Survival Analysis of Click Logs

Si-Chi Chin
Information Science
The University of Iowa
Iowa City, Iowa 52242
si-chi-chin@uiowa.edu

W. Nick Street
Department of Management Sciences
The University of Iowa
Iowa City, Iowa 52242
nick-street@uiowa.edu

ABSTRACT

Click logs from search engines provide a rich opportunity to acquire implicit feedback from users. Patterns derived from the time between a posted query and a click provide information on the ranking quality, reflecting the perceived relevance of a retrieved URL. This paper applies the Kaplan-Meier estimator to study click patterns. The visualization of click curves demonstrates the interaction between the relevance and the rank position of URLs. The observed results demonstrate the potential of using click curves to predict the quality of the top-ranked results.

Categories and Subject Descriptors: H.4 [Information Systems Applications]: Miscellaneous

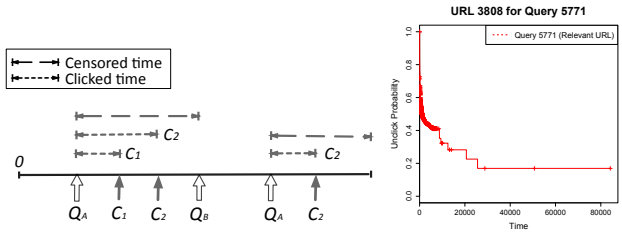
General Terms: Theory

Keywords: Clickthrough Logs, Kaplan-Meier Survival Analysis, Click Curve Visualization

1. INTRODUCTION AND DATASET

Ranking is a major concern to information retrieval applications such as document ranking on search engines. Implicit feedback from click logs can be used to approximate relevance judgements and improve rank results [7, 5]. This paper adopts Kaplan-Meier survival analysis [8] to examine clicking patterns, accounting for relevance labels and rank position of the URLs on a search engine result page (SERP). In medical research, the Kaplan-Meier estimator approximates the survival function, measuring the probability of patients surviving for different time points after treatment. We use the estimator to create click curves, showing the probability of URLs remaining unclicked across time. In both cases, the data is often right-censored, i.e. final outcomes are not observed for some instances. In click logs, censoring occurs when a search session ends or a query is replaced by the succeeding query. Figure 1(a) illustrates the definitions for clicked and censored time in one session. Click times are derived from the immediately previous query, and censored times include the time from the first Q_A to Q_B, and from second Q_A to the end of the session.

The visualization of click curves can show the interaction between relevance and position bias, making the knowledge of clicks more explicit. We expect that if a search engine returns a relevant URL on top of the ranked list, its click curve will drop drastically at the beginning and eventually approach zero, e.g., in Figure 1(b), URL 3808 is relevant to Query 5771. Our goal is to quantify and visualize the effects of relevance and ranking on the probability and speed of user clicks. To our knowledge, this is the first work that applying survival analysis to study click patterns.



(a) Clicked and Censored Times (b) URL 3808

Figure 1: Times and click curves.

We use the Yandex click log dataset provided by the Relevance Prediction Challenge and Workshop on Web Search Click Data (WSCD 2012). The goal of the challenge is to rank URLs by relevance based on user search behavior. The click log is grouped by sessions, containing query IDs, ranked URLs from the Yandex SERP for each query, and a sequence of click actions. Yandex assigned binary labels ("relevant" and "not relevant") to a subset of URLs appearing in the log. There are 5,191 unique queries in the original training set. We selected Query 5771 (Q5771), Query 222491 (Q222491), to test our approach. Q5771 and Q222491 each have 3468 and 2179 unique sessions, and have 14 and 13 judged URLs. As shown in Table 1, Q5771 has stronger performance in ranking than Query 222491. The top 10 ranked list in the table is sorted by the average rank for each labeled URLs displayed on SERP, showing the overall performance for the two queries. However, the top 10 URLs in actual session logs can be different from the table.

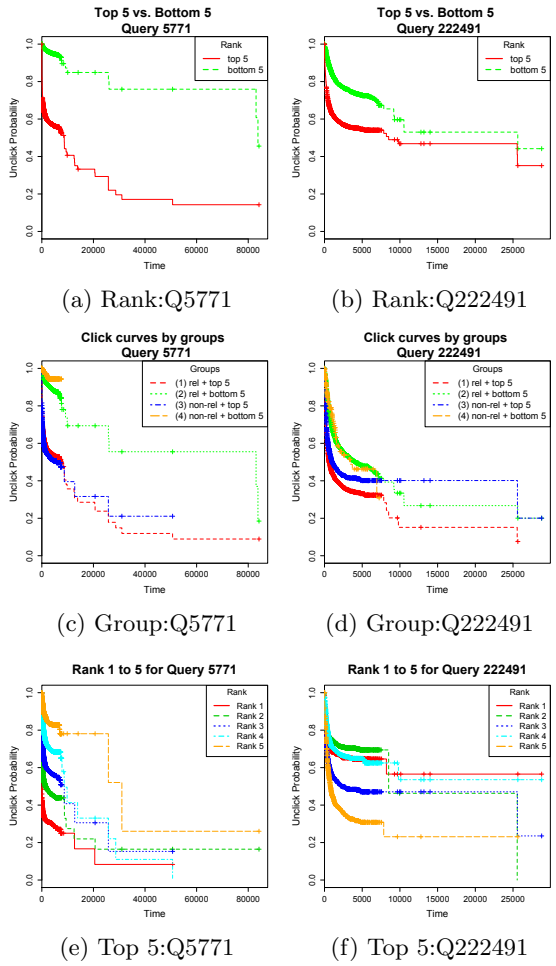
Table 1: Summary statistics for Q5771 and Q222491.

	Query 5771	Query 222491
Rank 1-10	{1,1,1,1,0,1,0,0,1,0}	{1,0,0,1,1,1,1,1,1,1}
NDCG@5	0.88	0.54
NDCG@10	0.73	0.69

2. CLICK CURVE ANALYSIS

Click curve analysis and visualization was performed using the R (http://www.r-project.org) survival package. Figure 2 shows the click curves for Q222491 and Q5771, grouping

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.



(a) Rank:Q5771 (b) Rank:Q222491

(c) Group:Q5771 (d) Group:Q222491

(e) Top 5:Q5771 (f) Top 5:Q222491

Figure 2: Click curves for various groupings of the URLs from the two queries.

URLs by rank and relevance. Log-rank tests ($p = 0.05$) were performed on all pairs of curves.

Both Figures 2(a) and 2(b) show significantly different clicking patterns for URLs in the top 5 on the SERP. However, the distance between the two curves diminishes as more relevant URLs were ranked lower for Q222491. The results indicate that the effect of position bias may vary among queries. To further study the interaction between relevance and rank, we divide URLs into four groups as shown in Figure 2(c) and 2(d). Group 1 contains relevant URLs ranked in the top 5; Group 2, relevant URLs in bottom 5; Group 3, irrelevant URLs in top 5; Group 4, irrelevant URLs in bottom 5. The pair-wise tests on groups are significant except in two cases: between Group 2 and 4 for Q222491, and between Group 1 and 3 for Q5771. Figure 2(c) indicates that the irrelevant URLs mixed in the top 5 can become hard to discover as Q5771 has evident position bias. On the other hand, since Q222491 has better ranking at the bottom, the differences between Group 2 and 4 diminish. In Figure 2(d), since the ranking on the top 5 for Query 222491 is imperfect, the click patterns for relevant and irrelevant URLs are distinct. Moreover, both Figure 2(c) and 2(d) show that, even given the position bias, the click curve for relevant URLs ranked at the bottom is still lower than irrelevant URLs on the top.

Figure 2(e) and 2(f) explore the effect of position bias on click curves at the position 1 to 5. As the ranked results are ideal for Q5771, we observed distinct clicking patterns on the top 5 results for Q5771 in Figure 2(e). Figure 2(f) shows a sharp drop on rank 5, implying the URL at rank 5 is more relevant than URLs at rank 2 and 3, which are judged not relevant to the query.

3. DISCUSSION AND CONCLUSION

Position bias is a primary concern in click logs. The probability of a click depends not only on its relevance, but on its position in the results page. Several papers formalized position bias using probabilistic models, such as the Cascade model [2], dynamic Bayseian network click model [1], Click Chain model [6], and user browsing model [3]. Unlike these models, we incorporate the time from a query to a click to model and visualize the unclicked probability of URLs. We believe that although top-ranked URLs may be clicked faster, a lower ranked but relevant URL may still be quickly discovered and clicked. The proposed click curve analysis exhibits distinct patterns between groups with various combinations of relevance and rank positions.

Although the time to first click was considered by Fox et al. [4], opportunities remain for exploring the use of activity time as a source of implicit feedback to judge the relevance of search results [5]. Compared to [4], we expand the scope from time to first click to time to *all* clicks. We adopted survival analysis to examine click patters in click logs, investigating the inter-related effect of relevance and rank positions.

This paper demonstrates the potential of applying survival analysis to determine the quality of ranked results. Visualizing click curves helps understand the interaction between the rank positions and the relevance of the retrieved URLs. The proposed approach can also be personalized to model individual users. Future work can focus on deriving functions for click curves to predict relevance labels for individual URLs. Click curves can also be included as an informative feature for learning to rank.

4. REFERENCES

[1] O. Chapelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking. In *WWW*, pages 1–10, Madrid, Spain, 2009.

[2] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM*, pages 87–94, Stanford, CA, 2008.

[3] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR*, pages 331–338, Singapore, 2008.

[4] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, Apr. 2005.

[5] X. Fu. Towards a model of implicit feedback for web search. *Journal of the American Society for Information Science and Technology*, 61(1):30–49, Jan. 2010.

[6] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, and C. Faloutsos. Click chain model in web search. In *WWW*, pages 11–20, Madrid, Spain, 2009.

[7] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161, Salvador, Brazil, 2005. ACM.

[8] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958.