What Queries are Likely to Recur in Web Search?

Dell Zhang
SCSIS
Birkbeck, University of London
Malet Street
London WC1E 7HX, UK
dell.z@ieee.org

Jinsong Lu
SEMS
Birkbeck, University of London
Malet Street
London WC1E 7HX, UK
jingsong.lu@gmail.com

ABSTRACT

We study the recurrence dynamics of queries in Web search by analysing a large real-world query log dataset. We find that query frequency is more useful in predicting collective query recurrence whereas query recency is more useful in predicting individual query recurrence. Our findings provide valuable insights for understanding and improving Web search.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement

Keywords

Web Search, Web Mining, Query Log Analysis, Personalisation.

1. INTRODUCTION

What queries are are likely to recur? This question is crucial to the effective and efficient design of Web search engines. In this paper, we study the recurrence dynamics of queries, i.e., how people reuse their past queries, by analysing a large real-world query log dataset. Although there exists a lot of work in the analysis of Web query logs (such as [6]), the pattern of query recurrence is neither well-understood nor well-studied.

2. DATA

We use the AOL query log dataset [9] that is provided to the research community by AOL search engine[1] for our analysis. In this paper, we focus on the search events happened within the first week of March 2006. Each search event is represented by a tuple (u, q, t) which means user u issued query q at time t, and we sort all the search events by their time. The queries have already been normalised through punctuation-removal and case-folding etc. The final dataset used in this paper consists of $1,908,135$ queries from $309,078$ users.

[1] http://search.aol.com/

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

3. ANALYSIS

We consider two settings of query recurrence:

- **collective** — a query from user u is regarded as a recurrence if it has been used by *any* user before.
- **individual** — a query from user u is regarded as a recurrence if it has been used by u herself before;

For each incident of query recurrence (happened at time t), we compute:

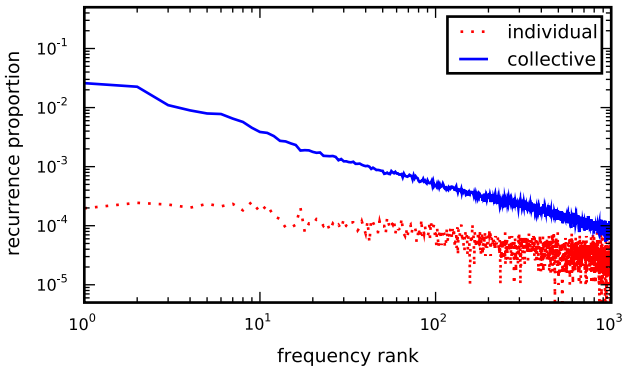
- the ***frequency*** of the query in the search history, i.e., how many times the query has been used before t;
- the ***recency*** of the query in the search history, i.e., how long ago the query was last used before t.

In each setting, we first rank all distinctive query recurrence incidents according to their frequency or recency values (in the collective or individual history), and then calculate the proportion of query recurrence incidents for each frequency or recency rank.

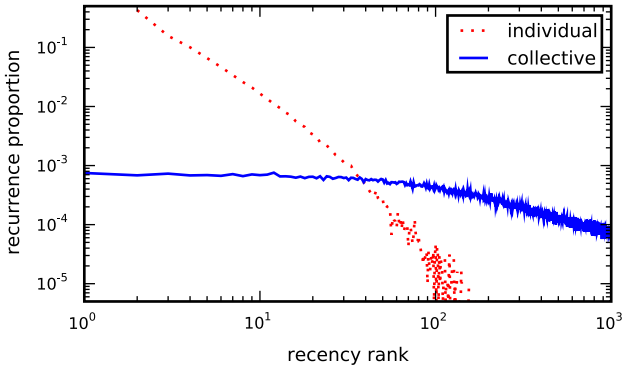
Figure 1(a) and 1(b) show the *log-log* plots of query recurrence proportion over query frequency and recency ranks respectively. We observe that in general, consistent with our intuition, (1) more *frequently* used queries are more likely to recur; and (2) more *recently* used queries are more likely to recur. More interestingly, the recurrence dynamics of queries in the two different settings exhibit drastically different characteristics: collective query recurrence is dominated by query frequency, whereas individual query recurrence is dominated by query recency. Furthermore, the two nearly-straight lines in the below log-log plots suggest that, the relationship between query recurrence proportion p and its frequency rank r (in the collective setting) or its recency rank r (in the individual setting) roughly follows the *power law* [8]: $p \propto 1/r^{\alpha}$, where the corresponding scaling-exponent α is shown in Table 1.

Table 1: The scaling-exponent α.

α	frequency	recency
collective	0.8996	——
individual	——	2.7523



(a)



(b)

Figure 1: The proportion of query recurrence over query frequency/recency.

We compute Kendall's rank-correlation coefficient τ [4] to quantitatively measure the utility of query frequency or recency in predicting query recurrence. Table 2 shows the values of τ calculated over the top 100 query frequency or recency ranks. We see that in the collective setting query recurrence is correlated more strongly with query frequency than recency, and in the individual setting query recurrence is correlated more strongly with query recency than frequency. Furthermore, query frequency and recency are indeed highly useful predictors for collective and individual query recurrence respectively, as reflected by the corresponding high τ values (close to the perfect correlation score 1).

Table 2: Kendall's rank-correlation coefficient τ.

τ	frequency	recency
collective	0.9395	0.8025
individual	0.4470	0.9367

A close inspection of the query log reveals that the better performance of query recency in predicting query recurrence on the individual level is mainly due to the greater burst and drift of user interests on the individual level.

4. CONCLUSIONS

In summary, according to our analysis based on the AOL query log dataset[2], we find that query frequency is more useful in predicting collective query recurrence whereas query recency is more useful in predicting individual query recurrence. Our findings provide valuable insights for the development of better *query suggestion* [7], *information re-finding* [10], and *result caching/prefetching* [5, 1, 2, 3] techniques etc. Most notably, we point out that temporal factors such as query recency, which have been largely overlooked in existing research, are actually very important to *personalised* Web search [6].

5. REFERENCES

[1] R. A. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. The impact of caching on search engines. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 183–190, Amsterdam, The Netherlands, 2007.

[2] T. Fagni, R. Perego, F. Silvestri, and S. Orlando. Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data. *ACM Transactions on Information Systems (TOIS)*, 24(1):51–78, 2006.

[3] Q. Gan and T. Suel. Improved techniques for result caching in web search engines. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*, pages 431–440, Madrid, Spain, 2009.

[4] M. Kendall and J. D. Gibbons. *Rank Correlation Methods*. A Charles Griffin Book, 5th edition, 1990.

[5] R. Lempel and S. Moran. Predictive caching and prefetching of query results in search engines. In *Proceedings of the 12th International World Wide Web Conference (WWW)*, pages 19–28, Budapest, Hungary, 2003.

[6] Q. Mei and K. W. Church. Entropy of search logs: How hard is search? with personalization? with backoff? In *Proceedings of the 1st International Conference on Web Search and Web Data Mining (WSDM)*, pages 45–54, Palo Alto, CA, USA, 2008.

[7] Q. Mei, D. Zhou, and K. W. Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 469–478, Napa Valley, CA, USA, 2008.

[8] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[9] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems (Infoscale)*, page 1, Hong Kong, 2006.

[10] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: Repeat queries in yahoo's logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 151–158, Amsterdam, The Netherlands, 2007.

[2]It would be necessary to confirm the findings on other query log datasets, but to the best of our knowledge all publicly available query log datasets (except the AOL one) do not contain *user-id* information because of privacy concerns.