Heads and Tails: Studies of Web Search with Common and Rare Queries

Doug Downey
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195
ddowney@cs.washington.edu

Susan Dumais & Eric Horvitz
Microsoft Research
Redmond, WA 98052
{sdumais,horvitz}@microsoft.com

ABSTRACT

A large fraction of queries submitted to Web search engines occur very infrequently. We describe search log studies aimed at elucidating behaviors associated with rare and common queries. We present several analyses and discuss research directions.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement

Keywords

Web search, Zipf distribution, Probabilistic Latent Semantic Analysis

1. INTRODUCTION

Queries submitted to Web search engines follow a heavy-tailed Zipf distribution, in which a large fraction of queries are issued infrequently [4]. Yet much regarding this well-known long-tail of rare queries remains unexplored. Do users behave differently on rare queries than on common ones? What portion of rare queries represent rare informational goals, versus atypical means of specifying common goals? How might answers to such questions guide research toward enhancing Web search experiences?

We present experiments motivated by these questions. First, we compare search behavior on rare queries with behavior on common queries. We present evidence suggesting that current search engines perform less well on rare queries. Then, we explore transitions among rare and common queries during sessions. Lastly, we present methods for predicting query reformulations and discuss how the predictions might be used to improve performance on rare queries.

2. RARE VERSUS COMMON QUERIES

We first investigate whether the behavior of users following the input of rare queries differs from behavior following common queries. We compiled a case library of Web browsing activity from a sample of over 250,000 users, collected via opt-in, client-side instrumentation as part of the Windows Live Toolbar download. We identified users' queries and result clicks on three popular search engines over a period of three weeks. The resulting dataset contained approximately 10 million query events.

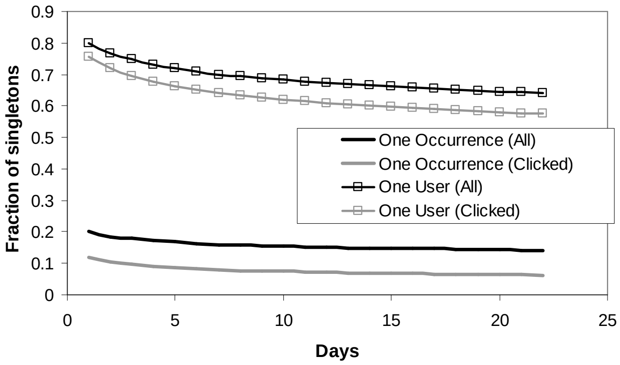

Figure 1: Fraction of singleton queries.

Of the rare queries, we distinguish *one-occurrence* queries, which are executed only once, from *one-user* queries, which are executed by a single user (potentially multiple times). For this analysis, a return to the results page or a request for the next page of results was recorded as a repeat of the original query. Figure 1 shows how the fraction of *singleton* queries (one-user or one-occurrence queries) decreases with the number of days of observation in the dataset. The graph shows that queries executed by a single user make up a significant fraction of total query executions, whereas queries executed only once make up a smaller fraction. The curves appear to asymptote near the end of the 22-day aggregation period, suggesting that the fraction of singleton queries is likely to converge on a large number as the observation period is extended in time or over a larger user base. Figure 1 also shows that the fraction of singleton queries of either type decreases when we consider only those queries that generated a result click at least once in the dataset ("Clicked"). If we take clicks as an indication of the success of a search engine in satisfying a user's information needs, it is clear that many singleton queries are not satisfied.

We next compare users' search activities following rare and common queries. For this analysis, we measure user

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

Query Type	Result Click	Re-Query	End Session
Tail	49.5%	44.8%	5.7%
Non-Tail	58.0%	33.4%	8.6%

Table 1: Behavior following tail and non-tail queries.

	To Tail	To Non-Tail
Tail	84.4%	15.6%
Non-Tail	49.9%	50.1%

Table 2: Probabilities of transitioning between tail and non-tail queries.

behavior in the second week of our dataset, and define *tail queries* to be those queries that did not appear in the first week of the dataset. This definition allows tail queries to be identified at query time, unlike the more global notion of singleton queries used in Figure 1.

Table 1 shows the probability distribution over the users' next actions (Click, Re-Query, End Session) following tail and non-tail queries. Users click less often following a tail query, suggesting that the results returned by search engines are not as valuable for tail queries. Searchers are also less likely to end their search session following a tail query than after a non-tail query. Users issuing tail queries more frequently issue a *reformulation* of their previous query.

Table 2 summarizes the nature of searchers' reformulations. We consider cases where users reformulated their original query into a new query, without clicking on a result from the original query. Users often reformulate non-tail queries into tail queries, and less often reformulate tail queries into non-tail queries. This finding is consistent with previous results showing that specializations are more common than generalizations [3]. The off-diagonal transitions also suggest an important distinction between a query's frequency and the commonality of the information need it serves: users often enter both rare and common queries in pursuit of the same goal. Distinguishing rare information needs from rare queries, based on clues like reformulations and destination URLs, is an intriguing direction for future work.

3. PREDICTING REFORMULATIONS

The data suggests that search engines are less effective on tail queries than on non-tail queries in that users are less likely to click results and more likely to reformulate, and that such reformulations are common. Predictive models of this reformulation behavior could improve search engine effectiveness, particularly on rare queries. Anticipating reformulations in advance could be used to support suggestions for query reformulations or improved ranking.

We constructed and evaluated a model that predicts query reformulations, given a current query and user. Using an approach based on Probabilistic Latent Semantic Analysis (PLSA) [2], we model each pair of consecutive queries that a user composes as generated by a single latent "topic" variable (z). The topic variable is generated from a multinomial specific to each user. Formally, the probability of a second query q_2 given an initial query q_1 and user U is:

$$P(q_2|q_1, U) = \alpha \sum_{z=1}^{T} P(q_2|z)P(q_1|z)P(z|U) \qquad (1)$$

where α is a normalization constant, and the number of topics T is a parameter (set to 200 in our experiments). We learned each of the multinomial distributions in Equation 1 using a training set of query transitions, and tested the model on held-out data. We examined two simpler clustering models: a PLSA model without the user information (*qq-PLSA*) and an PLSA model without query transition information (*qu-PLSA*). We also evaluated two non-clustering baseline models. The *Query* baseline predicts $P(q_2|q_1) = \lambda P_{train}(q_2) + (1 - \lambda)P_{train}(q_2|q_1)$, where P_{train} indicates maximum likelihood distributions learned from the training set, and λ is a learned parameter. The *Marginal* baseline simply predicts $P(q_2|q_1) = P_{train}(q_2)$.

To avoid assumptions about out-of-vocabulary terms, we evaluated our methods on only those test cases for which q_1 and q_2 appeared at least once in the training set. Because the output space of the models is large, we measured performance in terms of the perplexity measure commonly used in language modeling. A lower perplexity score indicates better performance. Table 3 shows that the PLSA-based techniques substantially outperform the Marginal and Query baselines, and the full PLSA model outperforms its simpler versions.

Although query clustering for reformulation suggestions has been investigated in prior research (*e.g.*, [1]), to our knowledge this is the first experiment on predicting specific reformulations. An advantage of the PLSA approach over previous techniques is that it can be readily augmented to incorporate new sources of information. In addition to the user and previous queries, the model can also include result URLs, individual query terms or phrases, or important relatedness indicators like the temporal delay between queries [3]. We are interested in exploring additional information sources in future work.

Marginal	Query	qq-PLSA	qu-PLSA	Full PLSA
109.3	39.7	22.3	24.3	14.0

Table 3: Performance of predicting query reformulations, measured by test set perplexity (in 000s).

4. CONCLUSIONS

We investigated several aspects of rare queries, including comparisons of search behavior following the input of common and rare queries. We identified differences suggesting that search engines perform less well on rare queries. We also studied transitions between rare and common queries during search sessions, highlighting the difference between the frequency of queries and information needs. Finally, we constructed and tested a probabilistic model to predict query reformulations given the preceding query and user.

5. REFERENCES

[1] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proc. of KDD*, 2000.

[2] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of UAI*, Stockholm, 1999.

[3] T. Lau and E. Horvitz. Patterns of search: Analyzing and modeling web query refinement. In *Proc. of UM*, 1999.

[4] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large altavista query log. Technical Report 1998-014, Digital SRC, 1998.