Identifying Modes of User Engagement with Online News and Their Relationship to Information Gain in Text

Nir Grinberg
Network Science Institute, Northeastern University
Institute for Quantitative Social Science, Harvard University
n.grinberg@northeastern.edu

ABSTRACT

Prior work established the benefits of server-recorded user engagement measures (e.g. clickthrough rates) for improving the results of search engines and recommendation systems. Client-side measures of post-click behavior received relatively little attention despite the fact that publishers have now the ability to measure how millions of people interact with their content at a fine resolution using client-side logging.

In this study, we examine patterns of user engagement in a large, client-side log dataset of over 7.7 million page views (including both mobile and non-mobile devices) of 66,821 news articles from seven popular news publishers. For each page view we use three summary statistics: dwell time, the furthest position the user reached on the page, and the amount of interaction with the page through any form of input (touch, mouse move, etc.). We show that simple transformations on these summary statistics reveal six prototypical modes of reading that range from scanning to extensive reading and persist across sites. Furthermore, we develop a novel measure of information gain in text to capture the development of ideas within the body of articles and investigate how information gain relates to the engagement with articles. Finally, we show that our new measure of information gain is particularly useful for predicting reading of news articles before publication, and that the measure captures unique information not available otherwise.

CCS CONCEPTS

• **Information systems → Information extraction**; **Multimedia and multimodal retrieval**; **Content analysis and feature selection**; *Document structure*; *Data encoding and canonicalization*;

KEYWORDS

User engagement, Online news, Information gain, Reading

ACM Reference Format:
Nir Grinberg. 2018. Identifying Modes of User Engagement with Online News and Their Relationship to Information Gain in Text. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186180

1 INTRODUCTION

Over the past two decades, our reading habits have turned from physical media (books, magazines and newspapers) to their digital counterparts (e-readers, websites, and apps). Pew research estimated last year that 38% of Americans *often* got their news online, almost twice the number of people who read it in print [31]. Where previously news publishers had to rely on gross sales numbers or small-scale surveys that took weeks or months to collect, they now have near real time information about individual readers engaging with news content on their website.

The shift to digital media creates new opportunities for publishers to better understand user engagement *within* an article page using client-side logging. Thus far, the dominant measure of post-click behavior has been dwell time, an estimate of the total time a user spent on the page. Dwell time is a useful measure for improving the results of search engines and recommendation systems [21, 42]. However, dwell time only provides partial information about the activity of a user on a page. Other client-side interactions such as cursor movement, scrolling, and highlighting provide additional information about the article relevance and the distribution of attention on a page [15, 23]. Although beneficial, these additional client-side measures incur substantial costs in terms of model complexity, network communication, and storage, thus making these measures difficult for news outlets to use in practice, especially at large scale.

Furthermore, there is a disconnect between measures of user engagement and *the structure* of news articles. Reading is a process that involves a sequence of decisions about how to direct one's attention to the text. Yet, existing engagement measures do not take into account how the development of ideas within the body of text may shape user engagement with it. Previous work linked measures of user engagement to visual and dynamic properties of a page such as layout, saliency of page elements, and presence of images or videos [14, 23, 24, 41, 42]. The relatively little work that examined user engagement and textual content concentrated on the general topic, sentiment and readability of the text [2, 21, 24]. Here, we are interested in exploring how the development of ideas within the text relates to user engagement.

Practical and informative measures of post-click user engagement can improve recommendations of news content and enable more informed editorial decisions. Distinguishing between different modes of enagement with an article, such as scan, skim, or in-depth reading, can enable recommendation systems to better match articles with potential readers based on their engagement profile. In addition, accurate predictions about engagement with an article prior to publication can guide editorial decisions, help journalists write higher quality content, and set expectations for

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186180

the reception of articles by their audience. Post-hoc examination of the extent to which readers engaged with articles can enable editors to better understand their audience interests, and inform both the coverage and writing style of future articles. The challenge we tackle in this work is to derive engagement measures that provide meaningful descriptions of post-click behavior in articles with as little storage and recording costs as possible. Moreover, we seek to better understand the relationship between articles' text and the engagement of readers with it.

In this work, we use compact summaries of user interaction with a news article to identify robust and interpretable modes of engagement. We propose a set of simple transformations that capture user engagement with news in relative terms: relative to the article being viewed and context of viewing it. We show that by using these measures we can identify prototypical modes of engagement that persist across different sites and browsing devices, and align with previous findings about reading obtained using much more granular data. In addition, we develop a novel measure of information gain within the text of news articles. We demonstrate that our measure of information gain is the single best predictor of reading engagement with news articles, and that the new measure captures unique information not available otherwise.
Therefore, our contributions are:

- A compact metric of user interaction that carries valuable information about post-click user engagement with a news article.
- A novel measure of information gain within the text of news articles.
- Empirical evidence linking information gain in text to user engagement as observed in a diversity of news sites, at large scale, and outside lab settings.

2 BACKGROUND AND RELATED WORK

In this section, we describe three lines of related work: the use of post-click engagement measures in systems' design, the relationship between content properties and user engagement, and the study of reading in Web settings.

Post-click user engagement has largely been studied in the context of information retrieval and recommendation systems. Early works in this area established dwell time, scrolling, and other post-click behaviors as useful proxies for users' subjective rating of web content [8, 32]. Based on these observations, numerous studies examined the utility of implicit feedback measures, and dwell time in particular, for improving ranking of search results [1, 12] and content recommendations [42]. Perhaps central to the success of dwell time is the simplicity of estimating it from server logs, and the fact that it is both a good proxy for user dissatisfaction (previously operationalized as time spent of 30 seconds or less) and a reasonable approximation for satisfaction [21]. Guo and Agichtein introduced a model that improves upon dwell time by using cursor movement and scrolling behaviors, demonstrating gains in both relevance judgments and search results ranking [15]. Perhaps underlying many of these metrics are the fundamental concepts from classical physics that can describe a user's navigation through a page using the position, speed, and acceleration over time. In practice, however, it is costly to store complete user sessions for the entire user population, even when sampled at a moderate rate over time. Therefore, in this work we quantify user engagement using summary statistics calculated over the entire user session and transformed to represent user behavior in more "natural" coordinates.

Several studies investigated the relationship between the content of web pages and user engagement with them. At the most basic level, the time spent on a page was shown to depend on the time it takes to load and render a page [28]. Then, the visual complexity and aesthetics of web pages demand different amounts of cognitive processing of visual information [40, 41]. Another important aspect is the visual saliency of page elements, which affects the distribution of attention on a page [6, 23]. In the context of online news, Yi et al. showed that total dwell time on a page is associated with longer articles, having more images and videos [42]. The work of Arapakis et al. established that emotional dimensions of news articles, such as sentiment and polarity, help predict user engagement, and that these emotional aspects vary considerably across different genres [2]. Kim et al. found similar dependency of dwell time on article length, and further demonstrated that dwell time also depends on the readability of the text and its general topic [21]. More recently, Lagun and Lalmas proposed a joint model for capturing the relationship between latent topics in text and user engagement, demonstrating superior predictive ability for latent topics informed by past user engagement [24]. Our work directly builds on previous studies by incorporating many of the content features associated with user engagement. Moreover, we extend this line of work by developing content features that focus on the development of ideas *within an article*, studying how the density of information within the text is associated with the sustained attention of readers.

A third line of work concerns assessing reading on the Web. Evaluating people's focus of attention and the extent to which they read content is a difficult task even in lab settings, let alone in Web settings. In fact, humans only master the metacognitive skill of evaluating their own reading when they reach a certain intermediate level of reading [13]. Traditionally, reading has been assessed in lab settings through comprehension tests, eye-tracking, and brain-imaging techniques (see [20, 37, 38] for representative examples). While reading rates generally vary from one person to another, the literature generally describes normal reading rates in the range of 200-300 words per minutes for native speakers of about average level of education and intelligence [30, 35]. Often, particularly in Web settings, people only skim articles by moving rapidly through the text, reading in-depth certain parts and skipping others [17, 29, 30]. According to Liao, people skim at a speed of up to three or four times faster than their normal reading speeds [27]. In terms of detecting reading on web pages, the works of Biedert et al. and Campbell and Maglio used eye-tracking in well-controlled lab settings to identify patterns of reading [3, 7]. While the dependency on eye-tracking can be alleviated to a degree through cursor-gaze correlations [18, 33], this approximation does not apply to reading on mobile devices, which occupy smaller portion of the visual field and does not involve positioning a cursor on the screen. Informed by these findings, the current work seeks to identify modes of engagement with news articles that *correspond* to reading, but not necessarily mean reading in a strict sense. Assessing reading with stronger guarantees is currently only possible in well-controlled environments, which trade off accuracy for generalizability and representativeness.

Closest to the current work is the work of Lagun and Lalmas [24], which similarly examined patterns of user engagement with news, and studied the relationship between the text and user engagement. The current work, however, seeks to learn post-click engagement patterns using much more compact representations (i.e. summaries rather than the full time series), utilizing both mobile and non-mobile engagement information, and focusing specifically on the relationship between the development of ideas within the text and user engagement with it. As the rest of the paper demonstrates, summaries of user interaction are able to retain substantial amounts of information about the underlying user behavior, and our new measure of information gain in text outperforms all previous measures in predicting, prior to publication, the fraction of page visits that will involve reading.

3 ENGAGEMENT DATA

We analyzed a large dataset that consists of 7.7 million page views of articles from seven major news publications. For each page view we have information about the article being viewed and three key measures that summarize the visitor's activity on the page as we describe later in the section. The raw dataset was obtained from Chartbeat[1], one of the leading web analytics companies for online publishers. The raw dataset consists of a random sample of over 8.7 million page views, which we further filtered to contain only news article pages. We identified article pages by crawling each page in the dataset and extracting the content using regular expressions, customized for each site. The page view data was collected both on mobile and non-mobile devices during the last two weeks of October 2014 and pages were crawled in February 2015. The sites chosen for the analysis cover a diverse set of topics (daily news, finance, sports, technology and science); target audiences (e.g. women[2], young adults), and include both short and long-form articles. As shown by Lehmann et al. [26], engagement metrics vary by site and thus it is important that we examine a diverse set of sites. In order to protect publishers' identity we only refer to sites by their differentiating characteristics (e.g. financial, technology, or magazine site).

Table 1 describes the resulting dataset that is analyzed in this work. The dataset consists of 66,821 news articles, viewed a total of 7.7 million times by over 4 million unique visitors. We consider the problem of associating page views by the same individual on multiple devices outside the scope of the current work.

Each data point in the sample is a summary of a page view as collected by Chartbeat's client-side logging system. The summary consists of three measures: dwell time, maximal depth, and active engagement. Similar to other work, dwell time is the total amount of time the page was visible on the user's screen. Maximal depth is the furthest position reached on the page during a page visit, measured in pixels (vertical), and active engagement is the amount of user interaction with a page in any form (mouse move, scroll, swipe, key press, etc.). Since user interaction tends to occur in short bursts, Chartbeat measures active engagement (or engagement for short) in units of seconds and smooths the signal over time using a sliding window of five seconds. In addition to these measures, we obtained the length of articles' content and the height of users' viewport in pixels[3].

Site	# Articles	# Visitors	# Page views
Financial	9,088	323,691	923,993
Tech	12,188	739,415	822,627
HowTo	13,297	532,354	556,408
Science	10,837	725,039	1,554,008
Women	9,807	639,261	1,629,765
Sports	7,659	937,631	1,660,367
Magazine	3,945	508,136	569,574
Total	66,821	4,405,527	7,716,742

Table 1: The number of articles, visitors, and page views analyzed from each site.

Next, we describe the transformations we apply to this dataset in order to study robust patterns of engagement with news that persist across different sites, audiences, and devices.

4 TRANSFORMED MEASURES

The raw measures provided by Chartbeat for each page view already provide useful information about engagement with news articles. They quantify the bulk amount of time spent on the page, how far down the page individuals scrolled, and the amount of interaction with the page. Prior work showed that these signals correlate with the relevance judgments of individuals and provide useful information for recommendation systems [8, 15, 21].

Despite the valuable information captured by measures of dwell time, scroll depth, and the amount of page interaction, these measures suffer from a few notable weaknesses. First, as absolute measures, these are not universal for different sites and not necessarily consistent across different pages of the same site (e.g. sections with different layouts). The differences in the absolute measures across sites can be seen in the top three rows of Figure 1, which differ in their mean, variance, and skew even after log-transformation. Furthermore, the raw measures are agnostic to the content being viewed and the context of viewing it. For example, a person reading a short article to completion on a desktop device may track similarly to an individual reading only a small portion of a long article on a mobile device. Previous work proposed ways to normalize the raw metric per group, but not for single page views [21]. Of course, content and context features can be included in recommendation systems along with engagment measures, but without directly modeling their relationship to user engagement the quality of model fit may suffer. Moreover, since dwell time, scrolling depth, and active engagement are all positively correlated, using these measures without modeling their inter-dependency misses valuable information, as we demonstrate in Section 5.

In order to address the aforementioned limitations we propose the following three relative measures for each page view. We define **Relative Depth (*Rel. Depth*)** and **Average Scrolling Speed**

[1] http://www.chartbeat.com

[2] Self-described in the publication's tagline.

[3] Vertical pixel length is sufficient for our purposes since all the included sites used non-responsive layouts, which make the length of articles effectively constant across screen resolutions.

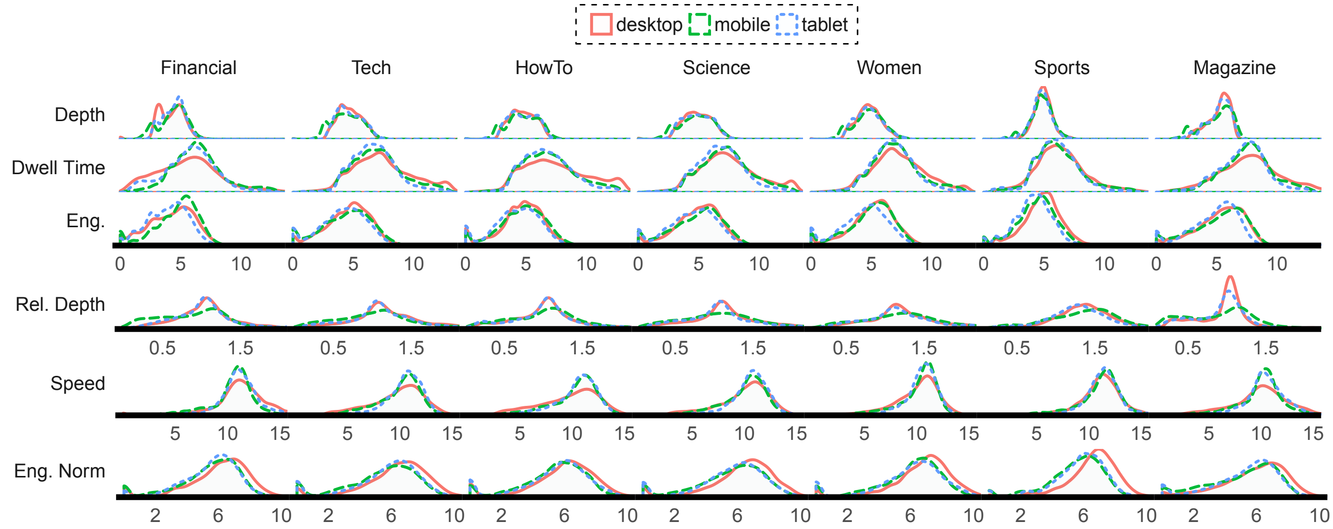

Figure 1: Distribution of log-transformed absolute and relative measures (rows) for the seven sites in our dataset (columns). Y-axes are scaled by the maximal value in each row.

(***Speed***) as:

$$\text{Rel. Depth} = \frac{d+h}{l} \qquad \text{Speed} = \frac{d+h}{t} \tag{1}$$

Where d is the maximal position reached on the page, h is the user's viewport height, l is the length of article content, and t is the total time an individual spent on the page. *Rel. depth* captures the proportion of an article that was visible on the user's screen. The measure of *Speed* records how quickly the user scrolled through the visible part of the page (on average). Together, the relative depth and speed provide information about the overall navigation through the article. Nevertheless, they do not capture activity while the page was static (e.g. highlighting or mouse movement) or the amount of energy spent in reaching the final position on the page.

Our third measure aims to complement previous measures by offering a normalized version of active engagement. As mentioned in Section 3, Chartbeat records active engagement as a smoothed number of seconds of page interactions of any form (mouse move, scroll, swipe, key press, etc.). The raw measure tends to increase with article length since longer articles provide more opportunities for interaction. In addition, reading the same article on different devices requires different amounts of interaction with the page. For example, reading on a mobile device generally requires more scrolls in order to cover an entire article, compared to desktop devices. To address these issues, we define **Normalized Engagement (*Norm. Eng.*)** as:

$$\text{Norm. Eng.} = \frac{e}{{}^{l}/_{h \cdot c}} \tag{2}$$

Where e is the raw measure of active engagement from Chartbeat, l and h are the length of the article and height of the viewport view as before, and c is a constant scaler. The fraction ${}^{l}/_{h}$ represents the number of screens it would take a visitor with viewport view of height h to cover an entire article of length l. Of course, not all page views cover the article in its entirety, and thus a constant c is useful for scaling and shifting the distribution away from its natural limit of zero engagement. We found empirically that $c = 10$ is sufficient for shifting values away from zero without decreasing the variance too heavily.

The three bottom rows of Figure 1 show the log-transformed distributions of the three relative measures we described in this section. Similar to the absolute measures (upper three rows), after the logarithmic transformation the distributions follow a bell shaped curve. Generally speaking, after the relative transformations, site differences are notably less prominent. One notable exception is the distribution of relative depth for the magazine site, which is bimodal and more concentrated around values of one. This suggests that people read articles to completion more often on this site.

Finally, we use both the absolute and relative measures to form a six-dimensional vector $\vec{v}$ that summarizes user engagement during a page view. $\vec{v}$ consists of the three absolute measures (dwell time, scrolling depth, and active engagement) as well as the three relative measures defined in Equations 1 and 2. Since all measures may produce zero or near-zero values, we add 1 to all measures prior to computing the logarithm (base two).

Next, we use the measures described in this section to identify different modes of engagement with news articles.

5 MODES OF ENGAGEMENT

Previous sections described some of the most common measures of engagement on the Web as well as transformations that are particularly tailored to capture engagement with news articles. In this section, we describe our approach for learning robust patterns of engagement with news articles, that are indicative of user behavior during a page visit. We first identify likely modes of behavior that are present in individual page views and then use this information to reflect on the collective behavior of users in articles.

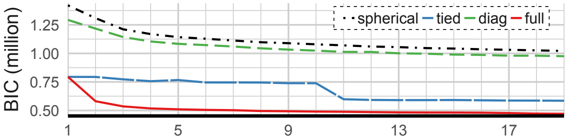

Figure 2: Baysian Information Criteria (BIC) for different number of components K and covariance matrix structures.

5.1 Characterizing page views

Central to our approach is the idea that the use of multiple summaries of user engagement with a page can "triangulate" and recover valuable information about the central tendencies of user behavior in the full time series. To test this idea, we model each summary vector $\vec{v}$ as the result of a probabilistic multivariate mixture model, and use the identified mixture modes to compare our findings to previous work, which had access to the full time series. In order to learn a model that generalizes better across sites and browsing devices, we train the model on a balanced dataset of 63,000 page views, where different sites and devices are equally represented.

We use a multivariate normal (MVN) mixture model for several reasons. First, a mixture of MVN model can capture multi-modal distributions and provide a reasonable approximation for the bell-curve distributions observed in Figure 1. Second, the covariates of the MVN components are flexible enough to capture interdependency between measures while keeping the model tractable and interpretable as a whole. Third, the model is simple enough so that publishers could easily adopt it[4]. The posterior probability of an MVN mixture model is:

$$P(\Theta|V) = \sum_{i=1}^{K} \pi_k \mathcal{MVN}(V|\mu_k, \Sigma_k) \tag{3}$$

where V is the set of page view summaries $\vec{v}$ in our data and $\Theta = \{\pi, \mu, \Sigma\}$ is the set of model parameters, which consist of K multivariate normals, each parameterized by μ_k and Σ_k and weighted by π_k. The covariance matrix Σ_k can be shared across different components (tied), restricted to a diagonal matrix (diag) or single value variance matrix (spherical), or be independent for each component (full). We consider all four variants of the covariance matrix structure when optimizing model parameters.

Before fitting the model, we leave out page views that are merely quick bounce backs. Similar to previous work [24], we consider bounce backs to be page visits where the page was visible for less than 10 seconds and/or had no user interaction (active engagement of zero). This enables the mixture model to focus on page visits where the user interacted with the content beyond just a quick glance (if any) of the first viewport view.

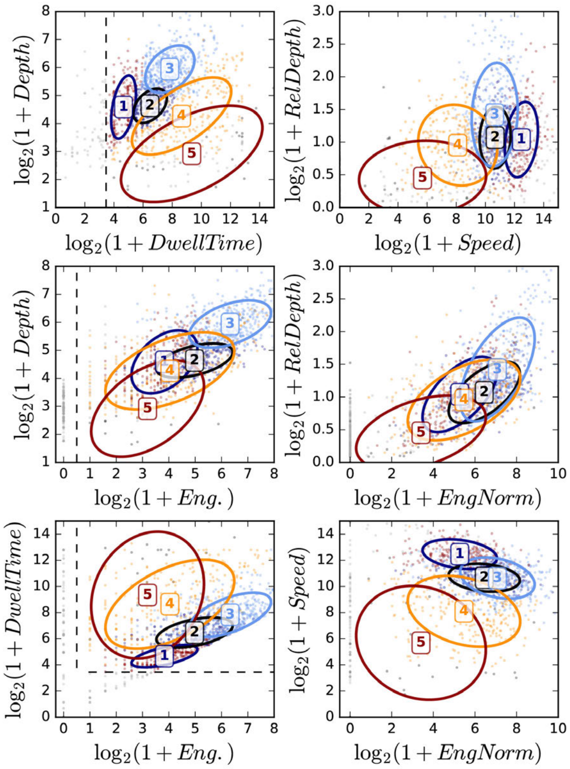

Figure 3: The multivariate normal components identified by a mixture model with $k = 5$ and projected in two dimensions.

#	Depth	Dwell	Eng.	Rel.	Speed	Eng.	Label
	[px]	Time	[s]	Depth	[$\frac{px}{m}$]	Norm	
		[s]		[%]		[s]	
1	2305	24	13	106	5671	37	Scan
2	2385	87	30	110	1622	86	Read
3	6038	226	79	168	1596	134	Read (long)
4	1824	398	15	94	273	43	Idle
5	656	647	8	37	50	9	Shallow

Table 2: Cluster means and labels.

Choosing the optimal K: Figure 2 shows the Bayesian information criteria (BIC) for values of K ranging from 1 to 19, for the four different types of covariance matrix described earlier. Model parameters were optimized using the EM algorithm with 100 random initializations for each value of K and covariance structure. It is clear from the figure that the full covariance matrix outperforms all other covariance structures, and that the improvements in BIC diminish considerably beyond $k = 5$. Manual examination of the resulting clusters with $k > 5$ showed that the additional components only further partitioned the clusters already found with $k = 5$. With $k < 4$ high-variance clusters emerged, covering the entire span of the data. Therefore, we use $k = 5$ for the rest of the analysis.

Interpreting the clustering results: The results of clustering page views using a MVN mixture model are in Figure 3, with cluster means in Table 2. Each point in the figure is a page view $\vec{v}$, colored according to its most probable cluster assignment based on posterior probability for that page view. Cluster means and variance are shown using ellipses, centered at the mean of each cluster, with a

[4] The trained mixture model is publicly available at https://github.com/nirg/mods_usr_eng.

95% confidence interval around it, as projected into two dimensions. For example, the top left panel shows how the five clusters vary in terms of scroll depth and dwell time (in logarithmic scale). One can see, for instance, that the light blue cluster (numbered “3”) has the highest average scrolling depth (top left panel) as well as the highest average relative depth (top right panel). Table 2 confirms these observations, showing cluster 3 with an average scrolling depth of 6038 pixels and average relative depth of 168%. The dashed lines on the left panels designate the thresholds used for delineating quick bounce backs as described earlier.

Using the clustering results, we can label the different engagement modes identified in the data. The top right panel of Figure 3 shows that the different components vary in their scrolling speeds and coverage of articles. Clusters 1, 2, and 4 cover roughly the same portion of articles (same *RelDepth*), but at vastly different speeds. Table 2 shows that page views in clusters 1, 2, and 4 covered around 100% of articles, but page views in cluster 1 did so in speeds 3-20 times faster. In order to compare these speeds to the ones documented in the literature, in words per minute, we multiply the number of words in an article by the relative portion viewed (Rel. Depth) and divide by Dwell Time (in minutes). The central 50% of reading speeds in cluster 1 ranges from 1035 to 1823 words/minute, which is well above in-depth or even skim reading speeds documented in the literature [27, 30]. Therefore, we consider cluster 1 as reflecting scanning behavior.

Clusters 2 and 3 have similar distributions of speed and normalized engagement distributions, but differ in all other aspects. While the absolute measures (left three panels) show cluster 3 as mostly distinct from other clusters, the relative measures (right three panels) show significant overlaps between the two clusters. This suggests that a similar underlying user behavior is present in both of these clusters, one that we could not identify based on the absolute measures alone. Page views in clusters 2 and 3 reach relatively deeply into articles, have high levels of user interaction, and translate to reading speeds of 200 to 600 words/minute in the central 50%. Page views in cluster 3 often reach past the article body and involve slightly more user interactions, potentially interacting with user-generated comments after the main body of articles. Based on these characteristics, we believe that it is likely that clusters 1 and 3 reflect reading behaviors. This interpretation is consistent with the literature, which describe in-depth reading at a speed of 200-300 words/minute and skimming at 3-4 times faster speed [27, 30]. Moreover, clusters 1 and 3 correspond to the “deep” (covering entire articles) and “complete” (reaching the content) behaviors identified in Lagun and Lalmas’s work using the full engagement time series [24]. Therefore, we use “read” and “read (long)” to distinguish between the two clusters, reflecting that extended engagement in cluster 3.

Cluster 4 is characterized by relative depth that is similar to other clusters, but at significantly slower speeds, with relatively little page interaction. Together with the fact that 75% of page views in this cluster correspond to reading speeds below 100 words/minute it seems likely that the cluster includes a short period of user activity coupled with a longer period of inactivity. We suspected that this cluster might capture user engagement with videos, but a random sample of articles in this cluster did not support this hypothesis. Hence, we refer to cluster 4 as involving some “idle” behavior.

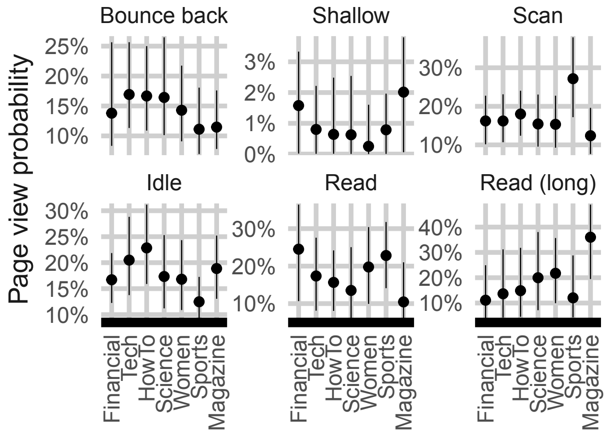

Figure 4: Median and central 50% of articles in a given engagement mode for each site.

The last cluster, numbered 5, is referred to as “shallow” due to its low values across all engagement measures. The cluster is present in less than 2% of page view and seems to serve as a buffer between the quick bounce backs and the richer modes of engagement with an article.

5.2 Characterizing engagement with news articles

Using the clusters identified in the previous section, we can now characterize the collective behavior in news articles.

We obtain an estimate for the engagement with articles by averaging the probability of assignment to different engagement modes across page views. Given a page view summary $\vec{v}_i$ and the trained mixture model parameters Θ, the conditional probability that $\vec{v}_i$ came from the k^{th} cluster is $P(z_i = k|\Theta, \vec{v}_i) \propto \pi_{z_i}\mathcal{MVN}(\vec{v}_i|\mu_{z_i}, \Sigma_{z_i})$, where z_i is a latent categorical variable representing the assignment to cluster k. Let $\vec{e}_i$ be a J-dimensional vector representing the probability of assignment of $\vec{v}_i$ to one of the K engagement modes identified in the previous section or being a quick bounce back (thus $J = K + 1$). We consider bounce backs as a “hard” assignment, meaning the when $\vec{v}_i$ meets the criteria for a bounce back the entire probability mass is assigned to this mode and $e_{i,j>0} = 0$. Otherwise, $e_{i,0} = 0$ and $e_{i,j\in 1..(k+1)} = P(z_i = j|\Theta, \vec{v}_i)$. We calculate the engagement vector $\vec{e}^{(a)}$ for article a as the mean assignment vector $\vec{e}_i$ across all page views of the article.

Figure 4 shows the different levels of engagement with articles from different sites. Each point is the median and the central 50% of articles according to the proportion of page views of a given engagement mode (different panels) and site (x-axis). For example, the top right panel shows that the median article on the Sports site has about 28% of page views that are merely scanning the article. This percentage is considerably higher than the rate of scanning on other sites, suggesting that people visit Sports articles with different intent, perhaps simply looking for a result of a recent sporting event.

A few interesting points emerge from Figure 4. The proportion of *Read (long)* (bottom right panel) shows that visitors to the Magazine articles are almost twice as likely to engage in longer reads than visitors on any other site. This runs counter to the belief that reading on digital devices, particularly of web pages content, diminishes people's ability to engage for long periods of time [4, 9, 29, 34]. Another interesting point is the relatively high percent of *Idle* engagements in How To articles. The few articles we examined from this site gave instructions for fixing, making, or doing something in the physical world. It is therefore plausible that people disengage from their digital devices to follow instructions in the physical world.

In summary, this section presented an approach for utilizing simple summary statistics of page visits to news articles to characterize likely user behaviors on the page. We identified modes that can enable publishers and recommendation system to distinguish between different modes of reading, scanning and other lighter forms of engagement, and showed that distribution among different modes of engagement varies across sites.

Next, we develop a measure of information gain within the text of articles and examine how this measure relates to user engagement.

6 SEMANTIC INFORMATION GAIN

The previous section characterized six prototypical modes of engagement with news articles. Of course, we expect individuals' reading decisions to be influenced by the content they read. Previous work explored the impact of non-textual elements (e.g. images, ads) and general properties of a text on users' engagement [14, 24, 42]. Our goal in this section is to develop a novel measure that captures the flow of information *within* the text of articles, and explains some of the variability in the way people engage with articles.

Inspired by ideas from the theory of Information Foraging [10, 11, 36], we develop a measure for semantic information gain with each paragraph of a text[5]. Paragraphs serve as natural units of analysis, which according to news writing guidelines should contain a single idea [39]. We further assume that reading happens linearly, one paragraph at a time, in the order it appears in the text. This assumption is somewhat supported by studies of eye-movement [10, 23], and is necessary in our case since we do not have the time series information about users navigating through pages.

We calculate the semantic information gain (SIG) of articles as follows. We train a 48-dimensional Doc2Vec model [25] based on the entire corpus of articles (N=66,821) with 20 iterations over the corpus, excluding stopwords and removing tokens that appeared in less than 10 documents. A simple validity test shows that the vector embeddings capture a significant amount of information about the articles – the cosine self-similarity of train documents to their inferred vector embedding after training is above 0.8 in 95% of articles. Let $\vec{d}^{(a)}$ be the vector embedding for article a, and $\vec{p}^{(a)}_{1..l}$ be the vector embedding inferred for the text accumulated from paragraphs 1 through l of the article. We define $SIG(l)$ for each article as the cosine similarity of $\vec{d}^{(a)}$ and $\vec{p}^{(a)}_{1..l}$.

The SIG as a function of ordinal position of paragraphs has two main drawbacks: it does not distinguish between paragraphs of different length and it is not monotonically increasing, as one might expect when comparing the full text to increasing prefixes of itself. To address both of these issues, we convert the SIG to a function of relative pixel depth in an article (as defined in Section 4) by replacing the ordinal indices of paragraphs with their relative pixel depth and interpolating the information gain between points. We fit a second-degree polynomial to the SIG of each article using quadratic programming, with constraints that restrict the polynomial to be mostly monotonic increasing, bounded from above by the maximum SIG value of one, and reach its peak within the bounds of the article. More formally, these constraints are a non-positive quadratic term ($a <= 0$), intercept $c <= 1$ and vertex $= -\frac{b}{2a} <= 1$. The quadratic approximation achieved an R^2 of 0.87, a 5% improvement over the unconstrained quadratic fit.

Figure 5 shows how the SIG at the beginning of articles (first 30%) varies across sites and engagement modes. For each site (panel), a line represents the average of the top 300 articles ranked by proportion of page visits of a given engagement mode. For instance, in the Magazine site (right panel) we see that the top articles that people read for longer periods of time (solid black line) start with about 4% more information than articles that people read for shorter period of time (dashed orange line). We observe the opposite trend in other sites, the SIG in articles that are read for longer periods of times open with relatively less information and develop more gradually. Overall, the figure clearly demonstrates that the amount of information conveyed in articles' text differs by site and is associated with different types of user engagement.

Next, we examine the extent to which engagement with news articles can be predicted using the measure of information gain we developed in this section, as well as other features.

7 PREDICTING ENGAGEMENT

In this section, we focus on predicting the distribution of page views among the six engagement modes for an out-of-sample article. Section 5 denoted this quantity as $\vec{e}^{a}$, a six-dimensional vector for each article, which represents the expected proportion of page views that are generated from each engagement mode. We use the measure of semantic information gain described in the previous section as well as other pre-publication features to examine the ability to predict engagement with previously unseen articles.

We use ten-fold cross validation, stratified by site, to assess the ability of regression model to predict engagement proportions. We restrict our analysis hereafter to articles with more than 10 page views (N=26,203) in order to improve the accuracy of proportions estimates we set out to learn. Since the dependent variable, $\vec{e}^{a}$, consists of proportions that sum up to one, a natural model selection is the Dirichlet Regression model [16]. In practice, however, we found that this model suffers from numerical instability and proceed with a linear regression model fit separately for each engagement mode. In addition, we fit a separate regression for each site to allow the regression model to better capture site differences. We test different feature sets as we describe next and report on the Pearson correlation between the ground truth values ($\vec{e}^{a}$) and model predictions for held-out documents.

The baseline for comparison consists of the length of text and the amount of non-textual elements in an article, which were associated

[5] Our notion of information gain is different than the one commonly used in decision trees.

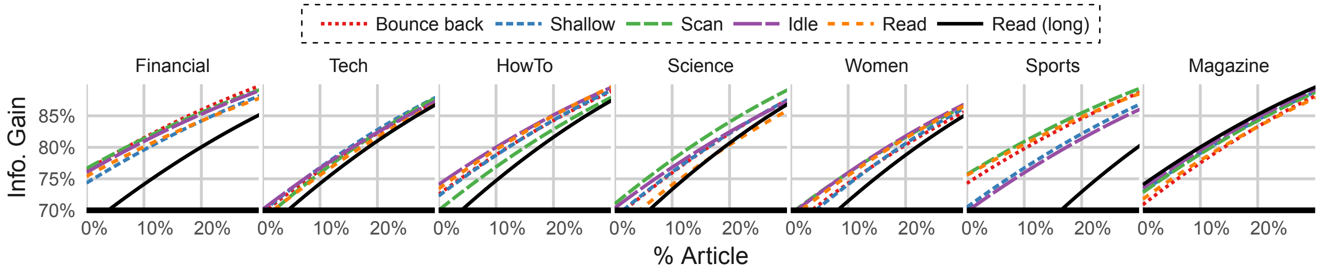

Figure 5: Semantic information gain for the top articles in each site by engagement mode.

to user engagement in previous work [14, 24, 42]. We measure the text length in number of words, excluding stopwords and infrequent words as described in Section 6. The amount of non-textual content is computed as the ratio of article visual length (i.e. in pixels) to number of words.

The next set of features includes the latent topics of the text. We run the standard LDA model [5] using Gibbs sampling on the same train-test folds of our cross validation and include the topic proportion predictions from LDA (θ_d) as features in the regression. We implement the TUNE model[6] proposed by Lagun and Lalmas [24], which jointly models the text and the proportions of engagement modes. From the TUNE model we extract the topic proportions of test documents and the predicted distribution among the engagement modes as features. Hyperparameters in both models were assigned to $\alpha = 50/T$ and $\beta = 0.1$ following the commonly used heuristic, and we experimented with the number of topics $T \in \{10, 20, 30, 40, 50\}$. We closely monitored convergence in both models using the log-likelihood. For brevity, we only report the results of these models with $T = 50$, which outperformed the results obtained using all other configurations.

In addition to broad topics of the text, we include features that describe the difficulty of the text, its sentiment, and of course the semantic information gain (SIG). We use sentence length and Flesch-Kincaid grade-level score as proxy for difficulty [22]. We compute the difficulty for both article lede (first three sentences) and of the entire article. We use the empirically-validated sentiment analysis tool VADER [19] to compute the average sentiment for the lede and the entire article. Finally, we include the three quadratic terms (a, b, and c) approximating the SIG in each article. We also include the raw SIG scores for the first three paragraphs (without any approximation) to better capture the information gained at the beginning of articles.

Lastly, we examine the ability of post-publication features to predict the distribution among different engagement modes. We use audience composition features that describe percentages of visitors on mobile devices, visitors from different referral sources (search, social media, other news sites), and visits at different times of the day. We compare the predictive power of post-publication averaged dwell time, which can be approximated on the server-side, with client-side averages of scrolling depth and active engagement.

Table 3 shows the ability of different feature sets to predict engagement with articles. Each entry is the Pearson correlation between the predicted level of engagement and the level observed in data. As mentioned before, the regression models were fit using ten-fold cross validation for each site and then averaged across sites. The baseline model, which consists of the visual and textual length of articles, confirms that length of articles is a better predictor for engagement modes that involve more of the content (reading and idling) than more shallow forms of engagement (bounce, shallow, or scan). Except for the textual difficulty and sentiment, all feature sets provide significant improvements over the baseline in all engagment mode ($p < 0.05$ in one-sided t-test using Fisher's transformation of correlation coefficients to z-values). Text difficulty and sentiment provide statistically significant improvements for the reading and bounce back modes, while the other modes are not significantly different than the baseline.

The two topic models (LDA and TUNE) outperform the baseline, both individually and jointly. LDA produces slightly better predictions than TUNE, but when features from the two models are combined the prediction improves across all engagement modes. This suggest that LDA and TUNE capture different type of topics that are complementary to each other.

SIG significantly improves the ability to predict reading. The six features of SIG obtain Pearson correlation of 0.434, a 0.103 improvement over the baseline and 0.035 improvement over the two topic models with over 100 features ($p < 0.001$). Text difficulty, sentiment and SIG add modest improvements of about of 0.01-0.02 to all other modes of engagement. Combining all the textual features (noted as *Text* in the table) results in substantial improvements over the baseline across all modes of engagement ($p < 0.001$), and in particular for reading due to the measure of SIG as highlighted in the table.

These findings provide supporting evidence for the hypothesis that information gain in text does indeed affect how people read, particularly in short articles. It is plausible that information gain does not achieve similar improvements in long articles since reading for extended periods of time depends more on inherent readers' interest than on a particular arrangement of information. This interpretation is also supported by the higher accuracy topic models achieved for longer reads than the model based on SIG.

Finally, Table 3 shows that post-publication information achieves the highest predictive accuracy when combining audience characteristics, client-side engagement information, and textual features.

[6] Our implementation of the TUNE algorithm is available at https://github.com/nirg/mods_usr_eng.

Feature set	Bounce	Shallow	Scan	Idle	Read	Read (long)
Baseline	.236	.080	.280	.370	.330	.679
Baseline+LDA(T=50)	.302	.132	.355	.410	.393	.710
Baseline+TUNE(T=50)	.286	.137	.330	.390	.367	.698
Baseline+LDA+TUNE(T=50)	**.311**	**.140**	**.362**	**.414**	**.399**	**.711**
Baseline+Difficulty+Sentiment	.250	.093	.290	.373	.347	.689
Baseline+SIG	.254	.099	.296	.386	.434	.689
Baseline+Difficulty+Sentiment+SIG	**.265**	**.105**	**.303**	**.388**	**.442**	**.696**
Baseline+Text	**.321**	**.145**	**.366**	**.426**	**.476**	**.717**
Baseline+Audience	.373	.205	.383	.446	.367	.694
Baseline+Avg. Dwell Time	.562	.188	.472	.578	.397	.713
Baseline+Avg. Engagement	.848	.478	.585	.657	.580	.864
Baseline+Audience+Avg.Eng+Text	**.858**	**.484**	**.608**	**.684**	**.634**	**.875**

Table 3: Pearson correlation (averaged across sites) between predicted and observed article engagement. Brackets indicate key prediction improvements described in the text.

The features that characterize the audience of an article predict bounce backs and shallow engagement better than the text-based models. Article-averaged dwell time, which can be approximated from server logs, is the single best predictor of non-reading engagement. It requires additional client-side information about the article, average scrolling depth and active interaction (indicated as *Avg. Engagement*), to obtain better estimates of reading. Combining all of these with the textual features provides further improvements, demonstrating that each feature category captures different and unique aspect of user engagement.

8 CONCLUSION AND FUTURE WORK

In this paper, we developed a scalable approach for capturing post-click user engagement with online news articles. We proposed a metric that enabled us to meaningfully compare summaries of user interaction across different news articles, sites, and mobile and non-mobile devices. Despite being compact, we showed that our metric retains sufficient information in order to distinguish between different modalities of post-click behavior, which mapped to behaviors identified in previous work using much more granular data. Using a multivariate normal mixture model we identified six prototypical modes of engagement that range from quick bounce backs to extended reading. We showed that the engagement modes are present, to varying degrees, in all seven sites in our dataset. Furthermore, we introduced a novel measure of semantic information gain in news articles that captures the development of ideas within the text of articles. We found that our measure of information gain is the single best predictor of reading engagement prior to publication, and that the measure remains valuable even after publication in the presence of information about the article audience and actual engagement averages.

The study highlighted several key findings. We observed that certain modes of engagement are more prominent in certain sites. In particular, we found substantially more scanning in Sports, more idling in “How To”, and more extensive reading for long-form magazine content. Extensive reading of long online content is particularly interesting since it runs counter to popular claims that online reading hinders people’s ability to focus for extended periods of time [4, 9, 29, 34]. Second, our findings suggest that the organization of ideas within the body of articles affects how people engage with it. We found that our measure of information gain in text is a good predictor for reading of articles, but less so for extended reading of long articles. We believe that this finding reflects an important limitation of good writing – it needs to be coupled with strong user interest in order to be read more fully, especially for long articles.

The overall approach laid out in this work provides a tangible way for digital publishers to adopt informative and cost-effective measures of user engagement with news articles. The summaries of user interaction we proposed are simple to track using client-side logging and compact enough to require only a minimal amount of additional storage per page view. Publishers can use the mixture model trained in this work to translate raw page views into more meaningful concepts for describing user engagement with news content, such as the proportion of likely reading or skimming events. These new measures can inform editorial decisions, which in turn could help improve the experience of readers.

There are several avenues for future work to extend the current study. The additional information extracted from page visits can be integrated into recommendation systems, which could lead to similar improvements derived from the use of dwell time [21, 42]. A more nuanced view of the behavior of visitors can enable better targeting of content to potential reader populations. In addition, combining measures of user engagment with text could prove useful in text summarization tasks. More generally, future work could investigate the origins of the engagement differences we observed in the current work, distinguishing between individual differences, differences in user intent across different genres, and properties of the content itself.

9 ACKNOWLEDGMENTS

The author would like to thank Josh Schwartz, Chief Data Scientist at Chartbeat, for his continued support of this work and many useful discussions about tracking in real-world settings. The author would also like to thank Dan Valente for his instrumental help in devising and optimizing the collection process of the raw data. Finally, the author is grateful to the anonymous reviewers for their thoughtful and insightful comments, which helped improve the quality of the paper.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proc. SIGIR 2006.*

[2] Ioannis Arapakis, Mounia Lalmas, B. Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M. Jose. 2014. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *JASIST* 65, 10 (2014), 1988–2005.

[3] Ralf Biedert, Jörn Hees, Andreas Dengel, and Georg Buscher. 2012. A robust realtime reading-skimming classifier. In *Proc. ETRA 2012.*

[4] Sven Birkerts. 2006. *The Gutenberg elegies: The fate of reading in an electronic age.* Macmillan.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *JMLR* 3, Jan (2003), 993–1022.

[6] Georg Buscher, Edward Cutrell, and Meredith R. Morris. 2009. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proc. CHI 2009.*

[7] Christopher S. Campbell and Paul P. Maglio. 2001. A robust algorithm for reading detection. In *Proc. PUI 2001.*

[8] Mark Claypool, Phong Le, Makoto Wased, and David Brown. 2001. Implicit Interest Indicators. In *Proc. of IUI 2001.* 33–40.

[9] Microsoft Canada Consumer insights. 2015. Attention spans. (2015). https://advertising.microsoft.com/en/WWDocs/User/display/cl/researchreport/31966/en/microsoft-attention-spans-research-report.pdf.

[10] Geoffrey B. Duggan and Stephen J. Payne. 2009. Text skimming: the process and effectiveness of foraging through text under time pressure. *Journal of Experimental Psychology: Applied* 15, 3 (2009), 228.

[11] Geoffrey B. Duggan and Stephen J. Payne. 2011. Skim reading by satisficing: evidence from eye tracking. In *Proc. of CHI'2011.* ACM.

[12] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.

[13] Ruth Garner. 1987. *Metacognition and reading comprehension.* Ablex Publishing.

[14] Daniel G. Goldstein, Siddharth Suri, R. Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz. 2014. The economic and cognitive costs of annoying display advertisements. *Journal of Marketing Research* 51, 6 (2014), 742–752.

[15] Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proc. WWW 2012.*

[16] Rafiq H. Hijazi and Robert W. Jernigan. 2009. Modelling compositional data using Dirichlet regression models. *Applied Probability & Statistics* 4, 1 (2009), 77–91.

[17] William Horton, Lee Taylor, Arthur Ignacio, and Nancy L. Hoft. 1996. The Web page design cookbook: all the ingredients you need to create 5-star Web pages. *New York: Wiley, 1996* (1996).

[18] Jeff Huang, Ryen W. White, Georg Buscher, and Kuansan Wang. 2012. Improving searcher models using mouse cursor activity. In *Proc. SIGIR 2012.*

[19] Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. ICWSM 2014.*

[20] Peter H. Johnston. 1983. *Reading comprehension assessment: A cognitive basis.* ERIC.

[21] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proc. WSDM 2014.*

[22] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Technical Report. DTIC Document.

[23] Dmitry Lagun and Eugene Agichtein. 2015. Inferring searcher attention by jointly modeling user interactions and content salience. In *Proc. SIGIR 2015.*

[24] Dmitry Lagun and Mounia Lalmas. 2016. Understanding User Attention and Engagement in Online News Reading. In *Proc. of WSDM 2016.*

[25] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. ICML 2014).*

[26] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. 2012. Models of user engagement. In *Proc. UMAP 2012.*

[27] Guoqiang Liao. 2011. On the Development of Reading Ability. *Theory and Practice in Language Studies* 1, 3 (2011), 302–305.

[28] Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proc. SIGIR 2010.*

[29] Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of documentation* 61, 6 (2005), 700–712.

[30] Michael E. Masson. 1982. Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8, 5 (1982).

[31] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. 2016. The Modern News Consumer, News Attitudes and Practices in the Digital Era. (2016).

[32] Masahiro Morita and Yoichi Shinoda. 1994. Information filtering based on user behavior analysis and best match text retrieval. In *Proc. SIGIR 1994.*

[33] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proc. WWW 2013.*

[34] Jakob Nielsen. 2008. How Little Do Users Read. (2008). http://www.nngroup.com/articles/how-little-do-users-read/.

[35] Christine Nuttall. 1996. *Teaching reading skills in a foreign language.* ERIC.

[36] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[37] Kenneth R. Pugh, Bennett A. Shaywitz, Sally E. Shaywitz, R. Todd Constable, Pawel Skudlarski, Robert K. Fulbright, Richard A. Bronen, Donald P. Shankweiler, Leonard Katz, Jack M. Fletcher, and John C. Gore. 1996. Cerebral organization of component processes in reading. *Brain* 119, 4 (1996), 1221–1238.

[38] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

[39] Carole Rich. 2015. *Writing and reporting news: A coaching method.* Cengage Learning. 259 pages.

[40] Alexandre N. Tuch, Javier A. Bargas-Avila, Klaus Opwis, and Frank H. Wilhelm. 2009. Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *International Journal of Human-Computer Studies* 67, 9 (2009), 703 – 715.

[41] Ou Wu, Yunfei Chen, Bing Li, and Weiming Hu. 2011. Evaluating the visual quality of web pages using a computational aesthetic approach. In *Proc. WSDM 2011.*

[42] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan N. Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *Proc. RecSys 2014.*