score of STR is proportional to the percent of positive labels. If a classifier makes fewer positive predictions but tends to get them right, this will increase the P_{Pres} scores drastically. We believe this is happening for the regression P_{Pres} scores.

Lastly, one important note should be made in the number and choice of features shown in Tables 3 and 4. Overall – with a few exceptions - the best accuracy is achieved with all features, as is the best P_{Abs} score. As the majority of intention labels are the negative/"absent" label, we take this to mean that using all features is good at making overall improvements in accuracy. The all-features classifier could artificially raise these scores by labeling all intentions as absent, but due to improvements over MFQ, we do not believe this to be the case. But more interestingly, our P_{Pres} results in Table 4 tell us that specific features are useful for indicating the presence of specific intentions, and pooling all features together into a linear classifier is insufficient to correctly determine that an intention is present. Content page behavior (dwell times) are important in determining whether a person is accessing a home page, evaluating duplicate information, or finding a known item. SERP and query behavior (SERP dwell times and query reformulation patterns) are most useful in determining whether a person is finding specific information. It also suggests a way of determining which intentions can be most reliably detected, given a set of available browser logging features.

4 CONCLUSIONS AND FUTURE WORK

We have shown that information seeking intentions can be predicted with a simple classification model using a linear combination of browser-based features. Our approach makes improvements over baselines based on priors. The suggested classifier features can be passively collected on a browser without the need for external equipment – such as an eye tracker – so can be deployed at a mass scale. Further, we have shown that while a linear classifier produces good overall performance, correctly predicting that an intention is present is better done by smaller subsets of features. Certain subsets – such as page dwell time, bookmarking behavior, and query reformulation behavior – are more predictive of some intentions than others. This can inform the implementer of a logging system how each logging capability affects the type of intentions they can capture and analyze. In the future, if it is shown for instance that the "find specific information" intention can help improve recommendations for specified queries or for optimizing NDCG@1 ranking, an interface that cannot capture this is inherently limited.

For future work, we would first like to incorporate additional features, which we have categorized into two types. The first type is a session-based feature. We treated each query segment in isolation in this work, but queries occur in the context of a session and a task. Simple session-based features include the number of previous queries, number of previous total bookmarks, and amount of time spent on the task. The second type of feature is cognitive features, such as eye tracking data. Cole et al. [2, 3] showed that there are consistent, distinguishable page-level and query segment-level reading patterns that can be used to differentiate between task types. Eye tracking features include statistics on reading speed and areas of interest. [2, 3] moreover showed that there is a relationship between eye-level reading patterns and browser-level page transition patterns. Since we found relationships between page-level features (e.g., dwell time) and intentions, it stands to reason that eye tracking behavior will also be useful for classification.

We can also explore interdependence between intentions. We found that there is moderate or less correlation among intentions within query segments. However, a user may commonly express two or more intentions together, or some sets of intentions may commonly occur together for one type of search task. Similarly, sequences of intentions can possibly identify "types" of searchers.

One of our ultimate goals is to understand the relationship between information seeking intentions, cognition, and task type to better understand and support search tasks. Can tasks be categorized by consistent sequences of intentions? How closely connected are intentions to browser-level activity and cognitive activity? And ultimately, how can these be used to improve search ranking and recommendation over current approaches? While this is all future work, we believe our work here to be a step in a positive direction towards task understanding and task-level recommendation.

5 ACKNOWLEDGMENTS

This work was supported through the National Science Foundation, grant #IIS-1423239.

REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International*

Intention	% Positive	ACC(feat)	ACC(STR)	ACC(MFQ)
Access Common (AC)	24.30	**0.761(CP)**	0.628	0.755
Access Page (AP)	10.80	**0.900(CP+BK)**	0.812	0.894
Access Specific (AS)	27.60	**0.731(ALL)**	0.598	0.721
Evaluate Best (EB)	19.80	**0.815(ALL)**	0.669	0.792
Evaluate Correctness (EC)	26.30	**0.754(ALL)**	0.610	0.735
Evaluate Duplication (ED)	7.50	**0.929(ALL)**	0.856	0.922
Evaluate Specific (ES)	23.80	**0.776(ALL)**	0.646	0.766
Evaluate Usefulness (EU)	25.30	**0.775(ALL)**	0.638	0.763
Find Characteristic (FC)	20.10	**0.806(ALL)**	0.675	0.797
Find Known (FK)	17.00	**0.832(ALL)**	0.705	0.825
Find Without Predefined (FP)	8.00	**0.926(ALL)**	0.858	0.922
Find Specific (FS)	57.10	**0.608(ALL)**	0.511	0.579
Identify More (IM)	37.50	**0.668(ALL)**	0.540	0.641
Identify Specific (IS)	29.00	**0.817(ALL)**	0.568	0.688
Keep Record (KR)	33.40	**0.714(ALL)**	0.551	0.659
Learn Database (LD)	16.20	**0.839(BK)**	0.729	0.837
Learn Domain Knowledge (LK)	33.20	**0.712(ALL)**	0.548	0.657
Obtain Part (OP)	18.90	**0.802(ALL)**	0.679	**0.802**
Obtain Specific (OS)	43.20	**0.645(ALL)**	0.513	0.581
Obtain Whole (OW)	8.30	**0.918(CP)**	0.850	0.917

Table 3: Results for accuracy (ACC, larger is better), and the best feature set (feat). Baselines STR and MFQ are also given. Nearly all classification results are significant (see Results section).

Intention	% Positive	P_{Abs}(feat)	P_{Abs}(STR)	P_{Abs}(MFQ)	P_{Pres}(feat)	P_{Pres}(STR)	P_{Pres}(MFQ)
AC	24.30	**0.763(CP)**	0.753	0.755	**0.771(CP+BK)**	0.248	0.000
AP	10.80	**0.899(CP+BK)**	0.895	0.894	**1.000(CP)**	0.108	0.000
AS	27.60	**0.735(ALL)**	0.722	0.721	**0.820(QU)**	0.277	0.000
EB	19.80	**0.818(ALL)**	0.791	0.792	**0.752(ALL)**	0.207	0.000
EC	26.30	**0.757(ALL)**	0.736	0.735	**0.828(CP+BK)**	0.261	0.000
ED	7.50	**0.933(ALL)**	0.922	0.922	**0.868(CP)**	0.083	0.000
ES	23.80	**0.780(ALL)**	0.768	0.766	**0.656(ALL)**	0.235	0.000
EU	25.30	**0.781(ALL)**	0.762	0.763	**0.833(SP+QU)**	0.235	0.000
FC	20.10	**0.809(ALL)**	0.795	0.797	**0.969(CP+BK)**	0.207	0.000
FK	17.00	**0.834(ALL)**	0.821	0.825	**0.782(CP)**	0.172	0.000
FP	8.00	**0.926(ALL)**	0.922	0.922	**0.987(CP+BK)**	0.077	0.000
FS	57.10	**0.967(BK)**	0.422	0.000	**0.623(SP+QU)**	0.577	0.579
IM	37.50	**0.677(ALL)**	0.640	0.641	**0.649(CP)**	0.360	0.000
IS	29.00	**0.810(ALL)**	0.682	0.688	**0.845(ALL)**	0.315	0.000
KR	33.40	**0.723(ALL)**	0.658	0.659	**0.689(SP)**	0.345	0.000
LD	16.20	**0.841(ALL)**	0.837	0.837	**0.411(ALL)**	0.163	0.000
LK	33.20	**0.720(ALL)**	0.656	0.657	**0.886(BK)**	0.342	0.000
OP	18.90	**0.807(ALL)**	0.799	0.802	**0.561(CP+BK)**	0.196	0.000
OS	43.20	**0.661(ALL)**	0.580	0.581	**0.660(CP)**	0.421	0.000
OW	8.30	**0.918(CP)**	0.918	0.917	**0.761(CP+BK)**	0.084	0.000

Table 4: Results for P_{Abs} (larger is better), and P_{Pres} (larger is better), as well as the best feature set (feat). Baselines STR and MFQ are also given. Nearly all classification results are significant (see Results section).

ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pages 19–26, New York, NY, USA, 2006. ACM.

[2] M. J. Cole, J. Gwizdka, C. Liu, R. Bierig, N. J. Belkin, and X. Zhang. Task and user effects on reading patterns in information search. *Interacting with Computers*, 23(4):346 – 362, 2011. Cognitive Ergonomics for Situated Human-Automation Collaboration.

[3] M. J. Cole, C. Hendahewa, N. J. Belkin, and C. Shah. User activity patterns during information search. *ACM Trans. Inf. Syst.*, 33(1):1:1–1:39, Mar. 2015.

[4] Y. Li. *Relationships among work tasks, search tasks, and interactive information searching behavior*. ProQuest, 2008.

[5] J. Liu, M. J. Cole, C. Liu, R. Bierig, J. Gwizdka, N. J. Belkin, J. Zhang, and X. Zhang. Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 69–78, New York, NY, USA, 2010. ACM.

[6] J. Liu, C. Liu, M. Cole, N. J. Belkin, and X. Zhang. Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1313–1322, New York, NY, USA, 2012. ACM.

[7] M. Mitsui, C. Shah, and N. J. Belkin. 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, pages 841–844, New York, NY, USA, 2016. ACM.

[8] E. Y. Rha, M. Mitsui, N. J. Belkin, and C. Shah. Exploring the relationships between search intentions and query reformulations. *Proceedings of the Association for Information Science and Technology*, 53(1):1–9, 2016.

[9] H. I. Xie. Patterns between interactive intentions and information-seeking strategies. *Information Processing & Management*, 38(1):55 – 77, 2002.

[10] X. Yuan and N. J. Belkin. Supporting multiple information-seeking strategies in a single system framework. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 247–254, New York, NY, USA, 2007. ACM.