On Correlation of Absence Time and Search Effectiveness

Sunandan Chakraborty*
New York University
New York, USA
sunandan@cs.nyu.edu

Filip Radlinski
Microsoft
Cambridge, UK

Milad Shokouhi
Microsoft
Cambridge, UK

Paul Baecke
Microsoft
London, UK

{filiprad, milads, pbaecke}@microsoft.com

ABSTRACT

Online search evaluation metrics are typically derived based on implicit feedback from the users. For instance, computing the number of page clicks, number of queries, or dwell time on a search result. In a recent paper, Dupret and Lalmas introduced a new metric called *absence time*, which uses the time interval between successive sessions of users to measure their satisfaction with the system. They evaluated this metric on a version of Yahoo! Answers. In this paper, we investigate the effectiveness of absence time in evaluating new features in a web search engine, such as new ranking algorithm or a new user interface. We measured the variation of absence time to the effects of 21 experiments performed on a search engine. Our findings show that the outcomes of absence time agreed with the judgement of human experts performing a thorough analysis of a wide range of online and offline metrics in 14 out of these 21 cases.

We also investigated the relationship between absence time and a set of commonly-used covariates (features) such as the number of queries and clicks in the session. Our results suggest that users are likely to return to the search engine sooner when their previous session has more queries and more clicks.

Categories and Subject Descriptors

Information Systems [**Information Retrieval**]: Evaluation of retrieval results—*Retrieval effectiveness*

Keywords

Absence time; Survival analysis; Search evaluation

1. INTRODUCTION

There are different metrics to measure the efficiency of a new treatment in web search engines. These metrics can be broadly classified into two groups; offline and online [10]. Offline metrics deal with the accuracy of the retrieval process and are usually measured before deployment. On the other hand online metrics try to evaluate a new treatment (e.g. new ranking algorithm) in an already deployed system. Online evaluation metrics are mostly based on the implicit feedback from the users, which measures how the users are interacting with the system.

Some of the popular online metrics currently used include, *clickthrough rate* (CTR) [13, 17], *queries per user* (QPU) [14] and *dwell time* [22]. Clickthrough rate records the number of clicks on each result link in the search result page. While simplicity and general effectiveness of CTR have made it a widely popular choice as a metric [13, 17, 18], clicks are subject to different biases and can be misleading [8, 15]. Furthermore, CTR cannot be used in cases where clicks are not necessary to satisfy the information need (e.g. various vertical results such as time, weather, currency conversion etc.). Similarly, while an increase in QPU can be regarded as a sign of user satisfaction and growing trust, a short-term boost in QPU can be also interpreted in the opposite way as users tend to submit more queries when they struggle to find the information they need [11]. Again, dwell time over a fixed threshold cannot always determine user satisfaction accurately. High dwell time can be due to the topic, readability and length of the text of the target pages[15]. So, dwell time as a metric might not give consistent outcomes.

Dupret and Lalmas [7] proposed a new metric to measure user engagement called *absence time*. Absence time of a user is defined by the time between two consecutive sessions of the user, where a session is a time period when the user has been continuously active with one or more information needs. This metric reflects how often and how soon a user is coming back to use the system again. By definition, lower absence time indicates higher user engagement and is an evidence of better satisfaction. Dupret and Lalmas evaluated the metric of absence time for the users using a particular version of Yahoo! Answers. Their findings showed that absence time could produce a stable evaluation of the system compared to other standard online evaluation metrics.

In this paper, we investigate the effectiveness of absence time metric in the context of web search evaluation. We explore absence time using the concept of *survival analysis* and Cox model [6, 7]. We apply this technique to compare control-treatment pairs in 21 different ranking experiments and find that it correctly detects the better run in 14 of the cases. We also explore various user activities that are related to absence time. For instance, we find that users are more likely to return when their last search session involved more queries or more clicks. Overall, we believe that our results help us to achieve a better understanding of absence time as a metric and learn about other covariants that can potentially influence it.

2. ABSENCE TIME

Absence time is the time interval when the user was absent from the system, where lower absence time signifies higher user engagement. The definition of the metric is based on the intuition that, when a user is satisfied, he/she will come back more frequently to

*Work done during internship at Microsoft Research Cambridge.

SIGIR '14, July 06 – 11 2014, Gold Coast, Queensland, Australia

http://dx.doi.org/10.1145/2600428.2609535.

meet his/her future information needs. Absence time has certain advantages over other metrics. For example, any click related metric fails when the search result page (SERP) itself has the information the user is looking for. This can happen when the text snippets contain the information, or when rich inline vertical answers are displayed for queries such as those related to stock prices or sports results. In such cases, the click information will erroneously underestimate user satisfaction. Similarly, dwell time can be a strong indicator of how good the results are, but can sometimes inaccurately overestimate the engagement as the dwell time can vary depending on a variety of factors [15, 22]. Since absence time is independent of such biases, it can uniformly judge any kind of activity on a search engine.

A user session is defined as the time period when the user has been actively interacting with the system to meet his/her information needs. User activities include, submission of a query, clicking on a result link or clicking on an ad, etc. The absence time is defined as the time between two such successive sessions. In our case, we followed the common definition [2] and have defined boundaries of user sessions where there has been no activity for at least 30 minutes. In other words, if there is a time gap of 30 minutes between two consecutive actions (e.g. query submission, result click, ad click etc.), session boundaries were drawn. Hence, in our case absence time will have a minimum value of 30 minutes. The analysis of absence time and its characterization has been done by adapting the concept of survival analysis and Cox models.

3. SURVIVAL ANALYSIS AND COX MODEL

Survival analysis investigates *event history* by modeling occurrences of an event and its dependence on various factors. This procedure is most commonly used in medical science to model survival rate of patients in response to different treatments [19] but its application is also common in other fields such as sociology [3], and economics [16]. In survival analysis, the survival function provides us with the chances that a subject will survive beyond a particular time *t*. On the other hand, hazard rate *h(t)*, describes the risk of a specified event occurring at time *t*, based on the subject's survival up to that time. $h(t)$ is defined as,

$$h(t) = \lim_{\Delta t \to 0} \frac{P[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} \tag{1}$$

The Cox model [6] is one of the most commonly used techniques in survival analysis. It models the hazard rate of an event based on various covariates. Here, covariates are dependent variables, which can potentially influence the hazard rate, in addition to the new treatment. Hazard rate $h(t)$ is a function of time, which examines the relationship of the different covariates on the hazard rate. Assuming the covariates vary linearly, the linear model of log hazard rate for a treatment condition i can be represented as,

$$h_i(t) = h_0(t) e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_k x_{ik}} \tag{2}$$

Here, xs are the covariates and the βs are the weights signifying the influence of a particular covariate on the hazard rate. The baseline hazard rate is denoted by $h_0(t)$, and it represents the hazard rate when all the x values are 0. The Cox model is a proportional hazard model, computed as the ratio of two hazard rates. Given two treatment conditions, i and j, the hazard ratio is given by,

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_k x_{ik}}}{h_0(t) e^{\beta_1 x_{j1} + \beta_2 x_{j2} + \ldots \beta_k x_{jk}}} \tag{3}$$

As the $h_0(t)$ term cancels out in the right hand side, this ratio becomes independent of time and the baseline hazard rate can remain unspecified. Using this model, we can test whether a new search engine feature results in more frequent visits by the user. In other words, if the absence time of the users decreases after the introduction of a new feature, it means that the change is a positive one.

In our setting, these tests were done simultaneously on two separate sets of users, one from the treatment group who were exposed to the new feature, and the other from the control group using the original version. Significant difference in the two group's absence time can indicate the effect of the new feature. The hazard rate is computed for the event of a user coming back to use the system. Consequently, increased hazard rate translates into more frequent returning visits by the user, i.e. reduced absence time. In the simplest case, our model assumes that this hazard rate (or, returning rate of users) is only dependent upon whether the user is coming from the treatment group or not. Hence, we have only one variable x_1 where, $x_1 = 1$ if the user is coming from the treatment group, and $x_1 = 0$ otherwise. The treatment hazard rate (τ) and the control ($\mathcal{C}$) hazard rate are defined respectively as,

$$h_\tau(t) = h_0(t) e^{\beta_1 x_1} = h_0(t) e^{\beta_1} \quad [x_1 = 1] \tag{4}$$

$$h_\mathcal{C}(t) = h_0(t) e^{\beta_1 x_1} = h_0(t) \quad [x_1 = 0] \tag{5}$$

Here, the control hazard rate coincides with the baseline hazard rate. Finally, proportional rate become,

$$h_\tau(t) = e^{\beta_1} h_\mathcal{C}(t) \tag{6}$$

If $e^{\beta_1} > 1$ ($\beta_1 > 0$) then the treatment hazard rate is higher, i.e. treatment group absence time is lower or the new feature is effective.

4. EXPERIMENTS AND RESULTS

We implemented absence time based on the Cox model to evaluate various online search treatments presented to live users. Given a new feature of the system, the model is designed to compare the absence time of the users who are using the new feature against a control group using the original version. The effectiveness of an online metric depends on its discriminative power (efficiency), and obviously on the fact that it can correctly determine the better system between control and treatment (precision). The effect of a new feature can be positive, negative or neutral. We conjecture that for the first two cases, the absence time should be considerably different between the treatment and control groups of users. Moreover, for a positive feature, absence time should be statistically significantly lower for treatment users. If the empirical results match these hypotheses, it can be claimed that the absence time is an effective metric for evaluation. In the remainder of this paper we put these hypotheses into test by computing the absence time over several search treatments collected over large groups of users.

4.1 Data

We ran our experiments on the search logs collected from 21 previously experiments testing different new features of the Bing search engine. The new features tested were on different areas, such as ranking algorithm, user interface modifications and changes related to ads. All the experiments were manually inspected and hand labelled by a group of human experts as *positive* or *negative* based on inspecting sampled sessions and reviewing the outcome of a large number of online A/B testing metrics, which we used as our ground truth. Each experiment was tested *online* against the live traffic of Bing search engine for a short period of 1-4 weeks (median 2 weeks). The users in each experiment were split randomly into two groups (control vs treatment), each group receiving a unique and consistent search experience throughout the testing period.

Table 1: Performance of absence time as a metric compared to expert labels over a set of 21 experimental control-treatment pairs. Positive (green) and negative (red) labels respectively represent cases where treatment run performs better and worse than control according to the Cox model or expert ground-truth. Statistically significant differences ($p < 0.05$) according to the likelihood ratio test are denoted by $*$.

Experiment seq no.	e^{β}	p-value	Cox model Label	Expert Ground-truth
1	0.999	0.149	negative	negative
2	1.000	0.145	positive	positive
3	1.009	0.007	positive*	negative
4	1.000	0.825	positive	positive
5	1.001	0.270	positive	positive
6	1.026	0.000	positive*	positive
7	1.000	0.816	positive	positive
8	0.999	0.556	negative	positive
9	0.999	0.718	negative	positive
10	0.997	0.000	negative*	negative
11	1.008	0.010	positive*	positive
12	1.006	0.010	positive*	positive
13	1.001	0.009	positive*	positive
14	0.997	0.011	negative*	positive
15	0.999	0.997	negative	positive
16	0.999	0.152	negative	positive
17	0.998	0.030	negative*	positive
18	0.999	0.824	negative	negative
19	1.000	0.027	positive*	positive
20	1.000	0.603	positive	positive
21	1.050	0.000	positive*	positive

4.2 Results

For each anonymous user, the search logs containing queries and clicks were processed to identify sessions and the corresponding absence times. As mentioned earlier, the session boundaries were drawn when there has been a gap in activity of 30 minutes. We considered the case of a user returning to search as an event and the time taken to return (i.e. the absence time) to compute the rate of that event occurring. These data were used to implement the Cox model as explained in Equations 4, 5 and 6. Initially, we use one feature (co-variate) only; whether the user is from the control group ($x = 0$) or from the treatment group ($x = 1$). Hence, the hazard rate of a group translates into the rate of coming back to use the system. For a particular group, higher hazard rate means that those users are coming back sooner and are more satisfied with the system.

We implemented our model and applied in 21 different experiments and compared the outcome with the labels assigned by human experts, to see in how many cases the model's outcomes agree with the expert opinion. The results of this experiment are summarized in Table 1. In 14 out of 21 cases the absence time agrees with the expert labels. Out of these 14 cases, the outcome was statistically significant in 7 cases ($p < 0.05$) according to likelihood ratio test. The majority of misclassified cases were false negatives (6 out of 7 in total, and 3 statistically significant), where positive treatments were classified as negatives.

Absence time as a metric of evaluation can be better understood by comparing it with the performance of standard metrics. We evaluated the same 21 experiments with several other metrics as baselines and compared their outcomes with the same ground truth. The metrics used were,

- Queries per User (QPU).
- Average Result Clicks per User (RCU).
- Average Ad Clicks per User (ACU).
- SAT clicks per user (SAT): A SAT click [9] is defined as a click which is not followed by another click within 30 seconds. A SAT click is an indication of user satisfaction.
- Quickback Clicks per User (QbCU): Quickback click is a click, where the dwell time is less than 30 seconds [21]. Quickback clicks are usually considered as negative feedback from users.

The most successful metric in this experiment was RCU, which could correctly identify 15 out of 21 experiments. Absence time and SAT clicks had the next best success with 14 correct identifications. This was followed respectively by QPU (7), QbCU (6) and ACU (3). Although, RCU demonstrated the best performance, it is important to note that RCU is not applicable in cases where information needs can be satisfied with no clicks.

For a finer analysis of absence time and its characteristics, we added more features (covariants) to the default Cox model. The purpose of this experiment was to investigate what influences absence time (apart from being in the treatment or control group). We assumed that the absence time between n_ith and n_{i+1}th session is only influenced by the activities of the user in the n_ith session. That is, duration of being absent between sessions is dependent upon the experiences of the user in the just concluded session. In this experiment, we used the more general version of the Cox model (Equation 2) to add these covariates or features of absence time. We used the following features of the previous session, which represent a summary and a description of users' activity in that session.

- Number of queries in the previous session.
- Number of clicks in the previous session.
- Was the session abandoned: A session is said to be abandoned if the session had no clicks [7, 17]. Abandoned session is a reflection of poor performance of the system.
- Was there any query reformulations: High number of query reformulations in a session is usually considered to be an indication of user dissatisfaction [7, 11, 14, 17].
- Was there any SAT click [9, 21].
- Was there any quickback clicks [21].

In this experiment, we used the same control-treatment pairs as in the previous experiment and applied Equation 2 to include all the features from the above list together as co-variates. The results of this experiment is summarized in Table 2. The numbers suggest that if there is an increase in the number of queries by 1 in a session, the absence time between that session and the next one decreases by 0.3%. Similarly, for page clicks the absence time decreases by 0.8% with every click on search result page. For the rest of the features, we observe that absence time decreases with satisfaction and increases with dissatisfaction, like, absence time tends to increase if the previous session was abandoned [7, 17] or had quickback clicks [21]. On the other hand, the absence time decreases if there was a SAT click [9, 21]. However, the presence of a reformulated query in a session has a counter-intuitive effect on absence time and decreases with reformulation queries. Previous work has shown that query reformulation is an indication of dissatisfaction [11, 17], hence absence time was expected to increase with query reformulation. However, a reformulated query can mean that user was not dissatisfied and used reformulation to disambiguate a query [14], which improved the search results. This can explain why reformulated queries can have a positive effect on absence time. The impact of all these covariants was identified by the likelihood ratio test as statistically significant ($p < 0.05$).

Table 2: Influence of user activities in a session on absence time. The third column represents the change brought in the users' absence time if there has been a change in the feature value (second column). The impact of all these covariants was identified by the likelihood ratio test as statistically significant ($p < 0.05$).

Feature	Change	Effect on absence time
No. of queries	+1	-0.3%
No. of page clicks	+1	-0.8%
Has reformulation	Yes	-1.4%
Is Abandoned	Yes	+0.6%
SAT Clicks	Yes	-1.2%
Quickback Clicks	Yes	+2.3%

5. RELATED WORK

The recent trends in evaluating different aspects of a search engine are based on how the users are interacting with the system [1]. These methods have certain advantages over earlier offline methods in terms of time and cost, and are easier to use in evaluating already deployed systems. Researchers in the recent past have explored a variety of techniques to model user interactions and engagement with the search engine. CTR is one of the most widely researched metrics to evaluate relevance and ranking of documents [4, 13, 17]. An alternative to CTR is the *pSkip* metric [20] that has been suggested to remove some of the biases imposed by user clicks. Interleaving [5] blends the results returned by two search systems (control versus treatment) and uses the collected clicks to decide which one is better. Other user activities which have been used for evaluation include, eye-tracking [13] and mouse cursor movement [12]. Metrics, which are related to time intervals, include dwell time [22] and time to first click [17].

6. CONCLUSIONS

In this paper, we investigated the effectiveness of absence time [7] for identifying differences in search quality. We applied absence time to evaluate 21 different control-treatment pairs of ranking experiments tested on the Bing search engine and found that the absence time could correctly identify the better run in 14 of them. Moreover, our experiments demonstrated that some activities during a session can influence the users' absence time just after the session concludes. We also compared the performance of absence time with some other standard metrics and found that absence time performed better than most of them. For a more minute assessment of absence time as a metric, future work can be directed at identifying the types of treatment in the search engine (e.g. ranking, user interface etc.), where absence time can be more effective. It would be also interesting to investigate the sensitivity of absence time for evaluation, i.e. how long an experiment should run or how many users/sessions it needs for a more accurate deduction.

References

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proc. SIGIR*, pages 19–26, Seattle, Washington, USA, 2006.

[2] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proc. SIGIR*, pages 185–194, Portland, OR, 2012.

[3] D. Borsic and A. Kavkler. Duration of regional unemployment spells in slovenia. *Managing Global Transitions*, 7(2):123–146, 2009.

[4] B. Carterette and R. Jones. Evaluating search engines by modeling the relationship between relevance and clicks. In *Proc. NIPS*, 2007.

[5] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.*, 30(1), Mar. 2012.

[6] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2), Mar. 1972.

[7] G. Dupret and M. Lalmas. Absence time and user engagement: Evaluating ranking functions. In *Proc. WSDM*, pages 173–182, Rome, Italy, 2013.

[8] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using click-through data and a user model. In *Proc. WWW*, Banff, Canada, 2007.

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

[10] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *J. Mach. Learn. Res.*, 10:2935–2962, Dec. 2009.

[11] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proc. CIKM*, pages 2019–2028, San Francisco, California, USA, 2013.

[12] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: Using cursor movements to understand and improve search. In *SIGCHI*, pages 1225–1234, Vancouver, BC, Canada, 2011.

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

[14] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2), Apr. 2007.

[15] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proc. WSDM*, pages 193–202, New York, New York, USA, 2014.

[16] M. J. LeClere. Preface modeling time to event: Applications of survival analysis in accounting, economics and finance. *Review of Accounting and Finance*, 4(4):5 – 12, 2005.

[17] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proc. CIKM*, pages 43–52, Napa Valley, California, USA, 2008.

[18] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proc. WWW*, pages 521–530, Banff, Alberta, Canada, 2007.

[19] S. L. Spruance, J. E. Reid1, M. Grace, and M. Samore. Hazard ratio in clinical trials. *Antimicrob Agents Chemother*, 48(8):2787–2792, 2004.

[20] K. Wang, T. Walker, and Z. Zheng. Pskip: Estimating relevance ranking quality from web search clickthrough data. In *Proc. SIGKDD*, pages 1355–1364, Paris, France, 2009.

[21] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang. Enhancing personalized search by mining and modeling task behavior. In *WWW*, pages 1411–1420, Rio de Janeiro, Brazil, 2013.

[22] S. Xu, H. Jiang, and F. C. M. Lau. Mining user dwell time for personalized web search re-ranking. In *Proc. IJCAI*, pages 2367–2372, Barcelona, Catalonia, Spain, 2011.