Yahoo! Labs Technical Report No. YL-2010-006

Editorial relevance judgments are commonly used to evaluate search engine success, but these judgments are expensive and hard to collect at scale. On-line proxy of editorial judgments that can be easily computed from server click log data, with no repeatedly required editorial component can be useful for that reason. We analyzed the quality of clicked documents as a function of the dwell-time (time between two consecutive events) and found that a clicked document was more than 65% likely to get a grade of "good" or higher from the editors on a five point scale, and the likelihood increases to 77% if we consider clicked documents with dwell-time greater than 100 seconds. We also discovered that presence of at least one long dwell-time clicks (AOLC) in a session is associated with higher rate of returning for another session within the next 24 hour period. We used survival analysis and computed the probability that the users return as a function of time. Finally, we present an application of using AOLC to compare summary generation algorithms.

TECHNICAL REPORT

YL-2010-006

ON THE USE OF LONG DWELL TIME CLICKS FOR MEASURING USER SATISFACTION — WITH APPLICATION TO WEB SUMMARIZATION

D. Ciemiewicz, T. Kanungo, A. Laxminarayan and M. Stone
Yahoo! Labs, 2821 Mission College Blvd., Santa Clara, CA 95054

September 1, 2010

Bangalore • Barcelona • Haifa • Montreal • New York
Santiago • Silicon Valley

Original date of creation: February 23, 2009

Yahoo! Labs Technical Report No. YL-2010-006

Yahoo! Labs Technical Report No. YL-2010-006

ON THE USE OF LONG DWELL TIME CLICKS FOR MEASURING USER SATISFACTION — WITH APPLICATION TO WEB SUMMARIZATION

D. Ciemiewicz, T. Kanungo, A. Laxminarayan and M. Stone
Yahoo! Labs, 2821 Mission College Blvd., Santa Clara, CA 95054

September 1, 2010*

1. Introduction

The goal of any changes to search results ranking or presentation is to improve user satisfaction with search results. The ultimate goal for a commercial search engine is increasing market share — presumably, increased satisfaction with search results will eventually lead to higher market share. It is common to rely on editorial (human rater) evaluation of both perceived and experienced search results relevance to approximate end-user satisfaction. These judgments are expensive and time-consuming to collect. For this reason, much effort has gone towards understanding online implicit markers of satisfying search experience. One common intuition is that click properties, particularly order and time-based properties, such as duration of a click, or time to a click, as well as click order contain such implied signals of relevance that can be exploited. Most previous work that tried to take click properties into account involved small-scale data collection that relied on elaborate logs obtained with browser plugins (myref). Also, the outcome of this work is mixed, with no one useful simple click property emerging that can be reliably used to increase the probability that clicks with that property resulted in better user satisfaction than clicks without this property. An additional practical constraint we faced is to be able to collect these measures at scale, using server side logs, rather than more extensive measures that can be collected via toolbar or other browser plug-ins. If such measures could be collected, more elaborate relevance feedback, besides simple clicks can be used for evaluation and training purposes. An even better outcome is to tie these measures to possible market share changes. [1] [5] [4] [3] [7]

2. Long Dwell Time Click

We collected a random sample of search results clicks from a major search engine over 24 hour period. We then obtained editorial judgments of relevance for each query-clicked document pair on a standard PEGFB scale. We also examined additional binary click features, such as whether or not this click was the last click on the page, last click in session, whether click duration was longest for that page, whether this was the last click on a given url and many more properties. For each discrete feature, we computed precision and recall for each value, and for continuous variables such as click duration, we plotted precision, recall and F value as a function of duration threshold.

*Original date of creation: February 23, 2009

In the plot 1 below demonstrates the relationship between click duration threshold, precision, recall, accuracy and f value for clicked documents with editorial judgments of "good" or better. As is clear from this graph, precision reaches 77% around 80 seconds, and then gradually increases to 80% over time. However, recall declines sharply, and while it is nearly 77% around 80 seconds, and goes down to nearly 65% by 200 seconds. The best balance of recall and precision based on optimizing the f value (geometric mean of precision and recall with alpha value of .5) appears to be around 100 seconds.

3. Probability of Not Returning

To examine session-level user return data, we coded sessions as containing no long clicks, only long clicks, and some long clicks and examined probability of return after each type of session over the 24 hour period over which clicks were sampled. We used survival analysis [6] to compute the probability that a user does not return as a function of time after sessions containing only short clicks, only long clicks, or at least one long click (AOLC). It is clear from this graph (see Figure 2) that users are more likely to *not* return after a session without any long clicks. In conclusion, long clicks are more likely for documents judged "good" or better, and the absence of long clicks in a session leads to a higher probability that the user will not return for another session within the next 24 hours.

4. Web Summarization

We used the notion of at least one long click (AOLC) discussed in the previous section to compare a control web summarization algorithm [2, 8] to multiple test web summarization algorithms. The one of test summarization algorithms (MAP15) was trained using human quality judgments (groundtruth) and MAP12 and MAP14 were increasingly degraded versions of MAP15. Click data was collected from a search engine server log for a period of one week and we computed the queries that resulted in AOLC for until each position on the SERP. In Figure 3 we see that relative to the control algorithm AOLC increases increases when the test summary algorithm's quality is improved. The AOLC metric thus provides us with a methodology to evaluate summarization algorithms using clicks and gives us a faster surrogate for the slower human evaluation process.

5. Conclusion

We show that long dwell-time clicks are highly correlated with more relevant documents. In addition we show that having at least one long click in a session can be used for predicting the probability that the user will return to the search engine. Finally we show how the AOLC criterion can be used for evaluating web summarization algorithm by deliberately degrading their quality.

References

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of ACM SIGIR Conference*, pages 19–26, 2006.

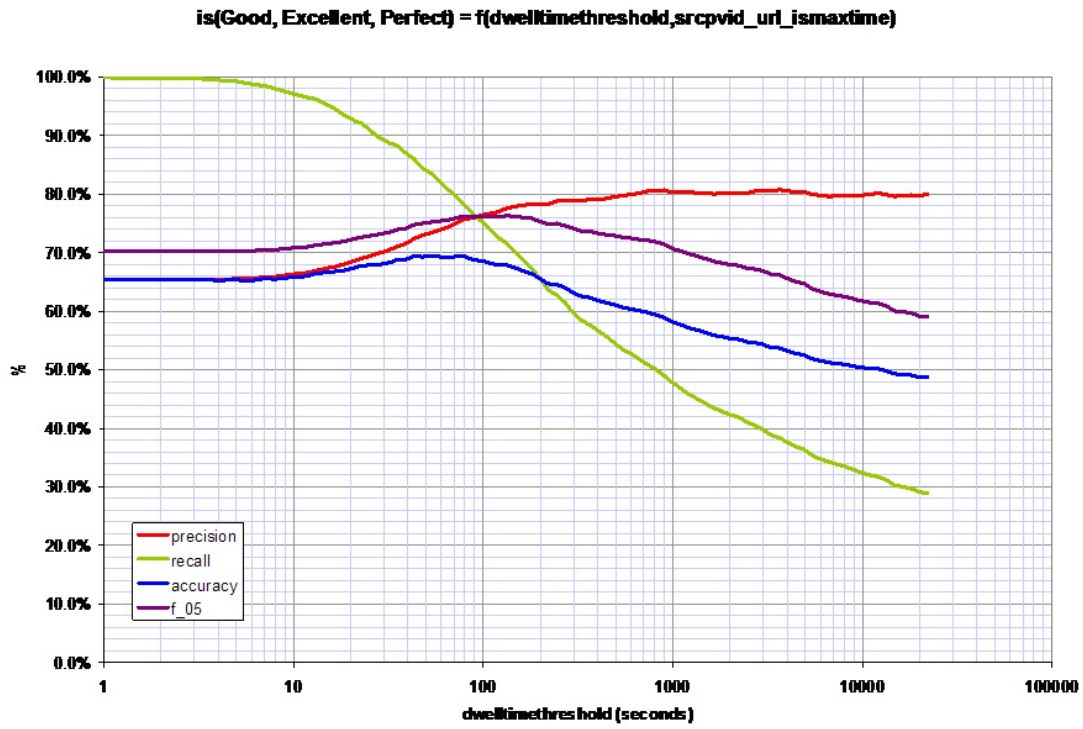

Figure 1: Precision, recall and accuracy and F value as a function of dwell time threshold.

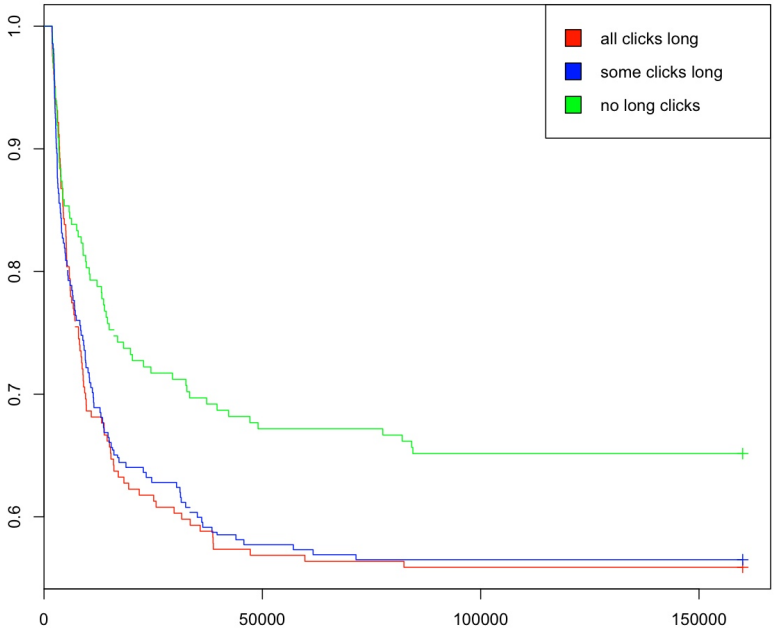

Figure 2: Non-return probability.

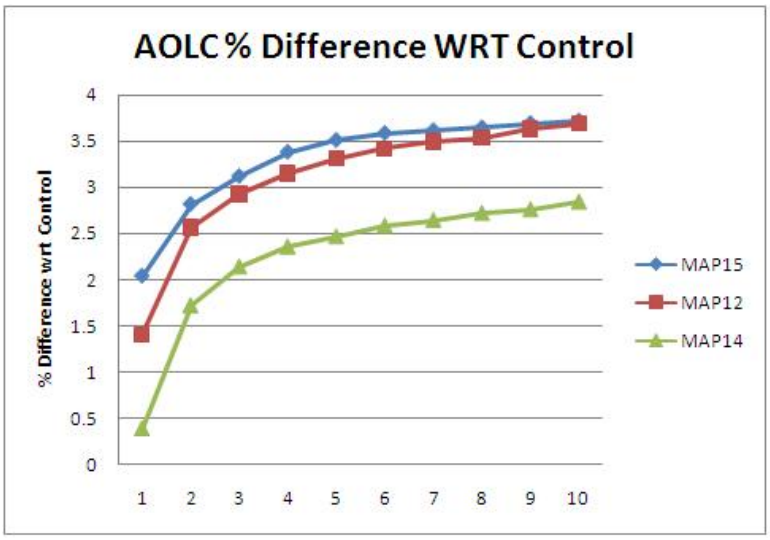

Figure 3: X-Axis is the position on the SERP. Y-Axis is percentage increase in number of queries for which there is at least one long click (AOLC) uptil position i on the SERP, with respect to control algorithm. Test summarization algorithm MAP15 has no degradation, MAP12 and MAP14 are increasingly degraded versions of MAP15.

[2] C. Clarke, E. Agchtein, S. Dumais, and R. White. The influence of caption features on click-through patterns in web search. In *Proc. of SIGIR*, 2007.

[3] E. Cutrell and Z. Guan. What are you looking for? an eye tracking study of information usage in web search. In *Proc. of SIGCHI*, 2007.

[4] S. Dumais, T. Joachims, K. Bharat, and A. Weigend. SIGIR 2003 workshop report: Implicit measures of user interests and preferences. *ACM SIGIR Forum*, 37:50–54, 2003.

[5] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23:147–168, 2005.

[6] D. Hosmer and S. Lemenshow. *Applied Survival Analysis*. John Wiley and Sons, Inc., 1999.

[7] J. Kamps, M. Koolen, and A. Trotman. Comparative analysis of clicks and judgments for IR evaluation. In *Proceedings of WSCD09*, 2009.

[8] T. Kanungo and D. Orr. Predicting readability of short web summaries. In *Proceedings of Second ACM Conference on Web Search and Data Mining*, 2009.