Understanding and Measuring User Engagement and Attention in Online News Reading

Dmitry Lagun*
Google
dlagun@google.com

Mounia Lalmas
Yahoo Labs
mounia@acm.org

ABSTRACT

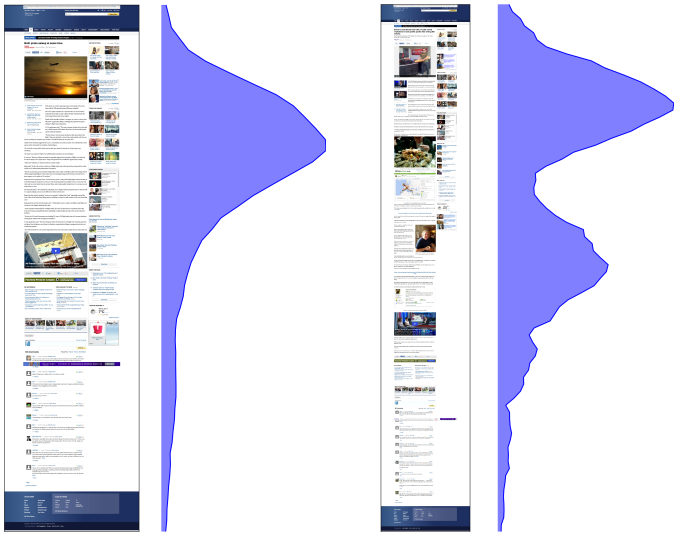
Prior work on user engagement with online news sites identified dwell time as a key engagement metric. Whereas on average, dwell time gives a reasonable estimate of user engagement with a news article, it does not capture user engagement with the news article at *sub-document* level nor it allows to measure the proportion of article read by the user.

In this paper, we analyze online news reading patterns using large-scale viewport data collected from 267,210 page views on 1,971 news articles on a major online news website. We propose four engagement metrics that, unlike dwell time, more accurately reflect how users engage with *and* attend to the news content. The four metrics capture different levels of engagement, ranging from bounce to complete, providing clear and interpretable characterizations of user engagement with online news. Furthermore, we develop a probabilistic model that combines both an article textual content and level of user engagement information in a joint model. In our experiments we show that our model, called TUNE, is able to predict future level of user engagement based on textual content alone and outperform currently available methods.

1. INTRODUCTION

User engagement has been coined as the "emotional, cognitive and behavioral connection that exists between a user and a resource" [5]. Online content providers such as news portals constantly seek to attract large shares of online attention by keeping their users engaged. A common challenge is to identify which aspects of the online interaction influence user engagement the most. We focus on one component of engagement with online content, "stickiness", concerned with users "spending time" on a content provider site.

This component of engagement is usually described as a combination of cognitive processes such as focused attention, affect and interest, traditionally measured using surveys [29]. It is also measured through large-scale analytical metrics that assess users' depth of interaction with the site. *Dwell time*, the time spent on a resource (e.g., a webpage) is one such metric, and has proven to be a meaningful and robust metric of user engagement in the context of web search [2, 6] and recommendation tasks [34].



(a) Most common (b) Less common

Figure 1: Two example pages showing different patterns of user attention: (a) shows the most common pattern when reader's attention decays monotonically towards the bottom of the article; (b) shows unusual distribution of attention indicating that content positioned closer to the end of the article attracts significant portion of user attention. The blue densities on the right side indicate the average amount of time users spent viewing a particular part of the article.

However, dwell time has limitations. For example, consider Figure 1 which shows examples of two webpages (news articles) of a major news portal, with associated distribution of time users spend at each vertical position of the article. We see two patterns. In (a) users spend most of their time towards the top of the page, whereas in (b) users spend significant amount of time further down the page, likely reading and contributing comments to the news articles. Although the dwell time for (b) is likely to be higher (the data shows this), it does not tell us much about user attention on the page, neither it allows us to differentiate between consump-

*Work done while a student at Emory University, and as part of a Yahoo Faculty Research and Engagement Program.

WSDM'16, February 22 - 25, 2016, San Francisco, CA, USA

ISBN 978-1-4503-3716-8/16/02...$15.00

DOI: http://dx.doi.org/10.1145/2835776.2835833

tion patterns with similar dwell time values. Using *viewport time* allows us to do exactly this.

In this work we build upon this observation and analyze patterns in online news reading using *viewport* data. Viewport is defined as the position of the webpage that is visible at any given time to the user. Such data allows us to measure aspects of user engagement with news articles that are *not measurable with dwell time*, such as the proportion of article read by the user or the amount of time spent at each part of the article. Furthermore, we employ viewport data to develop user engagement metrics that can measure to what extent the user interaction with a news article follows the signature of positive user engagement, i.e., users read most of the article and read/post/reply to a comment. Unlike dwell time, our metrics do not depend on the amount of textual content in the article but, instead, on the proportion of article read by users making it easier to compare articles with different amount of content.

We take one step further and develop a probabilistic model that accounts for both the extent of the engagement *and* the textual topic of the article. In contrast with previously explored text-only approaches, our model utilizes the viewport behavioral data, that enables the model to learn a joint mapping between textual topic and user engagement level. Through our experiments we demonstrate that such model is able to predict future level of user engagement with a news article significantly better than currently available methods. In addition, our model can be used, e.g. by the news editors, as an exploratory tool to investigate which textual topics correspond to higher engagement levels.

Our paper makes the following contributions:

- an analysis of large-scale viewport data in news reading, including the description of typical patterns of news reading by the online users of a large news portal;
- a family of user engagement levels that reflects user attention during news article reading; and
- a joint model of news article textual content and level of user engagement.

2. BACKGROUND AND MOTIVATION

Several works looked at the relationship between dwell time and properties of webpages. A study [27] asking participants to provide explicit feedback about the interestingness of news articles they read revealed a strong tendency to spend more time on interesting articles rather than on uninteresting ones, similarly to that reported in [10, 14]. However, only a very weak correlation between the length of articles and associated reading times was found, indicating that most articles were only read in parts, not in their entirety. Recently, [34] showed that article length correlated with dwell time, but only to some extent. They observed a large variance for articles longer than 1,000 words, suggesting that users may have a maximum "time-budget" to consume an article. The presence of videos and photos, and the article genre (e.g. politics versus food) had an effect on dwell time [34]. Works looking at webpage aesthetics showed that layout and textual features [33], and that a combination of content and dynamic features (e.g. page size or time to download all URLs) [22] had an effect on page dwell time. Finally, [18] successfully incorporated webpage readability level and search query topic into predicting dwell time.

The above and other works showed dwell time to be strong signal of user interest, an important component of user engagement [3, 4, 24]. Their aim was also to identify aspects of the webpage that make users spend time on it. However, dwell time does not capture where on the page users are focusing, namely the *user attention*. It just tells us that users spend time on it and this may be caused by the page properties such as its length, its genre, etc. Hence the suggestion of using other measurements to study user attention, an important component of user engagement [5].

Studies of user attention using eye-tracking provided numerous insights about typical content examination strategies, such as top to bottom scanning of web search results [23]. In the context of news reading, [4] showed gaze to be reliable indicator of interestingness and to correlate with self-reported engagement metrics, such as focused attention and affect. However, due to the high cost of eye-tracking studies, a considerable amount of research was devoted to finding more scalable methods of attention measurement, which would allow monitoring attention of online users at large scale. Mouse cursor tracking was proposed as a cheap alternative to eye-tracking, and the relationship between cursor and gaze was studied [28]. Mouse cursor position was shown to be aligned with gaze position, when users performed a click or a pointing action in many search contexts.

Mouse cursor movement has been studied to inform various types of user engagement with web content, in particular to infer user interests in webpages [15]. For instance, [31] found that the ratio of mouse cursor movement to time spent on a webpage was a good indicator of how interested users were in the webpage content, and explored the extent to which cursor tracking can inform about whether users are attentive to certain content when reading it, and what their experience was. Recently, [3] recorded the mouse cursor movements from users reading interesting versus non-interesting news articles, from which they generated cursor gesture patterns through unsupervised learning. They identified several significant correlations between cursor behaviour and the focused attention and affect engagement metrics, and could predict with high accuracy user interests on the news articles based on the cursor gestures. Finally, [16, 19] showed that mouse cursor movements outperform dwell time in their ability to predict relevance ratings.

However, despite promising results on using mouse cursor for measurement of user attention [28] in web search, it has been shown that the extent of coordination between gaze and mouse cursor depends on the user task [9], e.g. text highlighting, pointing or clicking. Moreover, it was found [9] that eye and cursor are poorly coordinated during cursor inactivity, hence limiting the utility of mouse cursor as an attention measurement tool in a news reading task, where minimal pointing is required. Thus, we propose to use instead *viewport time* to study user attention.

Using viewport to measure the amount of time users spend on each portion of the webpage is not new. It was used as an implicit feedback information to improve search result ranking for subsequent search queries [8], to help eliminating position bias in search result examination, detecting bad snippets and improving search result ranking [20] and in document summarization [1]. Viewport time was successfully used on mobile devices to infer user interest at the sub-document level [17]. More recently, it was used for accurate measurement of search result examination on a mobile

phone and was helpful for the evaluation of rich informational results that may lack active user interaction, such as click [21]. Our work adds to this body of works, and explores viewport time, as a coarse, but more robust instrument to measure user attention during news reading.

3. VIEWPORT DATA

Information about the amount of time online users spend reading a particular portion of a news article can be measured by tracking the position of user *viewport*. Similarly to [21], we define *viewport* as the position of the webpage portion visible to the user at any given time.

To collect viewport data for our study, we used JavaScript tracking instrumentation that allowed us to track user scroll positions, the size of the user browser viewport (width and height), as well as the position of key page elements, such as article header, body and comment blocks. This information allowed us to reconstruct the scrolling actions and calculate the amount of time users spend at given portion of the news article. Similar instrumentation was used in [1, 17, 20, 21].

We calculate the time a user spent viewing an article at a vertical position y, or simply viewport time, as:

$$ViewportTime(y) = \sum_{i=1}^{\#scrolls} t_i \cdot \mathcal{I}(y \in V_i)$$

where t_i is the time the user spent at the i-th scroll position, V_i is the viewport defined by the scroll offset and the size of the user browser window and $\mathcal{I}(\cdot)$ is an indicator function that evaluates to one if y falls inside of V_i, otherwise it evaluates to zero.

We analyze the *viewport* of a sample of data collected during one calendar month in 2013. We collected viewport data for 267,210 page views on an online news website from Yahoo!. These page views include visits to 1,971 unique news articles. We ensured that each individual news article received at least 10 page visits. Approximately 60% of the articles in our dataset have user comments.

4. NEWS READING BEHAVIOR

We take a holistic approach to analyze how users read online news. First, we analyze the overall pattern of news article examination represented by a distribution of the time users spend at each portion of the article. Then, we describe our findings on the effect of the article media components on the attention of a news reader. Finally, to analyze sequence of user actions (e.g., scroll), we apply a mixture of Markov chains model that enables us to identify common news article reading patterns from the viewport data.

4.1 Overall Pattern of User Attention

First, we analyze the overall pattern of news article examination measured with viewport time. Figure 2 shows the viewport time distribution computed from all page views in our data. It has a bi-modal shape with the first peak occurring at approximately 1000 px and the second, less pronounced peak at 5000 px. This suggests that most page views have the viewport profile that falls between cases (a) and (b) of Figure 1. This also shows that on average user spends significantly smaller amount of time at lower scroll positions – the viewport time decays towards the bottom of the page. The fact that users spend substantially less

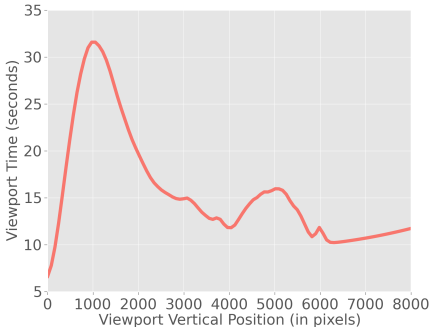


Figure 2: Distribution of viewport time averaged across all page views.

time reading seemingly equivalent amount of text (top versus bottom of the article) may also explain the weak correlation between article length and the dwell time reported in several works [22, 27, 34].

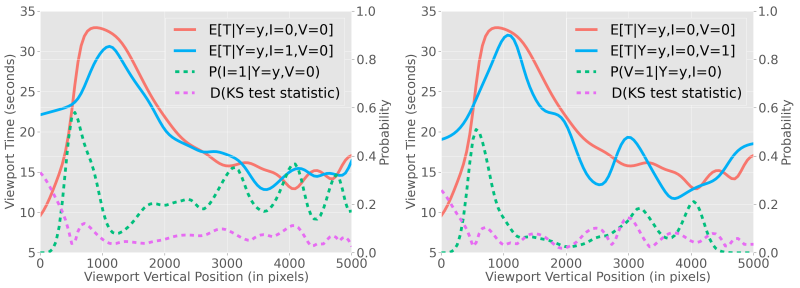
Figure 2 provides us with an overall picture on *how* users consume the news articles; however, it tells us nothing about article reading dynamics, i.e., "how often do users skip portion of the article". In Section 4.3, we attempt to answer such questions, and analyze how users navigate through the news article until they decide to leave the page.

4.2 Effect of Media Elements

We now analyze the possible impact of image and video on viewport time. We use non-parametric methods to estimate the distributions of *viewport* time conditioned on the presence of image or video element on the page. Using a Gaussian kernel density estimator we estimate the joint distribution of four variables: *viewport* time T, *viewport* vertical position Y, presence of image element I and presence of video element V. By fixing the values for some of the variables and ensuring proper normalization (by numerical integration) we can derive a conditional distribution of the other variable(s). For example, if we are interested in the effect of article images on viewport time at the same position, we can analyze the differences between the conditional distributions $P(T|Y = y, I = 0, V = 0)$ and $P(T|Y = y, I = 1, V = 0)$.

When expected viewport times are different across different conditions (e.g., video presence versus absence), we perform the Kolmogorov-Smirnov (KS) test to establish statistical significance of the observed difference $D = \sup_x |F_1(x) - F_2(x)|$, where D is the test statistic, $F_1(x)$ and $F_2(x)$ are cumulative distributions under comparison. We compute the empirical distribution of viewport time by evaluating the appropriate condition distribution at fixed number of points ($N = 200$) that reasonably covers the support of viewport time distribution ([0, 300] seconds).

Figure 3a shows the expected *viewport* time at various vertical positions of the viewport conditioned on the presence of an image. Two solid curves represent conditional means, i.e. expected values, of the viewport time given the presence of images. We also show the probability of an image displayed at each vertical position $P(I = 1|Y = y, V = 0)$. We see large difference in the expected *viewport* time at the very beginning of the page – users tend to stay almost twice longer at the first screen when the article starts with an image, compared to articles without an image on top. Due to



(a) Image (b) Video

Figure 3: Effect of media elements on viewport time.

the large sample size, the KS-test shows significant differences ($p<0.001$) at all vertical offsets. However, the peak value of the KS test statistic D is found near the article's top, which is consistent with the difference in distribution means. Figure 3b shows the same as above, but with respect to video. Overall, the presence of image or video significantly affects viewport time in initial viewport.

4.3 Markovian Analysis of News Reading

The *viewport time* analysis identified a strong positional bias in news article reading. However, it left unanswered several important questions, e.g. "How do users navigate (consume) the article during the page visit?", "Do users always read from top to bottom or are there deviations from this pattern?". To address these questions, we analyze article reading behavior through Markov chains.

Data Pre-processing: For this analysis we represent the user scrolling actions as states in a Markov chain. Specifically, using the position of the viewport and the article layout information we determine the portion of the article visible to a user during a given point in time. We divide an article into four areas of interest (AOI): *Top*, *Middle*, *Bottom* and *Comment*. The *Top*-*Bottom* AOIs are located in the beginning, middle (50%) and in the end of the article main content; the vertical dimension of each of the AOIs (except *Comment*) is equal to the height of the user viewport.[1] Then, given the page view and the associated sequence of viewport positions $\{y_i\}_{i=1}^{n}$ ordered by time, we construct a sequence of $\{v_i\}_{i=1}^{n^*}$ encoding portions of the article viewed by a user. To reduce noise in the viewport movement data we ignore viewport positions that lasted less than one second. Each viewport position is transformed into the Markov chain state by finding the closest AOI and making sure the recorded viewport position lays within one screen from the AOI margin (making sure the AOI was visible to the user). If no AOI meets this criteria, we discard this viewport position. Furthermore, as we are interested in studying the within article reading patterns, we eliminate self-transitions by collapsing adjacent states that correspond to the same portion of the article, hence the number of viewport positions n and the length of the Markov chain n^* might not always match. Finally, we augment the viewing sequences with initial *Start* and terminal *Leave* states, so our set of possible states becomes $\mathcal{O} = \{$*Top*, *Middle*, *Bottom*, *Comment*, *Start*, *Leave*$\}$.

[1]Alternative AOI definitions are possible. In this paper, we favor simplicity over other concerns.

Identifying Reading Patterns: The Markov chain models the article reading behavior using the following probabilistic distribution:

$$P(\mathbf{V_{1:n}}) = P(V_1 = v_1) \prod_{i=2}^{n} P(V_i = v_i | V_{i-1} = v_{i-1})$$

where V_i is a random variable representing the portion of the article (an AOI) viewed at the i-th time, v_i is the observed AOI viewed at the i-th time, and n is the number of times user transitioned from one portion of the article to another. The probabilities $P(V = v_i)$ and $P(V_i = v_i | V_{i-1} = v_j)$ can be directly estimated from the data, e.g., by counting the number of transitions from v_j to v_i and ensuring proper normalization ($\sum_{v \in \mathcal{O}} P(V = v | V_j) = 1$).

However, such approach will only allow us to analyze the *average* reading behavior, potentially ignoring the variability in article examination found in the data. Thus, we model the article reading behavior using a discrete mixture of Markov chains (MMC), which is able to account for different reading patterns present in the data. We choose MMC over more standard Hidden Markov Model (HMM [30]) on purpose, since our goal is to cluster entire page views, rather than state observations. The MMC models the data using the following distribution:

$$P(\mathbf{V_{1:n}}) = \sum_{k=1}^{K} \alpha_k P_k(V_1 = v_1) \prod_{i=2}^{n} P_k(V_i = v_i | V_{i-1} = v_{i-1})$$

where $P_k(V_1)$ and $P_k(V_i | V_{i-1})$ are mixture specific probability distributions, α_k is a mixture weight and k indexes the mixture component. Unlike the Markov chain model, MCC requires inference that could be done using the Expectation Maximization algorithm (EM [12]). The following equations are used to optimize the log-likelihood:

$$P^{(t)}(z_{j,k} = 1) = \frac{\alpha_k P_k^{(t)}(V_{j,1}) \prod_{i=2}^{n_j} P_k^{(t)}(V_{j,i} | V_{j,i-1})}{\sum_{k=1}^{K} \alpha_k P_k^{(t)}(V_{j,1}) \prod_{i=2}^{n_j} P_k^{(t)}(V_{j,i} | V_{j,i-1})}$$

$$\alpha_k^{(t+1)} = \frac{1}{N} \sum_{j=1}^{N} P^{(t)}(z_{j,k} = 1)$$

$$P_k^{(t+1)}(V_i = o_m | V_{i-1} = o_l) = \\ = \frac{\sum_{j=1}^{N} P^{(t)}(z_{j,k} = 1) \sum_{i=2}^{n_j} \mathbf{I}(v_i = o_m) \mathbf{I}(v_j = o_l)}{\sum_{j=1}^{N} P^{(t)}(z_{j,k} = 1) \sum_{i=2}^{n_j} \mathbf{I}(v_j = o_l)}$$

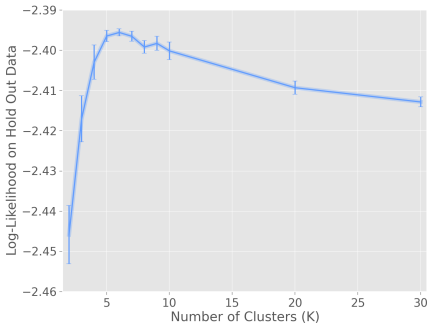


Figure 4: Log-Likelihood for different numbers of clusters (**k**). The optimal is achieved when **k**=6. The error bars shows the standard deviation of the mean (averaged over 24 random restarts).

where $z_{j,k}$ is a binary variable indicating whether j-th page view belongs to cluster k.

Choosing the Optimal Number of Mixture Components: To determine the optimal optimal number of clusters **K**, we experiment with different values of **K** and choose the one that maximizes the model generalization ability on a held-out data. That is, we split our data set in equal proportions, and use the first half of the data for training, while the second half of the data is used for assessing the model performance on an out of sample data. Due to the non-convex nature of the optimization problem, the EM algorithm is not guaranteed to find the globally optimal solution. Hence, we perform multiple restarts for the same value of **K** and report the average log-likelihood across all random restarts. In our experiments, the number of random restarts was equal to 24. Figure 4 shows the log-likelihood on the held-out data for different values of **K**. The log-likelihood is maximized at **K**=6, and the model starts to overfit with **K** larger than six, which negatively impacts its generalization ability.

Identified Reading Patterns: The reading patterns identified after running the Markov mixture model with **K=6** on the entire data set are summarized in Figure 5. We focus on the four largest clusters (ordered by their size), which account for the majority of the data.

Figure 5a shows the most probable sequences generated from the model together with their probabilities (shown on top). A considerable number of users with this pattern leave the page after viewing the article *Top* part. Although the transition diagram does not give us information on how much time users spend in the "Top" state, it clearly shows that users following this pattern are unlikely to scroll down the page. Interestingly, the second most probable sequence shows a different pattern, when users are starting reading the article at the *Top*, then transition to the *Middle*, then return back to the *Top*, followed by a page *Leave*. The third cluster (Figure 5c) is very similar to this cluster, except for the fact that users are more likely to return to the Top position from *Middle*, *Bottom* and *Comment*.

The second and fourth clusters are shown in Figures 5b and 5d. They both describe users that read the article entirely from *Top* to *Bottom* (*Top*, *Middle*, *Bottom*), and even likely to transition to *Comment* upon reaching the *Bottom* part of the article. Although these reading patterns are the most desirable goal for a news website, and are likely to be associated with high level of user engagement, to the best of our knowledge, they have not been used to study and optimize user engagement with online news. The second most probable sequence shows a shallower reading, i.e., when users leave the page upon reaching the article midpoint, which is more akin to the cascade examination of web search results [11].

Our results show that most users are likely to read article from top to bottom, and some are likely to scroll up before leaving the page. We found that the reading depth differs greatly ranging from deep engagement, when the entire article is read, to relatively short reading, when users leave after examining the first screen of the article. While some of these may seem obvious, in terms of the identified news reading patterns, to the best of our knowledge, this work is the first to validate them on a large-scale data using unsupervised sequence clustering. Clearly, by analyzig the dwell time alone we would not be able to distinguish between such patterns. More importantly, these results warrant a revision of user engagement metrics beyond *high* and *low* user engagement levels dictated by dwell time.

5. ENGAGEMENT LEVEL METRICS

Motivated by the reading patterns identified in the previous section, we propose a set of user engagement levels that more accurately reflect user engagement *and* attention with a news article, compared to current approaches using dwell time. Given the availability of viewport data, our proposed taxonomy classifies each individual page view into one of the four *level* categories: *Bounce*, *Shallow* engagement, *Deep* engagement and *Complete* engagement.

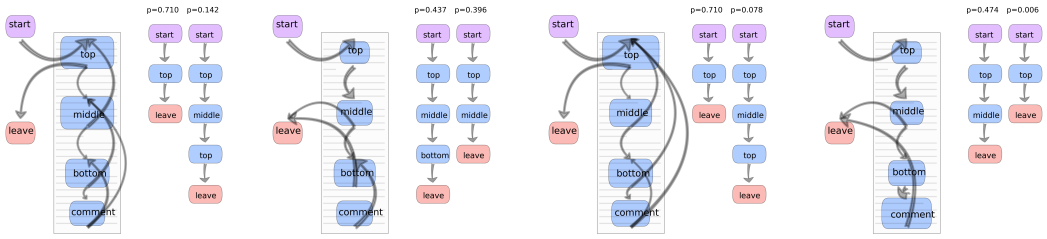
A *Bounce* indicates that users do not engage with the article and leave the page relatively quickly. We adopt 10 seconds dwell time threshold to determine a *Bounce*. Other thresholds can be used, for example accounting for genre (politics versus sport); we leave this for future work.

If the user decides to stay and read the article but reads less than 50% of it, we categorize such a page view as *Shallow* engagement, since the user has not fully consumed the content. The percentage of article read is defined as the proportion of the article body (main article text) having a *viewport* time longer than 5 seconds. Using 50% is rather arbitrary and used only to distinguish between extreme cases of shallow reading and consumption of the entire article. It is sufficient to demonstrate the insights brought with our proposed four levels of engagement.

On the other hand, if the user decides to read more than 50% of the article content, we refer to this as *Deep* engagement, since the user most likely needs to scroll down the article, indicating greater interest in the article content.

Finally, if after reading most of the article the user decides to interact (post or reply) with comments, we call such experience *Complete* engagement. The users are fully engaged with the article content to the point of interacting with its associated comments.

To understand what insights these engagement levels can bring, in particular in terms of modeling user attention, we group our data according to the proposed taxonomy, and compare each group with three sets of measures. Table 1 presents a summary of this comparison. We start with dwell time, *viewport* time broken down for the proposed engagement levels. We then report *viewport* time for article header



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4

Figure 5: Top four reading patterns identified with a mixture of Markov chain models.

Metric	*Bounce* (N=26542)	*Shallow* (N=63982)	*Deep* (N=164197)	*Complete* (N=12489)	p-value (ANOVA)
dwell	6.17 (0.02)	63.75 (0.37)	99.02 (0.22)	228.35 (1.48)	<0.001 (F=16091.6)
header time	2.99 (0.03)	15.39 (0.14)	18.48 (0.08)	17.41 (0.25)	<0.001 (F=1369.6)
body time	5.06 (0.02)	35.13 (0.21)	86.24 (0.20)	85.00 (0.70)	<0.001 (F=12229.3)
comment time	0.56 (0.01)	17.27 (0.23)	9.72 (0.07)	110.90 (0.89)	<0.001 (F=24012.4)
% header time	0.31 (0.00)	0.23 (0.00)	0.17 (0.00)	0.09 (0.00)	<0.001 (F=4784.3)
% body time	0.62 (0.00)	0.58 (0.00)	0.76 (0.00)	0.40 (0.00)	<0.001 (F=16377.2)
% comment time	0.07 (0.00)	0.20 (0.00)	0.07 (0.00)	0.51 (0.00)	<0.001 (F=20180.4)
% article read	0.12 (0.00)	0.23 (0.00)	0.83 (0.00)	0.84 (0.00)	<0.001 (F=318141.1)
# comment clicks	0.01 (0.00)	0.43 (0.01)	0.00 (0.00)	3.14 (0.03)	<0.001 (F=25351.6)

Table 1: Means and standard errors of the fine grained engagement measures for *Bounce*, *Shallow*, *Deep* and *Complete* engagement levels.

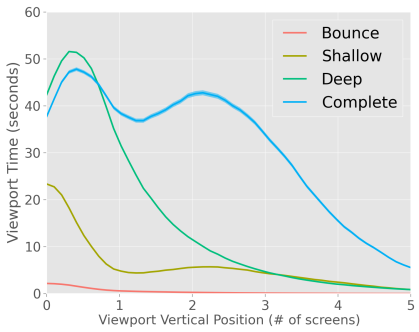


Figure 6: Mean viewport time at different viewport positions for each of the engagement levels: *Bounce*, *Shallow*, *Deep* and *Complete*. The thickness of a line corresponds to the standard error of the mean.

.

(usually a title which may include small image thumbnail), body (main body of the article), and comment block. We report the percentage of the total *viewport* time spent viewing one of these regions. The percentage of article read is defined according to the above definition. The comment clicks shows the number of clicks on the comment block.

For each measure, we report the mean and standard errors. The last column of the table shows the p-value and test statistics for the one way analysis of variance (ANOVA), allowing us to establish statistical differences in measured quantities among the four engagement levels. All of the ANOVA tests showed significant differences ($p<0.001$). Due to lack of space we omit the pos-hoc analysis of the pairwise differences between the four groups.

While we find that dwell time and *viewport* time on head, body and comment increase from *Bounce* to *Complete*, we note that the distribution of the percentage of time among these blocks changes in an interesting manner. The *viewport* time on head steadily decreases from 0.31 for *Bounce* to 0.09 for *Complete* indicating that users spend an increasing amount of time reading content deeper in the article. The percentage of article read steadily increases from *Bounce* to *Complete*, as expected. With respect to this measure, *Bounce* (12%) and *Shallow* (23%) clearly represent low levels of engagement with the article, since less than 25% of the article was read. On the other hand, *Deep* and *Complete* correspond to the situations when the majority (83%) of the article was read. The number of comment clicks is highest for the *Complete* class (3.14), followed by *Shallow* (0.43),[2] suggesting that users may engage with comments even if they do not read a large proportion of the article.

To complement this analysis, we show the average *viewport* time computed at varying vertical position in Figure 6. Each of the four curves corresponds to one of the engagement levels. The thickness of the line shows the standard error (only visible for *Complete*). We see that for the page views in the *Bounce* case, users rarely scroll down the page, whereas many users in the *Shallow* case spend approximately another 5 seconds of *viewport* time at lower scroll positions. *Deep* engagement is characterized by significant time spent on the entire article (peak at the first screen amounts to about 50 seconds) with a steady position decay of the *viewport* time towards the bottom. Interestingly, the *viewport* time profile for *Complete* engagement no longer monotonically decays with the position; instead it has a bi-modal shape. We believe this is due to a significant time users spend viewing and interacting with comments which are normally placed right after the main article content.

[2]The number of clicks for *Deep* is zero due to the definition of the *Deep* engagement class.

Symbol	Description
T	number of topics
D	number of articles
V	number of unique words
N_d	number of word tokens in article d
θ_d	the multinomial distribution of topics for article d
ϕ_z	multinomial distribution of words for topic z
ψ_z	dirichlet distribution of engagement rates for topic z
z_{di}	topic assignment for i-th word token in article d
w_{id}	i-th word token in article d
ε_d	user engagement profile for article d (multinomial)

Table 2: Notation used in TUNE model.

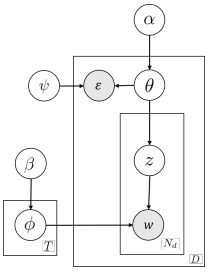


Figure 7: Topics of User Engagement (TUNE) model.

We put forward four user engagement levels that characterize how users *attend* to articles they have decided to read, as they landed on the article page. We recall that we are not attempting to capture scrolling behaviors, but to exploit these to understand which parts of a news article users engaged with, and to map these to the four proposed engagement levels. Our analysis shows that these levels are intuitive, and bring more refined insights about how user engage with articles, than using dwell time alone. The engagement levels were derived using the viewport time information, which can be computed through scalable and non-intrusive instrumentation. Next, we study how viewport time can be used to predict these levels of engagement based on the textual topics of a news article.

6. MODELING ARTICLE CONTENT AND LEVEL OF USER ENGAGEMENT

The four user engagement level metrics derived in the previous section provide clear criteria for measuring user engagement in the context of news reading at sub-document level, thus accounting for user attention. In this section we investigate whether we can model the distribution of user engagement levels with an article purely from the article textual content. If we are able to do this, this would allow us (and news content providers) to optimize online news content more effectively, compared to current methods.

The primary source of information that online users interact during news reading is the actual article text. Thus, we choose to model article text in conjunction with engagement levels observed with viewport data. To this end, we adopt a topic modeling approach, because of its intuitive structure and great ability to model textual data. Among many latent topic models, Latent Dirichlet Allocation (LDA) [7] gained significant popularity due to its relative simplicity and good empirical results. Hence, we adopt LDA for our task. Unlike the original LDA model, which uses word co-occurrence within a document to form topics, we design a model that uses both sources of information – word co-occurrence and level of user engagement – to define the topics. We call our model TUNE for Topics of User Engagement with News.

We briefly review the basic LDA model. The notation is summarized in Table 2 and the plate diagram is shown in Figure 7. LDA is a Bayesian network that generates a document (a news article in our context) using a mixture of topics [7]. In its generative process, for each document d, a multinomial distribution θ_d over the topics is sampled from a Dirichlet distribution with parameter α. Then, to generate each word a topic z_{di} is chosen from this topic distribution, and a word w_{di} is chosen by random sampling from a topic multinomial distribution $\phi_{z_{di}}$. An efficient inference allows to greatly improve the robustness of the model by "integrating out" nuance parameters θ and ϕ.

To represent a user level of engagement with the article we adopt the taxonomy introduced in Section 5. The fact that some popular articles in our dataset may have large number of page views will make the proposed model to focus on modeling of such popular articles. To avoid this, we normalize for article popularity by computing an article user engagement profile – a multinomial distribution over the four engagement levels calculated over all page views for the article. More specifically, using the engagement levels introduced earlier, we classify each individual page view into the engagement level and calculate the article user engagement profile as follows:

$$\varepsilon = \left(\frac{\#Bounce}{N}, \frac{\#Shallow}{N}, \frac{\#Deep}{N}, \frac{\#Complete}{N} \right)$$

where N is total number of page views for the article.

There are several ways user engagement information can be incorporated into the LDA model. Since its introduction, LDA model has been extended to include document author information [13], the time the document was generated [32] and many other types of information (e.g. [25]). Generalization of these was introduced in [25] allowing conditioning topics on arbitrary features of the document. As our goal is to predict future user engagement level based on an article text, we want to be able to derive tractable inference for ε. The Topic over Time (TOT) model in [32] meets this criterion, thus, we adopt its structure to incorporate the user engagement levels into the LDA model. The graphical structure of our model is summarized in Figure 7 and its generative process is described as follows:

1. Draw T multinomial distributions ϕ_z from Dirichlet prior β, one for each topic z
2. For each article d, draw a multinomial distribution θ_d from Dirichlet prior α. Then for each word w_{di} in article d:
   (a) Draw a topic z_{di} from multinomial θ_d
   (b) Draw a word w_{di} from multinomial $\phi_{z_{di}}$
   (c) Draw a user engagement level ε_{di} from Dirichlet $\psi_{z_{di}}$

This generative process follows an alternative view of TOT model, which is more suitable for performing Gibbs inference. In this view, a user engagement profile is generated for each word, while in practice it is only measured once per entire article. In our experiments, we assign the same profile of user engagement to all words within the same article. As shown in the process, the posterior distribution over topics depends on both the text and the level of user engagement. More precisely, the conditional distributions in the TUNE model are defined as follows:

$$\theta_d|\alpha \sim Dirichlet(\alpha)$$
$$\phi_z|\beta \sim Dirichlet(\alpha)$$
$$z_{di}|\theta_d \sim Multinomial(\theta_d)$$
$$w_{di}|z_{di} \sim Multinomial(\phi_{z_{di}})$$
$$\varepsilon_{di}|\psi_{z_{di}} \sim Multinomial(\psi_{z_{di}})$$

An exact inference is intractable in the TUNE model, as in other LDA like models. We therefore employ collapsed Gibbs sampling to perform approximate inference. Using conjugate priors (Dirichlet) for the multinomial distributions allows us to "integrate out" nuance parameters θ and ϕ and not sample them during the inference. For simplicity and speed we estimate the parameters of the Dirichlet distribution $\psi_{z_{di}}$ using fixed point iterations proposed in [26]. Instead of estimating hyper-parameters we employ commonly used heuristics and set $\alpha = 50/T$ and $\beta = 0.1$.

In the Gibbs sampling procedure above, we need to compute the conditional distribution:

$$P(z_{di}|\mathbf{w}, \varepsilon, \mathbf{z}_{-di}, \alpha, \beta, \Psi) \propto (m_{dz_{di}} + \alpha_{z_{di}} - 1)$$
$$\times \frac{n_{z_{di}w_{di}} + \beta - 1}{\sum_{v=1}^{V} n_{z_{di}} + \beta - 1} \frac{\Gamma(\sum_{j=1}^{4} \psi_{z_{di},j})}{\prod_{j=1}^{4} \Gamma(\psi_{z_{di},j})} \prod_{j=1}^{4} \varepsilon_j^{\psi_{z_{di},j}-1}$$

where n_{zv} is the number of tokens of word v that are assigned to topic z, m_{dz} represents the number of tokens in article d that were assigned to topic z, ε_j refers to the j-th component of the article engagement profile. The last factor in this conditional probability distribution corresponds to the user engagement model captured with the Dirichlet distribution over the engagement levels. We omit the detailed derivations for the rest of the formula and the Gibbs sampling procedure, since they are identical to the TOT model and can be found in the original paper [32].

Using this model and the Gibbs sampling procedure we can predict the future level of user engagement with an article. Given the trained model (estimated collection of probabilities) and a test article, we infer simultaneously the topic assignment for the words in this article together with the level of user engagement by running Gibbs sampling until the topic assignment and ε converge. Note that at the test stage we sample topic assignments and engagement levels for words in the test article, and that the same quantities for words in the training data remain unchanged, allowing the model to condition its prediction on the training data.

Our model goes beyond an ordinary LDA and extends it by incorporating viewport based user engagement information into the model. Such approach allows us to learn a joint mapping between text content and user engagement, which can be used to either predict future level of user engagement or to better understand which topics attract user interest and lead to higher level of engagement. To the best of our knowledge, this is the first attempt to combine these types of data into a joint model. While the model extension presented in this paper is a relatively simple one, it provides a clean way to combined both sources of information. We nonetheless plan to investigate more sophisticated ways to combine textual and viewport data in future work.

7. PREDICTING ENGAGEMENT LEVELS

To investigate whether our approach (TUNE) is able to model user level of user engagement from article text we experiment with the prediction of the engagement levels (*%Bounce*, *%Shallow*, *%Complete* and *%Deep*) for held-out articles. Our experiments are based on the set of articles described in Section 3. Since the engagement rates are numeric, we use the Pearson correlation coefficient to evaluate the quality of the prediction. To ensure that our estimates of the model performance are not over-optimistic, we perform a ten-fold cross validation over our data. Thus, we repeatedly train ten models and use them to obtain the predictions for each of the held-out set among the ten folds. We report the Pearson correlation computed from the ten test folds combined.

We compare our approach to several baseline variants. All models in our experiments perform a linear regression from a set of features to one of the engagement levels, and differ only in the exact features used to train linear regression. We also report the results for predicting dwell time.

The *NumWords* feature encodes the number of words in the article. The features in the *Media* group (9 in total) include dimensions and vertical position of the largest image/video element on the page, as these were shown to have some effect in Section 4.2 in viewport bias and in the work of [34] for predicting dwell time. It also includes the number of media (image or video) elements on the page. Other features such as article genre, text sentimentality are left for future work.

The features in the *LDA* group represent the article topic probabilities computed using the standard LDA model that does not account for the level of engagement, where T denotes the number of topics. Finally, the features in the TUNE group represent the predicted engagement levels (four values). Note that these features are obtained by training/testing ten TUNE models in order to not leak the label data between the cross validation folds. Table 3 reports the performance of the regression models with the various feature sets.

The baseline model using the number of words in the article is able to predict *%Shallow* with 0.494 and *%Deep* with 0.37 correlation, but fails to predict *%Bounce* or *%Complete*. This suggests that the length of an article text has no effect on user decision to either bounce (to hardly read the article) nor to fully engage with the article, such as reading and/or posting comments. The number of words provides some indications on how long users will spend on the article (as shown with *Dwell*, which is not new) but when *they decide* to actually read the article.

Adding the article media features improves the performance for *Dwell* and all engagement levels, except *%Bounce*. This confirms, as already reported in [34], that media elements entice users to spend time on an article (e.g. watching the video in addition to reading the text). However, they have little effect on users deciding to read an article, suggesting that it is not the presence/absence of media elements that makes users bounce.

The features extracted with the original LDA model substantially improve the prediction quality and achieve almost two-fold improvement in *%Bounce* rate. This shows that the topic of the article has a clear effect in enticing the users to start reading the article. It should be noted that users who landed on the article, thus leading to a page view, in principle intended to read the article. However, when landing on the article page, they decided to not read it. There can be many reasons for this, for instance the way the story was

Feature Set	Dwell	%Bounce	%Shallow	%Deep	%Complete
NumWords	0.420	0.063	0.494	0.370	0.017
NumWords + Media	0.465	0.071	0.571	0.410	0.185
NumWords + Media + LDA (T=5)	0.528	0.119	0.597	0.466	0.328
NumWords + Media + LDA (T=10)	0.528	0.110	0.606	0.497	0.379
NumWords + Media + LDA (T=20)	0.543	0.15	0.626	0.531	0.402
NumWords + Media + LDA (T=50)	0.547	0.143	0.629	0.538	0.405
NumWords + Media + TUNE (T=5)	0.476	0.079	0.648	0.544	0.282
NumWords + Media + TUNE (T=10)	0.526	0.311	0.713	0.660	0.400
NumWords + Media + TUNE (T=20)	0.537	0.349	0.724	0.682	0.409
NumWords + Media + TUNE (T=50)	0.545	0.333	**0.742**	**0.697**	0.428
NumWords + Media + LDA + TUNE (T=50)	**0.572**	0.334	0.730	0.696	**0.442**
Dwell	1.000	**0.392**	0.203	0.128	0.351

Table 3: Comparison of regression models with different feature sets. The table reports Pearson correlation coefficient between predicted engagement rates and the actual levels.

presented, i.e. its scope and the topics covered. It is for the editors to understand what causes this behavior.

Not surprisingly, the TUNE features enable even higher performance lift. This clearly suggests that it is possible to predict the level of engagement with respect to the topics of the article (how much of the article will be consumed and also as a consequence the time spent on it). Note that as T increases, the predictions get better, with the exception of *%Bounce* for which TUNE with T=20 performs slightly better than TUNE with T=50, although the difference is not statistically significant (p-value>0.1, two tailed t-test). On *%Bounce*, *%Shallow* and *%Deep* TUNE (T=50) performs significantly better (p-value<0.01, two tailed t-test) than a baseline approach denoted as NumWords + Media + LDA(T=50); the performance of TUNE with T=50 is not significantly different from the baseline (p-values>0.1) for *Dwell* and *Complete* rate predictions.

Finally, the model that combines all the feature groups performs best on Dwell time and *%Complete* rate. It achieves 0.334 for *%Bounce*, 0.73 for *%Shallow*, 0.696 for *%Deep* and 0.442 *%Complete* engagement levels prediction. The TUNE model with T=50 performs reasonably well showing highest correlation with ground truth values for *%Shallow* and *%Deep* engagement levels.

Our results would have been incomplete, if we would not try to use dwell time to predict the engagement classes. Some may argue that our four engagement classes can be distinguished by setting appropriate thresholds for dwell time. However, this is not the case. The last row of Table 3 shows to what extent average dwell time is able to predict specific class of user engagement. We find that dwell time provides a strong signal for identifying *%Bounce*, which is reasonable, given our definition of *Bounce*.[3] However, dwell time does not match the performance of other approaches for the rest of the classes. This is likely due to the inability of dwell time to capture how users attend to the articles they are engaging with, e.g., amount of content read, etc. It is possible that combining dwell time with other content or behavioral features could lead to higher predictive performance. However, doing so would undermine the purpose of our work – *using content only features to predict level of user engagement.*

These results show that combining user engagement *and* article text information into a joint model clearly benefits the quality of future user engagement prediction. Our proposed *joint model* is able to provide accurate predictions about future user engagement levels for a news article, which can be used by news editors to fine tune the article content and optimize the user experience with it more effectively.

[3]Note that Bounce rate and average dwell time are different quantities; hence, accuracy of prediction using average dwell time is far from perfect.

8. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel way to measure user engagement in the context of online news reading. We focus on *viewport* time, which is the time a user spends viewing an article at a given position and which can be instrumented at large scale, to derive measurement of engagement that accounts for how users *attend* to the content they are reading. We analyzed the viewport data of a sample of 267,210 page views on a total of 1,971 news articles from a major online news website from Yahoo.

First, the viewport time analysis, which to our best knowledge is the first large scale analysis of the kind, identified a strong positional bias in news article reading, which by itself is not new. However, although users often remain in the upper part of an article, some users do find the article interesting enough to spend significant amount of time at the lower part of the article, and even to interact with the comments. Thus, some articles entice users to deeply engage with their content.

Second, we carry out an analysis of viewport data, employing a mixture of Markov chains, to identify how users read articles, and accounting for the depth of the article consumption. We identified four main patterns, clearly showing that most users read article from top to bottom, and some users scroll up before leaving the page. We found that the reading depth differs greatly ranging from deep engagement, when the entire article is read, to relatively short reading, when users leave after examining the first screen of the article. These patterns inspired us to propose a taxonomy for user engagement, with four levels, namely *Bounce*, *Shallow* engagement, *Deep* engagement and *Complete* engagement. Our analysis shows that these levels are intuitive, and bring more refined insights about how user engage with articles, than using dwell time alone.

Third, we described a probabilistic approach that allowed us to incorporate these four levels of engagement in the modeling of the topics covered by a news article. We do this by employing LDA, and develop what we call Topics of User Engagement with News (TUNE). We show that the level of user engagement could be successfully incorporated in the topic modeling using the LDA approach. We carried out experiments to investigate how article features such as their

length and the presence of media elements affect the prediction of our proposed metrics of engagement levels. We found that TUNE, compared to two baselines (using article length only and using the original LDA) leads to improved performance not only with respect to the proposed metrics of engagement levels, but also dwell time.

We also obtained several interesting insights. The presence of media elements, although enticing users to dwell longer on the page, has little effect on users deciding whether to actually read the article (i.e. the effect on *Bounce* was small). The same was observed with respect to article length. This indicates that what matters most is the actual content of the article *and* the way the story it is covering is presented. Although, in itself, this statement may be obvious, we could demonstrate this effect by incorporating the four engagement levels within the LDA topic modeling approach.

In the future, we want to experiment further with TUNE, and look at more features of articles, including with respect to aspects related to their genre, layout, and the presence of advertisements on the article page. This will bring more extensive insights into user engagement with online content, that are usually only possible through small-scale experimentation using surveys and eye-tracking. Furthermore, in this paper, the engagement patterns and the resulting modeling were with respect to article pages, and not users. It is very likely that users engage in different ways with online content, and whether this has an effect on our proposed four engagement levels should be investigated. Finally, this work was applied in the context of news reading on desktop. The next step would be to deploy the same instrumentation and corresponding measurement methodology in the context of tablet and smartphone.

9. REFERENCES

[1] M. Ageev, D. Lagun, and E. Agichtein. Improving search result summaries by using searcher behavior data. In *Proc. of SIGIR*, pages 13–22. ACM, 2013.

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

[3] I. Arapakis, B. Cambazoglu, and M. Lalmas. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proc. of CIKM*. ACM, 2014.

[4] I. Arapakis, M. Lalmas, B. B. Cambazoglu, M.-C. Marcos, and J. M. Jose. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *JAIST*, 2014.

[5] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski. Towards a science of user engagement (position paper). In *WSDM Workshop on UMWA, year=2011*.

[6] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *Proc. of WWW*, pages 51–60, 2008.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[8] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: a comparison to eye tracking. In *Proc. of SIGIR*, pages 67–74, 2009.

[9] M. C. Chen, J. R. Anderson, and M. H. Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI extended abstracts*, pages 281–282. ACM, 2001.

[10] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proc. of IUI*. ACM, 2001.

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

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, pages 1–38, 1977.

[13] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *In Proc. of PNAS*, 101:5220–5227, 2004.

[14] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *TOIS*, 23(2):147–168, 2005.

[15] J. Goecks and J. Shavlik. Learning users' interests by unobtrusively observing their normal behavior. In *Proc. of IUI*, pages 129–132.

[16] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proc. of WWW*, pages 569–578. ACM, 2012.

[17] J. Huang and A. Diriye. Web user interaction mining from touch-enabled mobile devices, 2012.

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

[19] D. Lagun, M. Ageev, Q. Guo, and E. Agichtein. Discovering common motifs in cursor movement data for improving web search. In *Proc. of WSDM*, pages 183–192. ACM, 2014.

[20] D. Lagun and E. Agichtein. Viewser: enabling large-scale remote user studies of web search examination and interaction. In *Proc. of SIGIR*, 2011.

[21] D. Lagun, C. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proc. of SIGIR*. ACM, 2014.

[22] C. Liu, R. W. White, and S. Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proc. of SIGIR*, pages 379–386, 2010.

[23] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan. Eye tracking and online search: Lessons learned and challenges ahead. *JASIST*, 59(7):1041–1052, 2008.

[24] L. McCay-Peet, M. Lalmas, and V. Navalpakkam. On saliency, affect and focused attention. In *Proc. of CHI*, pages 541–550, 2012.

[25] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

[26] T. Minka. Estimating a dirichlet distribution, 2000.

[27] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proc. of SIGIR*, pages 272–281, 1994.

[28] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proc. of WWW*, pages 953–964, 2013.

[29] H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *JASIST*, 61(1):50–69, January 2010.

[30] L. Rabiner and B.-H. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[31] B. Shapira, M. Taieb-Maimon, and A. Moskowitz. Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. In *Proc. SAC*, pages 1118–1119, 2006.

[32] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proce. of KDD*, pages 424–433. ACM, 2006.

[33] O. Wu, Y. Chen, B. Li, and W. Hu. Evaluating the visual quality of web pages using a computational aesthetic approach. In *Proc. of WSDM*, pages 337–346, 2011.

[34] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan. Beyond clicks: dwell time for personalization. In *Proc. of RecSys*, pages 113–120, 2014.