

Action Modeling: Language Models That Predict Query Behavior

G. Craig Murray, Jimmy Lin
College of Information Studies
University of Maryland, College Park
{gcraigm,jimmylin}@umd.edu

Abdur Chowdhury
AOL Search
America Online, Inc.
cabdur@aol.com

ABSTRACT

We present a novel language modeling approach to capturing the query reformulation behavior of Web search users. Based on a framework that categorizes eight different types of “user moves” (adding/removing query terms, etc.), we treat search sessions as sequence data and build n -gram language models to capture user behavior. We evaluated our models in a prediction task. The results suggest that useful patterns of activity can be extracted from user histories. Furthermore, by examining prediction performance under different order n -gram models, we gained insight into the amount of history/context that is associated with different types of user actions. Our work serves as the basis for more refined user models.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *query formulation, search process* H.1.2 [Models and Principles]: User/Machine Systems – *human factors*.

General Terms: Experimentation, Human Factors, Languages, Theory

Keywords: Web search, user models, query modeling, query reformulation

1. INTRODUCTION

Information-seeking is a fundamentally iterative process. During search, user’s queries are generally a compromised expression of what they are looking for—a trade-off between a tacit information need and what the user knows about the solution space [3]. Our analysis of AOL Search logs reveals that 28% of all search queries are reformulations of previous queries. Therefore, accurate models of search behavior will require a component for query reformulation [1,4]. Users’ query behaviors are linear in nature. As such, they can be modeled as sequence data using n -grams. We applied a simple language modeling technique to predict types of query changes. Our work tests a framework of user’s information-seeking behavior and can inform user models of interactive search.

2. METHOD AND RESULTS

When a user issues a query against a collection of documents, the system returns a set of results and the user typically examines only some subset of those results (e.g., the top 10). When the examined subset does not contain enough relevant documents, the

user usually modifies the query. The initial query and the reformulated queries that follow can be conceptualized as a sequence of signs intended by the user to signify what that user knows about the information search space, and how the user is trying to change that knowledge. Beyond the semantics of the query terms, there is a syntagmatic relationship between two queries in a sequence; one query follows the other as a particular type of action to adjust the original expression.

We constructed a typology of user actions, independent of the semantic content of the queries. We identified a set of eight “user moves” that capture the generic ways Web users modify their queries. These can be seen as a vocabulary of move types:

1. **new** – a new query that is not at the start of a new session.
2. **repeat** – a query that is identical to the immediately preceding query (usually a request for next page of results.)
3. **return** – a return to an earlier query in the session but *not* a repeat of the immediately preceding query.
4. **add_to_prev** – a query that includes the entirety of the immediately preceding query plus new characters/words.
5. **remove_from_prev** – a query that is a substring of the immediately preceding query (i.e., a partial deletion.)
6. **edit_longer** – a query that includes some but not all of the previous query, and is longer than previous query.
7. **edit_same_length** – a query that includes some but not all of the previous query and is the same length as previous.
8. **edit_shorter** – a query that is partly made up of words from the previous query and is shorter.

Similar taxonomies have been used elsewhere [5]. Naturally, semantic features must *also* be modeled, but they should be integrated into a higher level syntactic framework. Our work provides a first step toward that framework. We explore the use of n -gram models of query moves and demonstrate that they can account for much of the observed patterns of query behavior.

We took a sample of 8.3 million queries issued by AOL Search users over a period of three months and labeled them with the “move types” described above. We trained n -gram models on the sequences of moves within users’ query sessions. Users were randomly sorted into 10 groups. We trained our models on 8 groups of users at a time, held one group out, and evaluated our models using the remaining group. This “leave one out” approach reduces over-fitting and allows for parameter tuning in the future. We repeated our process 10 times to cross-validate our results. Models were evaluated by predicting next moves based on preceding moves. To assess the effect of using different amounts of history/context we evaluated performance with n -gram models of varying length. We used the SRILM toolkit [2] to build Bayesian n -gram mixture models of move sequences. The models were smoothed using the Turing method and Katz back-off

weighting. Thus each model was a mixture of n -grams of a maximum size and smaller order n -grams. For a baseline comparison, we also implemented a weighted random guess model. This baseline is essentially a context-free unigram model that “guesses” moves with a frequency distribution directly proportional to moves in the training set.

We evaluated each n -gram mixture model for its ability to predict the next move in a series of moves, given the context of previous moves. From each of the 10 random groups we randomly selected 10,000 user sessions that contained 10 or more queries. From each of these sessions we randomly picked one of the moves in the session as a target for prediction. This gave us 10 evaluation test sets of 10,000 prediction targets each, for a total of 100,000 prediction targets. We then evaluated each test set against five language models: 2-gram, 3-gram, 4-gram, 5-gram and the unigram baseline. Using the sequence of moves leading up to each target move, we predicted what the next move (the target) would be based on the perplexity of the sequence of moves when compared to the each of the five language models. To quantify the accuracy of our prediction models we use traditional IR measures of precision and recall. In our case, precision is calculated as the number of times a move was correctly predicted, divided by the number of times it was predicted overall. Recall is calculated as the number of times a move was correctly predicted divided by the number of times that move *should* have been predicted. Results are presented in Tables 1 and 2. We compared the results of the four different context-based models to the results that would be obtained from the baseline. For all but two of the moves, our language modeling technique achieves results that are significantly better than the baseline (using a 2-tailed paired t-test,

$\alpha=0.01$). The gray cells in Tables 1 and 2 indicate the max values for each row; i.e., they indicate the best model for that move.

Nearly all of the non-zero values for n -gram models’ precision and recall are statistically significant when compared to the random baseline. With only two exceptions (2-gram/edit_longer and 3-gram/return) all of the n -gram models make consistent successful predictions for the moves repeat, new, edit_longer, return, and edit_same_length. 2-grams models successfully predicted edit_longer but with less consistency across the 10 folds of the cross validation. Predictions for add_to_prev are better than the baseline in two conditions (4-gram and 5-gram) but the difference is not statistically significant. The same is true of 3-gram models and the return move. The data shows that for most moves, a bi-gram model achieves the highest precision. For some moves, a trigram model is more accurate. For all eight moves, 4- and 5-gram models are overkill.

3. CONCLUSION

Language modeling techniques are useful for predicting types of changes applied to web search queries. Our perplexity based prediction models outperformed a weighted random guess for six out of eight query move types. These results indicate that sequences of previous moves *are* predictive of the query changes users will make. The fact that shorter n -gram models outperformed longer ones suggests that the predictive influence of preceding context is very local. For some move types, a little bit of context goes a long way toward prediction, but for most, adding a lot more context just adds noise. We anticipate that adding semantic features to the model(s) will improve performance. However, we also believe that syntagmatic abstractions such as query moves will be an important component in a larger framework.

4. REFERENCES

- [1] Shen, X., Tan, B., and Zhai, C. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM2005)* (Bremen, Germany, Oct. 31-Nov. 5, 2005). ACM Press, New York, NY, 2005, 824-831.
- [2] Stolcke, A. SRILM—An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)* (Boulder, Colorado, USA, October, 2000). 901-904.
- [3] Taylor, R. Question negotiation and information seeking in libraries *College & Research Libraries*, 29 (1968), 178-194.
- [4] Teevan, J., Dumais, S. T., and Horvitz, E. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR* (Salvador, Brazil, August, 2005). ACM Press, New York, NY, 2005, 449-456.
- [5] Vakkari, P., Pennanen, M., and Serola, S. Changes of search terms and tactics while writing a research proposal. *Information Processing & Management*, 39 (2003), 445-463.

Table 1. Precision measure for move predictions

Move	2gram	3gram	4gram	5gram	Baseline
repeat	0.644*	0.590*	0.603*	0.590*	0.464
new	0.373*	0.369*	0.365*	0.331*	0.140
edit_longer	0.198**	0.286*	0.274*	0.273*	0.108
edit_shorter	0.246*	0.310*	0.253*	0.296*	0.091
return	0.275*	0.269*	0.190*	0.224*	0.086
edit_same_length	0.423*	0.552*	0.505*	0.535*	0.059
add_to_prev	0	0.033	0.096	0.079	0.040
remove_from_prev	0	0	0	0	0
Overall	0.490*	0.520*	0.506*	0.497*	0.279

* significant at $\alpha=0.01$

** significant at $\alpha=0.05$

Table 2. Recall measure for move predictions

Move	2gram	3gram	4gram	5gram	Baseline
repeat	0.745*	0.880*	0.856*	0.824*	0.497
new	0.371*	0.377*	0.310*	0.351*	0.131
edit_longer	0.176**	0.148**	0.153*	0.136**	0.102
edit_shorter	0.283*	0.204*	0.146*	0.189*	0.087
return	0.244*	0.028	0.116*	0.107**	0.070
edit_same_length	0.434*	0.371*	0.410*	0.393*	0.061
add_to_prev	0	0.002	0.012	0.011	0.037
remove_from_prev	0	0	0	0	0
Overall	0.490*	0.520*	0.506*	0.497*	0.279

* significant at $\alpha=0.01$

** significant at $\alpha=0.05$