Deep Learning of Human Information Foraging Behavior with a Search Engine

Xi Niu
University of North Carolina at Charlotte
Charlotte, North Carolina
xniu2@uncc.edu

Xiangyu Fan
Endgame360 Inc.
Asheville, North Carolina
xyfan@alumni.unc.edu

ABSTRACT

In this paper, a two-level deep learning framework is presented to model human information foraging behavior with search engines. A recurrent neural network architecture is designed using LSTM as the base unit to explicitly consider the temporal and spatial dependencies of information scents, the key concept in Information Foraging Theory. The target is to predict several major search behaviors, such as query abandonment, query reformulation, number of clicks, and information gain. The memory capability and the sequence structure of LSTM allow to naturally mimic not only what users are perceiving and performing at the moment but also what they have seen and learned from the past during the search dynamics. The promising results indicate that our information scent models with different input variations were better, compared to the state-of-the art neural click models, at predicting some search behaviors. When incorporating the knowledge from a previous query in the same search session, the prediction of current query abandonment, pagination, and information gain has been improved. Compared to the well known neural click models that model search behaviors under a single search query thread, this study takes a broader view to consider an entire search session which may contain multiple queries. More importantly, our model takes the search result relevance pattern on the Search Engine Results Pages (SERP) as a whole as the information scent input to the deep learning model, instead of considering one search result at each step. The results have insights on the impact of information scents on how people forage for information, which has implications for designing or refining a set of design guidelines for search engines.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**;

KEYWORDS

Information Foraging Theory; online search behavior; deep learning; information scent;

ACM Reference format:
Xi Niu and Xiangyu Fan. 2019. Deep Learning of Human Information Foraging Behavior with a Search Engine. In *Proceedings of The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval, Santa Clara, CA, USA, October 2–5, 2019 (ICTIR '19),* 9 pages.
https://doi.org/10.1145/3341981.3344231

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344231

1 INTRODUCTION

Nowadays information technologies have created more information than any of us can realistically process. In order to cope with vast amount of information available online, many people rely on their perception of information value or just their intuition to guide them click on, read through, follow up, or allocate attention on different information sources. Such information seeking behavior is modeled by the well-known Information Foraging Theory [32], which borrowed concepts from animals' foraging for food in Biology to understand information seekers' acquisition of information. This Theory uses the concept of information scent to mimic the concept of food odor that animals use when foraging for food. Most previous studies, e.g., [8, 37], quantified information scent of each individual information item in a "flat" or "cumulative" way without considering the order effect of information scent along the information seeking process. However, temporal information was shown important to improve performance of user click models for Web searches [19]. Compared to the well-known neural click models, for example [5, 6, 38], which have adopted the "micro-level" view, focusing on user interactions under one query, we will design a two-level deep learning framework to model and predict both query-level and session-level interactions, by encoding information scent along the process under the guidance of Information Foraging Theory. In addition, instead of predicting just a binary click behavior, we will extend the target variables to other major search behavioral variables, either binary or numeric, such as query reformulation, number of clicks, and information gain.

Specifically, we operationalize the concept of information scent as the perceived relevance of a search result on Search Engine Results Pages (SERP). We designed a recurrent neural network architecture using LSTM as the the base unit to explicitly consider the temporal and spatial aspects of the relevance scores on a SERP. Our target is to predict several major search behaviors, such as query abandonment, number of clicks, and information gain. The memory capability and the sequence structure of LSTM [13] allow us to naturally mimic not only what users are perceiving and performing at the moment but also what they have seen and learned from the past during the search dynamics. In this architecture, a recurrent neural net is used to encode the SERP results' relevance scores and their positions. We then used another recurrent net to capture the learning effect of users. We showed that our model "remembers" and "forgets" like a human: the model memories the influence of a scanned item from past and that knowledge fades

away over time. Prior work models the learning effect as frequency counts such as number of clicks, which does not take into account the "remembering" and "forgetting" effect during a sequence.

The contribution of the study is: to our best knowledge, this study is the first to design a two-level (both query level and session level) neural network architecture to test the Information Foraging Theory.

2 LITERATURE REVIEW

This study connects three threads of work, Information Foraging Theory and its role in understanding search behavior, deep learning models for human behavior, and well-know click models.

2.1 Information Foraging Theory and Search Behavior

Information foraging theory (IFT) by the study [32] is a framework borrowed from the optimal foraging theory in Biology to understand information seekers' "stay" or "leave" decisions when facing today's flux of information. The goal of the decision is to maximize the information gain and minimize the cost of the forager. During the navigation between information patches, imperfect information at intermediate locations is used to make such a decision. Such intermediate imperfect information is a key concept in IFT, called *information scent*. For example, on a web page, information scent may be represented by link descriptors, images, preceding headings, and page arrangements.

The concept of information scent has been suggested to explain a user's web search behavior on SERPs [8, 9, 37]. Card et al. [7] developed the Web Behavior Graphs, suggesting the role played by information scent as a driver in the process of information seeking. Cutrell and Guan [9] found that positions of relevant search results influenced search behavior and suggested the use of information scent for future work. Following their suggestions, Wu et al. [37] conducted a controlled user study to understand the effect of information scent on search behavior on a desktop search system. SERPs with different information relevance levels and patterns were presented to the users. They found that participants viewed documents in lower positions when more relevant search results were present. The participants also abandoned their search earlier if relevant search results were only shown later on the SERPs. They also found that search behavior also depended on personality, such as the level of NFC (Need for Cognition), a cognitive scale that measures the extent to which a person enjoys a challenging task that requires cognitive thinking. As a follow-up study, Ong et al. [30] developed a user study with similar manipulations on information scent levels and patterns to compare search behavior on mobile devices and desktop search systems. It showed that search behaviors on mobile and desktop were measurably different. For example, mobile participants achieved higher search accuracy than desktop participants for tasks with increasing numbers of relevant search results. Overall, both an increased number and better positioning of relevant search results improved the ability of participants to locate relevant results on both desktop and mobile. Participants spent more time and issued more queries on desktop, but abandoned less and saved more results on mobile devices. In another recent study [2] that tried to measure the quality of a SERP, a measure based on Information Foraging Theory was proposed: experienced utility, which accounted for the heterogeneity of search results and naturally connected how to model search with how to evaluate search. Through an experiment with 1,000 popular queries issued to a major search engine, the measure of experienced utility has demonstrated more accurate reflection of observed behaviors.

These studies have inspire this study of using information scent to predict search behavior. Most of these studies are based on user studies where the number of users, the search scenarios, and the possibility of information scent manipulations are limited. In this study, we will go beyond to work on a large amount of search logs collected from a well-known search engine that represents a natural search behavior "in the wild". With assistance of deep learning models, we are able to test Information Foraging Theory in a scalable and natural way for human information foraging process.

2.2 Deep Learning Models to Understand Human Behavior

Based on the recent success on deep learning techniques in machine translation and speech recognition, there has been a few studies that leverage deep learning techniques to model human interactions and predict some human behavior. Most of these models have demonstrated better performance than the traditional machine learning models that require sophisticated feature engineering process. For example, Li et al. [18] in Google Research designed a neural network architecture based on LSTM model to predict how much time was needed for a user to find an item from a menu on a user interface (UI). In the online education area, Jo et al. [15] used the students' click logs harvested from the Massive Open Online Courses (MOOCs) platform, and exploited the temporal dynamics of student behaviors via a LSTM neural network in order to first encode student behavior and then predict their final learning outcomes in the course. LSTM model has also been applied in the e-commerce area to encode customers' shopping actions in a sequence and predict the purchase outcome of the sequence. Using clickstream data generated during live shopping sessions, Toth et al. [35] adopted LSTM to predict three shopping outcomes: purchase, abandoned shopping cart, and browse-only. Their experiments have shown better prediction performance than Markov models. In summary, these studies have marked milestones of applying deep learning algorithms in modeling and predicting human behavior, especially those behaviors that are of sequence nature. Inspired by these studies, we will design a neural network structure using LSTM as the basic unit to encode and decode human search behaviors with search engines.

2.3 Well-Known Click Models

In recent years there are a few IR studies experimenting with deep learning techniques to model sequential dependency in search activities and predict search behavior. For example, Zhang et al. [39] used a RNN structure to predict users' advertisement click behavior in a sponsored search system under the hypothesis that users' ads click probability depends on their ads browsing behavior in the past, such as the timing of the first click on an ads and the dwell time on the ads. Williams and Zitouni [36] adopted LSTM to model the number of user interactions and the nature of those interactions

to distinguish the good and bad query abandonment. Their model performed significantly better than other baselines.

For click behavior, Borisov et al. [5] proposed a LSTM-based neural click model to predict user clicks on search engine results. They represented users' information need (query) and the information available to the user with a vector state, which was initialized with a query, and then iteratively updated based on subsequent interactions. This neural click model demonstrated better performance than the traditional probabilistic graphical model (PGM) that needed a predefined set of rules. In a follow-up study in 2018 [6], they further designed a click sequence model (CSM) with an attention mechanism that aimed to predict the order in which a user would click the search results. The goal was one step forward compared to most machine learning models that were to predict an un-ordered set of results a user would click on. As the result, CSM has achieved a comparable results to those standard machine learning models for predicting the un-ordered set of results, but with the additional prediction of the order of those clicks. Built on Borisov et al's models, Yu et al. [38] proposed an approach to representing query and documents in a lower-dimensional space, and deployed different weight matrices at each layer of the RNN model to address the so-called position bias [10, 16]. They achieved better model performance as well as less computational cost.

These studies' positive results suggest the effectiveness of deep learning models, especially the LSTM-based RNN structure, in modeling and predicting search behaviors in the information retrieval field. As mentioned in the Introduction section, this work built on and extend these previous studies by constructing neural network models for a search sessions that may contain multiple queries. Also, instead of just using click behavior pattern as input, this work considers result relevance pattern, as the information scent, to input into the deep learning models, under the guidance of Information Foraging Theory. In addition, instead of predicting just a binary click behavior, this work will extend the target to other major search behavioral variables, either binary or numeric, such as query reformulation, number of clicks, and information gain.

3 DEEP LEARNING OF INFORMATION FORAGING PROCESS

We propose a two-level deep learning framework to describe human information foraging process. The "goodness" of the description is represented as the model performance of using information scent sequence as input to predict some important search behavioral variables.

3.1 Target Behavioral Variables

The six search behavioral variables we are interested in are:

- **Query Abandonment (qry_abandon)**: a binary variable; describing whether a user abandons the query without any interaction with the SERP returned by the query
- **Query Reformulation (qry_reform)**: a binary variable; describing whether a user re-issues another query after an original query
- **Pagination (pagination)**: a binary variable; describing whether a user goes beyond the first SERP and paginates to the next page
- **Number of Clicks (num_clicks)**: a numeric variable; the total number of items the user has clicked for a query
- **Click Depth (click_depth)**: a numeric variable; the deepest rank that the user has clicked for a query
- **Information Gain (info_gain)**: a numeric variable; the DCG (Discounted Cumulative Gain) value based on the clicked results

These six behavioral variables are commonly used in describing search behaviors [4, 14, 24–27, 29, 33, 34, 37]. The former three possess some "sequential" nature, reflecting a action based upon not only what the user is seeing at the moment but also what they have seen in the past. In contrast, the latter three have some "cumulative" nature, representing the aggregate "memory capability" of the search behavior . Both groups of behaviors capture the heuristics that the Information Foraging Theory describes. We want to further investigate which group of these variables our LSTM neural network is able to better predict.

3.2 Proposed Deep Learning Framework

We propose a two-level deep learning modeling framework as presented in Figure 1, to represent the information foraging process. Both levels follow a encoder-decoder structure. The lower-level captures the user-information interaction under each query, including the user search style, query formulation, scanning the SERP returned by the query, clicking on the result items, deciding on whether to reformulate another query, etc. The upper level describes the user's session-level actions. It is worth mentioning that a search session is defined as a series of search activities driven by one search goal. One search session may contain one or multiple search query threads.

The framework architecture builds on two important capabilities of LSTM [12]. First, it is capable of "reading in" a sequence of input and encoding it as a certain representation. This encoding process considers the temporal and spatial features of the input which are believed useful in predicting search behavior than the "flat" features. Second, the model is capable of mimicking users' "remembering" and "forgetting" effects of previous learning on the current move. Such learning effect is believed to play an important role in human behavior. The details of the two levels are introduced in below.

3.2.1 Modeling and Predicting Query-Level Interactions. Query-level interactions include the behaviors under one query, which is part of a search session that may contain multiple queries. Understanding the query-level interactions is essential for understanding the entire search session. Our neural network structure follows the encoder-decoder structure to encode the information scents on a SERP and predict the query-level interactions.

Encoder. The aim of the encoder is to represent the search results' information scents on a SERP as well as other important information, and pass the representation to the decoder.

We describe three sets of representations, each built on the previous. Set 1 (SCENT) will encode only the information scent on the first SERP to predict search behavior, as the idea of most of the lab-based user studies. Set 2 (SCENT+QUERY) extends Set 1 by considering the query intent of the searcher. Set 3 (SCENT+QUERY+SEARCHER) extends Set 2 by incorporating the person's search style through

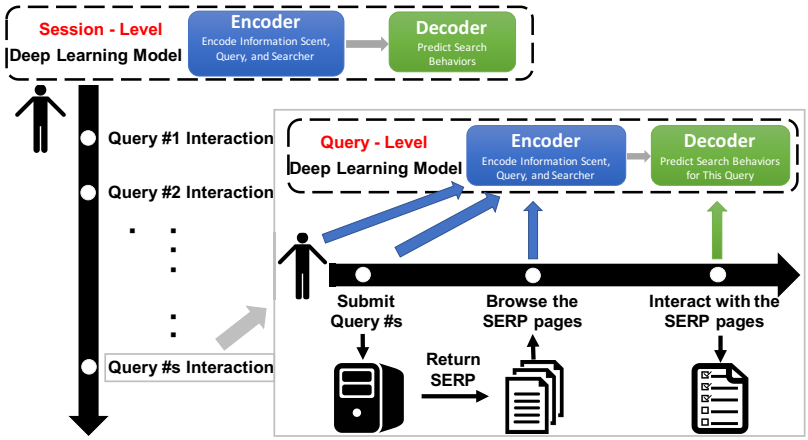

Figure 1: Proposed deep learning framework to represent information foraging process

examining all their historic behaviors recorded in a large amount of log dataset.

Set 1 (SCENT): As in the encoder structure shown in Figure 2, the input of the encoder is the sequence of the raw relevance scores of the first SERP returned by the query. r_i represents the relevance score of the search result at position i. The sequence of $(e_s^1, e_s^2, ...e_s^N,)$ represent the LSTM chain for the encoder, where each e_s^i is one LSTM computation unit. h_i denotes the hidden state of each LSTM computation unit after reading in r_i as well as the output from previous LSTM unit e_s^{i-1}. N denotes the number of search results on the SERP. The final hidden state h_N is then used to represent the overall information scent distribution on this SERP and is passed to the decoder for behavior prediction.

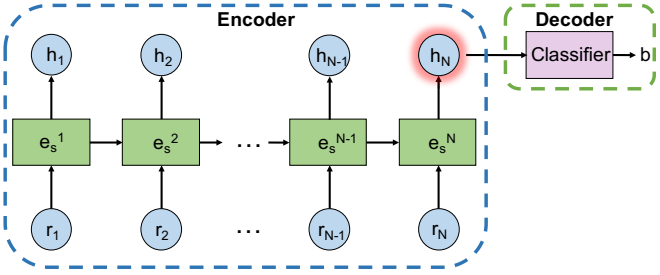

Figure 2: Deep learning model with Set 1 input at the query level

Set 2 (SCENT+QUERY): The second set of representation extends Set 1 by considering the query intent, as in Figure 3. The representation used in Set 1 ignore potentially useful information about this particular query. Set 2 incorporates such query intent information by aggregating all the interactions by different users but under this same query q. In another words, we represent the intent of this query q by its aggregate click patterns observed on SERPs generated by the query q, especially the click patterns on the first three ranks, which are the most important positions on a SERP. Since there are $2^3 = 8$ possible click patterns for the first three positions, we represent the query q with a vector of size 8. In each component of the vector $\mathbf{q}$, we store the number of times a particular click pattern was observed on SERPs generated by the query q.

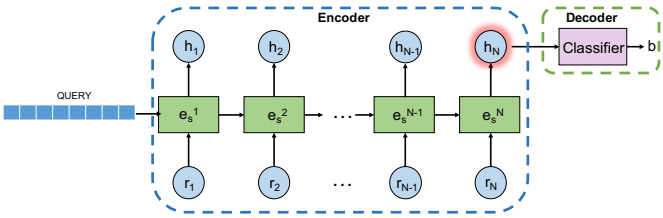

Figure 3: Deep learning model with Set 2 input at the query level

Set 3 (SCENT+QUERY+SEARCHER): The third set of representation extends Set 2 by considering the searcher's search behavior style information, as in Figure 4. For example, people were found to exhibit their Web navigation style as Web returner or Web explorer [3]. Aula et al. [1] categorized information seekers into "economic" or "exhaustive" styles based on their approach to evaluating search results on Web Pages. In a similar way, the depth-first and breadth-first strategies have been observed by Klockner et al. [17] to understand individual differences. These notions of differentiating people's search behaviors are quite useful for this study, and therefore we will encode such information as input of the deep learning models.

Set 3 allows us to collect behavioral information of this particular user from a large amount of search sessions issued by this person. Similar to the representation method of the query q, we represent the searcher u by their click patterns,observed on SERPs under this user u, especially the click patterns on the first three positions. Since there are $2^3 = 8$ possible click patterns, we represent the user u with a vector of size 8, the same size of $\mathbf{q}$. In each component of the vector $\mathbf{u}$, we store the number of times a particular click pattern was observed on SERPs requested by this user u.

Decoder. The decoder is to predict a probability for a behavior given an information scent distribution. If the behavior is described by a binary variable, like query abandonment, the decoder will label

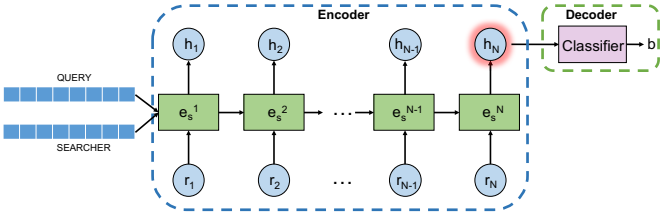

Figure 4: Deep learning model with Set 3 input at the query level

a "yes" or "no" based on whether the predicted probability is above or below 50%. If the behavior is represented by a numeric variable, such as number of clicks, the decoder will predict a value for this variable.

We will implement the decoder using a three-layer fully connected neural network. The result of the three-layer neural network will be passed to a Softmax layer, which calculates the probability or the value for each behavioral variable.

Since we have six target behavioral variables with two types (binary and numeric) in this study, we propose to use two different loss functions. For the binary behavioral variables, we will the cross-entropy loss function. The overall loss for the training dataset is defined as the average of the loss value for each record. For the numeric behavioral variables, we will use Root Mean Square Error (RMSE) as the loss function.

3.2.2 Modeling and Predicting Session-Level Interactions. Modeling and predicting session-level interaction is to connect the models of query-level interactions as building blocks.

Encoder. With each query-level encoder LSTM chain e_s, we are able to concatenate them as the input of a session-level encoder. In that sense, a series of LSTM chains, $(e_1, e_2, e_3, ..., e_s)$, as presented in Figure 5, will be fed into an encoder, represented as $(Q_1, Q_2, ..., Q_s)$, where each Q_s is a LSTM computation unit, similar to e_s^i in the query-level encoder. The outcome of each Q_s will be passed to a classifier for behavior prediction.

Decoder. Similar to the query-level decoder, this session-level decoder is also to predict a user's query-level behavior, but to put them in a sequence context. Specifically, in order to predict the current query-level behaviors, the decoder reads in not only the hidden state generated by the current query's encoder, but also the hidden state and actions from a previous query encoder. In that sense, the behaviors under the current query are actually under the cumulative influence of the previous queries, which more precisely reflects the nature of user interaction in a multi-query search session, than the traditional neural click models. The loss function in the session level is defined as a sum of the loss function for each query-level classifier.

4 EXPERIMENTS

In this section, we will introduce the dataset used in this work as well as our deep learning model implementation details.

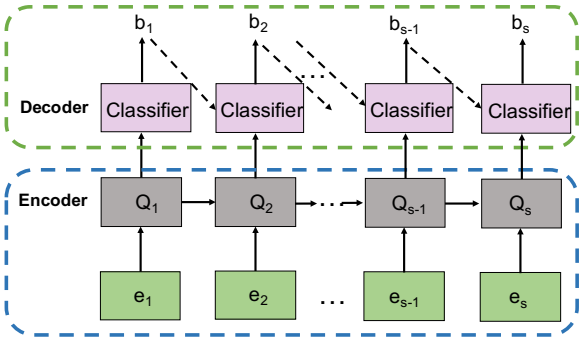

Figure 5: Deep learning model at the session-level

4.1 Search Log Dataset

Data used for training the deep learning framework is from Yahoo!. It contains one-month search logs. The search logs contain two parts, as summarized in Table 1: 1) queries, corresponding SERPs, and clicks, and 2) relevance judgments on the first ten results for each query. The relevance judgments by the corporate research team have added the research value of this dataset. The two parts are joined by the combination of a query and a search result.

Table 1: Description of the search log dataset

	Data Field	Description
PART 1	**query**	The query (de-identified as a code) submitted by a searcher
	cookie	An anonymized version of a user identifier
	time stamp	The UNIX time when the query was submitted
	url (10 columns)	The top 10 result links of that query
	number of clicks	The total number of clicks during that search session
	clicked position (19 columns)	The position of each click
	elapsed time (19 columns)	The elapsed time between the query submission time and a click time
PART 2	**query**	The same with the query in Part 1
	url	a result link
	judgment	five-grade judgment on relevance: 0 means very irrelevant and 4 means very relevant

The joined data contains 80,779,266 records. After removing the records whose judgment scores were incomplete for the top ten results, there were 1,930,932 records left. These were the final dataset on which our deep learning models were constructed. As descriptive statistics, 13.20% of the queries were abandoned, 2.44% were reformulated, and 3.93% led to pagination beyond the first page. These low percentages suggest the unbalanced distribution of the positive and negative cases for each binary behavioral variable. All the distributions for the number of clicks, click depth, and information gain roughly follow a power law distribution, suggesting that most searchers clicked very few results at fairly high ranks with a little information gain, in line with other studies, such as [31].

The search logs have provided us with the cookie information, which is the anonymized user identifiers. The recorded interactions of different sessions interleaved together and they were ordered

according to their time stamps. In this study, a session is defined as a series of actions conducted by one person (identified by one cookie identifier) at certain continuous period of time, and different sessions are assumed independent of each other in this study. Generally, researchers use consecutive actions that fall into a time range between 5 to 60 minutes as a session [21, 22]. This study follows the idea in [20, 25, 28] to separate a session as a series of consecutive actions from the same user identifier with no period of inaction greater than 30 minutes.

After grouping the 1,930,932 records into sessions, we obtained 1,873,942 search sessions. The majority (98.01%) of the search sessions only contained one single query, suggesting that search sessions "in the wild" tended to be brief.

4.2 Model Configuration & Hyper-Parameters

For the query-level encoder, N took the value of 10 since each SERP of the big name search engine contained 10 search results. Since the search sessions with more than 2 queries were too few (less than 0.1%) to train a deep learning model, s took the value of 2 for the session level encoder.

Both neural network models (query-level and session-level) were trained by minimizing the defined loss functions L_{binary}, $L_{numeric}$, and $L_{session}$ respectively. ADAM optimizer was adopted with the learning rate of 0.01. To achieve best model performance, we applied the activation function to the first layer neurons in the decoder net to do the non-linear transformation. After several rounds of experiments, we found that ReLU [23] usually did better in classification models and Tanh in regression models for our tasks. Therefore we applied ReLu for the binary variable prediction and Tanh for the numeric prediction. Each model was trained 1,000 epochs and the one with the lowest loss value was saved for future test.

As common practice in machine learning, we randomly sampled 80% of the dataset as the training set and the remaining 20% as the test set. The unbalanced distribution of those binary target behavioral variables was a potential problem for training the model, since most common prediction algorithms would minimize the overall error rate rather than paying special attention to the minority cases. In this study, for those binary target behavioral variables, we adopted the method of under-sampling the majority cases [11] to make both types of cases roughly balanced at each query level in the training dataset. The test set remained untouched. For those numeric behavioral variables, we did not have the unbalanced distribution problem and therefore we split the original dataset into the 80% for training and 20% for test.

5 RESULTS

Model results are reported at both query and session levels. Accuracy and F-measure were selected as the model performance metrics for the binary behavioral variables whereas Pearson correlation coefficient and RMSE (Root Mean Square Error) were selected for model performance for those numeric behavioral variables. In addition to our proposed information scent models (IFMs), we also selected two state-of-the-art deep learning models as the baseline: $RBNN*$ from [38], which is a rank-biased neural network model and has the best performance in several variations of such models; and NCM^{LSTM}_{QD+Q+D} from [5], which is a neural click model and considers complete information from a SERP including query document pairs, the query, and the documents.

5.1 Query-Level Model Performance

The model performance for binary query-level behaviors is summarized in Table 2. Overall speaking, at least one variation of our information scent models (IFM) has improved the prediction accuracy and F-measure compared to those baseline models, suggesting the effectiveness of IFMs. Focusing on those three variations of IFMs, we find that incorporating the information scent, query, as well as searchers' behavioral style information (Set 3) has helped performance on both query abandonment and pagination. However, IFM including only the information scent and query information (Set 2) has the best performance for query reformulation, indicating people's decision on query reformulation does not rely much on each individual's search behavioral style. If the F-measure is broken down by precision and recall, we find that all the five models are better at precision than recall, resulting in a moderate value of F-measure. The finding is in line with other click models that are better at recognition than being complete [5, 6].

It is worth noting that information scents alone as input without any other information could achieve reasonable accuracy and F-measure, especially for query abandonment and pagination. The reasonable performance speaks to the fact that information scent matters in directing people's search behavior. Another interesting finding is that all the five models performed relatively better for query abandonment and pagination, but relatively worse for query reformulation, suggesting that query reformulation is more elusive or hard to predict.

Table 2: Query-level model performance for binary behavioral variables

Model for Predicting qry_abandon	Accuracy	F-Measure
IFM_{SCENT}	0.7066	0.6109
$IFM_{SCENT+QUERY}$	0.8027	0.6844
$IFM_{SCENT+QUERY+SEARCHER}$	**0.8315**	**0.7035**
NCM^{LSTM}_{QD+Q+D}	0.7377	0.6223
$RBNN*$	0.7954	0.6757
Model for Predicting qry_reform	**Accuracy**	**F-Measure**
IFM_{SCENT}	0.5313	0.4224
$IFM_{SCENT+QUERY}$	**0.7224**	**0.5856**
$IFM_{SCENT+QUERY+SEARCHER}$	0.7015	0.5049
NCM^{LSTM}_{QD+Q+D}	0.5549	0.4566
$RBNN*$	0.6027	0.4677
Model for Predicting pagination	**Accuracy**	**F-Measure**
IFM_{SCENT}	0.6218	0.3089
$IFM_{SCENT+QUERY}$	0.7344	0.4905
$IFM_{SCENT+QUERY+SEARCHER}$	**0.8128**	**0.5027**
NCM^{LSTM}_{QD+Q+D}	0.8080	0.4727
$RBNN*$	0.7822	0.4545

As to the numeric behavioral variables, the two baseline models have better performance than our IFM models for number of clicks and click depth, probably because those baseline models are click models especially for predicting whether a searcher will click on each result, to which the number of clicks and click depth are directly related. However, when predicting information gain, represented as the discounted cumulative gain of clicked results for each

query, our IFM model with Set 3 (SCENT+QUERY+SEARCHER) input has the best performance, because it considers not only the click behavior, but also the relevance level of each clicked result.

All the Pearson correlation coefficients are significant at the 0.05 level. The medium effect (0.2478 to 0.4223) of Pearson correlation and the relatively small RMSEs suggest that deep learning approach is able to achieve a reasonable performance in predicting the "cumulative" search behaviors beyond a "isolated" binary click behavior.

Table 3: Query-level model performance for numeric behavioral variables

Model for Predicting num_clicks	Pearson Correlation	RMSE
IFM_{SCENT}	0.1379**	1.2877
$IFM_{SCENT+QUERY}$	0.1730**	1.0112
$IFM_{SCENT+QUERY+SEARCHER}$	0.1959**	**0.9742**
NCM^{LSTM}_{QD+Q+D}	**0.2478****	0.9889
$RBNN*$	0.2326**	2.1355
Model for Predicting click_depth	**Pearson Correlation**	**RMSE**
IFM_{SCENT}	0.1801**	2.0001
$IFM_{SCENT+QUERY}$	0.1900**	1.4496
$IFM_{SCENT+QUERY+SEARCHER}$	0.2859**	1.4545
NCM^{LSTM}_{QD+Q+D}	0.3085**	1.6654
$RBNN*$	**0.3212****	**1.3725**
Model for Predicting info_gain	**Pearson Correlation**	**RMSE**
IFM_{SCENT}	0.1953**	1.4444
$IFM_{SCENT+QUERY}$	0.3305**	1.2857
$IFM_{SCENT+QUERY+SEARCHER}$	**0.4223****	**0.9797**
NCM^{LSTM}_{QD+Q+D}	0.3210**	0.9996
$RBNN*$	0.3118**	1.2538

Note: ** denotes that the correlation coefficient is significant at the 0.05 level.

5.2 Session-Level Model Performance

We will report the model performance broken down by the two queries $Q1$ and $Q2$. It is worth attention that for the IFM models, the prediction was sequential, meaning the prediction for the actions under $Q2$ was built on the knowledge of information scent and actions on $Q1$. We will investigate whether the cumulative knowledge that obtained from placing $Q1$ and $Q2$ in a temporal sequence will help improve the prediction of the actions under $Q2$.

Since the two baseline models NCM^{LSTM}_{QD+Q+D} and $RBNN*$ do not have the structure to incorporate session-level prediction, we will only compare the results from the three variations of the IFM models. The model performance for the binary behavioral variables is presented in Table 4. Overall, we have seen great improvement of the model performance on Query 2, especially for query abandonment and pagination. For the challenging behavioral variable qry_reform, only $IFM_{SCENT+QUERY+SEARCHER}$ has improvement on Query 2, and the improvement is dramatic. The finding indicates the advantage of this "sequential" modeling, especially when incorporating information scent, query, and searcher information, on the prediction of these binary search behavioral variables based on knowledge of a previous query thread.

The model performance on the numeric variables is presented in Table 5. Our IFM models are able to have better performance for Q2 for information gain, but not for number of clicks or click depth, suggesting once again the usefulness of previous query thread knowledge is limited to situations considering not only clicks but also relevance levels on the clicked results.

Table 4: Session-level model performance (s = 2) for binary behavioral variables

	Accuracy		F-Measure	
Model for Predicting qry_abandon	Q1	Q2	Q1	Q2
IFM_{SCENT}	0.6733	**0.7233**	0.6103	**0.6230**
$IFM_{SCENT+QUERY}$	0.7946	**0.8255**	0.6533	**0.6972**
$IFM_{SCENT+QUERY+SEARCHER}$	0.8022	**0.8754**	0.6884	**0.7172**
	Accuracy		**F-Measure**	
Model for Predicting qry_reform	Q1	Q2	Q1	Q2
IFM_{SCENT}	**0.5800**	0.5514	**0.4876**	0.3960
$IFM_{SCENT+QUERY}$	**0.7745**	0.7068	0.5044	**0.6060**
$IFM_{SCENT+QUERY+SEARCHER}$	0.6236	**0.8023**	0.4765	**0.5528**
	Accuracy		**F-Measure**	
Model for Predicting pagination	Q1	Q2	Q1	Q2
IFM_{SCENT}	0.5267	**0.6843**	0.2241	**0.4496**
$IFM_{SCENT+QUERY}$	0.7055	**0.7868**	0.3877	**0.5211**
$IFM_{SCENT+QUERY+SEARCHER}$	0.6638	**0.8854**	0.4069	**0.5762**

Table 5: Session-level model performance (s = 2) for numeric behavioral variables

	Pearson Correlation		RMSE	
Model for Predicting num_clicks	Q1	Q2	Q1	Q2
IFM_{SCENT}	0.0905**	**0.1875****	1.3654	**1.1369**
$IFM_{SCENT+QUERY}$	**0.2007****	0.1553**	**0.9998**	1.4401
$IFM_{SCENT+QUERY+SEARCHER}$	**0.2346****	0.1280**	1.2945	**0.6754**
	Pearson Correlation		**RMSE**	
Model for Predicting click_depth	Q1	Q2	Q1	Q2
IFM_{SCENT}	**0.1933****	0.1600**	2.4566	**1.9340**
$IFM_{SCENT+QUERY}$	**0.2098****	0.1166**	**1.2265**	1.5273
$IFM_{SCENT+QUERY+SEARCHER}$	0.1777**	**0.3258****	**1.1098**	1.8904
	Pearson Correlation		**RMSE**	
Model for Predicting info_gain	Q1	Q2	Q1	Q2
IFM_{SCENT}	0.1368**	**0.2045****	2.0561	**1.2389**
$IFM_{SCENT+QUERY}$	0.2477**	**0.3945****	1.9766	**1.1068**
$IFM_{SCENT+QUERY+SEARCHER}$	0.4106**	**0.4533****	1.3766	**0.5623**

Note: ** denotes that the correlation coefficient is significant at the 0.05 level.

6 DISCUSSION AND CONCLUSION

This study designed a two-level deep learning framework to model users' two-level action sequences: micro-level (query-level) and macro-level (session-level). We used LSTM as the base model because of its memory capability and the sequence structure that allow for natural mimic of not only what users are perceiving and performing at the moment but also what they have seen and learned from the past during the search dynamics. As the result, at the micro level, the information scent models were able to outperform the baseline neural click models in predicting the behaviors that require sequential scanning of search results. However, they have not performed as well as the baseline models in predicting the "cumulative" behaviors that are directly related to clicks. IFM models are also better at predicting information gain which considers not only clicks but also relevance scores of the clicked results.

At the macro level, by placing the current query in a sequence of queries, and incorporating the knowledge of the previous query's information scent and actions, the current prediction has seen great improvement for query abandonment, pagination, and information gain. The improvement has not been seen on the elusive behavioral variable, query reformation, as well as the "cumulative" click-related variables such as number of clicks and click depth.

One drawback of the deep learning model is that it is difficult to interpret what the model has learned, such as what kind of information scent distribution tends to lead to a query abandonment, query reformulation, or more information gain. Therefore in near future we plan to do some follow-up analysis on a simulated test dataset with all possible permutations of the information scent distribution on the first ten search results. For each permutation, we will apply the trained deep learning model to predict a probability for those behavioral variables. The average predicted variable probabilities or values for different information scent levels and patterns could offer some interpretability of our IFM models.

REFERENCES

[1] Anne Aula, Päivi Majaranta, and Kari-Jouko Räihä. 2005. Eye-tracking reveals the personal styles for search result evaluation. In *IFIP Conference on Human-Computer Interaction*. Springer, 1058–1061.

[2] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 605–614.

[3] Hugo S Barbosa, Fernando B de Lima Neto, Alexandre Evsukoff, and Ronaldo Menezes. 2016. Returners and Explorers Dichotomy in Web Browsing BehaviorâĂŤA Human Mobility Approach. In *Complex Networks VII*. Springer, 173–184.

[4] Nicholas J Belkin, Diane Kelly, G Kim, J-Y Kim, H-J Lee, Gheorghe Muresan, M-C Tang, X-J Yuan, and Colleen Cool. 2003. Query length in interactive information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 205–212.

[5] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A neural click model for web search. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 531–541.

[6] Alexey Borisov, Wardenaar Martijn, Maarten Ilya, Markov, and Rijke de. 2018. A click sequence model for web search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM.

[7] Stuart K Card, Peter Pirolli, Mija Van Der Wege, Julie B Morrison, Robert W Reeder, Pamela K Schraedley, and Jenea Boshart. 2001. Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 498–505.

[8] Ed H Chi, Peter Pirolli, Kim Chen, and James Pitkow. 2001. Using information scent to model user information needs and actions and the Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 490–497.

[9] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 407–416.

[10] Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 478–479.

[11] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[13] Minlie Huang, Qiao Qian, and Xiaoyan Zhu. 2017. Encoding syntactic knowledge in neural networks for sentiment classification. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 26.

[14] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2009. Patterns of query reformulation during Web searching. *Journal of the american society for information science and technology* 60, 7 (2009), 1358–1371.

[15] Yohan Jo, Keith Maki, and Gaurav Tomar. 2018. Time Series Analysis of Clickstream Logs from Online Courses. *arXiv preprint arXiv:1809.04177* (2018).

[16] Thorsten Joachims, Laura A Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Sigir*, Vol. 5. 154–161.

[17] Kerstin Klöckner, Nadine Wirschum, and Anthony Jameson. 2004. Depth- and breadth-first processing of search result lists. In *Conference on Human Factors in Computing Systems: CHI'04 extended abstracts on Human factors in computing systems*, Vol. 24. Citeseer, 1539–1539.

[18] Yang Li, Samy Bengio, and Gilles Bailly. 2018. Predicting Human Performance in Vertical Menu Selection Using Deep Learning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 29.

[19] Yiqun Liu, Xiaohui Xie, Chao Wang, Jian-Yun Nie, Min Zhang, and Shaoping Ma. 2017. Time-aware click model. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 16.

[20] Cory Lown. 2008. A transaction log analysis of NCSU's faceted navigation OPAC. (2008).

[21] Gary Marchionini. 2002. Co-evolution of user and organizational interfaces: A longitudinal case study of WWW dissemination of national statistics. *Journal of the Association for Information Science and Technology* 53, 14 (2002), 1192–1209.

[22] Mazlita Mat-Hassan and Mark Levene. 2005. Associating search and navigation behavior through log analysis. *Journal of the Association for Information Science and Technology* 56, 9 (2005), 913–934.

[23] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[24] Xi Niu, Xiangyu Fan, and Tao Zhang. 2019. Understanding Faceted Search from Data Science and Human Factor Perspectives. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 14.

[25] Xi Niu and BM Hemminger. 2011. Beyond text querying and ranked lists: Faceted search in library catalogs. In *Proceedings of the 74th ASIS&T Annual Meeting*, Vol. 48.

[26] Xi Niu and Bradley Hemminger. 2015. Analyzing the interaction patterns in a faceted search interface. *Journal of the Association for Information Science and Technology* 66, 5 (2015), 1030–1047.

[27] Xi Niu and Diane Kelly. 2014. The use of query suggestions during information search. *Information Processing & Management* 50, 1 (2014), 218–234.

[28] Xi Niu, Cory Lown, and Bradley M Hemminger. 2009. Log based analysis of how faceted and text based search interact in a library catalog interface. In *Proceedings of Third Workshop on Human-Computer Interaction and Information Retrieval*.

[29] Xi Niu, Tao Zhang, and Hsin-liang Chen. 2014. Study of user search activities with two discovery tools at an academic library. *International journal of human-computer interaction* 30, 5 (2014), 422–433.

[30] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk Scholer. 2017. Using information scent to understand mobile and desktop web search behavior. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 295–304.

[31] Casper Petersen, Jakob Grue Simonsen, and Christina Lioma. 2016. Power law distributions in information retrieval. *ACM Transactions on Information Systems (TOIS)* 34, 2 (2016), 8.

[32] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[33] Soo Young Rieh et al. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management* 42, 3 (2006), 751–768.

[34] Amanda Spink, Bernard J Jansen, and H Cenk Ozmultu. 2000. Use of query reformulation and relevance feedback by Excite users. *Internet research* 10, 4 (2000), 317–328.

[35] Arthur Toth, Louis Tan, Giuseppe Di Fabbrizio, and Ankur Datta. 2017. Predicting Shopping Behavior with Mixture of RNNs. In *Proceedings of the SIGIR 2017 Workshop on eCommerce (ECOM 17)*.

[36] Kyle Williams and Imed Zitouni. 2017. Does That Mean You're Happy?: RNN-based Modeling of User Interaction Sequences to Detect Good Abandonment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 727–736.

[37] Wan-Ching Wu, Diane Kelly, and Avneesh Sud. 2014. Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 557–566.

[38] Hai-Tao Yu, Adam Jatowt, Roi Blanco, Joemon M Jose, and Ke Zhou. 2019. A Rank-biased Neural Network Model for Click Modeling. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, 183–191.

[39] Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. 2014. Sequential Click Prediction for Sponsored Search with Recurrent Neural Networks.. In *AAAI*, Vol. 14. 1369–1375.