Measuring Inter-Site Engagement

Elad Yom-Tov*, Mounia Lalmas†, Ricardo Baeza-Yates†, Georges Dupret†, Janette Lehmann† and Pinar Donmez‡
**Microsoft Research*
†*Yahoo! Labs*
‡*Banjo Inc.*

***Abstract*—Many large online providers offer a variety of content sites (e.g. news, sport, e-commerce). These providers endeavor to keep users accessing and interacting with their sites, that is to *engage* users by spending time using their sites and to return regularly to them. They do so by serving users the most relevant content in an attractive and enticing manner. Due to their highly varied content, each site is usually studied and optimized separately. However, these online providers aim not only to engage users with individual sites, but across all sites in their network. In these cases, site engagement should be examined not only within individual sites, but also across the entire content provider network. This paper investigates *inter-site engagement*, that is, site engagement within a network of sites, by defining a global measure of engagement that captures the effect sites have on the engagement on other sites. As an application, we look at the effect of web page layout and structure, which we refer to as *web page stylistics*, on inter-site engagement on Yahoo! properties. Through the analysis of 50 popular Yahoo! sites and a sample of 265,000 users and 19.4M online sessions, we demonstrate that the stylistic components of a web page on a site can be used to predict inter-site engagement across the Yahoo! network of sites. Inter-site engagement is a new big data problem as overall it implies analyzing dozen of sites visited by hundreds of millions of people generating billions of sessions.**

***Keywords*-downstream engagement; inter-site engagement; predicting engagement; page layout and structure; stylistics;**

I. Introduction

User engagement [1], which has been addressed as the *the emotional, cognitive and behavioral connection that exists between a user and a resource*, is the result of trustworthy, quality, relevant and interesting content. In this paper, we evaluate one type of engagement, "stickiness" also referred to as *site engagement*, which is concerned about users "spending time" on a site. Certainly, the success of a site largely depends on itself, but also on how it is connected on the web and how web traffic arrives to it. This is particularly important for large online providers, such as AOL, MSN and Yahoo!, which offer a variety of content sites (e.g. news, sport, e-commerce).

Due to their highly varied content, each site is usually studied and optimized separately, for example, by serving users the most relevant content in an attractive and enticing manner, in particular what relates to the layout and structure of the content. However, these large online providers aim not only to engage users with each individual site, but across all sites in their network, as sites can (and do) link to each other. For example, if a site does not have any links on its pages to other sites of the same provider, users will find it difficult to navigate to them, creating an engagement barrier. Conversely, linking to relevant content of the same provider will improve engagement.

On the other hand, users spend more and more of their online sessions multi-tasking, e.g. emailing, reading news, accessing a social network and generally navigating between sites. Online multi-tasking has implications when looking at the network of sites offered by online providers, as several of the provider sites can be accessed during a single session. Therefore, site engagement should be examined not only within individual sites, but also across sites, that is, the entire content provider network. We refer to the site engagement with a network of sites as *inter-site engagement*.

This paper investigates inter-site engagement, by defining a global measure of engagement that captures the effect sites have on the engagement on other sites within the same online browsing session. Intuitively, our global measure, which we name *downstream engagement*, measures the fraction of time users spent on a content provider's sites without leaving out of the entire online session time. Our study makes use of 50 Yahoo! sites and a sample of 265,000 users and 19.4M online sessions taken during one month of interaction data. Yahoo! operates a multitude of sites, including sites on such diverse topics as finance, sports, celebrities, and shopping, and as such provides the perfect forum for our study of inter-site engagement.

As an application, we look at the effect of page layout (how the page is displayed) and page structure (the links and their positions), which we refer to as *web page stylistics*, on inter-site engagement, as these are commonly used to influence site engagement in the online industry. Anecdotally, these stylistic properties indeed change over time: Figure 1 shows the changes in the number of links (structure) and the number of list items (layout) on the pages of three Yahoo! websites over time. As the figure shows, these changes are significant and uncorrelated with the time of day. This observation motivated the outline of our study. The main contributions of this paper are the following:

1) To the best of our knowledge, we introduce a new big (usage) data problem in the Web, as this is the first large-scale study that investigates the interactions between the different sites of a content provider and

gives insight on inter-site engagement.

2) We define a global measure of inter-site engagement, which we call *downstream engagement*, that measures the effect that sites have on engagement in sibling sites.
3) We show that stylistic attributes, i.e., the elements of a web page, such as the links, tables, of each site can be used to predict and influence inter-site engagement. This shows that web interconnections matter more than what some people may expect.

This paper is organised as follows. Section II provides background information and positions our work. Our measure of inter-site engagement is described in Section III. In Section IV we describe the data used in this work. Our results are reported in three parts. First, we discuss the outcomes related to downstream engagement in Section V. We build on this to understand how to influence downstream engagement in Section VI. Third, we relate our findings to another measure of engagement in Section VII. We end with conclusions and thoughts for future work.

II. Background

In the online industry, site engagement or "stickiness" is mostly measured through behaviour measures aiming at assessing users' depth of interaction with a site. Widely-used measures include click-through rates, number of page views, time spent on a site (dwell time), how often users return to a site and number of users per month. Dwell time has proven to be a meaningful and robust measure of site engagement over the years; for example in the context of web search [10], where it is used to improve retrieval [2], [5]. Several white papers and reports contain studies on existing engagement measures and their usage [9], and proposals for a uniform measure of engagement based on several metrics [15].

Several works investigated how users access the web and how they navigate between sites or web pages. From these and other studies, several user navigation models were developed [16], [7], [14], for example accounting for the usage of bookmarks, back buttons and teleportation. These models, based on formalisms such as branching processes, aimed to understand how users access sites and pages within them, and its effect on, for instance, site popularity [16], [14], and loyalty to a site [8], [4], but not the effect of the stylistics of a web page to the engagement of further web pages or sites.

Online behavior measures have been used for many years by the web-analytics community and Internet marketing research companies (e.g. comScore). Because they are scalable to millions of users, they are commonly employed as a *proxy* for site engagement: the higher and the more frequent the usage, the more engaged the user. Although these measures cannot explicitly explain why users engage with a site, the fact that, for example, two million users choose to access a site daily is a strong indication of a high engagement with that site. Furthermore, by varying specific aspects of the site, e.g. structure and layout, and assessing the effect on online behavior, these measures can provide implicit understanding on why users engage with the site. Our work extends online behavior measures with a measure defined to capture inter-site engagement.

How long users spend on a provider set of sites (a provider network) from a given site is how we propose to measure inter-site engagement. We therefore study the effect of each site in promoting engagement on Yahoo! network of sites. Previous work looking at aesthetics, accessibility and engagement based on dwell time include [18] who showed that layout and textual features affect dwell time; and [13] who showed that a combination of content and dynamic features (e.g. page size or time to download all URLs) had also an effect on page dwell time. Following this line of work, this paper attempts to relate stylistic elements (e.g. layout and structure) of a web page, more particularly the main page of a site, and inter-site engagement.

III. Definitions

A site is an entity made of web pages put together to form a service. In the context of Yahoo! these include sites like Yahoo! News, Yahoo! Sports and Yahoo! Mail. Other examples include Facebook (chat, apps) or Google (search, Gmail, scholar). We define a session as all the pages visited by a user within 30 minutes or less from the first interaction in the session. This definition captures over 95% of session boundaries as defined in [11]. A *provider* session (in our case a Yahoo! session) corresponds to all contiguous pages of the provider sites visited within a session. Thus a session is composed of one or more provider sessions.

There are several ways we could measure inter-site engagement. In this paper, we propose the following measure of inter-site engagement: *the total time spent on a contiguous sequence of provider sites from the next site until the end of the provider session, divided by the total remaining session time*. We refer to this measure as *downstream engagement*. By definition, if the next site in the session is not a provider site, the downstream engagement is zero. Intuitively, downstream engagement measures the fraction of time users spent on a content provider's sites, without leaving, out of the entire (remaining) time they had available to spend online (the total remaining session time).

Formally, we index the sequence of sites S visited during a session by $i \in 1, 2, \ldots, n$. To compute the downstream engagement of site S_i, we introduce a binary indicator $\mathbb{1}_j$ for $j > i$ that is 1 if site S_j belongs to Yahoo! *and* if the user did not visit any site not belonging to Yahoo! between his or her visit from sites S_i to S_j. It is 0 otherwise. Let $t(S_i)$ be the time spent (also referred as dwell time) on site S_i. The downstream engagement of site S_i denoted $E(S_i)$ for

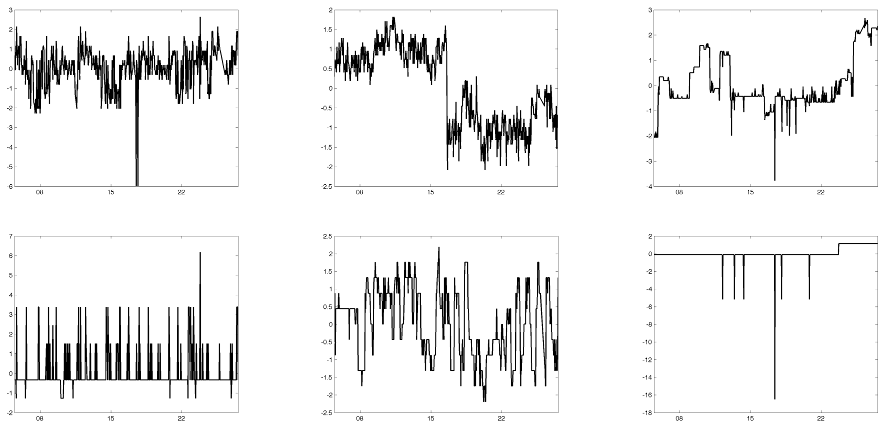

Figure 1. Change in the number of links (top) and the number of list items (bottom) on the pages of three Yahoo! websites over time. Values are normalised to zero mean and unit variance. Each sample represents the value at a particular hour and date. From left to right: a country main page, a news website, and a women-interests site.

sessions that contain S_i is:

$$E(S_i) = \frac{\sum_{j>i} t(S_j)\mathbb{1}_j}{\sum_{j>i} t(S_j)} \qquad (1)$$

Consider a session where a user visits 7 sites out of which only the second, the third, the fourth and the seventh sites belong to the provider of interest; for the Yahoo! session starting at site number 2, the indicator vector is given by:

$$(\mathbb{1}_1, \mathbb{1}_2, \mathbb{1}_3, \mathbb{1}_4, \mathbb{1}_5, \mathbb{1}_6, \mathbb{1}_7) = (0, 1, 1, 1, 0, 0, 0). \qquad (2)$$

Note that $\mathbb{1}_7 = 0$ because sites that did not belong to the provider were visited between S_2 and S_7. Then, the downstream engagement for S_2 is

$$E(S_2) = \frac{t(S_3) + t(S_4)}{t(S_3) + t(S_4) + t(S_5) + t(S_6) + t(S_7)} \qquad (3)$$

The provider session for S_2 contains S_3 and S_4. Downstream engagement of sites S_1 and S_i with $i \geq 4$ are by definition 0.

Our measure was defined specifically to capture inter-site engagement. Several other definitions were also attempted, such as the fraction of time users spent on the provider network out of the total session time. Since these two definitions, and others, were found to lead to highly correlated measures of inter-site engagement, we report our investigation with the measure defined as above, and leave the investigation of alternative measures of inter-site engagement for future work.

We also report results using *dwell time*, a common measure of engagement, allowing us to position our proposed measure. Dwell time is the total time a user spends visiting contiguous pages within a specific site.

IV. Data

Our study required collecting both user behavior data for computing the three measures described in the previous section, and site data, for analyzing page stylistics (page layout and structure).

To collect site data, we extracted the main page of 50 popular Yahoo! sites (in terms of page views) once every hour during the month of May 2011. At the time of writing, relatively few major internet content providers have more than 50 different content sites (some well-known ones are yahoo.com, google.com, msn.com and wikipedia.org), thus 50 was a number high enough to carry out our study. Although we collected the main page of these sites, they are representative of most other pages of the site because they use similar stylistic guidelines. For reason of confidentiality we denote each site by one of the following six categories: country main page (e.g., uk.yahoo.com), e-commerce (e.g., auctions.yahoo.com), informational (e.g., weather.yahoo.com), leisure (e.g., games.yahoo.com), news (e.g., news.yahoo.com), specific interests (e.g., omg.yahoo.com) or sports (e.g., eurosport.yahoo.com).

The site data was used to generate two types of attributes as described in Table I, for each site at each time and date. The attributes at the top were general site parameters, including site ID, time of day, date, etc. The attributes at the bottom determine the stylistics of a page, which are themselves of two types, layout and structure. Layout refers to how the web page is displayed, its images, tables, etc., whereas the structure is concerned with the links and their positions on the web page. The top set of attributes were included to study temporal and site-specific effects on downstream engagement, as opposed to stylistic effects; for example, performing some typical online activities at specific times of the day and on specific days of the week.

User data was collected from a sample of users who gave their consent to provide browsing data through the Yahoo! toolbar. A total of 19.4M sessions were recorded from approximately 265,000 users. These users represent a

Table I
ATTRIBUTES ASSOCIATED WITH SITE PAGES – GENERAL AND STYLISTIC (STRUCTURE AND LAYOUT) SEPARATED.

General attributes
+ Site ID;
+ Time of day (GMT);
+ Is the day of the week a workday or weekend;
Structure attributes
+ Number of links, partitioned by: – Location (title or body of the HTML) – Text (e.g., an HTML page), pictures (e.g., specific images), or video (e.g., an embedded video); – Linking to the same site, to other Yahoo! sites, or to non-Yahoo! sites;
+ Average rank of links in the page, i.e. whether the links located mostly on the top of the page, or near the bottom; + Unique links;
Layout attributes
+ Number of script tags, i.e. used to define a client-side script, such as a JavaScript; + Number of all HTML classes (one feature per class).

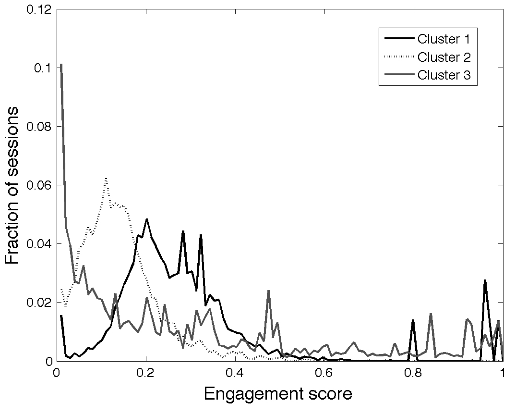

Figure 3. Average engagement scores distributions per cluster. Cluster 1 consists of 9 members, cluster 2 of 28 members, and cluster 3 of 13 members.

sample of the Yahoo! user base, who spend a certain fraction of their online time on Yahoo! sites, and our goal is to study inter-site engagement on the Yahoo! network. To protect user privacy, no user identifiers were recorded, and only the top level domain was used in our analysis. We stripped the URLs to the last 3 components (for example mail.yahoo.com). This was also done to reduce the effect of sub-domains containing similar content (e.g., health in Yahoo! Lifestyle). However, for the 50 sites we analyzed, the large majority of browsing events happened on the main third-level domain, and not in its sub-domains. The user data was used to measure average dwell time and average downstream engagement for each of the sites at each time and date. Note that the last page in each session cannot be taken into account because we cannot record the time at which a user leaves the site.

The user and main page datasets were joined by site, date and time, such that for each site and each date and time combination there was a measure of the average downstream engagement and dwell time, as well as a vector of corresponding style attributes collected around the time that site engagement was measured.

V. DOWNSTREAM ENGAGEMENT

We study now how site characteristics (layout, structure and general) correlate with (and are predictive of) inter-site engagement. First, we present some general statistics of downstream engagement for the 50 sites under study. We then study the effect of the general, link and layout attributes on downstream engagement, also focusing on three case studies. We end with a number of further insights.

A. Variations in Downstream Engagement

We first report the distribution of the downstream engagement scores – "engagement scores" for short. Figure 2 shows the engagement scores distribution (average: 0.18, standard deviation: 0.14). Most scores lie around 0.2. This may be due to those sessions corresponding to sites that people check on a regular basis during a typical online session. Indeed, the sites corresponding to these sessions are visited, on average, an order of magnitude more frequently than the average, and as a result the time spent on them will tend to be short.

We also report the average engagement score of the 50 sites under study in Figure 2 (left). As the figure shows, there is significant variance in engagement over the period we tested, and some sites have a much stronger average downstream engagement than others. The sites with the highest average engagement are leisure and sport sites (sites to which users come to spend time on, e.g., to play or read), whereas those with the lowest average engagement are e-commerce and informational sites (sites to which users come for specific purposes, e.g., to purchase or check something). Overall, the sites with the highest scores can often be associated with browsing sessions, whereas those with the lowest scores can be considered as goal-oriented sessions. We return to this in Section VI.

To identify types (if any) of engagement we clustered the sites accordingly to their engagement profiles. The distribution of engagement scores per site was binned according to quantiles, and the resulting vectors of quantiles used to characterize the engagement on a given site. Using the k-means algorithm and the L2 norm for distances, we identified three clusters, which was the maximal number that resulted in non-empty sets. We did not find a priory characteristics of sites which would suggest their cluster membership, i.e., not all country-specific main pages were assigned to the same cluster, indicating that each of these sites and their respective audiences has its own characteristics.

The average distribution of each of the clusters is shown in Figure 3. As the figure shows, two of the clusters have a lognormal-like distribution, differentiated by the average engagement score. The last cluster has an

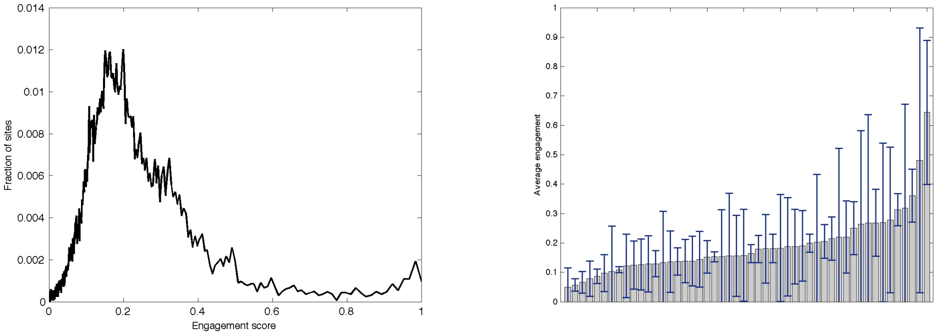

Figure 2. Distribution of downstream engagement scores (left) and the distribution of average downstream engagement scores (right). Each bar represents a single site. The error bars show one standard deviation across time.

exponential-like distribution with most engagement scores close to zero. Thus, different sites are characterized by typical engagement patterns, with some sites exhibiting a wide range of engagement ranges, while others are more likely to generate little or no downstream engagement. In Section V-C, we focus on three case studies, one per cluster, and discuss how different stylistic and other elements are correlated with each of these engagement styles.

B. *Stylistic Attributes and Engagement*

It is known that the style of a page is correlated with dwell time. In this section we test whether style attributes are correlated with engagement, by attempting to predict downstream engagement (i.e., inter-site engagement) given the attributes of a page. We used decision trees for classification, and 10-fold cross-validation to reduce the chance of an over-fit, for this purpose. We trained site-specific classifiers by setting the threshold for significant engagement as one standard deviation above the average engagement for each site separately. As discussed in Section IV, we include non-stylistic attributes in our prediction, to compare their effect to that of stylistic page attributes on downstream engagement. The average performance is reported.

The ten sites with the highest AUC obtained an average AUC of 0.67 (min: 0.60, max: 0.80), and an average f-measure of 0.41 (min: 0.31, max: 0.54). The ten sites include sport, special interest, and e-commerce sites, and are distributed across the three set of clusters identified in Section V-A. We noted that some sites do not lend themselves to prediction of engagement. These includes several of the informational, leisure and country main pages. Using more sophisticated classifiers, or looking at the Yahoo! sites offered by each country[1] may improve this result.

Examining the top ten sites by popularity, the five most frequently chosen features (in descending order of frequency) were:

1) Time of day
2) Number of (non-image and non-video) links to Yahoo! sites in the body of the HTML
3) Average rank of Yahoo! links on the page
4) Number of (non-image and non-video) links to non-Yahoo! sites in the body of the HTML
5) Number of span tags, which are tags that allow adding style to content or manipulating content, for example using JavaScript.

Obviously, the time of the day is highly influential. People spend longer time online during particular times of the day, thus increasing downstream engagement during those times, and this independently of the site in question.

The above list also suggests that link placements (as given by the average position in the ranking of Yahoo! links), as well as adding more Yahoo! links can influence downstream engagement. Link placement and number of links have already been used as a means to keep users engaged within a site. It is interesting that these seem to act similarly with respect to site engagement across sites, where different services are offered (e.g., various types of content).

Interestingly, links to non-Yahoo! sites have a positive effect on downstream engagement. Note that content providers usually make it clear that such links will bring users to external sites. Our hypothesis is that users see a richer page when such links abound, and, when faced with such abundance, decide to focus their attention on a central repository of information, rather than visiting a multitude of external sites. This hypothesis, however, requires further investigation through user-centred studies.

C. *Case Studies*

We focus on three case studies to gain further insights on the effect of page stylistics on downstream engagement. We investigated three sites, randomly selected from one for each of the clusters identified in Section V-A (Figure 3).

[1] A internal study of 800 Yahoo! sites worldwide showed that traffic between sites varies per country, with no particular patterns.

Table II
PERFORMANCE OF PER-SITE CLASSIFIERS FOR THREE SPECIFIC SITES.

Site	AUC	f-measure
e-commerce	0.63	0.38
news	0.65	0.37
women-interests	0.72	0.43

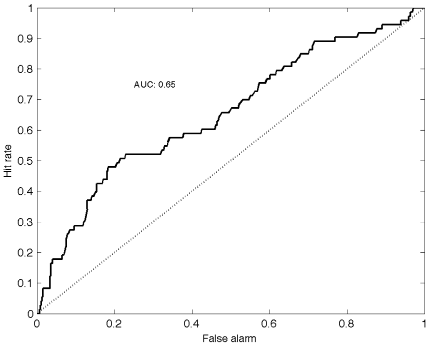

Figure 4. Precision-Recall graph for downstream engagement in a news website.

These are an e-commerce site, a news site, and a women-interests site. The average engagement for those sites was 0.26 (±0.31), 0.15 (±0.02), and 0.21 (±0.06), respectively.

We used a linear classifier because it gave better performance than the decision trees. The AUC and f-measures obtained are shown in Table II. Note that the accuracy of the classifiers is lower than that obtained for the entire set of sites, which is likely due to the sparsity of the training data, compared to using the entire data set from all sites. Nonetheless, interesting insights were derived.

The ROC for the news site is shown in Figure 4. We can identify 50% of the samples above the threshold at a cost of approximately 20% of false detections. This suggests that the best features in the classifier could be useful for improving downstream engagement, and that most instances of high engagement can be correctly inferred at a relatively low false-detection rate.

Using sequential forward feature selection, we found that the best features for each classifier were as shown in Table III. There are mainly stylistic attributes in this list. First, stylistic attributes such as links to non-Yahoo! sites feature prominently for both the news and women-interests sites, but not for the e-commerce site. This may be due to the different nature of these sites, where in the former users expect to see more external content. A low average rank of links to Yahoo! sites, i.e., if they are displayed towards the top of the page, is important for increasing downstream engagement in the e-commerce and women-interests sites, possibly because structurally, the news changes less and thus this attribute has a smaller effect. Additional links to video content on the news site increase engagement, while external image links have this effect on the women-interests site.

Non-stylistic attributes include time of day and day of the week for the e-commerce site, which are related to the way people interact with the site (specific time of the day as well as specific day of the week). Note that for the women-interest site, no non-stylistic attributes were prominent in predicting downstream engagement.

Here we have shown, through three case studies, that the style elements of a site can predict downstream engagement. Some of our findings (in terms of the influential features) are less intuitive than others, nonetheless they provide interesting insights to the effect of a site on downstream engagement. In addition, using a larger set of data (here we are using one month only), is likely to provide further as well as more significant insights.

D. Textual content versus style

This paper focuses on stylistic attributes of websites (their main pages). We now study the effect, if any, of the actual content of the sites on downstream engagement, as compared to stylistic attributes. That is, we explore the actual page content. We extracted the text from each crawled page and modeled it in two ways. First, we used a vector-space model, where each word is an attribute. Second, since raw text may be too sparse, we also modeled the text by counting the number of entities of each type that appeared in the text, as identified using the Stanford Named Entity extractor.[2]

We measured the distances between two adjacent instances of the same site using the cosine similarity. For comparison purposes, we did the same for the stylistic (layout and structure) features extracted as described above. We found that in 86% of cases, the distance between the adjacent texts was greater than that of the stylistic attributes. Thus, textual content changes faster than stylistics, and may require faster sampling to enable downstream engagement prediction. Next, we trained classifiers using the same procedures used for the stylistic attributes, as described above. Both text representation methods obtained a worse prediction accuracy: 2% lower in AUC for the word-based representation (not statistically significant) and 7% lower for the entity-based representation ($p < 0.01$, sign-test). Moreover, using combinations of stylistics and text did not improve the classification performance.

There could be several reasons for this lower performance. First, as noted above, text changes faster than stylistic attributes, so the predictor may be suffering from a too low sampling rate. Second, the text is sparser than stylistic attributes and is therefore harder to generalize from. However, given the poor results of the entity-based representation, this seems to be a weaker explanation. Finally, text may be inherently a worse predictor of downstream engagement, because it represents a poorer cue for users to browse other

[2] http://nlp.stanford.edu/software/CRF-NER.shtml

Table III
MOST INFLUENTIAL ATTRIBUTES FOR PREDICTING ENGAGEMENT IN THREE SPECIFIC SITES (SIGNS IN PARENTHESIS INDICATE POSITIVE OR NEGATIVE SPEARMAN CORRELATION WITH DOWNSTREAM ENGAGEMENT).

Site type	Best attributes
e-commerce	1. Time of day (+) 2. Weekend (-) 3. Number of unique Yahoo! links (-) 4. Average rank of Yahoo! links on page (-) 5. Number of paragraph tags (+)
news	1. Number of unique Yahoo! links (-) 2. Number of (non-image and non-video) links to Yahoo! sites in the body of the HTML (+) 3. Number of (non-image and non-video) links to non-Yahoo! sites in the body of the HTML (+) 4. Number of video links within the page (+) 5. Time of day (+)
women-interest	1. Number of image links to non-Yahoo! sites in the body of the HTML (-) 2. Number of table elements (-) 3. Average rank of Yahoo! links on page (-) 4. Number of (non-image and non-video) links to non-Yahoo! sites in the body of the HTML (+) 5. Number of Java scripts on the page (-)

sites, i.e., a link on Yahoo! News to Yahoo! Finance is likely a better engagement 'incentive' than a text (a news article) about a financial story.

VI. INFLUENCING ENGAGEMENT

Some sessions may be more sensitive to changes aimed at increasing downstream engagement level than others. For example, a user who accesses a site to check for emails may be less influenced by additional browsing opportunities from posted links, than a user who is browsing for leisure. We call the first kind of sessions goal-specific, and distinguish between goal-specific sessions and all other sessions in the following manner. We use the following procedure to distinguish between the two types of sessions. The objective is not to identify a session type as it develops, or whether a user belongs to one category rather than another (we leave this for future work), but to identify the sites where stylistic changes are more likely to have an influence on engagement.

For each user, the five most common sites he or she visited during the data collection period were identified. For information, in our data set, the most frequent common sites were related to mail, search and social networks. We hypothesis that sessions in which 50% or more of the visited sites belonged to the five most common sites (for that user) were classified as goal-specific. Goal-specific sessions accounted for 38% of the sessions. Approximately 92% of users had sessions with both kinds of sites. The average downstream engagement in goal-specific sessions was 0.16. This is to be contrasted with a higher average downstream engagement of 0.2 during other sessions.

If engagement of non goal-specific sessions is more influenced by stylistic changes, it should be easier to predict using style attributes. Following the procedure described in Section V-B, we were able to predict engagement with an AUC of 0.76 for goal-specific sessions, and an AUC of 0.81 for other sessions. This implies that when users do not have specific goals in mind, they may be more ready to accept suggestions (more links, more images, etc.) for additional browsing, and we can influence their engagement in this way. The fact that there is an effect of task type on engagement measures (such as dwell time) is in itself not new (see early work in [12]). Here we are re-confirming this fact, and show that downstream engagement patterns are also influenced by some given task types.

Additional work is required to identify more fine-grained session classes (besides goal-specific versus others here and hypothesized here) and corresponding task types (e.g. entertainment versus fact-finding). The user browsing history and external factors like the time of the day or the day of the week are likely to be useful. Also, using mouse tracking recording such as in [6] (applied to search) may help identifying finer-grained online tasks.

VII. RELATIONSHIPS BETWEEN ENGAGEMENT AND DWELL TIME

Dwell time is the time spent contiguously on a site, and is a popular measure of site engagement. It is often used as one of the official, standard measures by the web analytics community and Internet market research companies such as comScore. The distribution of dwell time for the Yahoo! 50 sites under study in this paper, shown in Figure VII (left), follows approximately a lognormal distribution. Dwell time per site, shown in Figure VII (right), varies much from site to site, ranging from less than 5 seconds to more than 15 minutes. The sites with the shortest dwell time are e-commerce and sports sites. The highest average dwell time is recorded on leisure and home page country sites.

Interestingly, the Spearman correlation between dwell time and downstream engagement is small and negative: $\rho = -0.05$ ($p < 10^{-5}$), somewhat indicating that more engagement downstream is associated with a slight reduction in dwell time. Users who spend more time on some site, will thus tend to spend less time elsewhere during their session, and vice versa. Finally, we look at the relation between stylistic characteristics and dwell time using the same decision tree procedure as in Section V-B. The corresponding precision-recall and ROC graphs are shown in Figure 6. The prediction accuracies are similar to those of downstream engagement.

We conclude that stylistic attributes may be used to predict site dwell time as well as downstream engagement. The (extremely small) negative correlation between these two measures seems to indicate that the provider (here Yahoo!) can decide which of these measures it wishes to optimize, i.e., engagement with respect to a particular site, or/and across sites.[3]

This section shows that downstream engagement is clearly measuring something different, which, we argue, is inter-site engagement. Interestingly, style attributes are associated with each of these measures, in a comparable manner. However, they may differ across sites and measures, as demonstrated in [19], where the effect of links (to the same site, another Yahoo! site, or a site outside the Yahoo! network) was shown to differ depending on the site itself.

VIII. Conclusions and Future work

This paper is concerned with measuring inter-site engagement. This is a relevant big (usage) data problem as many of today's online content providers operate multiple sites, each optimized separately. It is even more relevant at the global web level, to understand how inter-site engagement is generated at large. To this end, we propose a measure, *downstream engagement*, which calculates the percentage of time spent within sites from the provider in a contiguous fashion (a provider session) from the total time spent online (total session time), for a given site. To characterize the site for which the downstream engagement is calculated, and as an application, we look at the web stylistic features of the site, e.g. the layout and the structure of the main page of a site, for 50 popular Yahoo! properties.

Our extensive experiments showed that the stylistic features associated with the main page of many of the sites investigated in this paper can be used to "predict" downstream engagement with good accuracy. Further work reported in [19] showed the impact and importance of links and their type in inter-site engagement, depending on the site. Our investigation of the impact of the content of a site, as opposed to its stylistic features, found that the former was not a good predictor of downstream engagement. Finally, we showed that downstream engagement is different from dwell time, a common measure of engagement.

Our analysis is retrospective in nature, which makes it harder to establish that downstream engagement can be predicted using stylistic attributes also implies that such attributes can influence engagement. However, several findings suggest that such a causal effect exists. First, previous studies have demonstrated this relationship for dwell time. Second, our findings demonstrate the predictive ability as a function of session type. Third, our analysis shows that changes in style over time, though uncorrelated with it, have an effect on downstream engagement. Thus, although careful randomized studies are required to positively prove a causal effect, our results suggest that this effect exists.

This work provides new insights into site engagement. Although, relating page style and engagement (e.g. dwell time) in itself is not novel [13], previous work did not look at the effect of the network of sites on site engagement. Our results open up a new line of research, that of measuring inter-site engagement. We are now looking at how these findings can be used by web-masters and editors, among others, to improve inter-site engagement, here at Yahoo!.

In future work we will study more sites, including those for which we obtained relatively poor results for predicting engagement from their stylistics, and use larger data sets (e.g. several months), as well as investigating more closely the effect of time and day. We will analyze in more detail the data that we have already collected to find out other interesting relations that can influence inter-site engagement. User studies are also planned to obtain a complementary (qualitative) understanding of user behavior in terms of inter-site engagement. The latter may lead to alternative definitions of downstream engagement, deemed more appropriate to typical or task-related user behavior on a site, e.g. star-like behavior.

Our ultimate aim is to understand how to improve the engagement with users across a network of sites. Studying inter-site engagement is important for companies offering a diverse range services, such as Yahoo! or AOL, but also for those with a more limited range of services (e.g. LinkedIn, Facebook), or with one main service with several instances of it (e.g. Amazon, Wikipedia). For example, Amazon does not just sell books, but also clothes, furniture, etc; each of these categories can be viewed as a specific site, and the same methodology can be used to predict downstream engagement across these categories. The Wikipedia project can also benefit from this work. Indeed, Wikipedia offers different functionalities, and each of them can be viewed as a site on its own. It will be important to investigate how our findings generalize to any of the above and extend the analysis to larger data samples.

References

[1] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski. Towards a science of user engagement (Position Paper) WSDM Workshop on User Modeling for Web Applications, 2011.

[3] We should emphasize that a small negative but significant correlation – as is the case here – when dealing with millions of user sessions is meaningful, as it can, for instance, impact revenues (e.g. obtained through advertising).

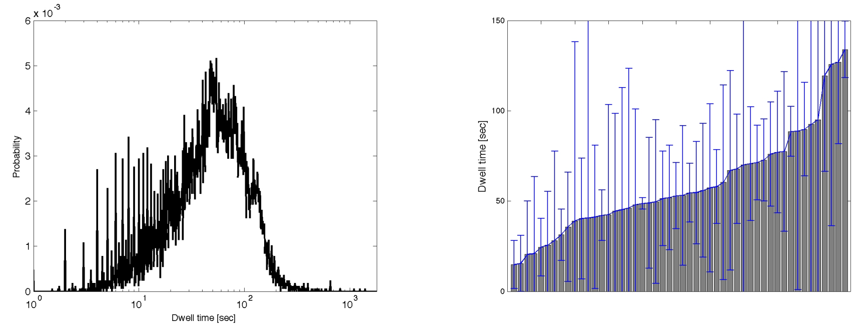

Figure 5. Dwell time distribution (left) and Average dwell time per site (right) for the top 50 Yahoo! sites (each bar represents a single site).

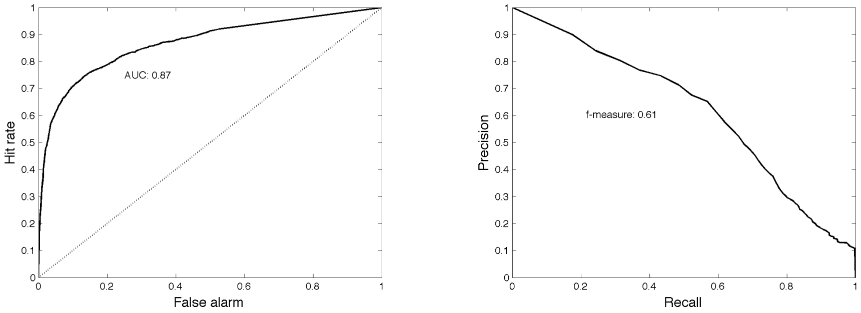

Figure 6. Receiver Operating Curve (left) and Precision-Recall curve (right) for dwell time

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

[3] Bandari, R., Asur, S., and Huberman, B. A. The pulse of news in social media: Forecasting popularity. In *ICWSM* (2012).

[4] Beauvisage, T. The dynamics of personal territories on the web. In *Hypertext* (2009).

[5] Bilenko, M., and White, R. W. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *WWW* (2008).

[6] Buscher, G., White, R. W., Dumais, S. T., and Huang, J. Large-scale analysis of individual and task differences in search result page examination strategies. In *WSDM* (2012).

[7] A. Chmiel, K. K., and Hołyst, J. A. Scaling of human behavior during portal browsing. *Phys. Rev. E 80* (2009).

[8] Cockburn, A., and Mckenzie, B. What do web users do? an empirical analysis of web use. *Journal of Human-Computer Studies* (2000).

[9] Forrester-Consulting. How engaged are your customers? Forrester Research, 2008.

[10] Jansen, B. J., and McNeese, M. D. Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *JASIST 56*, 14 (2005).

[11] Jones, R., and Klinkner, K. L. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKMt* (2008).

[12] Kelly, D., and Belkin, N. Display time as implicit feedback: understanding task effects. In *SIGIR* (2004).

[13] Liu, C., White, R. W., and Dumais, S. Understanding web browsing behaviors through Weibull analysis of dwell time. In *33rd international ACM SIGIR conference on Research and development in information retrieval* (2010).

[14] M. Meiss, B. Goncalves, J. R. A. F., and Menczer, F. Agents, bookmarks and clicks: A topical model of web navigation. In *Hypertext and hypermedia* (2010).

[15] Peterson, E. T., and Carrabis, J. Measuring the immeasurable: Visitor engagement. Tech. rep., WebAnalytics Demystified, 2008.

[16] Simkin, M. V., and Roychowdhury, V. P. A theory of web traffic. *Europhys. Lett. 82* (2008).

[17] Upton, G., and Cook, I. *Oxford dictionary of statistics*. OUP, 2002.

[18] Wu, O., Chen, Y., Li, B., and Hu, W. Evaluating the visual quality of web pages using a computational aesthetic approach. In *WSDM* (2011).

[19] E. Yom-Tov, M. Lalmas, G. Dupret, R. Baeza-Yates, P. Donmez and J. Lehmann. The Effect of Links on Networked User Engagement. In *WWW* (2012) (Poster).