Absence Time and User Engagement: Evaluating Ranking Functions

Georges Dupret
Yahoo! Labs Sunnyvale
gdupret@yahoo-inc.com

Mounia Lalmas
Yahoo! Labs Barcelona
mounia@acm.org

ABSTRACT

In the online industry, user engagement is measured with various engagement metrics used to assess users' depth of engagement with a website. Widely-used metrics include clickthrough rates, page views and dwell time. Relying solely on these metrics can lead to contradictory if not erroneous conclusions regarding user engagement. In this paper, we propose the time between two user visits, or the `absence` time, to measure user engagement. Our assumption is that if users find a website interesting, engaging or useful, they will return to it sooner – a reflection of their engagement with the site – than if this is not the case. This assumption has the advantage of being simple and intuitive and applicable to a large number of settings. As a case study, we use a community Q&A website, and compare the behaviour of users exposed to six functions used to rank past answers, both in terms of traditional metrics and `absence` time. We use Survival Analysis to show the relation between `absence` time and other engagement metrics. We demonstrate that the `absence` time leads to coherent, interpretable results and helps to better understand other metrics commonly used to evaluate user engagement in search.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing

General Terms

Measurement

Keywords

User engagement, time between visits, metrics, ranking evaluation, interleaving, search engine, clickthrough data.

1. INTRODUCTION

In the online industry, *user engagement* refers to the quality of the user experience with a website. Various *engagement metrics* used to assess users' depth of engagement have been proposed. Widely-used metrics include clickthrough rates, page views, time spent on a site ("dwell time"), or more generally "activity" metrics which relate to the user behaviour during an online session. Another class of metrics, "loyalty" metrics, are concerned with how often users return to a site[1].

Dwell time has proven to be a robust measure of user engagement over the years, for example in the context of web search where it is used to improve retrieval performance [2, 3]. The same holds for clickthrough rates because they provide a clear signal that users were attracted to the content of a site or to a search result. They have been used to compute relevance scores or to personalise websites [7, 8, 16].

Relying solely on user clicks and time spent can, however, lead to contradictory if not erroneous conclusions regarding user engagement, as they do not necessarily relate to users being engaged. In particular, with the current trend of displaying rich information on web pages in special modules or inserts called "Direct Displays[2]", for instance the phone number of restaurants or weather data in search results, users do not need to click to access the information and the time spent on a website is shorter.

In this paper, we propose the time between two user visits, or the `absence` time, to measure user engagement. Our intuition is that if users find a site interesting, engaging or useful, they will return to it sooner. The `absence` time is simply the time elapsed between two sessions of a user. More precisely, it is the time between the end of a session and the start of a new session with respect to X where X is a site, a set of sites (e.g. Yahoo! network of services), but also any part of a page, for example a particular module. We use the term `absence` time instead of return time to avoid ambiguity, since a return can occur during a session of a multitasking user, for example.

The `absence` time measures the time it takes a user to decide to return to the site of interest to accomplish a new task. Taking a news site as an example, a good experience associated with quality articles might motivate the user to come back to that news site on a regular basis. On the other hand, if the user is disappointed (the articles were not interesting, the site was confusing) he or she may return less often and even switch to an alternative news provider. Another example is a visit to a community questions and answers website. If the questions of a user are well and promptly answered, the odds are that he or she will be enticed to raise

WSDM'13, February 4–8, 2013, Rome, Italy.
Copyright 2013 ACM 978-1-4503-1869-3/13/02 ...$15.00.

[1] In this paper, we use the terms site, website, application and service interchangeably.

[2] Also sometimes called "Answers".

new questions and return to the site soon. *In summary, we assume that engaged users come back sooner, and hence their* `absence` *times are shorter.* This assumption has the advantage of being simple and intuitive and applicable to a large number of settings.

The aim of this work is to identify observable correlations between the `absence` time and user engagement. However, since user engagement is not directly observable, we study the relation between the `absence` time and various established "activity" metrics such as clicks and page views. We show that these metrics complement each other and that the `absence` time adds significant insights to the interpretation of activity metrics.

Using `absence` time to measure use engagement has three potential issues. A user might decide to return sooner or later to a website due to reasons unrelated with the previous visits (being on holidays for example). The signal is therefore expected to be noisy, so it is important to have a large sample of interaction data to detect coherent signals and to develop methods to take systematic effects into account. Another issue is how to identify session boundaries. It is sometimes difficult to decide whether a particular user action initiates a new session or belong to a previous one, and this decision has potentially a large impact on the `absence` time estimates. Finally, changes that affect the user experience, especially visible one such as a new interface, can have a transient impact. We must wait until the "novelty" effect shades away before studying `absence` time, or we must model it. In this paper, we account for the first two issues.

As a case study, we use a community querying and answering website hosted by Yahoo! Japan. We compare the behaviour of users exposed to six functions used to rank past answers both in term of traditional metrics and of `absence` time. We organise the rest of this paper as follows. Section 2 describes related work. Section 3 presents the application chosen in this study and limitations with common measurement approaches. We show in Section 4 how Survival Analysis and the `absence` time can be used to measure engagement. Section 5 shows how the `absence` time relates to different measures of engagement and provides additional insights into user interaction with search results. We end with conclusions and plans for future work.

2. RELATED WORK

One main approach for measuring user engagement makes use of the record of users online behaviour. We alluded above to such widely used metrics: clickthrough rates, page views, time spent on a site, how often users return to a site and number of users per specified time span. Although these metrics cannot explain why users engage with a service, they have been used for many years by the web analytics community and Internet market research companies such as comScore as proxy for online user engagement. Major web sites and online services are compared on their basis.

Two of the most widely employed engagement metrics are clickthrough rates and dwell time, in particular for services where user engagement is about clicking, for example in the context of search where presumably users click on relevant results, and/or spending time on a site, for example consuming content in the context of a news portal. In search, both have been used as indicator of relevance, and together with other metrics have been exploited to infer user satisfaction with their search results [3, 5, 10, 17].

However, how to properly interpret the relations between these metrics and retrieval quality and in the long term user engagement with the search application is not straightforward. For instance, in [19], metrics such as abandonment rate, reformulation rate, and clicks per query were shown to not reflect retrieval quality in a "significant, easily interpretable, and reliable way". We reach similar conclusions in Section 3. As a consequence, alternative approaches have been sought. One is interleaving, an evaluation method that performs paired-wise comparisons of two rankings [6]. However, this approach is only applicable to search and can compare a maximum of two rankings at a time. A second one, which can be used across applications, is based on tracking mouse movement, for example on the search result page [11, 12]. The use of mouse tracking brings additional signals on how users interact with the site, but faces the same issues regarding how to interpret this signal. Our work follows another direction, through the proposal and experimentation of the `absence` time, which is shown to bring complementary insights about user behaviour with an application.

New insights are important because user engagement possesses different characteristics depending on the web application. For instance, how users engage with a mail tool or a news portal is very different. Using several metrics to evaluate user engagement can cater for the diversity of experiences as demonstrated in [15], where a large-scale study led to the identification of patterns of user engagement. These patterns were characterised by engagement metrics related to popularity (e.g. number of users or clicks per day), activity (e.g. time spent or number of clicks per visit) and loyalty (e.g. how often users return to a site). Dwell time and clickthrough rates are activity metrics whereas our proposed `absence` time relates to loyalty. Visit activity depends on the sites, e.g. search sites tend to have a significantly shorter dwell time than sites related to entertainment (e.g. games). Loyalty per application differs as well. Media (news, magazines) and communication (e.g. messenger, mail) have many users returning to them much more regularly, than sites containing information of temporary interests (e.g. buying a car). Overall, the work described in [15] showed that activity and loyalty metrics capture different aspects of engagement.

The analysis of visit frequency and the method based on `absence` time presented in this paper are related. First, the visit frequency is simply the inverse of the `absence` time if we do not distinguish `absence` time and "return" time as we are doing in this paper. This difference is nevertheless important and makes the interpretation of the results possible. Second, our proposed method is different. Survival Analysis takes a longitudinal view of the data, and we attempt to relate the experience of individual users and their activity on the site of interest to their `absence` time. In other words, our proposed method allows us to relate "activity" metrics to the `absence` time at a more fined grained level, without sacrificing the large-scale character of the analysis that is possible with the record of users online behaviour.

3. MOTIVATION AND CONTEXT

In this section, we motivate the use of the `absence` time by looking at the limitations of other widely used measures of engagement. We do this by carrying out a two-week experiment in the context of Yahoo! Answers[3], a popular service in

[3] `http://chiebukuro.yahoo.co.jp/`

Table 1: DCG with respect to hand on a set of 653 queries chosen randomly from the clickthrough data.

	emlr	attr	util	attrc	satis
DCG@1	11.93%	-0.26%	0.74%	3.35%	2.35%
DCG@5	10.73%	1.16%	1.54%	3.93%	4.39%

Japan. This service is similar to other Q&A systems available in different countries where users are given the possibility to ask questions about any topic of their interest. Other users may respond by writing an answer. In Yahoo! Answers, these answers are recorded and can be searched by any user through a standard search interface.

3.1 Ranking Functions

We compare the user interaction data for six ranking functions deployed on Yahoo! Answers. During a period of two weeks, a subset of the users were randomly distributed towards six distinct "buckets" based on their browser cookie, one for each ranking function. This paper focus is not the ranking functions themselves but a method to compare them, so their description is intentionally succinct:

- hand: a baseline function that is the result of very carefully human hand-tuning over several years,
- emlr: a state-of-the-art machine learned ranking function trained on an extensive set of editorial labels.

The remaining functions are based on the click models described in [8]:

- attr: an attractiveness based model,
- util: a utility based model,
- attrc: an attractiveness model with extra click features,
- satis: utility & attractiveness combination model with click features.

In Table 1 we report the DCG [13] of these functions relative to the baseline hand. The performance of emlr stands out, but this not surprising as this function is trained to learn the editorial labels while the click models are learned without labels. We investigate next whether emlr higher DCG performance translates into better user engagement.

3.2 Sessions

We study the actions of approximately one million users during two weeks, with "one million" being large enough to separate between noisy and non-noisy signals. A user "action" happens every time a user interacts in some way with the Yahoo! Answers site, which happens every time he or she issues a query or clicks on a link, be it a answer, an ad or a navigation button. A "view" is defined as a page of search results (SERP) served to a user. A session is a set of user actions and views that belong together. A session is defined as the set of views and actions that are the consequences of a user decision to use Yahoo! Answers to meet one or more information needs, or to be entertained. Within a session, user might leave Yahoo! Answers to for example access other sites for a limited amount of time (multitasking), as long as the activity on Yahoo! Answers remains the main one.

To draw boundaries between sessions, we simply look into how the time between actions distributes (more sophisticated methods exist [4]). We plot in Figure 1 the histogram and the empirical cumulative distribution of the time between two consecutive actions of a user. The vast majority of actions occurs during the first 15 minutes, with only 2% happening between 15 and 30 minutes from the previous action, and 3.6% between 15 and 60 minutes.

Table 2: Number of session per bucket for different session threshold times (in minutes).

	emlr	attr	util	attrc	satis	hand
15	418299	414416	413103	415223	417805	413129
30	390001	386448	385884	387358	390196	385641
60	365604	362457	362048	363060	366117	361845

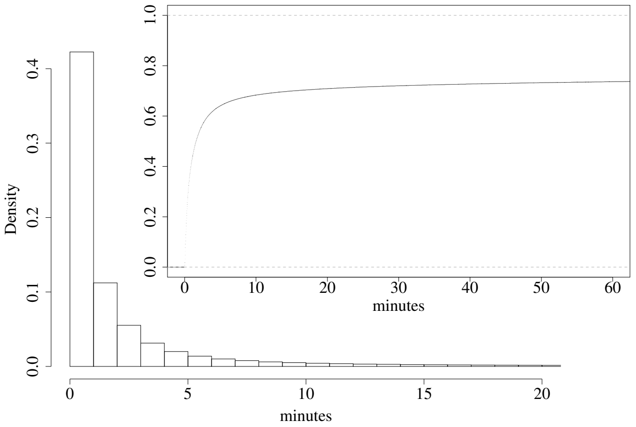

Figure 1: Histograms and cumulative distribution of time between consecutive user actions.

A heuristic commonly used to separate sessions is to define a "session threshold" beyond which two user actions are assigned to two distinct sessions. A common value is 30 minutes [14], and we see that this is a reasonable choice here as well. We nevertheless use different thresholds in this work to investigate how they impact the results. When these are not significantly different – in practical, not statistical terms, we only report one of them.

Table 2 reports the number of sessions per bucket, as a function of the thresholds taken as 15, 30 and 60 minutes. Even with a threshold of only 15 minutes between user actions, some sessions contain a very large number of views and/or clicks. We considered these as outliers and we identified the browser cookies associated with sessions containing more than 30 views. Approximately 15,000 browser cookies matched this condition and were removed from our data set. This had an almost negligible effect on the data set size. The results we report from now on are based on this cleaned data set. We report in Table 3 a summary of the number of views per session as we increase the "session threshold" and in Table 4, we do the same exercise with the number of clicks per session. Finally, Table 5 describes the distribution of the number of clicks per view.

3.3 Clickthrough Rates

It is common practice to use clickthrough rate (CTR) to compare the online performance of ranking functions. It is also commonly accepted that the CTR at position 1 is deemed particularly important and is related to the ability to place in first position a "relevant" result for a given query, i.e. one that is clicked by users. This is somewhat blurred by the

Table 3: Distribution of the number of views per session for different session thresholds (in minutes). The last column is the percentage of sessions with no more than 10 views. The maximum number of views per session is 30 by design.

	Min.	1st Qu.	Median	Mean	3rd Qu.	% ≤ 10
15	1	1	2	3.545	4	93%
30	1	1	2	3.710	4	93%
60	1	1	2	3.851	5	92%

Table 4: Distribution of the number of clicks per session for different session thresholds (in minutes).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15	0	0	2	3.467	4	216
30	0	0	2	3.646	4	224
60	0	0	2	3.807	5	224

Table 5: Distribution of the number of clicks per view. These numbers are identical for the 15, 30 and 60 minutes session thresholds. There is a maximum of 10 clicks because this is the maximum number of results presented on a result page.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0.8333	1.0310	1.5000	10

fact that users tend to click on the first result by default, even if it is not a priori very good [19]. Typically also, if the total number of clicks for two functions are similar, one function is better if clicks tend to occur at earlier positions.

In Figure 2 we plot the `CTR` of five of our ranking functions with respect to the reference `hand` (baseline) function (a carefully human hand-tuned function). According to the criteria listed above, all ranking functions but `util` are better than the reference. We also observe that `attrc` dominates `attr`. The comparison between `emlr` and both `attrc` and `attr` on the other hand is less clear cut. On one hand, `emlr` `CTR`@1 is clearly higher, but the overall `CTR` is lower. Also, `emlr` is dominated everywhere but at position 1. In view of this we can argue either that users find what they need at the first position more often with the `emlr` function and hence need not click further, or that `attrc` and `attr` offer more interesting results, and hence compel users into examining more results. If `emlr` receives less clicks than `attrc` and `attr` because users are satisfied with their first click, then we should observe more sessions with one click and less with several clicks. Table 6 reports the percentage of sessions with a given number of clicks for each ranking function[4]. We see that the data does not support this assumption. The percentage of sessions with exactly one click is quite stable across ranking functions, and, with 0.993 times the number of sessions in `hand`, `emlr` has even one of the lowest proportion of single click sessions.

This discussion illustrates the inherent ambiguity associated to interpreting clicks. `CTR` comparisons generally ignores that only part of the clicks are "good" clicks, leading to a good user experience. We note for example that `util` is a function derived from a model that attempts to identify "bad" clicks from "good" ones. Hence we expect by design a lower `CTR` and we indeed observe this in Figure 2. It is also worth mentioning the sharp contrast between the `DCG` performance reported in Table 1 and the conclusions we draw from comparing `CTR`.

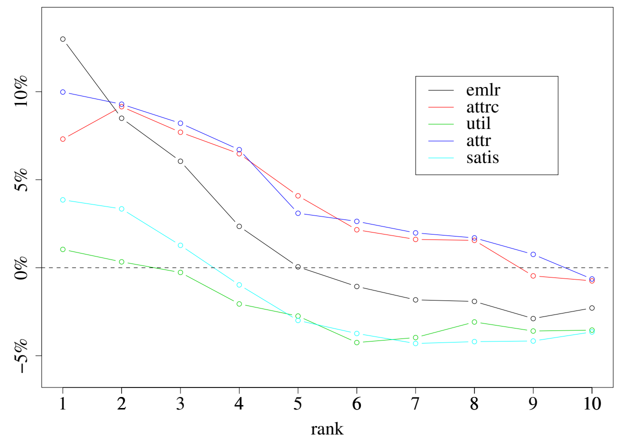

Figure 2: Clickthrough rate relative to the `hand` function.

Table 6: Percentage of single-view sessions with a given number of clicks for each ranking function, normalised by the same proportion for `hand`.

	`emlr`	`attr`	`util`	`attrc`	`satis`	`hand`
0	0.968	0.969	0.994	0.965	0.980	1.000
1	0.993	0.987	0.995	0.979	0.989	1.000
2	1.044	1.039	1.020	1.037	1.037	1.000
3	1.080	1.076	1.028	1.108	1.062	1.000
4	1.107	1.129	1.022	1.164	1.076	1.000
5	1.066	1.095	0.996	1.167	1.028	1.000
more	1.179	1.215	0.971	1.246	1.096	1.000

3.4 Query Reformulation

A higher number of query reformulations might suggest that users are not satisfied with the current search results. Therefore, if the click pattern observed with `emlr` (receives less clicks than for `attrc` and `attr` overall) reflects users finding sooner the answers they want (i.e. in the first position), then we should observe less query reformulation with that ranking function. In fact we observe the opposite. Table 7 contains the number of distinct queries issued by a user in the course of a given session for all six ranking functions. We see that `emlr` has a lower proportion of sessions with only one query than `attr` and `attrc`, suggesting that users reformulate their queries slightly more often.

3.5 Abandonment Rates

Finally, we compute the proportion of abandoned sessions, defined as sessions without a click, no matter how many views and how many reformulations the session contains. The abandonment rate is often used in evaluating search results, where a high abandonment rate suggests that poor search results were returned to users who then give up. Table 8 reports its value normalised by the abandonment rate on `hand` taken as the reference. Again, the conclusions are not in favour of `emlr` which shows a slightly higher abandon-

[4] We restricted the sessions to those with only one view and normalised by the same proportion in `hand` for privacy reasons.

Table 7: Percentage of distinct queries in a session (columns add up to 100%).

	`emlr`	`attr`	`util`	`attrc`	`satis`	`hand`
1	61.76	61.90	61.68	61.95	61.69	61.54
2	18.35	18.31	18.40	18.23	18.37	18.34
3	8.41	8.30	8.45	8.42	8.44	8.47
4	4.33	4.39	4.39	4.31	4.37	4.46
5	2.49	2.45	2.47	2.45	2.47	2.54
more	4.66	4.66	4.60	4.65	4.66	4.65

Table 8: Abandonment rate relative to `hand` for different session time thresholds. The choice of a threshold has little impact on the conclusions.

	`emlr`	`attr`	`util`	`attrc`	`satis`	`hand`
15'	0.975	0.969	1.007	0.970	0.986	1.000
30'	0.978	0.969	1.007	0.971	0.987	1.000
60'	0.980	0.972	1.008	0.973	0.989	1.000

ment rate than `attr` and `attrc`. However, Yahoo! Answers users see part of the answers on the `SERP`, which makes the interpretation of abandonment rate as a sign of failure not always accurate.

In this section we showed that in a ranking context two well known metrics of user satisfaction (`CTR` and abandonment rate) as well as `DCG`, the corner stone of web search evaluation, are not clearly and unambiguously related to an interpretation of user behaviour. In the remainder of this paper, we show how our proposed `absence` time brings additional perspectives, not accounted for by the above metrics, and that together with them, lead to a more intuitive understanding of search quality and long term user engagement.

4. SURVIVAL ANALYSIS

We use Survival Analysis[5] [1] to study `absence` time. Survival Analysis has many applications, the most important one is concerned with the death of biological organisms who have received different treatments. The latter are controlled by variables that can potentially alter the death rate. An example is throat cancer treatment where patients are administered one of several drugs and the practitioner is interested in seeing how effective the different treatments are. The survival of a particular patient might be influenced by his or her smoking habits, in which case a "confounding" or "control" variable associated with smoking is created, and treatment is administered once at the beginning, i.e. at time 0.

The analogy with our analysis of Yahoo! Answers `absence` times is unfortunate but nevertheless useful. We associate the user exposition to one of the ranking functions as a "treatment" and his or her survival time as the `absence` time. In other words, a Yahoo! Answers user "dies" each time he or she visits the site, but hopefully "resuscitates" instantly as soon as his or her visit ends.

Related to Survival Analysis is the Survival curve such as shown in Figure 3 where the percentage of users (or patients) still in the experiment (*y-axes*) is reported as a function of time. For example, we observe that 40% of the users return to Yahoo! Answers later than 5 days after their last visit. One such curve is drawn for each one of the six buckets. The differences are minimal, but a close look shows that the `absence` time of the users presented with the `emlr` ranking function is lower, implying that they return to Yahoo! Answers earlier. We also observe that `hand` is associated with a longer `absence` time, hinting at a lower performance (and hence associated user satisfaction with the search results).

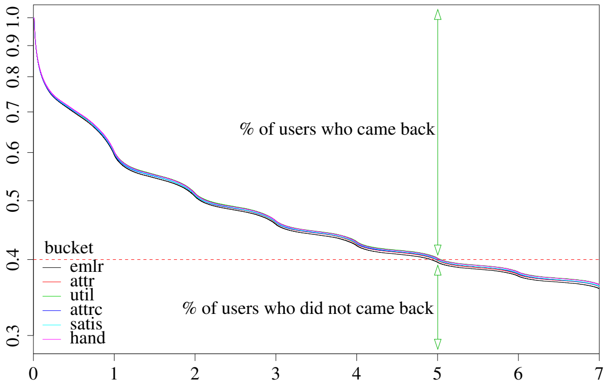

Figure 3: Proportion (log scale) of users who did not return to Yahoo! Answers after a given number of days. The `absence` time has been multiplied by a constant for confidentiality reasons.

The survival curves exhibit waves of an approximately 24 hours periodicity. This most probably reflects that user have habits regulated by whole day periods. In Section 5 we draw upon Survival Analysis to analyse the `absence` time in more detail and show that we can control for the 24 hour periodicity and quantify it. We also show that we can isolate different aspects such as the number of clicks, number of views, reformulations to obtain a better understanding of user engagement and to more accurately distinguish which ranking function performs best.

In the rest of this section we describe survival analysis in its classical usage, which has three main components, namely survival function, hazard rate, and Cox model. The analogy with the `absence` time is made in Section 5,

4.1 Survival Function and Hazard Rate

We define the survival function at time t as the percentage of users who survive past time t as $S(t)$. This is directly related to the probability $P(T \leq t)$ that a user dies at or before time t: $S(t) = 1 - P(T \leq t) = P(T > t)$. It happens that modelling the *hazard rate* rather than the survival function has several advantages. We therefore introduce the latter and describe its relation with $S(t)$.

The *hazard rate* $h(t)$ is the instant probability that a user dies at time t. Formally, this is:

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | \, T \geq t). \qquad (1)$$

It can be *very loosely* understood as the speed of death of a population of patients at a given moment t whereas the survival function $S(t)$ is the proportion of users who survived until t. The hazard rate and the survival function are closely

[5] See also the Wikipedia article at `http://en.wikipedia.org/wiki/Survival_analysis` for a short introduction.

related. Without demonstration:

$$h(t) = -\frac{S'(t)}{S(t)} \tag{2}$$

where $S'(t)$ is the derivative of $S(t)$, or, by integration

$$S(t) = \exp\{-\int_0^t h(s)ds\} \tag{3}$$

This relation shows that if the hazard rate of throat cancer patients administered with –say– `Drug` is higher than the hazard rate of patients under the `Placebo` treatment since the treatment was administered, then `Placebo` patients have a higher probability of surviving until time t. Nothing in the model prevents this situation to be reversed at a later time, depending on how both hazard rates evolve with time t. Overall, a higher hazard rate implies a lower survival rate.

4.2 Cox Model

The Cox model is a parametrisation of the previous model where the hazard rates under study are constrained to be proportional, thus allowing us to quantify their relations. Suppose that the `Drug` hazard rate is proportional to the `Placebo` hazard rate. We can then write:

$$h_{\texttt{Drug}}(t) = \alpha\, h_{\texttt{Placebo}}(t) \tag{4}$$

This does not entail that the hazard rates are constant.

The above (simple) Cox model can be extended by supposing that α is a function of any number of variables, as long as these variables are independent of the time t. We write $\alpha = \exp(\beta^T \mathbf{x})$ where $\mathbf{x} = (x_1, x_2, \ldots)$ is a vector of features and $\beta = (\beta_1, \beta_2, \ldots)$ are parameters or weights that control the influence of the corresponding variable. This is referred to as the Cox Model of Proportional Hazard.

Returning to our example of `Drug` and `Placebo`, we can set variable x_1 to be 0 if the observation comes from a user exposed to the `Placebo` cohort, and $x_1 = 1$ if it comes from the `Drug` cohort. The hazard rate of `Placebo` becomes:

$$h_{\texttt{Placebo}}(t) = h_0(t)\exp(\beta_1 x_1) = h_0(t)\exp(\beta_1 0) = h_0(t)$$

In this case, the baseline coincides with `Placebo`. For `Drug`, we have:

$$h_{\texttt{Drug}}(t) = h_0(t)\exp(\beta_1 x_1) = h_0(t)\exp(\beta_1 1) = h_0(t)\exp(\beta_1)$$

that is, if $\exp(\beta_1) > 1$ or, equivalently $\beta_1 > 0$, then the `Drug` hazard rate is higher than the baseline $h_0(t) = h_{\texttt{Placebo}}(t)$ and hence the `Drug` treatment is detrimental.

More generally, an arbitrary number of variables can be included in the Cox model. This is useful among other things to remove the effect of undesirable factors. For example, the number of smoking patients might be larger among the patients administered with the `Drug`. This higher number might be enough to explain the poor performance of the treatment. The multivariate equivalent of the model presented above can be rewritten:

$$\begin{aligned} h(t) &= h_0(t)\exp(\beta^T \mathbf{x}) \\ &= h_0(t)\prod_i \exp(\beta_i x_i) \\ &= \underbrace{h_0(t)}_{\text{baseline hazard}} \quad \overbrace{\exp(\beta_1 x_1)}^{\text{multiplicative effect of } x_1} \quad \exp(\beta_2 x_2)\ldots \end{aligned} \tag{5}$$

Table 9: Distribution of `absence` time per bucket with a session threshold of 15 minutes. (The results are normalised by the corresponding `hand absence` time for confidentiality reasons.)

	1st Qu.	Median	Mean	3rd Qu.	Max.
`emlr`	0.977	0.951	0.986	0.981	0.998
`attr`	0.978	0.957	0.988	0.990	1.003
`util`	0.996	0.989	0.997	1.002	0.997
`attrc`	0.969	0.970	0.992	0.989	0.999
`satis`	1.002	0.973	0.993	0.987	0.999

If we now set x_2 to be an indicator variable that is 1 if the patient is a smoker, then if β_2 is positive, then the effect of smoking on the hazard rate is to increase it. Another effect could be that a different estimate of β_1 can actually reverse the conclusions about which is best of `Drug` or `Placebo`. The treatment of categorical and continuous variables is similar. For example, x_2 could be redefined to represent the number of daily cigarette or the daily amount of nicotine.

Depending on whether the sign of β_i is positive or negative, a true value for x_i will contribute to, respectively, an increase or a decrease of the hazard rate and consequently will indicate whether x_i is associated with the survival of the patient to be, respectively, shorter or longer. In our case study comparing different ranking functions, a positive β_i and a large hazard rate translate into a short `absence` time and a prompter return to Yahoo! Answers, which itself can be considered as a sign of higher engagement.

5. CASE STUDY

We apply the survival analysis of the `absence` time to the six ranking functions described in Section 3. The aim of this section is to demonstrate the additional insights gained from the Cox models on specific aspects of user engagement. We use the R [18] statistical software and more specifically the Survival package [20] and the Survplot package [9] to compute the various Cox models.

We first present in Table 9 some statistics relating to the distribution of the `absence` time in the different buckets. All reported times are relative to the `hand` bucket, so for example, the median `absence` time in the `emlr` bucket is 0.951 times the `absence` median time in `hand`. We observe that `emlr` has the shortest median time while the shortest first quantile corresponds to `attrc`. This suggests that the `absence` times are spread, and that `attrc` results quality varies more than that of `emlr`. We study this further using the Cox model.

In Table 10, we show the Cox model parameter estimates associated with the 15 minutes session thresholds. The value of the β parameter for each bucket is reported together with its exponential, i.e. the coefficient that multiplies the baseline hazard $h_0(t)$. The baseline coincides with `hand`, i.e. the hand tuned ranking function. We see for example that the hazard rate of `emlr` for a 15 minutes threshold is $h_{\texttt{emlr}}(t) = \exp(0.01589)\ h_0(t) = 1.016\ h_0(t) = 1.016\ h_{\texttt{hand}}(t)$. This means that users exposed to `emlr` are returning faster to Yahoo! Answers. Moreover the *p-value* of H_0, i.e the null hypothesis $\beta = 0$, is 1.9e-8, which means that the value of β is statistically significantly different from zero. We also observe that `attr` and `attrc` are better, i.e. have a higher

Table 10: Cox model results with "bucket" as the independent variable. The baseline h_0 coincides with `hand` and is not reported

	β	$\exp(\beta)$	se(β)	z	*p-value*
`emlr`	0.01589	1.016	0.00283	5.619	1.9E-08
`attr`	0.00768	1.008	0.00284	2.706	6.8E-03
`util`	-0.00147	0.999	0.00284	-0.516	6.1E-01
`attrc`	0.00779	1.008	0.00284	2.745	6.0E-03
`satis`	0.00504	1.005	0.00283	1.780	7.5E-02

hazard rate, than the baseline `hand`, and these differences are statistically significant. On the other hand, neither `util` nor `satis` are significantly different from the baseline. The se(β) column reports the standard deviation of the corresponding β and z is the value of β after transformation into a standard normal variable under H_0.

The `DCG` values reported in Table 1 also singled out `emlr` as the best performing function, but it also predicted that `attrc` was significantly superior to `attr`, which contradicts the above findings. More striking, `satis` has the second best performance in terms of `DCG` but this clearly does not translate in users returning to Yahoo! Answers as often as for `attrc` or `attr`.

Using larger threshold values does not change the conclusions substantially (we tried 30 and 60 minutes). The main difference is that most parameters cease to be significantly different from zero. The estimated β parameters on the other hand retain their sign, and their numerical values remain surprisingly stable. For example, the factors $\exp(\beta)$ associated with `emlr` goes from 1.016 when the threshold is 15 minutes to 1.015 when it is one hour. Similarly, `attr` goes from 1.008 to 1.006 and `util` from 0.999 to 0.998. This is a hint that the choice of a specific threshold does not impact the qualitative conclusions.

In the remainder of this section, we focus on specific insights derived from our proposed survival analysis of the `absence` time.

5.1 Taking Periodicity into Account

In Section 4, we already noted that the time of the day and the day of the week influence user behaviour. This is also apparent in Figure 3. In this section, we study this quantitatively.

For example, the next session to an evening session will probably not start within 8 hours simply because most users sleep during the night. Also, behavioural patterns change during the weekend [15] and naturally influence the `absence` time. To control for this effect we introduce a categorical variable for both the hour of the day and the day of the week. The results can be found in Table 11 for a threshold of 15 minutes. Interestingly, the coefficients associated with the buckets turn out to be remarkably similar to the previous experiment where neither time nor week day were taken into account. An `anova` analysis nevertheless shows that the model fit is significantly better (*p-value* of 2.2e-16).

In the interest of space, we do not report the numerical values associated with each of the 24 hours in a day. Instead we represent them graphically in Figure 4. They are all statistically significant and we clearly observe a daily trend.

Table 11: Cox model summary with "bucket" as the independent variable. The hour of the day at the start of the visit and the weekday of the visit are included as control variables. The coefficients associated to hours are not represented to save space. The baseline h_0 coincides with `hand` on Sunday at hour 0.

	β	$\exp(\beta)$	se(β)	z	*p-value*
`emlr`	0.01561	1.016	0.00283	5.517	3.4E-08
`attr`	0.00780	1.008	0.00284	2.749	6.0E-03
`util`	-0.00191	0.998	0.00284	-0.671	5.0E-01
`attrc`	0.00795	1.008	0.00284	2.801	5.1E-03
`satis`	0.00499	1.005	0.00283	1.761	7.8E-02
hours	...	...	...	...	...
Mon	0.05349	1.055	0.00315	16.970	0.0E+00
Tue	0.08549	1.089	0.00329	25.948	0.0E+00
Wed	-0.01017	0.990	0.00302	-3.365	7.7E-04
Thu	0.04059	1.041	0.00304	13.352	0.0E+00
Fri	0.07100	1.074	0.00307	23.149	0.0E+00
Sat	0.02719	1.028	0.00308	8.823	0.0E+00

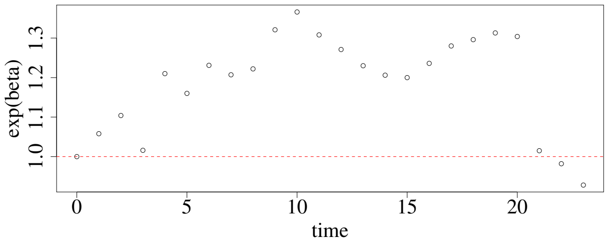

Figure 4: The influence of (GMT) time on the hazard rate as estimated by the $\exp(\beta)$ coefficients.

5.2 Relation between Activity & Engagement

We have explored in Section 3 several measures often used as indirect ways of assessing engagement: various guises of `CTR`, number of query reformulations and abandonment rates. In this section we add other measures and investigate how all these measures relate to the `absence` time. We show that Survival Analysis leads to a more nuanced interpretation of user interactions, as well as unifying them into a coherent framework.

5.2.1 Number of Clicks in a Session

Here, we investigate the common assumption that a higher number of click is a reflection of a higher user satisfaction and/or engagement with the search results. Table 12 shows the analysis of sessions with a single view. For ease of interpretation, we represented the number of clicks based on 10 binary variables $I_n, n = 0, \ldots, 10$ with I_n set to *true* if the sessions has more than n clicks. For example, a session with three clicks will have I_0, I_1 & I_2 set to true and $I_n, n > 2$ set to false. This has the advantage that each I_n represents the individual contribution of the n^{th} click to the hazard rate.

Interestingly we observe that up to 5 clicks, each new click is associated with a higher hazard rate, but the contributions from the third click are weak. The contributions of the fourth and fifth clicks are not statistically significant, suggesting that the effect on `absence` time of a session with three, four or five clicks is essentially equivalent. From the sixth click, the contribution is negative ($\beta < 0$ and hence

Table 12: The impact of the number of clicks on the hazard rate of sessions with a single view.

	β	exp(β)	se(β)	z	*p-value*
`emlr`	0.02271	1.023	0.00489	4.646	3.4E-06
`attr`	0.01158	1.012	0.00491	2.356	1.8E-02
`util`	0.01105	1.011	0.00491	2.249	2.5E-02
`attrc`	0.01468	1.015	0.00490	2.995	2.7E-03
`satis`	0.01754	1.018	0.00489	3.583	3.4E-04
clicks					
> 0	0.32701	1.387	0.00339	96.413	0.0E+00
> 1	0.10594	1.112	0.00451	23.507	0.0E+00
> 2	0.01528	1.015	0.00634	2.411	1.6E-02
> 3	0.01631	1.016	0.00877	1.860	6.3E-02
> 4	0.01177	1.012	0.01227	0.959	3.4E-01
> 5	-0.05710	0.944	0.01742	-3.277	1.0E-03
> 6	-0.06359	0.938	0.02476	-2.569	1.0E-02
> 7	-0.01657	0.984	0.03542	-0.468	6.4E-01
> 8	-0.08001	0.923	0.05020	-1.594	1.1E-01
> 9	-0.15279	0.858	0.06893	-2.216	2.7E-02

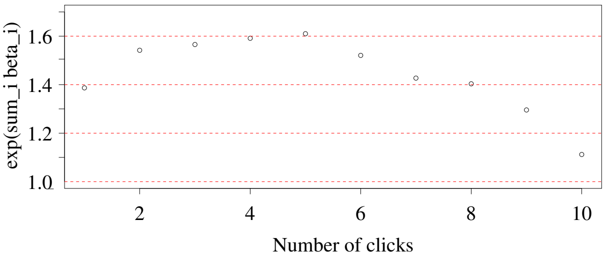

Figure 5: Influence of the number of clicks on the hazard rate, compared to no clicks (for which the value is 1.0).

exp(β) $<$ 1) and the hazard rates decreases slowly. This suggests that on average, clicks after the fifth one reflect a poorer user experience. The experience is nevertheless better than when there are no clicks at all. Indeed, in Figure 5 we present graphically the coefficient multiplying the hazard rate as a function of the number of clicks (i.e. the exponential of the cumulative sum of the β coefficients).

As for the influence of the number of views and distinct queries, the survival analysis provides additional insights. In particular, it makes clear that more clicks is not always better, which makes sense. Carrying the same analysis for sessions with more views while controlling for the number of views and distinct queries led to similar conclusions, and hence are not reported in this paper.

5.2.2 *Click Position*

We investigate whether the position at which a click occurs has an effect on the hazard rate. Table 13 shows the results obtained for sessions with one view and one click, where our baseline is a session with one view and one click at rank 1. Interestingly, the hazard rate is larger for ranks 2, 3 and 4, the maximum arising at rank 3. For lower ranks, the results are not statistically significant, but the trend is toward decreasing hazard. Only the click at rank 10 is statistically significant and clearly less valuable than a click at the first rank. We thus report in Table 14 the percentage of clicks at a given rank for a session with one view and one click. We observe more click at rank 10 than at ranks 8 and 9. A possible explanation is that users unhappy with the snippets at earlier ranks simply click on the last displayed result, for no apparent reason apart for it being the last one on the `SERP`.

Table 13: Influence on the hazard rate of the position of the click in sessions with one view and a single click.

	β	exp(β)	se(β)	z	*p-value*
`emlr`	0.04989	1.051	0.00885	5.641	1.7E-08
`attr`	0.03586	1.037	0.00891	4.026	5.7E-05
`util`	0.03006	1.031	0.00890	3.379	7.3E-04
`attrc`	0.03765	1.038	0.00891	4.227	2.4E-05
`satis`	0.04489	1.046	0.00887	5.063	4.1E-07
hours	...				
weekdays	...				
click position					
2	0.02359	1.024	0.00712	3.314	9.2E-04
3	0.05054	1.052	0.00874	5.783	7.4E-09
4	0.03822	1.039	0.01032	3.704	2.1E-04
5	0.02012	1.020	0.01184	1.699	8.9E-02
6	0.00047	1.000	0.01355	0.035	9.7E-01
7	-0.01963	0.981	0.01506	-1.304	1.9E-01
8	0.01632	1.016	0.01641	0.994	3.2E-01
9	-0.01130	0.989	0.01693	-0.667	5.0E-01
10	-0.06625	0.936	0.01621	-4.087	4.4E-05

Table 14: Percentage of clicks at a given rank for a session with one view and one click.

rank: 1	2	3	4	5	6	7	8	9	10
44.9	18.0	10.3	7.0	5.1	3.9	3.1	2.5	2.4	2.7

It appears then that a click at position 3 for example is associated with a higher engagement than a click at the first position. Clicking lower in the ranking suggests a more careful choice from the user, which might be an explanation, while clicking at the bottom of the ranking might be a sign that the overall ranking is of low quality.

5.2.3 *Time to the First Click*

Although it would have been interesting to compare dwell time on a `SERP` and `absence` time, the dwell time of the whole search session is not available because the time when a user leaves the Yahoo! Answers site is generally not known. Instead we look at the relation between the time of the first click and the `absence` time. In other words, we want to see if the time a user takes to decide which search result to click first has an effect on the `absence` time. The results are reported in Table 15. We see that the faster the decision (shorter time between the search results of a query being displayed and the first click), the higher the hazard rate. The trend seems to reverse for time longer than 300 seconds. However, five minutes seems a very long time to select one result from a list of 10, which calls for further investigations not carried out in this paper.

5.2.4 *Number of Views and Queries in a Session*

We now investigate the relation between the number of views and the number of distinct queries during a session and the hazard rate. (One query can have several views if the user clicks on the next button.) The results are reported in Ta-

Table 15: Relation between time to click and hazard rate for one view sessions, one click sessions. The baseline is the `hand` bucket with a click within 5 seconds.

	β	exp(β)	se(β)	z	*p-value*
buckets…					
$[5, 10)$	-0.20314	0.816	0.00704	-28.867	0.0E+00
$[10, 30)$	-0.39477	0.674	0.00680	-58.074	0.0E+00
$[30, 60)$	-0.65160	0.521	0.01071	-60.850	0.0E+00
$[60, 300)$	-0.67722	0.508	0.01494	-45.319	0.0E+00
> 300	-0.32264	0.724	0.02999	-10.758	0.0E+00

ble 16. We included the six ranking functions in the model and a binary variable reporting whether the number of views during the session was higher than the number of distinct queries.

The baseline is a session with a unique view (and a single query). Overall, the hazard rates associated with more views and more queries are significantly different, both in practical and statistical terms. For instance, the hazard rate of a session with two views and one query is exp(0.34982 + 0.05910) = 1.505[6] times the hazard rate of a session with a unique view, while the hazard rate of a session with two views and two queries is exp(0.34982 + 0.11716) = 1.595 times the baseline, all this for a given ranking function.

A similar computation can be carried out for all the combinations of number of views and number of distinct queries. We can conclude that having more views than distinct queries is associated on average with a longer `absence` times and hence a worse user experience. Without the `absence` time, it would have been harder to decide whether more page views is a sign of improved engagement or the opposite because we would have needed to understand why users decided to see more results, for example, whether they found the results interesting and wanted more of them, or browsed more because they did not find what they were looking for.

5.3 Click Value

Finally, we investigate whether a click "value" in terms of engagement depends on the bucket where it is observed (the deployed ranking function). We speculate that a click on the `SERP` of a better ranking function leads to a "better" result, and hence a shorter `absence` time.

In Table 17, we show the analysis of sessions with one view and a single click. We observe an effect of the ranking function. For example we see that a click originating on the `emlr` function is associated with a higher hazard rate when compared to the baseline (`hand`). This contradicts to some extent the "interleaving" hypothesis according to which all clicks have the same value [6]. A similar remark could have been done in Sections 5.2.1 and 5.2.2.

If we compare the values of the coefficients, we conclude that the `emlr` function is the best, followed by `attrc` and `satis`, respectively. On the other hand, if we study sessions with one view and two clicks (not reported here), we also observe a significant effect, but the functions' performances ranking is different; now `satis` turns out to be the best.

To decide which function to deploy on large-scale, it is best to compare overall performances as reported in Table 16 but it is nevertheless interesting to analyse the importance of clicks at this level of detail. For example, `satis` better performance when there are two clicks might reflect a better ability at showing diversified results. A possible way to verify this hypothesis would be to classify queries according to whether they would benefit from diversification and compare the performance on the two classes.

Table 16: The combined importance of the number of views and distinct queries on the hazard rate.

	β	exp(β)	se(β)	z	*p-value*
`emlr`	0.01845	1.019	0.00283	6.524	6.9E-11
`attr`	0.00859	1.009	0.00284	3.028	2.5E-03
`util`	-0.00041	1.000	0.00284	-0.146	8.8E-01
`attrc`	0.00969	1.010	0.00284	3.417	6.3E-04
`satis`	0.00618	1.006	0.00283	2.181	2.9E-02
views					
2	0.34982	1.419	0.00362	96.733	0.0E+00
3	0.41630	1.516	0.00468	88.990	0.0E+00
4	0.44909	1.567	0.00541	83.076	0.0E+00
5	0.46877	1.598	0.00597	78.503	0.0E+00
> 5	0.54433	1.723	0.00566	96.166	0.0E+00
queries					
2	0.11716	1.124	0.00292	40.114	0.0E+00
3	0.13529	1.145	0.00384	35.252	0.0E+00
4	0.15573	1.169	0.00472	32.975	0.0E+00
5	0.18158	1.199	0.00573	31.702	0.0E+00
> 5	0.29641	1.345	0.00472	62.808	0.0E+00
views $>$ queries					
TRUE	0.05910	1.061	0.00334	17.700	0.0E+00

Table 17: Influence of the originating bucket on the hazard rate of a single click session.

	β	exp(β)	se(β)	z	*p-value*
`emlr`	0.02951	1.030	0.00636	4.640	3.5E-06
`attr`	0.01931	1.019	0.00641	3.014	2.6E-03
`util`	0.01410	1.014	0.00637	2.215	2.7E-02
`attrc`	0.01966	1.020	0.00641	3.069	2.1E-03
`satis`	0.02180	1.022	0.00635	3.433	6.0E-04
hours, weekdays, views, queries, etc.					

6. DISCUSSION AND FUTURE WORK

In this paper we presented new insights in measuring and interpreting user engagement. We proposed to use between-visit or "`absence`" time to measure user engagement, motivated by the fact that it is easy to interpret and often less ambiguous than the "activity" metrics commonly used. We used a community querying and answering website hosted by Yahoo! Japan to demonstrate the benefits associated with the use of "`absence`" time.

We explored the relations between `absence` time and various "activity" metrics such as abandonment rates, click-through rates, number of views, etc. We found reasonable interpretations for what we observed and were able to quantify the relation between some activity metrics and engagement. For example, we saw that while observing a click is on average better than observing no click, a click at the first position of the ranking is a weaker indicator of success than

[6] The second term comes from the "views > queries" variable being true.

a click at the third position. While these experiments have been carried in the context of Yahoo! Answers in Japan, we believe they are representative of the results we would obtain for other ranking applications.

In addition, we compared six ranking functions deployed on Yahoo! Answers, one of them being hand tuned and the other learned either on a set of editorial labels or from clicks. In such settings, comparing the performance of these ranking functions using `DCG` is difficult because this metric is biased by construction in favour of the editorial ranking function. We showed that analysing the `absence` time and the user interaction data can lead to a more levelled comparison, making the case that some of the click learned functions [8] were in fact on par with the editorially based function.

It should be straightforward to extend this study to other web applications besides algorithmic search as long as we are confident that the `absence` time reflects user engagement. Of particular interest is the fact that the analysis can be carried out when no clicks or other record of user interaction are observed as is the case with "Direct Displays". In addition, we can also go beyond basic Survival Analysis, where only the last user experience is taken into account and instead generalise towards a complete longitudinal analysis where each interaction with a site is considered as a "treatment" of some kind that can potentially have an impact on a user engagement over time.

This research opens more questions than can be addressed in this paper regarding the relation between the user behaviour during a session and user decision to return to the site and their long term engagement, but it provides a direction on how to proceed with this challenge.

7. REFERENCES

[1] Aalen, O., Borgan, O., and Gjessing, H. *Survival and Event History Analysis: A Process Point of View.* Statistics for Biology and Health. Springer, 2008.

[2] Agichtein, E., Brill, E., and Dumais, S. Improving web search ranking by incorporating user behavior information. In *29th annual international ACM SIGIR conference on Research and development in information retrieval* (2006), pp. 19–26.

[3] Bilenko, M., and White, R. W. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *17th international conference on World Wide Web* (2008), pp. 51–60.

[4] Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., and Vigna, S. The query-flow graph: model and applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (2008), pp. 609–618.

[5] Carterette, B., and Jones, R. Evaluating search engines by modeling the relationship between relevance and clicks. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems* (2007).

[6] Chapelle, O., Joachims, T., Radlinski, F., and Yue, Y. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst. 30*, 1 (2012), 6.

[7] Chapelle, O., and Zhang, Y. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain* (2009).

[8] Dupret, G., and Piwowarski, B. A user behavior model for average precision and its generalization to graded judgments. In *Proceedings of the 33st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR* (2010), pp. 531–538.

[9] Eklund, A. *survplot: Plot survival curves with number-at-risk*, 2011. R package version 0.0.5.

[10] Hassan, A., Jones, R., and Klinkner, K. L. Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining* (2010), WSDM '10, pp. 221–230.

[11] Huang, J., White, R. W., Buscher, G., and Wang, K. Improving searcher models using mouse cursor activity. In *35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)* (2012).

[12] Huang, J., White, R. W., and Dumais, S. T. No clicks, no problem: using cursor movements to understand and improve search. In *CHI Conference on Human Factors in Computing Systems* (2011), pp. 1225–1234.

[13] Järvelin, K., and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst. 20*, 4 (2002), 422–446.

[14] Jones, R., and Klinkner, K. L. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *17th ACM conference on Information and knowledge management* (2008), pp. 699–708.

[15] Lehmann, J., Lalmas, M., Yom-Tov, E., and Dupret, G. Models of user engagement. In *20th conference on User Modeling, Adaptation, and Personalization* (2012).

[16] Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM* (2011), pp. 297–306.

[17] Liu, Y., Gao, B., Liu, T.-Y., Zhang, Y., Ma, Z., He, S., and Li, H. Browserank: letting web users vote for page importance. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2008), pp. 451–458.

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[19] Radlinski, F., Kurup, M., and Joachims, T. How does clickthrough data reflect retrieval quality? In *Proceeding of the 17th ACM conference on Information and knowledge management* (2008), pp. 43–52.

[20] Therneau, T., and original Splus->R port by Thomas Lumley. *survival: Survival analysis, including penalised likelihood.*, 2011. R package version 2.36-9.