Understanding Web Browsing Behaviors through Weibull Analysis of Dwell Time

Chao Liu
Microsoft Research
One Microsoft Way
Redmond, WA 98052
chaoliu@microsoft.com

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052
ryenw@microsoft.com

Susan Dumais
Microsoft Research
One Microsoft Way
Redmond, WA 98052
sdumais@microsoft.com

ABSTRACT

Dwell time on Web pages has been extensively used for various information retrieval tasks. However, some basic yet important questions have not been sufficiently addressed, *e.g.*, what distribution is appropriate to model the distribution of dwell times on a Web page, and furthermore, what the distribution tells us about the underlying browsing behaviors. In this paper, we draw an analogy between abandoning a page during Web browsing and a system failure in reliability analysis, and propose to model the dwell time using the Weibull distribution. Using this distribution provides better goodness-of-fit to real world data, and it uncovers some interesting patterns of user browsing behaviors not previously reported. For example, our analysis reveals that Web browsing in general exhibits a significant "negative aging" phenomenon, which means that some initial screening has to be passed before a page is examined in detail, giving rise to the browsing behavior that we call "screen-and-glean." In addition, we demonstrate that dwell time distributions can be reasonably predicted purely based on low-level page features, which broadens the possible applications of this study to situations where log data may be unavailable.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: Models and Principles – User/Machine Systems

General Terms

Algorithms, Measurement, Human Factors

Keywords

Weibull analysis, User behaviors, Web browsing, Dwell time

1. INTRODUCTION

Real-world information retrieval (IR) heavily relies on effective usage of implicit feedback, which comes in various forms such as document clickthrough, viewing, scrolling, and bookmarking. Many researchers have studied the correlations between implicit feedback and document relevance (*e.g.*, [6, 21, 5, 10]), and revealed that document dwell time (*i.e.*, the length of time a user spends on a document), is generally the most significant indicator of document relevance besides clickthrough, although the extent of the relationship may vary depending on the information seeking task [14, 15]. Because of the correlation between dwell time and document relevance, dwell time has been successfully used in various applications, such as learning to rank [2, 3], query expansion [5], and inferring query-independent page importance [19]. Specifically, Agichtein et al. [2, 3] demonstrate that user browsing features, a major component of which is Web page dwell time, significantly improve the retrieval performance of a competitive search engine, even with the presence of other important features such as BM25 and search-result clickthrough. Although post-query browsing is intuitively more relevant to IR than general browsing, general browsing is still an important component in information seeking [20, 25]. Indeed, some of aforementioned studies (*e.g.*, [19]) and real-world search engines also leverage the general browsing activities for improved efficacy and coverage.

Although dwell time has been extensively studied, some important questions have not been sufficiently addressed. For example, what distribution is appropriate to model the dwell time t on a Web page[1] d across all visits, *i.e.*, what does $Pr(t|d)$ look like? Furthermore, how does the distribution depend on the features of d? And finally, what does the distribution tell us about users' general browsing behaviors? These questions are not only interesting in themselves, but are also useful for various IR applications, as we now explain.

First, accurately modeling $Pr(t|d)$ would enable the construction of generative models involving dwell time for Web text analysis. For example, when dwell time is properly modeled, topic discovery can be guided by considering both $Pr(t|d)$ and content. Second, $Pr(t|d)$ can readily answer questions such as "what is the probability that a user will stay longer than t_1 on the page?" (answer: $Pr(t \geq t_1|d)$) or "what is the expected remaining time that a user will spend on a page that he has dwelled on for t_1?" (answer: $E(t|t \geq t_1, d)$). Answers to such questions could help publishers optimize advertising and content placement. Third, understanding $Pr(t|d)$ would help us gain insights into user browsing behaviors that can help inform the design of search

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

[1] We use page, Web page, URL and document interchangeably in this paper

and advertising technologies, as we will explain later in the paper.

Precise modeling of dwell time is not straightforward since duration on a Web page depends on many factors, some of which may not even be fully captured by log data (*e.g.*, the mood of the user). In addition, the distribution family may vary with different information seeking tasks in different settings (*e.g.*, time of day). As the first step towards precise modeling of dwell time, we choose to model the overall distribution of user dwells on each Web page, which we believe will help us better understand the dwell time distributions in general across all users.

In this paper, we draw an analogy between abandoning a page during Web browsing and a system failure in reliability analysis, and use Weibull analysis techniques which are commonly used in reliability engineering [1] to characterize general browsing behaviors. Furthermore, we demonstrate that it is possible to predict the dwell time distribution based on page-level features. We make the following contributions in this study:

- **Weibull analysis of Web dwell time data**: To the best of our knowledge, this is the first time an analogy has been drawn between abandoning a Webpage and a system failure, which leads to a principled way of analyzing dwell time data. The same or similar analogies can be made for other kinds of temporal data on the Web (*e.g.*, time-to-first-click on search result pages and session length).
- **Discoveries about user browsing behaviors**: Our analysis leads to some interesting new insights regarding users' Web browsing behaviors. Specifically, we find that Web browsing exhibits a significant "negative aging" phenomenon (*i.e.*, the rate of Web page abandonment decreases over time), and that this effect is stronger for less entertaining pages. These discoveries, together with the application of Weibull analysis to a new domain, enhance our understanding of user browsing behaviors.
- **Predicting dwell time distribution**: We demonstrate that the dwell time distribution (in Weibull parametric form) can be effectively predicted from low-level page features. Not only does this broaden the applicability of dwell time data, but it also reveals what page features correlate with the dwell time distribution.

The remainder of this paper is organized as follows. We first discuss the related work in Section 2, and then examine the goodness-of-fit of the Weibull distribution in Section 3. We present the Weibull analysis results in Section 4, and elaborate on the predictive model in Section 5. With extensions and future work discussed in Section 6, Section 7 concludes this study.

2. RELATED WORK

This paper is related to work on implicit feedback within IR. Research on implicit feedback has sought to address the high cost of soliciting explicit feedback from users by unobtrusively observing their natural interactions and building models for activities such as query expansion and user profiling [17]. Although implicit feedback may be less accurate than explicit feedback [22], it is available in significantly greater quantity than explicit feedback. Implicit measures include document retention (*e.g.*, printing, saving, bookmarking) and document interaction (*e.g.*, viewing, scrolling, dwell time) [21, 6, 14]. Morita and Shinoda [21] measured the relationship between dwell time, saving, following-up and copying of a document and users' explicit ratings, and showed that there was a relationship between dwell time and interest, but no relationship between interest and any other measures. Claypool et al. [6] examined mouse clicks, scrolling, dwell time, and requested explicit ratings, and found that dwell time and the amount of mouse scrolling had a strong positive correlation with explicit ratings. Studies by Kelly and Belkin [14, 15] further found that special attention is needed to interpret dwell time as relevance because of the implications of different tasks. On the application side, besides being incorporated into learning to rank for Web IR [2, 3, 19], dwell time is also widely used in other information seeking tasks (*e.g.*, [16, 5]). In this paper, we do not focus on particular search and retrieval tasks as done in most previous work, but instead try to model the dwell time distribution across all users engaged in general Web browsing.

This work is also related to online user behavior modeling, which has been attracting significant attention in recent years (*e.g.*, [24, 25, 9]). There are two main complementary approaches to uncovering user behavior models: one based on controlled user studies (*e.g.*, [12, 13, 24]), and the other based on large-scale log analysis (*e.g.*, [25, 9, 26, 4, 19]). This work falls into the second category, and is mostly related to BrowseRank [19], which tries to infer a query-independent score for each page from page dwell time in general browsing. In particular, BrowseRank assumes (mainly for tractability) that the dwell time for a given page follows an exponential distribution. In this paper, we show that the Weibull distribution is more versatile than the exponential distribution used in [19]; it better fits the real-world dwell time data and provides insights on browsing behaviors.

This work also relates to research on Weibull analysis, which has been extensively and successfully applied in nearly all scientific disciplines, such as biological, demographical, reliability sciences (c.f. [1, 23]). This paper therefore adds a new application area to the rich literature of Weibull analysis, and meanwhile introduces a disciplined method for analyzing temporal data on the Web, *e.g.*, time-to-first-click on search result pages and the session length in time, in addition to the page dwell time as studied here.

3. MODEL FITTING AND COMPARISON

In this section, we fit the dwell time data with exponential and Weibull distributions (Section 3.2), and compare their goodness-of-fit in Section 3.3. The data used for this comparison and throughout the paper are discussed in Section 3.1.

3.1 Experimental Data

We collected two-weeks of log data from a popular Web browser plug-in operating in the English (US) market, which records the searches and browsed pages for opted-in users. The log data is organized in sessions, each of which is defined as a series of Web page visits that extends until either the browser is closed or a period of 30-minutes of inactivity. Based on the visit time of consecutive page visits within sessions, the dwell time of each page visit is calculated. We do this for all pages apart from the last page in the session,

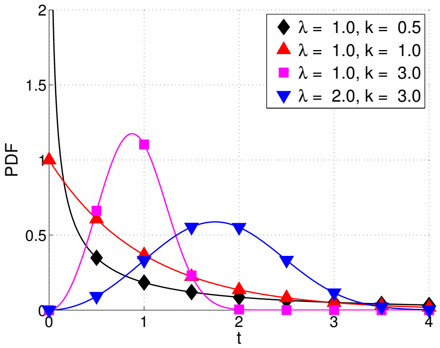


Figure 1: Example Weibull Distributions

which is then discarded from the analysis because we do not have a succeeding page visit from which to calculate its dwell time. For accurate parameter estimation, only pages with 10,000 or more visits are used. This results in a set of 205,873 URLs, each of which is accompanied by at least 10,000 dwell time observations.

3.2 Model Fitting with Maximum Likelihood

The probability density function (PDF) of Weibull distribution is given by

$$f(t|k,\lambda) = \frac{k}{\lambda}\Big(\frac{t}{\lambda}\Big)^{k-1} exp\Big\{-\big(\frac{t}{\lambda}\big)^{k}\Big\} \qquad t \geq 0, \qquad (1)$$

with $E(t|\lambda, k) = \lambda\Gamma(1+1/k)$. λ and k are the scale and shape parameters, respectively. Figure 1 plots the PDFs of some typical parameterizations, which exemplify the versatility of the Weibull distribution, and how parameters λ and k affect the scale and shape of the distribution, respectively.

When $k = 1$, the Weibull distribution reduces to the exponential distribution with PDF

$$f(t|\lambda) = \frac{1}{\lambda} exp\Big\{-\frac{t}{\lambda}\Big\} \qquad t \geq 0$$

and $E(t|\lambda) = \lambda$.

Given a sample of n observed dwell time for a page, $\{t_i\}_{i=1}^{n}$, we choose to fit the model through maximum likelihood estimation (MLE), and denote the fitted model with Weibull by M_W and that with exponential by M_E. While fitting M_E is as simple as

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} t_i}{n}, \qquad (2)$$

fitting M_W is nontrivial because the MLE of (λ, k) has no closed form. Instead, we need to use an iterative approach proposed in [7]. For completeness, we briefly outline the estimation.

Given the likelihood function as

$$L(t_1, t_2, \cdots, t_n|k,\lambda) = \prod_{i=1}^{n} \frac{k}{\lambda}\big(\frac{t_i}{\lambda}\big)^{k-1} e^{-(\frac{t_i}{\lambda})^k},$$

we set the partial derivative w.r.t. λ and k to 0, *i.e.*,

$$\frac{\partial ln(L)}{\partial k} = \frac{n}{k} - nln(\lambda) + \sum_{i=1}^{n} ln(t_i) - \sum_{i=1}^{n}\big(\frac{t_i}{\lambda}\big)^k ln\big(\frac{t_i}{\lambda}\big) = 0, \quad (3)$$

$$\frac{\partial ln(L)}{\partial \lambda} = -\frac{kn}{\lambda} + \sum_{i=1}^{n} k\frac{t_i^k}{\lambda^{k+1}} = 0. \qquad (4)$$

Eqn 4 gives

$$\lambda^k = \frac{\sum_{i=1}^{n} t_i^k}{n}, \qquad (5)$$

which, once plugged into Eqn 3, renders

$$\frac{n}{k} + \sum_{i=1}^{n} ln(t_i) - \frac{n\sum_{i=1}^{n} t_i^k ln(t_i)}{\sum_{i=1}^{n} t_i^k} = 0, \qquad (6)$$

which only involves k and can be solved through Newton-Raphson iterations. Specifically, let

$$g(k) = \sum_{i=1}^{n} t_i^k + k\frac{\sum_{i=1}^{n} t_i^k \sum_{i=1}^{n} ln(t_i)}{n} - k\sum_{i=1}^{n} t_i^k ln(t_i),$$

then

$$g'(k) = \frac{\sum_{i=1}^{n} ln(t_i)}{n}\Big(\sum_{i=1}^{n} t_i^k + k\sum_{i=1}^{n} t_i^k ln(t_i)\Big) - k\sum_{i=1}^{n} t_i^k ln^2(t_i).$$

Then the MLE of k, denoted by $\hat{k}$, is obtained by

$$\hat{k}^{(m+1)} \leftarrow \hat{k}^{(m)} - \frac{g(k^{(m)})}{g'(k^{(m)})} \qquad m = 1, 2, \cdots$$

We terminate the iterations when the change of k is less than 10^{-6}. Once $\hat{k}$ is obtained, $\hat{\lambda}$ immediately follows from Eqn 5. We try initial value $k^{(1)} = 0.1, 1, 10$, and choose the final $(\hat{\lambda}, \hat{k})$ with the largest likelihood. Readers interested in a thorough treatment of parameter estimations (besides MLE) for Weibull distributions are referred to [23]. The above estimation can be trivially parallelized across URLs for distributed computing, which affords Web-scale dwell time data analysis.

3.3 Goodness-of-fit Comparison

We use the log-likelihood (LL) and the Kolmogorov-Smirnov distance (KS-distance) [8] to evaluate the goodness-of-fit of M_W and M_E. In general, a better fit corresponds to a bigger LL and/or a smaller KS-distance.

The KS-distance as defined below

$$KS(F^*, S) = sup_x|F^*(x) - S(x)| \qquad (7)$$

is the test statistic for Kolmogorov goodness-of-fit test, which tests whether a random sample $X_1, X_2, \cdots, X_n$, whose empirical cumulative distribution function (eCDF) is described by $S(x)$, comes from a completely specified hypothesis distribution whose cumulative distribution function (CDF) is given by $F^*(x)$.

Because the exponential distribution is a special case of the Weibull distribution, M_W is guaranteed to be no worse than M_E if fitted and evaluated on the same dataset. To be fair, the dwell time observations for each page are randomly split into training and testing portions with a ratio of 4:1. In other words, the data are split within the dwell time observations for each page rather than across pages. M_W and M_E are fit using the training portion and evaluated on the testing portion for each page. LL and KS-distance on the testing portion are used to determine which model wins. The number of pages on which a model wins are listed in Table 1, which clearly shows the superiority of M_W to M_E. Specifically, M_W outperforms M_E on more than 85% of the pages in terms of both metrics, and a sign test for each result gives a p-value that is very close to zero.

	Log-Likelihood	KS-distance
M_W Wins	176,242	178,892
M_E Wins	29,631	26,981

Table 1: Comparison of Goodness-of-Fit

4. WEIBULL ANALYSIS OF DWELL TIME

In this section, we discuss the implications of a fitted Weibull distribution for understanding user browsing behaviors (Section 4.2). We first provide a brief introduction to Weibull analysis in Section 4.1.

4.1 A Primer on Weibull Analysis

Weibull analysis dates back to 1937 when Waloddi Weibull invented the Weibull distribution. It has been successfully applied to nearly all scientific disciplines, such as biological, environmental, health, physical and social sciences, but, to the best of our knowledge, not in the Web data analysis domain. By fitting time-to-failure data to Weibull distributions, Weibull analysis enables principled failure interpretation, risk assessment, failure forecasting, and planning of corrective actions. Since a full introduction to Weibull analysis is neither realistic nor necessary here, we will highlight those aspects that pertain to our analysis and referring interested readers to [23] and [1] for a thorough treatment of Weibull analysis and applications.

The most popular characteristic function of a Weibull distribution is the Hazard function, which is defined as

$$h_t(x) = \lim_{\delta \to 0} \frac{Pr(x \leq t < x + \delta | t \geq x)}{\delta}.$$

If an item that has survived time x is called an x-survivor, the hazard function gives the probability that an x-survivor fails immediately at time x, and it is also known as the *instantaneous failure rate* or the *hazard rate*. Usually, the hazard rate is interpreted as the amount of risk associated with an x-survivor at time x in reliability study and as the force of mortality in demography and actuarial science.

The hazard function of a Weibull distribution is given by

$$h_t(x) = \frac{k}{\lambda^k} x^{k-1}, \tag{8}$$

whose first-order derivative is

$$h'_t(x) = \frac{k}{\lambda} \frac{k-1}{\lambda} \left(\frac{x}{\lambda}\right)^{k-2}. \tag{9}$$

When $k \in (0, 1)$, the first-order derivative, $h'_t(x)$, is less than 0, so the hazard rate monotonically decreases w.r.t. x. This phenomenon is often termed "*negative aging*," which means that the longer one survives, the less likely it would fail instantaneously. Since the hazard rate is high at the onset, it is also called the "infant mortality" phenomenon. In abstract terms, negative aging means that a screening is taken place at the early stage so that weak items with hidden defects are sorted out while leaving robust and healthy ones in the population, or as Lehman [18] suggests "So once the obstacle of early youth have been hurdled, life can continue almost indefinitely." We will reveal the implication of negative aging for Web browsing in Section 4.2.

In contrast, $k > 1$ corresponds to the "*positive aging*" phenomenon, which means that the longer one survives, the more likely it fails instantaneously. Finally, $k = 1$ results in a constant hazard function, indicating a constant failure rate, which is the physical model of the exponential distribution.

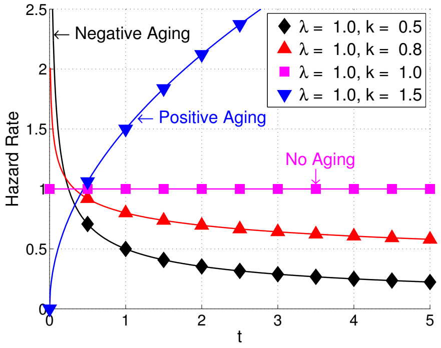
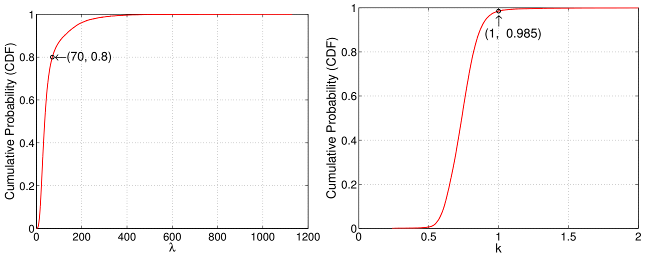


Figure 2: Example Weibull Hazard Functions



(a) eCDF of λ (b) eCDF of k

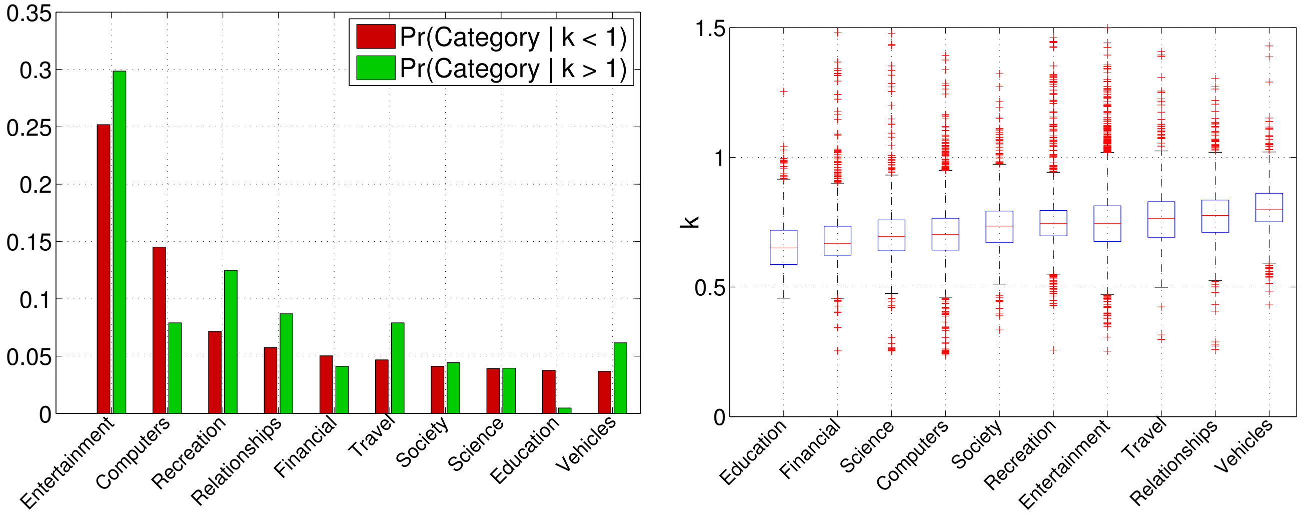
Figure 3: Distributions of the Fitted λ and k Values

The hazard functions of some example Weibull distributions are plotted in Figure 2, which illustrates different types of aging. Note that when $k \in (0, 1)$, we see negative aging, or a decrease in the failure rate over time. In the context of Web browsing this would mean a decrease in Web page abandonment rate over time. Conversely, when $k > 0$, we see positive aging, or an increase in the failure rate over time.

4.2 Weibull Analysis on Dwell Time

Using the data set as described in Section 3.1, we now examine the fitted λ and k values on the training portion for each page. Figure 3 plots the empirical cumulative distribution function (eCDF) for the fitted λ and k values. Figure 3(a) shows the eCDF for the scale parameter λ of the estimated dwell time distribution. We see that the dwell time is no more than 70 seconds on 80% of the 205,873 pages, which gives us an overall estimate of the dwell time scales across pages. Figure 3(b) shows the eCDF for the shape parameter, k. We see that k is less than 1 on 98.5% pages. Recalling that $k < 1$ indicates a negative aging effect. Thus, Figure 3(b) suggests that Web browsing exhibits a strong "negative aging" phenomenon, that is, some "screening" is carried out at the early stage of browsing a page, and the rate of subsequent abandonment decreases over time.

This discovery agrees well with the intuition about how a user browses a page: upon landing on a Web page, the user would first skim through the page, assessing the potential benefit of further reading, before delving into it and gleaning needed information. During the screening, the probability of abandoning the page is high (*i.e.*, a high hazard rate), but once the page survives the screening (*e.g.*, is regarded



(a) $Pr(Category \mid k < 1)$ vs. $Pr(Category \mid k > 1)$

(b) $Pr(k|Category)$

Figure 4: Relationship between Categories and Aging Effect as Characterized by k

as useful by the user), the abandonment rate decreases. We therefore suspect that users do in general adopt a "screen-and-glean" type browsing behavior, which gives rise to the dwell time distribution showing the observed negative aging effect.

We now examine whether and how the "negative aging" phenomenon relates to the topic of the page, as defined by category membership, *i.e.*, do people impose equal screening on pages of different categories? For this purpose, we employed a proprietary document classifier that assigned each page into one of 23 top-level categories in a taxonomy similar to dmoz[2]. The categorization succeeded on 136,395 pages. We then analyzed how category information relates to the aging effect from two complementary aspects: first, we compare $Pr(Category|k < 1)$ with $Pr(Category|k > 1)$, and second, we examine $Pr(k|Category)$ for different categories. In order to have sufficient data in each category, only the top-10 categories were retained, which included 106,169 (77.8%) pages.

Figure 4(a) compares the category distributions for Web pages with $k < 1$ and $k > 1$. We show the category distributions for each of these two types of pages. We see that *Entertainment*, *Recreation*, *Relationships*, *Travel* and *Vehicles* have a *proportionally* greater presence when $k > 1$ than when $k < 1$ (recall the sets of pages with $k > 1$ and $k < 1$ are highly imbalanced). Thus pages exhibiting positive aging ($k > 1$) are more likely to fall into these categories, which we can characterize as more entertaining, than those showing negative aging effect. Conversely, we see that the presence of *Computers* and *Education* is stronger in $k < 1$ than $k > 1$, indicating that people are more likely to screen pages in these two categories before examining them in more details. This observation leads to a hypothesis that negative aging is more common on less-entertaining pages than on fun pages, which in turn suggests that people tend to screen less-entertaining pages more harshly.

Figure 4(b) shows boxplots of $Pr(k|Category)$ for the 10 categories. The line in the middle of each box is the median of the data, and the lower and upper lines of the box represent the 25^{th} and 75^{th} percentiles of the data, respectively. The categories are ordered in ascending order of the median values from left to right, which median value of 0.6506 for *Education* and a median value of 0.7979 for *Vehicles*. Again, we observe that less-entertaining categories appear on the left of the figure, supporting the hypothesis that less-entertaining pages may be more harshly screened.

5. PREDICTING DWELL TIME DISTRIBUTION

In this section, we investigate the feasibility of predicting dwell time distribution from page-level features. A successful prediction from page features will not only enable third-parties without access to browsing logs to use dwell time information, but will also provide us with an opportunity to identify page features that are most related to dwell time distributions. We describe the experimental setup and page features in Section 5.1, report on the prediction results in Section 5.2, and inspect the learned model in Section 5.3.

5.1 Experimental Setup

We randomly sampled 5000 pages from the set of pages whose test KS-distance is less than 0.05 from the experiments in Section 3.3. By choosing pages with a high goodness-of-fit to the Weibull distribution (small KS-distance), we can provide good training examples to the classifier. For each sampled page, the λ and k values fitted on the training portion of data are taken as the learning labels. In order to extract page features, we crawled these pages using a dynamic crawler, which employs an Internet Explorer object to execute all dynamic components (*e.g.*, flash, javascript, *etc.*) and download the final rendered page. Pages containing the term "login" are excluded because these login pages are usually automatically loaded through a timeout redirection. This, together with failed crawling, gave us a set of 4,771 pages, which are randomly partitioned into training, validation and testing sets with a ratio of 7:1:2.

Since we want to inspect the learned model we use Mul-

[2] `http://www.dmoz.org/`

Feature	Description
PageSize	Size (in bytes) of the rendered page
PageHeight	Height of the rendered page
PageWidth	Width of the rendered page
DownloadCount	Number of total downloaded URLs
DownloadTime	Time to download all URLs
SecDownloadCount	Number of secondary URLs
SecDownloadTime	Time to download secondary URLs
ParseTime	Time to parse all URLs
RenderTime	Time to layout and render the page

Table 2: Details of Dynamic Features

tiple Additive Regression Trees (MART) [11], which provide good interpretability and high accuracy. We used the validation set to locate the optimal parameters, which include the maximum number (L) of leaf nodes of the base learner tree and the shrinkage parameter (v). We varied $L \in \{2, 3, 4, 6, 11, 21, 25\}$, $v \in \{1, 2^{-1}, 2^{-2}, \cdots, 2^{-6}\}$, and recorded the number (m) of iterations that achieved the minimum error. The tuple (L, v, m) that achieved the lowest error on the validation set was used in the final testing phase.

We constructed the following three sets of features, ranked in ascending order based on their closeness to what users would actually experience when viewing that page in a Web browser:

- HtmlTag: The frequency of each of the 93 HTML tags (obtained from `http://www.quackit.com/html/tags/`) is taken as an independent feature, comprising the first set. These features represent the underlying elements that are determinants of page formatting and layout but are not visible to users.
- Content: We leverage the "6of12 list" of the English words, which contains 32,153 most commonly used English words that "approximates the common core of the vocabulary of American English."[3] The top-1000 most frequent terms that appear in the training set of pages, together with one more dimension about the document length, are taken as the second set of features. They correspond to the most frequent words users would see. The value of each feature, except the document length, is the word frequency in each page.
- Dynamic: We also recorded nine measures during the dynamic crawling of each page. The dynamic crawler first downloads the backbone page, parses it, downloads any secondary URLs (*e.g.*, javascript, flash, image, *etc.*) if any, calculates the page layout, and finally renders the page. The nine features based on these measures are listed in Table 2. Because the crawler executes the scripts and renders the page, these features are meant to closely estimate users' browsing experience with the page.

We intentionally chose *not* to include advanced features such as PageRank, number of inlinks and any log-based features in the feature set, because these features are generally not available to researchers outside search engine companies. By restricting to page-level features, anyone can crawl the page, construct the features, reproduce the result, and more importantly, utilize the predictive model to asses dwell time for any pages that can be downloaded.

5.2 Prediction Results

In order to determine how different sets of features interact, we tested seven feature configurations as listed in Table 3. Also listed in the table are the optimal parameters determined by the validation set. In particular, the MART parameters for predicting λ and k are denoted by $\lambda(\cdot)$ and $k(\cdot)$, respectively.

Log-likelihood (LL) and KS-distance are again used as the evaluation metrics. For each test page, we evaluate the LL and KS-distance on the test portion data with the predicted λ and k values, and compare it with the baseline model, which returns the mean value $(\bar{\lambda}, \bar{k})$ across all training pages, which resembles, and is stronger than, the exponential model (c.f. Eq. 2).

Results are presented in Table 3, in which "Predict Win" means that the predictive model achieves a higher LL (or a smaller KS-distance) than the baseline model. As the two metrics give very similar results, we will focus on the result based on LL in the following.

First, we see that the prediction model outperforms the baseline method on all seven configurations with statistical significance. Sign tests for $\mathcal{H}_0 : \textit{predict} \leq \textit{baseline}$ all return p-values that are very close to zero for the seven configurations. This result shows that low-level page features do carry some prediction power that can be leveraged for effective dwell time prediction.

Second, HtmlTag is as effective as Dynamic when used individually, and when combined, they bring further improvement. This observation indicates that the nine Dynamic features are as predictive as the 93 HtmlTag features, and their prediction power is complementary. Conversely, Content in itself outperforms Content+Dynamic, and adding HtmlTag to Content only provides some marginal improvement.

Finally, the best performance is actually achieved by Content+Dynamic. This is reasonable in that Dynamic represents what users would experience immediately after clicking through to a page while Content corresponds to what content users would see once the page is loaded. Note that adding HtmlTag does not provide much benefit. Given the promising results from Content+Dynamic and the fact that only the top-1,000 frequent words are used in Content, we expect further improvements from better feature engineering, for example, by choosing words with high inverse document frequencies rather than simply the most frequent ones, or by including the number of graphics or tables in the page. We will explore how to fully utilize the content together with other kinds of features in future work. For now, let us inspect the learned models to understand what page features are the most useful for dwell time prediction.

5.3 Feature Importance

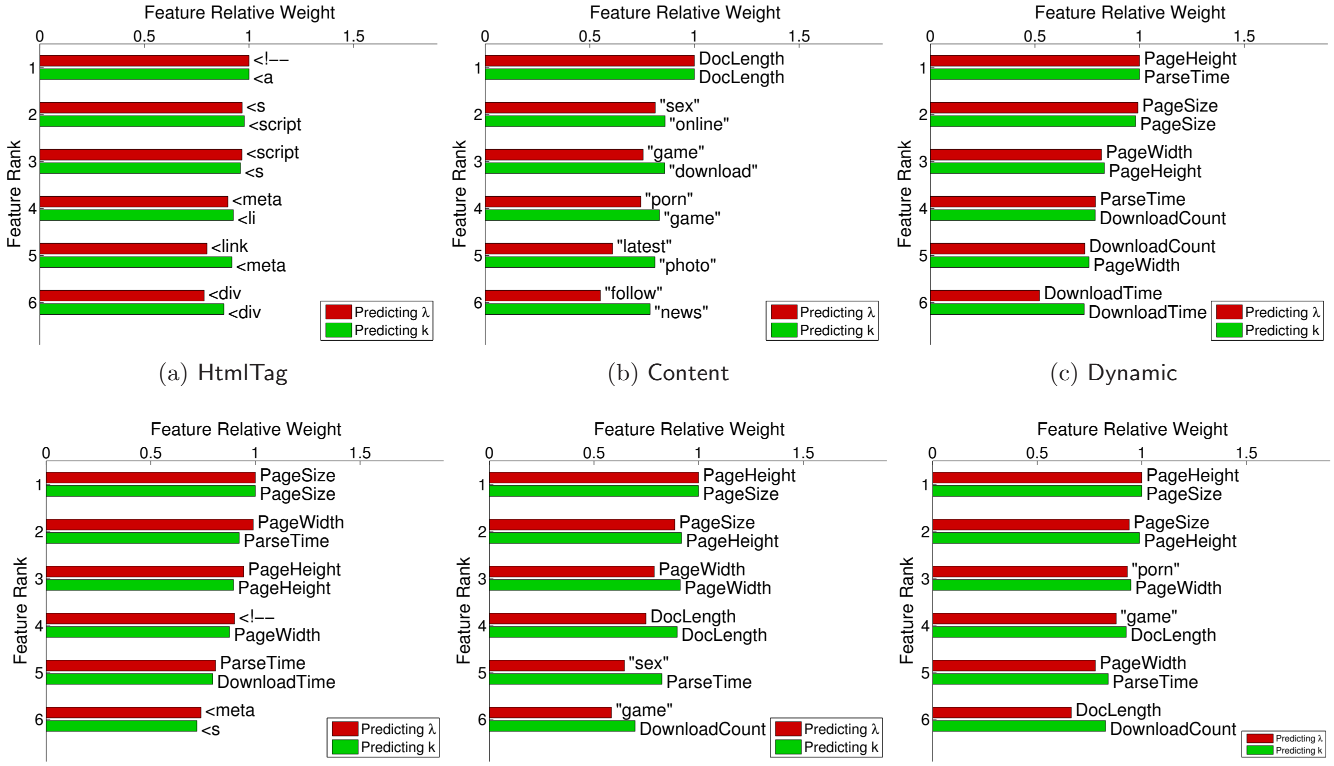
By virtue of the interpretability of MART, we could estimate the importance of each feature and sort them in descending order of the estimated importance. Figure 5 depicts the six most important features for predicting λ and k respectively under each configuration. The figure for "HtmlTag+Content" is dropped due to space constraints.

Figure 5(a) shows that Html tags about scripts and links are the most important ("`<!--`" is for comments in Html). In Figure 5(b), it is unsurprising to see that the document length is the most relevant feature, followed by words related to pornography, games and news. This looks reasonable as the dwell times for those topics are likely very different, since

[3] `http://wordlist.sourceforge.net/`

Features	Training & Validation						Test by Log-Likelihood		Test by KS-Distance	
	$\lambda(L)$	$\lambda(v)$	$\lambda(m)$	$k(L)$	$k(v)$	$k(m)$	Predict Win	Baseline Win	Predict Win	Baseline Win
HtmlTag	25	2^{-6}	113	25	2^{-4}	244	654	301	684	271
Content	25	2^{-5}	67	25	2^{-3}	159	702	253	727	228
Dynamic	25	2^{-6}	124	6	2^{-2}	186	653	302	685	270
HtmlTag+Content	25	2^{-6}	126	21	2^{-3}	199	706	249	724	231
HtmlTag+Dynamic	25	2^{-5}	65	25	2^{-3}	120	669	286	701	254
Content+Dynamic	25	2^{-6}	123	25	2^{-4}	195	724	231	727	228
HtmlTag+Content+Dynamic	25	2^{-6}	133	21	2^{-4}	198	717	238	725	230

Table 3: Prediction Efficacy with Different Feature Configurations



(a) HtmlTag

(b) Content

(c) Dynamic

(d) HtmlTag+Dynamic

(e) Content+Dynamic

(f) HtmlTag+Content+Dynamic

Figure 5: Feature Importance in Different Feature Configurations

users may interact with pages on these topics in different ways. Similarly, in Figure 5(c) we see that the height of the rendered page is the top feature for λ, followed by the page size and width. Interestingly, the time to parse the page is the most relevant feature for predicting k in Figure 5(c). This may suggest when parsing takes a comparably long time, the page will have a lower chance to survive users' screening. Finally, for the remaining three figures involving feature combinations, we see that Dynamic, although only comprising nine features, always appears near the top. This confirms our belief that the nine dynamic features are strong predictors; but because of the limited number of features, complementary support from Content is necessary for the best performance.

6. DISCUSSION AND FUTURE WORK

This paper presents the first step in Weibull analysis of Web page dwell time data, which can be extended in both breadth and depth. In breadth, there are many characteristic functions/quantities for Weibull analysis besides the hazard function, *e.g.*, the cumulative hazard rate function and the mean residual life, each of which has a natural correspondence to interesting aspects of Web browsing. In depth, it is interesting to investigate how sophisticated models (*e.g.*, mixture of Weibulls) would bring better goodness-of-fit and more insights into understanding user browsing behaviors.

The predictive models as presented here demonstrate the possibility of predicting Web page dwell time distributions. While better feature engineering and algorithm improvements would likely further improve performance, the current approach has an inherent shortcoming: it predicts λ and k separately whereas it is their combination that determines the goodness-of-fit of the predicted model. So instead of predicting λ and k separately, a more principled approach could be to optimize the likelihood directly, which would Likely provide a much better goodness-of-fit.

The Weibull analysis in this paper reveals some implications for understanding the browsing behaviors of all users. Alternatively, the user population can be partitioned along explicit dimensions such as time-of-day and geographical locations or implicit dimensions such as user intent and domain expertise estimates. For the latter, we can partition the dwell time based on how users reach the page, *e.g.*, through a search clickthrough, an advertisement clickthrough, or a link from a general Web page. In this way, we would gain more detailed understanding of user browsing dwell time in different scenarios.

7. CONCLUSION

This paper has drawn an analogy between abandoning a browsed page and the failure of a system, and presented the first Weibull analysis on Web page dwell time data. We found that general Web surfing exhibits a significant “negative aging” phenomenon, suggesting that users adopt a “screen-and-glean” browsing behavior where they vet the page prior to more detailed examination. This study brings a new approach to analyzing implicit feedback involving dwell time, complementing previously conducted user studies in that area. We have proposed some directions for building more sophisticated dwell time models and presented some implications for understanding user browsing behavior. Future work will build on our application of Weibull analysis, as well as the numerous successes of it in other application domains, to improve search and advertising.

8. ACKNOWLEDGEMENTS

The authors would like to thank Yutaka Suzue for the dynamic crawler, Krysta Svore and Qiang Wu for the MART implementation, Wen-tau Yih and Alice Zheng for the discussion on analyzing dwell time, and Xiaoxin Yin for the proofreading and helpful comments. The authors also appreciate the anonymous reviewers who not only offered us detailed review comments but also insightful suggestions on future work.

9. REFERENCES

[1] R. Abernethy. *The New Weibull Handbook*. fifth edition, 2006.

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

[3] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR*, pages 3–10, 2006.

[4] J. Attenberg, S. Pandey, and T. Suel. Modeling and predicting user behavior in sponsored search. In *KDD*, pages 1067–1076, 2009.

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

[6] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *IUI*, pages 33–40, 2001.

[7] A. C. Cohen. Maximum likelihood estimation in the weibull distribution based on complete and on censored samples. *Technometrics*, 7(4):579–588, 1965.

[8] W. J. Conover. *Practical Nonparametric Statistics*. Wiley, third edition, 1998.

[9] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers’ queries and information goals. In *CIKM*, pages 449–458, 2008.

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

[11] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:579–588, 1999.

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

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transaction on Information System*, 25(2):7, 2007.

[14] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *SIGIR*, pages 408–409, 2001.

[15] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *SIGIR’04*, pages 377–384, 2004.

[16] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *JCDL*, pages 74–75, 2002.

[17] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[18] E. Lehman. Shapes, moments and estimators of the weibull distribution. *IEEE Transactions on Reliability*, 12:32–38, 1963.

[19] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. BrowseRank: letting web users vote for page importance. In *SIGIR*, pages 451–458, 2008.

[20] G. Marchionini and B. Shneiderman. Finding facts vs. browsing knowledge in hypertext systems. *Computer*, 21(1):70–80, 1988.

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

[22] D. Nichols. Implicit ratings and filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, pages 31–36, 1997.

[23] H. Rinne. *The Weibull Distribution: A Handbook*. Chapman & Hall, first edition, 2008.

[24] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI*, pages 415–422, 2004.

[25] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *WWW*, pages 21–30, 2007.

[26] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *CIKM*, pages 87–96, 2009.