Understanding Within-Content Engagement through Pattern Analysis of Mouse Gestures

Ioannis Arapakis
Yahoo Labs
Barcelona, Spain
arapakis@yahoo-inc.com

Mounia Lalmas
Yahoo Labs
London, UK
mounia@acm.org

George Valkanas
University of Athens
Athens, Greece
gvalk@di.uoa.gr

ABSTRACT

The availability of large volumes of interaction data and scalable data mining techniques have made possible to study the online behaviour for millions of Web users. Part of the efforts have focused on understanding how users interact and engage with web content. However, the measurement of within-content engagement remains a difficult and unsolved task. This is because of the lack of standardised, well-validated methods for measuring engagement, especially in an online context. To address this gap, we perform a controlled user study where we observe how users respond to online news in the presence or lack of interest. We collect mouse tracking data, which are known to correlate with visual attention, and examine how cursor behaviour can inform user engagement measures. The proposed method does not use any pre-determined concepts to characterise the cursor patterns. We, rather, follow an unsupervised approach and use a large set of features engineered from our data to extract the cursor patterns. Our findings support the connection between gaze and cursor behaviour but also, and more importantly, reveal other dependencies, such as the correlation between cursor activity and experienced affect. Finally, we demonstrate the value of our method by predicting the outcome of online news reading experiences.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors ; I.5.0 [**PATTERN RECOGNITION**]: General

Keywords

Mouse gestures; user engagement; online news; pattern recognition; prediction

1. INTRODUCTION

Central to most computer-mediated tasks and online activities is the ability to navigate through, and interact with, a digital environment. In most cases, this involves the use of a pointing device, such as mouse or trackball, that requires the execution of visually-guided movements (e.g., selecting, positioning, clicking). Usage of the mouse device can be thought of as consisting of a series of moves, aka *gestures*. Each such gesture is a specific and continuous physical process that is initiated and concluded by the user. Therefore, capturing and analysing cursor behaviour arises as a low-cost and scalable alternative, which can be easily deployed in an online setting. Unlike other tracking technologies, recording of the cursor position is simple to implement, can be performed in a non-invasive manner, and without removing the users from their natural setting.

Several works in this area have provided evidence indicating that the mouse cursor can act as a weak proxy of gaze [13, 18, 34] and offer an inexpensive alternative to eye tracking. The utility of mouse tracking has been demonstrated for a number of applications, such as understanding search result page examination [18, 19, 20], improving results ranking [7, 38], and performing relevance predictions [14, 16]. Although the importance of mouse tracking data in web search is now evident, very little is known about within-content engagement. An in-depth analysis on the interplay between web content and the quality of the user experience, based on reliable ground truth and validated engagement measures, has been missing so far.

The major caveat of existing techniques that study cursor behaviour is that they target very specific tasks. This limits to some extent the utility and applicability of these approaches to broader and more heterogeneous contexts. For example, a number of studies [4, 16] examined cursor interactions with search engine results pages (SERP) and reported cursor-related measures that are specific to particular areas of interest (AOI), e.g., the document rank position. Similarly, in [7, 15, 19, 20], the authors analysed cursor activity in the context of SERPs to understand and improve search. In these studies, the content layout may have introduced a bias to the cursor behaviour that is linked to the structured presentation of search results.

In this paper, we are interested in a more generalisable solution to measuring within-content engagement. Our work is motivated by the fact that millions of users interact with online content without providing any explicit feedback about the quality of their experience. Therefore, deducing in an online, implicit and scalable manner how they feel is considered a high-value task. Given this, our main objective is to understand how cursor behaviour can inform us about well known user engagement measures [29], such as affect, attention, and interest.

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661909.

To this end, we conduct a small-scale, controlled user study, and record the cursor data of users that interacted with interesting and non-interesting web content like online news. We obtain and segment the cursor movements based on an anatomical interpretation that goes beyond a simple mapping of *xy* coordinates. In particular, we consider cursor behaviour irrespective of the page layout, the elements it contains, or their relative position. Therefore, our method can be seen as a baseline regardless of the context of application. We then engineer a large set of features and use various clustering techniques, combined through a robust aggregation framework, in order to conduct our analysis. Through it, we identify how the frequency of such patterns changes over different degrees of engagement. We also derive a taxonomy of cursor patterns, which may generally characterize mouse behaviour. Finally, we demonstrate the value of our approach in a user engagement scenario, where we predict the outcome of online news reading experiences.

Overall, this paper makes the following contributions:

- We propose a scalable and non-invasive methodology for measuring within-content engagement, based on mouse gestures, that is applicable to different contexts.
- We deliver a taxonomy of mouse gestures using an unsupervised learning approach.
- We evaluate our methodology through a controlled user study, and identify tangible associations between certain types of mouse gestures and engagement metrics.
- As a validation step, and real-life use case, we consider the task of predicting within-content engagement (interestingness), solely on the basis of cursor interactions.

The rest of the paper is organized as follows. Section 2 presents related work. In section 3 we thoroughly describe our experimental design, followed by Section 4 where we present our analysis and main findings, which we integrate in a real application scenario, given in Section 5. In Section 6 we discuss our findings and their implications, followed by Section 7 which lays the ground for future work.

2. RELATED WORK

In the online industry, user engagement is defined as the amount of attention and time users are willing to spend on a website and how likely they are to return to it. Engagement is usually described as a combination of cognitive processes (in this paper we focus on affect, focused attention and interest) traditionally measured using surveys. User engagement is also measured at large-scale through analytic tools assessing users' depth of interaction with a website, which include metrics such as dwell time and clickthrough rate. It has been argued that "within-page activity could inform [...] about the quality of the content on a page", which could be measured by tracking mouse movements [9].

Previous work on mouse tracking has mainly focused on two problem domains: understanding web page usability and understanding user interactions with search engines. The latter problem domain has been addressed by studies that have investigated the eye-mouse alignment, prediction of eye gaze from mouse movements, and using mouse movements to estimate relevance and improve ranking.

One of the earliest studies in mouse and gaze alignment in search is from [34], which identified multiple patterns of eye-mouse coordination: the mouse following the eye in the X and Y axes, marking a result relevant, and remaining stationary while the eye inspected results. Further works showed that eye-mouse coordination [13] and gaze position [18] could be predicted to some extent by mouse movements, also in the context of non-linear layout SERPs [25]. Outside the search domain, [5] showed in a debugging task that mouse cursor behaviour was a significant indication of the level of difficulty and performance, and the observed cursor patterns were found comparable with the gaze patterns. These works demonstrate that gaze can be approximated with relative accuracy by mouse movements. Given that gaze has been shown to correlate with other engagement measures [1], mouse tracking becomes an attractive and scalable alternative to predicting user engagement. In this paper, we go beyond the prediction of eye-mouse coordination to inform directly about engagement measures.

Mouse tracking has been also used to understand web search behaviour (e.g., relevance, search intent, search success). For instance, hovering over a search result is highly indicative of relevance and can distinguish between good and bad abandonments [20]. In unsuccessful search sessions, mouse movements were shown to be more spread-out than in the successful ones and gravitated towards the lower part of the result page [15]. Accounting for pre- and post-click mouse movements led to substantial improvements in estimating search relevance and re-ranking search results [15, 19]. Mouse movements were also shown to be good predictors of general search intent [12], to be sensitive to the position and relevance of the search results [25], but also to estimate searcher attention on novel SERPs [7].

Within the same context, [21] proposed the use of frequent *motifs* (cursor movement patterns) extracted using the Dynamic Time Warping distance, which they heavily optimized to scale up. Their work focused on evaluating SERPs relevance, whereas ours aims to understand and measure user engagement. Our methodologies also differ, as we extract a large number of features and employ clustering, instead of relying on motifs obtained under a fixed-size sliding window and a single distance function. Motifs could be an additional feature in our case and and their contribution to predicting user engagement is left for future work.

Cursor movements as implicit indicators of interest on web pages has been also explored. For instance, in [35], the ratio of mouse movement to reading time was shown to be a better indicator of page quality than mouse travel distance and dwell time [35]. In the e-commerce domain, users' second choice could be determined as the link on which they hesitated longest before clicking their first choice [23]. This further validated using mouse movements to measure user engagement, and in particular user interest.

Finally, mouse tracking has been used to study other cognitive processes. For instance, when hand motion was tracked by mouse movement, slow and arched mouse trajectories were shown to indicate ambiguous state of mind during decision-making [11]. In the context of the Web, mouse movements have been shown to predict with reasonably high accuracy whether a user was distracted, frustrated or had an unpleasant experience [24]. All these suggest that mouse movements can reveal important cognitive processes across domains. However, recent work has shown that mouse movement give no or little indication of user attention during relevance assessment tasks [36], indicating that not all cognitive processes can be modelled by mouse movement.

This paper expands on previous work presented in [1], where we investigate the effect of sentimentality and polarity of news articles on user engagement. We used a collection of online news articles and examined their variation in terms of the sentimentality and polarity of their content. We also demonstrated how gaze behaviour and attention differ across news articles of varying interestingness, through a controlled user study. In the current work, we take a more scalable approach to measuring engagement and look at the effect of "interestingness" on cursor behaviour. We also examine how cursor behaviour is linked to subjective measures of engagement such as affect and focused attention. Finally, we validate our approach against gaze metrics and reliable qualitative ground truth.

3. EXPERIMENTAL SETUP

There are several approaches to carrying out a study in our context: bucket testing, log analysis, and controlled user study. The first two methods allow the analysis of real-life or recorded data at large scale, but offer little flexibility for introducing new parameters. In addition, it is not easy to control certain parameters and observe the actual user experience. On the other hand, user studies are typically much smaller in scale [2], but a wider range of parameters can be explored in a controlled manner. The downside is the difficulty of generalising the findings. In our work, we choose to conduct a small-scale, controlled user study, and record the cursor data of users interacting with web content of varying interestingness. We use a collection of online news with relatively unstructured, heterogeneous content. In what follows, we provide an outline of our experimental setup and refer the reader to [1] for additional details.

3.1 News Dataset

Our dataset contained 383 news articles crawled from Yahoo News US over a period of two weeks, from three different genres: *crime and law*, *entertainment and lifestyle* and *science*. All news articles were presented in the same format. We kept the news articles that had between $300-600$ words to mitigate any effects due to the uneven article length. We then randomly selected 40 articles from each genre and asked twenty-four human judges to rate them on a 5-point interestingness scale. The reported scores allowed us to pre-rank the news articles and narrow down our selection to three interesting and three uninteresting candidates per article genre, prior to conducting our study.

3.2 Participants

There were 22[1] participants (9 females, 13 males), free from any obvious physical or sensory impairment, through a campus-wide ad. Participants aged from 18 to 47 and were of mixed ethnicity. The majority (54.54%) had a master's degree or some college degree (45.45%). They were primarily pursuing further studies while working (40.90%), although there were a number of students (40.90%) and full-time employees (18.18%). Participants were all proficient with the English language (18.18% intermediate level, 68.18% advanced level, 13.63% native speakers). To avoid any adverse effects because of language-specific bias, we evaluated their English language fluency during the tutorial.

[1]Originally, in [1], we report 57 participants; however, we recorded mouse tracking data for 22 participants only.

3.3 Design

The experiment had a mixed design with two independent variables: *article genre* (three levels: "crime and law", "entertainment and lifestyle", "science") and *article interestingness* (two levels: "interesting", "uninteresting"). The primary dependent variable was participants' online behaviour as determined by the mouse and eye tracking data. Other dependent variables were participants' pre- and post-task affect, level of focused attention, as well as reported interestingness of the news articles.

The study consisted of two news reading tasks: one involving an interesting news article and one involving an uninteresting news article. The article interestingness was determined by asking the participants to rank 18 news titles (six per article genre) and assign the most interesting news title to the first position, the next most interesting news title to the second position, and so on. From each participant's ranking, we selected the top-ranked news title for the interesting task and the bottom-ranked news title for the uninteresting task, thus personalising "interestingness". Moreover, the participants were asked to read the news articles as they would normally do in their natural setting, allowing them to stop reading at any point in time they felt like. To control the order effects, the news article genre and the task assignment were counterbalanced using a Latin Squares design.

3.4 Procedure

The participants sat in a quiet room, facing the computer used to perform the news reading tasks. The session began by informing the participants about the purpose of the study, addressing privacy issues, and outlining the experimental procedure. They were then asked to complete an entry questionnaire and proceed with the news reading task. Throughout the study the participants were presented with two web browser windows: a window showing the news article and a window indicating the steps to follow along with the main questionnaire. Participants were instructed to read the news article at their own pace and for as long as they wanted. Upon finishing reading the news article, they had to switch to the questionnaire to answer the relevant section. The same procedure was repeated for the second task.

3.5 Measures of Engagement

A psychometric scale was used to capture the hedonic and cognitive aspects of user experience: the User Engagement Scale (UES) [30]. The UES items pertain to positive and negative affect, users' felt involvement, and focused attention during the task. Affect refers to the emotion mechanisms that influence our everyday interactions and can act as the primary motivation for sustained engagement [37]. Focused attention [29] refers to the feeling of energised focus and total involvement, often accompanied by loss of awareness of the outside world and distortions in the subjective perception of time. We also tracked cursor activity and gaze.

PANAS: The Positive and Negative Affect Scale [37] was used to measure the affect before and after each task. PANAS includes 10 items measuring positive affect (PAS) and 10 items measuring negative affect (NAS). Participants were asked to respond on a 5-point Likert scale their agreement to the statement: "You feel this way right now, that is, at the present moment", for each item. Affect was also mea-

sured by asking the participants to respond to the question "Overall, did you feel positive or negative while completing the news reading task?".

Focused Attention: A 9-item focused attention subscale [29], was adapted to the context of the news reading task. The focused attention scale has been used in past work [28] to evaluate users' perceptions of time passing and their degree of awareness about what was taking place outside of their interaction with a task. For our news reading task, participants were instructed to state on a 5-point Likert scale their agreement to each item (e.g., "I was so involved in my news task that I lost track of time").

Interest: To validate the effectiveness of our experimental manipulation, we measured the perceived article interest at post-task by asking the participants to state on a 5-point Likert scale their agreement to questions such as "I found the news article interesting to read", etc. The reported scores were converted into binary judgments by assigning disagreement or neutral opinion to the uninteresting condition and agreement to the interesting condition. The binary judgments were then compared against the pre-task labelling of the news article interestingness. The Chi-Square test revealed a significant association ($\chi^2 = 29.52, p < .001$) between the two measures, with a strong positive relationship ($\phi = .518, p < .001$), which confirms the effect of our experimental manipulation. The follow-up analysis is based on the levels of perceived article interestingness reported at post-task and consists of slightly imbalanced classes; 23 interesting and 21 uninteresting instances.

Eye Tracking: The importance of gaze in the assessment of engagement lies in the fact that, although looking might appear to be a process that is under voluntary control, conscious and deliberate control of fixation happens infrequently. Therefore, gaze is considered a strong indicator of attention [10] and the utility of eye tracking in information processing tasks like reading [3] and micro-blogging [6] is well known. In our study, eye movements were recorded using a Tobii 1750 eye tracker integrated into a 17" TFT monitor with a 1280×1024 resolution. The pupil locations were extracted at a rate of 50 Hz and were mapped to gaze locations on the screen. We compute the eye metrics reported in [1].

Mouse Tracking: We used smt2 [22], an open source, client-server architecture mouse tracking tool. The smt2 uses JavaScript to log mouse and browser-related events at a configurable frequency, and stores the data at fixed-time intervals. This process does not interfere with the user's browsing experience or introduce delays associated with data capture. We set the recording rate at 40 ms, which provides a reasonable tradeoff between data quantity and granularity of the recorded mouse events [20]. Our relatively high recording rate allows us to pick up micro-pauses or ballistic movements, resulting in clearly-defined trajectories.

4. MOUSE GESTURE RECOGNITION

Our mouse gesture recognition approach is based on an anatomical interpretation of cursor movements that removes the effects of location and considers instead features stemming from rotation, speed, acceleration, spectral analysis, and other. In that respect, we analyse cursor data irrespective of the page layout and the relative position of the elements it contains. We follow an unsupervised learning approach and cluster together cursor movements that share similar properties, and then evaluate the results of our clustering. The best performing clustering setup is used to characterise the cursor data as mouse gestures.

4.1 Cursor Data

A mouse move is a continuous physical process: the cursor, controlled by the user, begins at a position x_0, y_0 with no motion, is accelerated in some direction, moving at non-zero speed for a time period Δt, and is finally brought to a halt at some position x_1, y_1. This trajectory has a specific beginning and ending and corresponds to a *mouse gesture*. Our aim is to identify segments of the cursor data corresponding to individual physical processes of this type. The data used in our analysis consists of uneven time series of cursor coordinates, which we cast into individual movements of the mouse. The assumption is that every movement has an intent, and that by tracking the movement and the manner of movement we may be able to discern that intent.

First, we need to split the recorded data into meaningful, deliberate movements performed by the user. Due to the nature of the cursor data, segmentation is not a self-evident task. Therefore, we perform this as follows. The data parser accepts a stream of x_i, y_i, t_i coordinates, indicating the cursor position on the X and Y axes (relative to the top-left corner of the browser) at time t_i, measured in milliseconds (ms).

We use a sliding window of fixed size (three pairs of x, y coordinates) over the data and examine two parameters: Δt and Δs: Δt measures time difference (in ms), whereas Δs measures distance in pixels. On occasion, e.g. out-of-focus events, smt2 may interrupt the recording, thus making it possible for $\Delta t > 40ms$, our sampling rate. Cases like that indicate that the browsing activity was interrupted, marking the end of the current mouse gesture. Furthermore, if $\Delta s < 5px$ (the cursor has been moved less than 5 pixels away from its previously logged position), we consider that a *rest*, also signifying the end of the current mouse gesture. In either case, we start a new mouse gesture, and proceed with the remaining data, until we reach the end of the input stream. The resulting gestures are therefore *subsequences* of the original data, consisting of two or more consecutive points, excluding *rests*. Our dataset contains 176,550 cursor positions, segmented into 2,913 mouse gestures, collected during the 44 news reading tasks of our 22 participants.

4.2 Features

Our task is to predict within-content engagement and, in particular, the interestingness of online news content. To this end, we explore a large number of features engineered from the cursor data. Our features, presented in Table 1, are computed for each mouse gesture as a whole and for all consecutive pairs of points found in it, and are categorised under nine main headings depending on how or where the feature is obtained from.

We do not apply any heuristics to characterise our gestures or pre-determine, for example, whether they represent a horizontal or vertical scroll; nor do we account for the time the cursor spent in pre-defined AOI or web page elements. We, rather, follow an unsupervised methodology to identify latent structures in our cursor data and develop a representative "vocabulary" that is more generalisable.

To measure the importance of each feature with respect to our goals, we perform a preliminary correlation analysis. Given that the predicted class is binary (interesting, uninteresting) we compute the point-biserial correlation coefficient (r_{pb}). We ignore the sign of the correlation which depends on the way we encode the variables. Table 1 shows several significant small-to-medium size correlations between our features and news article interestingness. The effect sizes from our correlation analysis appear to be in line with those reported in previous studies [14, 15]. We now introduce the features we use, and comment on significant results.

Time: Previous works [16, 24] have shown that accounting for the temporal characteristics of mouse interactions can improve the predictive power of a model. Similarly, we measure the duration of each gesture in milliseconds.

Coverage: These features include the number of x, y points observed in a gesture and the sum of their intra-distances, which indicates how compact or dispersed a gesture is. An interesting observation about coverage is its medium-size correlation, which suggests that the size of the surface that a mouse gesture traverses on is related to the interestingness of the content the user is interacting with.

Type: This feature describes the type of gesture, i.e. move or rest. This information, however, is not used in the unsupervised learning task because we do not want to force the clustering algorithm to learn a gesture concept that derives from a heuristic.

Distance: These features include the total distance that the cursor has traveled, the maximum, minimum, average, and standard deviation of the distances of all consecutive pairs in a gesture. They are computed using the Euclidean distance and the pixel distance travelled on the X and Y axes. As seen from Table 1, this category of features are significantly correlated with news article interestingness. Although we lack knowledge about the directionality of this relation (i.e. positive or negative), it is still evident that the distance traversed by the mouse cursor indicates how much the user interacted with the news article and, to some extent, how interesting the latter was perceived.

Speed: As suggested in [14], the speed of cursor movements can characterise mouse interactions and discriminate between cursor patterns. Slow gestures may indicate that the cursor is resting while the user is engaged in a cognitive demanding task like careful reading, while ballistic movements might suggest that the user is performing a quick scan to locate an information of interest in the text. We measure the speed for the total distance that the cursor has traveled, the maximum, minimum, average, and standard deviation of the speeds of all consecutive pairs in a gesture. We compute these features using the Euclidean distance and the pixel travel distance on the X and Y axes. We note that value of these features is supported by several significant correlations between the speed features and news article interestingness.

Acceleration: Similar to speed, we compute the acceleration for the total distance that the cursor has traveled, as well as the maximum, minimum, average, and standard deviation of the accelerations of all consecutive pairs in a gesture. We compute these features using the Euclidean distance and the pixel distance travelled on the X and Y axes.

Direction: For each consecutive pair in a mouse gesture we determine the direction of the movement and normalise for all range of angles. This feature is important to learning the mouse gestures because it discriminates vertical and horizontal scrolls from other kinds of cursor movements.

Rotations: We count the number of clockwise or counterclockwise rotations performed by the cursor. To identify a rotation we examine the sequence of x and y coordinates in a gesture and for every window of three consecutive points $A_i = (x_i, y_i)$, $A_{i+1} = (x_{i+1}, y_{i+1})$, and $A_{i+3} = (x_{i+3}, y_{i+3})$

Table 1: The features used in the clustering analysis and Pearson's correlations to the predicted class of news article interestingness

Category	Feature	r_{pb}
Time	gestureDuration	.015
Coverage	noPnts	.015
	dispersal	−.040*
Type	gestureType	.000
Distance	eucDistTotal	−.027
	eucDistMin	. − .039*
	eucDistMax	−.022
	eucDistAvg	−.040*
	eucDistSD	−.004
	pxlDistTotalX	−.013
	pxlDistMinX	−.037*
	pxlDistMaxX	.007
	pxlDistAvgX	−.020
	pxlDistSDX	.016
	pxlDistTotalY	−.015
	pxlDistMinY	−.034
	pxlDistMaxY	−.024
	pxlDistAvgY	−.038*
	pxlDistSDY	−.010
Speed	eucDistVelTotal	−.041*
	eucDistVelMin	−.039*
	eucDistVelMax	−.022
	eucDistVelAvg	−.040*
	eucDistVelSD	−.004
	pxlDistVelTotalX	−.020
	pxlDistVelMinX	−.037*
	pxlDistVelMaxX	.007
	pxlDistVelAvgX	−.020
	pxlDistVelSDX	.016
	pxlDistVelTotalY	−.038*
	pxlDistVelMinY	−.034
	pxlDistVelMaxY	−.024
	pxlDistVelAvgY	−.038*
	pxlDistVelSDY	−.010
Acceleration	eucDistAccMin	−.022
	eucDistAccMax	.010
	eucDistAccAvgSqRt	−.006
	pxlDistAccMinX	−.019
	pxlDistAccMaxX	.017
	pxlDistAccAvgSqRtX	.013
	pxlDistAccMinY	−.022
	pxlDistAccMaxY	.008
	pxlDistAccAvgSqRtY	−.012
Direction	angle0_15	−.022
	angle15_30	.021
	angle30_45	.008
	angle45_60	.003
	angle60_75	−.034
	angle75_90	−.013
	angle90_105	.013
	angle105_120	.013
	angle120_135	−.000
	angle135_150	−.000
	angle150_165	−.005
	angle165_180	−.000
Rotations	rotClckWise	−
	rotCntClckWise	−
FFT	FFTVelEucl	−.018
	FFTVelPxlX	−.010
	FFTVelPxlY	−.005
	FFTAccEucl	.010
	FFTAccPxlX	.015
	FFTAccPxlY	−.005

*Correlation is significant at the 0.05 level (2-tailed).

we calculate the signed angle $\measuredangle(\vec{A_iA}_{i+1}, \vec{A_{i+1}A}_{i+2})$. We then sum the signed angles using the sign of the vectors' $\vec{A_iA}_{i+1} \times \vec{A_{i+1}A}_{i+2}$ cross product. If the sum exceeds 360° we count a full rotation and, depending on the sign of the sum, we label it either as counterclockwise (positive sign) or as clockwise (negative sign).

Fast Fourier Transform: We apply a spectral analysis to velocity and acceleration to identify the dominant component frequencies in our cursor data. We use the fast Fourier transform (FFT) since it is a more efficient way to compute the discrete Fourier transform (DFT). Given a sequence x_n (point time domain signal) that has length N and is assumed to have a period of N, the FFT computes two $N/2+1$ point frequency domain signals, i.e.

$$X_\kappa = \sum_{n=0}^{N-1} x_n e^{-i2\pi\kappa\frac{n}{N}}, \kappa = 0 : N-1$$

The two signals in the frequency domain are the real part and the imaginary part, and hold the amplitudes of the cosine waves and sine waves respectively. This frequency representation tells how much of the variability of the data is comprised of low frequency waves and how much is due to high frequency patterns. We use the most powerful frequency (with respect to velocity and acceleration) of each mouse gesture as a feature.

4.3 Preprocessing

We apply two types of transformation: *normalisation* and *Principal Components Analysis* (PCA). We normalise each feature to the $[0, 1]$ range, to avoid having attributes in greater numeric ranges dominating those in smaller numeric ranges. We use PCA for dimensionality reduction, but select enough eigenvectors to account for some percentage of the variance in the original data; in our case 95%. This resulted in 4 datasets: *original*, *normalised*, *PCA*, and *normalised+PCA*. We ran the following analysis on each dataset, but report on the best performing one, which is the *original*, without transformations.

4.4 Unsupervised Learning

A key objective for us is to derive generic mouse movements which are associated with user engagement measures. By generic, we mean that we need to abstract away from gestures characteristics of an individual to gestures shared by the study's participants. Therefore, we employ clustering, to group together similar mouse gestures and movement patterns. Given that we target more cognitive processes, we would like a robust analytical framework. Therefore, instead of relying on a single clustering algorithm, which could be easily biased, we use multiple techniques and reach a consensus through *rank aggergation*, which is known to be effective in removing noise [8].

4.4.1 Clustering Algorithms

We perform unsupervised learning using the feature set presented in Section 4.2 and the datasets discussed in Section 4.3. We use five clustering methods: `K-Means`, `EM`, `Cobweb`, `Agglomerative Hierarchical Clustering`, and `Spectral Clustering` [39]. Weka[2] contains implementations of the first four, whereas we used the implementation for R[3] for the last one. Table 2 summarizes the various clustering parameters that we used. Most notably, when the algorithm accepts as input the number of clusters to produce, we ranged that value from 1 to 40, given that we have slightly over 40 different reading tasks, and it is reasonable to expect some consistency within each one. We remind the reader, however, that clustering is performed on mouse gestures, which are far more than the reading tasks. Overall, we consider $45,654$ different clustering configurations.

Table 2: Clustering Algorithm Parameters

Algorithm	Feature	Value
K-Means	`Distance`	*Euclidean, Manhattan*
	`#Clusters`	[1,40]
EM	`#Clusters`	[1,40]
CobWeb	`a(acuity), c(cutoff)`	Exhaustive grid search
Hierarchical	`Distance`	*Manhattan, Euclidean, Chebyshev*
	`Linkage`	*single, complete, centroid*
	`#Clusters`	[1,40]
Spectral	`Kernels`	*Radial Basis, Laplacian, Hyperbolic tangent k*
	`Kernel Parameters`	*Exponential growing*
	`Other Parameters`	See [26]
	`#Clusters`	[1,40]

4.4.2 Cluster Validity

Given that clustering is an unsupervised method, without any prior ground truth, we need a way to quantify the quality of the produced output. The process of evaluating clustering results is known as *cluster validity* [17], and three approaches exist for this purpose: external criteria, internal criteria, and relative criteria. We use internal criteria, thereby relying on quantities and features inherent to our data. We compute Index, Ball-Hall, C index, Calinski-Harabasz, Davies-Bouldin, Gamma, G+, $\log(BGSS/WGSS)$, McClain-Rao, PBM, Point biserial, SD Scat, SD Dis, Silhouette, $\mathrm{Tr}(W)$, all present in the *clusterCrit*[4] package for R.

4.4.3 Missing Value Substitution

For certain internal quality criteria that we consider, some clustering configurations were not assigned a value. Missing values are not uncommon in statistical analysis, yet they pose a problem as we cannot produce a complete ranking of the data. To address this, we perform missing value imputation, which is better than removing such entries [31]. For each configuration with a missing value, we select its k-nearest neighbours from the set of complete configurations (without any missing values). We then replace the missing value with the average one of the k-nearest neighbours for this dimension (measure). We can then induce a total ordering for each of our measures.

4.4.4 Downsizing the Validity Measures

We have discussed how to measure the quality of the produced clusters. The numerous measures can slow down rank aggregation, and could affect result quality. To address this issue, we downsize them to a subset that shows high consistency among the produced rankings. We compute all $\frac{n(n-1)}{2}$ pairwise Spearman correlations of the orderings. This gives a complete weighted graph $G = (V, E)$: vertices are the validity measures and each edge $e = (v_i, v_j)$ is weighted with the correlation value between v_i and v_j. We normalize the values in the range $[0, 1]$, which does not affect our solution, as we are interested in high correlation. Given G, we select the k vertices, such that the resulting k-clique has the maximum edge-weight. The maximum edge-weight clique

[2] `www.cs.waikato.ac.nz/ml/weka/`

[3] `cran.r-project.org/web/packages/kernlab/index.html`

[4] `cran.r-project.org/web/packages/clusterCrit/index.html`

Algorithm 1 `Select the internal validity measures with highest overall agreement.`

Input: Weighted Graph $G = (V, E)$, int k
Output: Max-Edge Weight Subgraph G_k, $|G_k| = k$
$G_k \leftarrow \emptyset$
$e_1 \leftarrow$ Select edge with maximum weight.
Add vertices of e_1 to G_k
while ($|G_k| < k$) **do**
  for $v_i \in (G\ G_k)$ **do**
    $s_i \leftarrow \sum_{v_j \in G_k} weight(v_i, v_j)$
  end for
  Select v_j with max s_i
  $G_k \leftarrow (G_k \cup v_j)$
end while
return G_k

problem is NP-Hard [33]. Therefore, we use the greedy approximation shown in Algorithm 1. The output contains the k validity measures with highest overall agreement.

4.4.5 Rank Aggregation

Given the k validity measures with high consistency, we want to identify the n clustering configurations that are better (ranked higher) than the rest. In other words, we need those configurations that have better standing in as many lists as possible. This is a typical instance of *multiple-winner voting problem*, where we are given a set of ordered preferences as input (ordered configurations per measure in our case), and we want to select n winners.

Various approaches exist from social theory, but we opt for *Rank Aggregation* [8], which is better at filtering out noise. Also, despite of the problem's NP-Hardness, there are efficient approximation implementations in many statistical packages. We use an implementation of *Rank Aggregation*[5] for R with the configuration suggested in [32]. Rank aggregation derives a single ranked list $\mathcal{L}'$ that has the minimum distance (e.g., Kendall τ) from a given set of ranked input lists $\mathcal{L} = \{\mathcal{L}_1, \mathcal{L}_2, \cdots, \mathcal{L}_m\}$. Formally, it is defined as follows:

$$\mathcal{L}' = arg\min(\frac{1}{m}\sum_{i=1}^{m}\mathcal{D}(\mathcal{L}_i, \mathcal{L}'))$$

The top positions of the list are the clustering configurations that performed better than others, for the input measures.

4.4.6 Towards a Taxonomy

We report cluster coverage for the mouse gestures we identified in the unsupervised learning. Using the method discussed in Section 4.4.5, we compute the aggregated ranking for all internal quality measures. The top-ranked clustering configuration is the `Spectral Clustering` for the original dataset, with hyperbolic tangent kernel, and for $k = 38$. Table 3 shows the distribution of mouse gestures per cluster. We note that the majority of mouse gestures were allocated to `c6`. Out of the $1,352$ mouse gestures in `c6`, over 70% are rests, suggesting that most of the observations allocated in this cluster are characterised by inertia, while the remaining clusters contain mouse gestures characterised by more activity (e.g., `c2`, `c9`, `c22`). Figure 1 shows the normalised coverage of mouse gestures per user and per task (interesting versus uninteresting), for all 38 clusters. The distribution of clusters validates the prominent presence of `c6` mouse gestures but also suggests the presence of other types of mouse gestures, which vary per condition. The diversity of patterns observed indicates that mouse gestures can be a good predictor of user engagement, something that is further demonstrated in Section 5 by the performance of our prediction models (if an article is interesting or not).

[5] `cran.r-project.org/web/packages/RankAggreg/`

4.5 User × Task Interactions

We collected questionnaire data on the experienced affect, focused attention, and degree of interest from 44 news reading tasks, carried out by 22 participants. A 5-point Likert scale was used in all questionnaires with high scores representing a stronger agreement and low scores representing a weaker perception with the given statement. Participants responses to the 10-item PAS, 10-item NAS, and 9-item focused attention scale were summed to obtain the final scores.

The Wilcoxon Signed-Rank test was applied to determine the significance of the variance observed in the frequency distribution of mouse gestures between interesting and uninteresting news articles. The results indicated a statistically significant difference ($z = -3.817, p = .000, r = -0.171$) between the counts of different categories of mouse gestures, suggesting an interaction effect between news article interestingness and cursor behaviour. We note that as interestingness we regard the level of interestingness reported by the participants at post-task.

When examining the eye metrics, the Wilcoxon Signed-Rank test reveals differences in gaze behaviour. More specifically, we observe that participants took significantly less time to perform their first fixation (`TFF`) on the news article in the interesting condition ($z = -2.718, p = .007, r = -0.377$), while they fixated more times on other elements (`FB`) when the news article was uninteresting ($z = -2.925, p = .003, r = -0.405$). Furthermore, participants performed significantly more fixations (`FC`) and for longer periods (`FD`, `TFD`) when reading an interesting news article ($z = -2.117, p = .034, r = -0.293$; $z = -2.076, p = .038, r = -0.288$; $z = -2.258, p = .024, r = -0.313$). Finally, they looked more times (`VC`) at the body of article AOI, and the duration of each individual visit lasted longer (`VD`), when reading an interesting news article ($z = -2.312, p = 0.020, r = -0.320$; $z = -2.258, p = 0.024, r = -0.313$).

These findings are in line with those reported in previous work [1], where we demonstrate the impact of news article interestingness on the gaze behaviour of male and female participants. For example, in [1], we observe that male participants perform their first fixation faster on a news article when the content is interesting and, in addition, perform significantly less fixations outside of the body of the news article. No significant differences were observed for female participants. We note, however, that in this paper we do not group our participants according to gender since we do not account for it in our cursor data analysis. The significant differences discovered in gaze behaviour establish a connection to user engagement but also suggest common gaze patterns across both genders. This finding suggests that knowledge of the user gender is not as important for gaze and, as an extension, to its proxy, the cursor. In regards to mouse tracking, this is a positive finding in the sense that predicting the user's gender would be a non-trivial task.

To determine if and to what extent the identified categories of mouse gestures are associated with experienced affect, focused attention, and gaze behaviour, we run a correlation analysis. To compute the correlation coefficients, we

Table 3: Distribution of mouse gestures across all clusters for top-ranked clustering configuration

Cluster	#	Cluster	#	Cluster	#	Cluster	#	Cluster	#	Cluster	#
c1	54(1.85%)	c8	38(1.30%)	c15	15(0.51%)	c22	22(0.76%)	c29	47(1.61%)	c36	49(1.68%)
c2	104(3.57%)	c9	137(4.70%)	c16	60(2.06%)	c23	43(1.48%)	c30	39(1.34%)	c37	37(1.27%)
c3	88(3.02%)	c10	14(0.48%)	c17	9(0.31%)	c24	54(1.85%)	c31	62(2.13%)	c38	11(0.38%)
c4	7(0.24%)	c11	16(0.55%)	c18	75(2.57%)	c25	14(0.48%)	c32	29(1.00%)		
c5	50(1.72%)	c12	55(1.89%)	c19	65(2.23%)	c26	28(0.96%)	c33	35(1.20%)		
c6	1352(46.41%)	c13	36(1.24%)	c20	28(0.96%)	c27	9(0.31%)	c34	18(0.62%)		
c7	22(0.76%)	c14	18(0.62%)	c21	64(2.20%)	c28	39(1.34%)	c35	70(2.40%)		

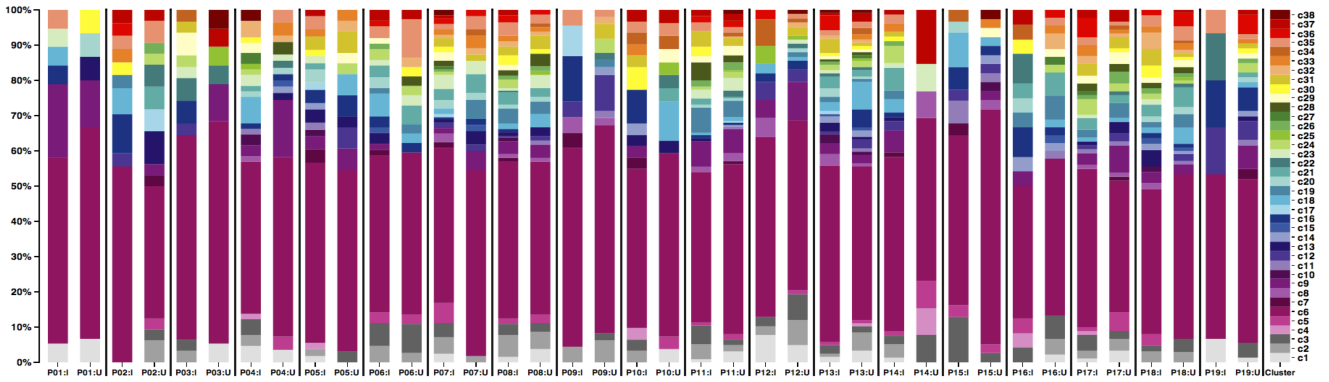

Figure 1: Coverage of mouse gestures per participant and per cluster for interesting (I) and uninteresting (U) news

consider the frequency of occurrence of the mouse gestures across all tasks and users, and compare it with the eye metrics and the questionnaire information we collected at pre- and post-task. Given the non-normal distribution of our data, we opt for the Spearman's rho non-parametric test. Table 4 shows several statistically significant correlations. Of interest are the medium-size, positive correlations between the frequency of certain types of mouse gestures (e.g., c6, c9, c12, c20, c3) and eye metrics like time to first fixation, duration of fixation, fixation count, and total visit duration. Although this finding is not new [18, 24, 25], it does connect our approach to analysing cursor behaviour with gaze, especially under the light of the findings presented in earlier studies, and shows that this correlation exists beyond a simple mapping of x and y coordinates.

We also report for the very first time several significant correlations between certain types of mouse gestures and preNAS, prePAS, postNAS, postPAS, affect, and focused attention. These correlations indicate that cursor behaviour can go beyond measuring frustration to inform us about the positive and negative valence of an interaction. We discuss this further in Section 6.

5. PREDICTING INTERESTINGNESS

In the previous section we presented our approach to extracting mouse gestures from cursor data and identified connections to gaze behaviour, affect and attention. In this section, we demonstrate the value of our method by predicting the outcome of online news reading experiences. Our objective is to identify how cursor behaviour can help us predict content interestingness and, in particular, define a taxonomy of cursor patterns as a basis to automatic classification. To this end, we perform pattern analysis and learn models using the large set of features we discussed in Section 4.2. We use the best performing clustering setup (`Spectral Clustering`, original dataset, hyperbolic tangent kernel, and for $k = 38$) to predict the class (interesting, uninteresting) of the news article. We train the following classifiers: (i) 1 Nearest-Neighbor (`1NN`), (ii) Support Vector Machines with a polynomial kernel using the Sequential Minimal Optimization algorithm (`SMO`), (iii) Random Forest (`RandomForest`), and (iv) Stacked Generalization (`Stacking`) for combining the above classifiers using the `Real Adaboost` method as a meta-classifier. Additionally, we perform feature selection using the *ClassifierAttributeEval*[6] of Weka, in combination with the *Ranker* search method that sorts the attributes by their individual evaluations.

We note that our training data consist of slightly imbalanced classes; 23 interesting and 21 uninteresting instances. Therefore, we include a baseline classifier `Baseline` that always predicts the majority class in the training data for comparison purposes. We report the classification performance in terms of weighted average precision, recall, f-measure and accuracy across classes. The reported results are obtained by performing cross-validation with ten folds. As shown in Table 5, the performance of each classifier model individually is encouraging, but not optimal. However, when we combine `1NN` and `SMO` using the stacked generalisation method, we improve further the performance of our model. This is anticipated, since `Stacking` combines the output from different classifiers and can increase the predictive performance over a single model. Overall, the model trained with the stacked generalisation method outperforms the baseline and introduces a notable improvement of 23% in accuracy.

6. DISCUSSION & CONCLUSIONS

In this paper, we presented a generalisable solution to measuring within-content engagement using mouse tracking data. Our work is motivated by the fact that millions of users interact daily with online content without providing any explicit feedback about the quality of their experience. Therefore, any effort towards developing a more nuanced

[6]Evaluates the worth of an attribute by using a user-specified classifier.

Table 4: Correlation matrix of clusters (categories of mouse gestures), FA, PANAS, and eye metrics

	FA	preNAS	prePAS	postNAS	postPAS	affect	TFF	FB	FFD	FD	TFD	FC	VD	TVD	VC
c1	−.177	.358**	.19	.325**	.113	.11	.118	.207	−.076	.086	.355*	.314*	.279	.355*	.022
c2	−.456**	.203	.083	.203	.083	−.159	−.077	−.128	−.279	.189	.305*	.261	.264	.305*	.003
c3	−.299**	.307**	.158	.259*	−.004	.003	−.064	−.094	−.332*	.095	.283	.271	.22	.283	−.015
c4	.098	.261*	.029	.178	−.017	.03	−.017	.02	−.091	.027	.135	.184	.081	.135	−.024
c5	−.097	.182	.06	.106	.055	.049	.032	.077	−.168	.057	.113	.037	.329*	.113	−.21
c6	−.405**	.354**	.207	.316**	.108	−.069	−.031	−.01	−.239	.273	.521**	.485**	.251	.521**	.135
c7	−.123	−.007	.153	.023	.067	−.117	−.199	−.269	−.221	−.263	.143	.26	.09	.143	.008
c8	−.415**	.244*	.038	.081	.065	.149	.028	.082	−.095	.201	.315*	.254	.223	.315*	.057
c9	−.393**	.231*	.145	.286**	.074	.04	.119	.091	−.207	.141	.349*	.375*	.283	.349*	.043
c10	−.089	.03	−.025	.115	−.013	.165	−.014	.025	.169	−.051	.262	.367*	−.182	.262	.360*
c11	−.196	.262*	.128	.114	.108	.066	−.011	−.014	−.055	−.162	.141	.282	−.088	.141	.145
c12	−.250*	.181	.138	.081	.057	.002	.01	.082	−.266	.12	.461**	.403**	.212	.461**	.148
c13	−.107	.068	−.114	.250*	.018	−.293**	−.102	−.1	−.157	.059	.002	.002	.186	.002	−.173
c14	−.051	.243*	.215*	.16	.106	−.022	−.082	−.021	−.132	.097	.358*	.376*	−.069	.358*	.248
c15	−.088	.258*	.154	.379**	.223*	−.169	−.164	−.144	−.183	.318*	.163	.042	.039	.163	.077
c16	.045	−.089	−.038	.078	−.049	.003	−.219	−.236	−.005	−.159	.115	.206	−.137	.115	.176
c17	−.053	.044	.163	−.046	.095	−.055	.048	.023	.004	−.117	.057	.129	.064	.057	−.042
c18	−.131	.153	−.035	.142	.032	−.058	−.131	−.089	−.158	.182	.255	.165	.1	.255	.104
c19	−.14	.312**	.169	.197	.137	−.141	−.176	−.113	−.051	.104	.082	.064	.16	.082	−.063
c20	−.221*	.196	.175	.269*	−.034	−.064	−.351*	−.339*	−.256	.077	.314**	.329*	.105	.314*	.067
c21	−.125	.321**	.089	.306**	.151	−.198	−.214	−.169	−.184	.115	.108	.035	.191	.108	−.134
c22	.065	.155	.012	.095	.099	.066	.440**	.456**	.153	.169	.008	−.034	.065	.008	−.102
c23	−.244*	.089	.021	.074	.032	.078	.033	−.086	−.156	.141	.167	.178	.274	.167	−.073
c24	−.018	.248*	.102	.244*	.105	−.014	−.01	−.018	−.258	.09	.257	.249	.162	.257	.004
c25	−.289**	.18	−.042	.195	−.001	−.077	.304*	.341*	−.032	.387**	.306*	.158	.155	.306*	.117
c26	−.128	.153	.111	.11	.092	−.034	−.005	−.004	−.201	.137	.184	.12	.256	.184	−.086
c27	.066	.446**	.299**	.374**	.298**	.064	−.055	.025	−.289	.082	.338*	.225	.165	.338*	.07
c28	−.243*	.233*	.066	.178	.054	.089	−.047	−.015	−.153	.224	.202	.125	.244	.202	−.078
c29	−.088	.233*	.033	.283**	.089	.002	.171	.148	−.084	.18	.268	.197	.201	.268	.008
c30	−.234*	.043	−.032	.01	−.078	.083	−.288	−.285	−.111	−.006	.126	.16	−.079	.126	.117
c31	−.374**	.14	.082	.167	−.016	−.072	.004	−.05	−.24	.155	.381*	.401**	.297	.381*	.034
c32	−.186	.401**	.184	.442**	.184	.12	−.054	−.013	−.209	.161	.263	.217	.094	.263	.058
c33	−.19	.250*	.227*	.315**	.161	−.214*	−.308*	−.236	−.193	.038	.287	.291	.057	.287	.11
c34	−.308**	.193	−.123	.148	.036	−.297**	.157	.173	.005	.312*	.085	−.05	.171	.085	−.114
c35	−.125	.091	.085	.066	.144	−.084	−.182	−.178	−.075	.151	.18	.144	.179	.18	−.035
c36	−.104	.296**	.345**	.254*	.292**	−.032	−.197	−.172	−.193	.28	.295	.159	.178	.295	.072
c37	.165	.275**	.153	.17	.086	−.047	−.025	.025	.054	.008	.092	.083	.266	.092	−.193
c38	−.309**	.282**	.115	.251*	.004	.312**	.327*	.348*	−.087	.175	.288	.24	.186	.288	.019

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Table 5: Performance metrics for the classifier models using 10-fold cross-validation

	Performance metrics			
Classifier	Precision	Recall	F-Measure	Accuracy
`Baseline`	0.273	0.523	0.359	0.522
`1NN`	0.664	0.659	0.659	0.659
`SMO`	0.700	0.682	0.678	0.681
`RandomForest`	0.727	0.727	0.727	0.727
`Stacking (1NN+SMO)`	0.751	0.750	0.750	0.750

understanding of user online behaviour is considered a high-value task. Mouse tracking can address this need in a low-cost and scalable manner, and without removing the users from their natural setting.

To this end, we demonstrated in detail a *rigorous* methodology for extracting purposeful mouse gestures from cursor coordinates; a high-level representation of cursor interactions. More, specifically, we conducted a small-scale, controlled user study and recorded the cursor data of users that interacted with interesting and non-interesting web content. From that data, we engineered a large set of features and used unsupervised learning to build a taxonomy of cursor patterns that share similar properties. In our analysis, we considered cursor behaviour independently of the page layout, the type of elements it contains, or their relative position. Finally, we demonstrated the value of our approach in a user engagement scenario, where we predict the outcome of online news reading experiences.

Our analysis of cursor interactions provides several insights into the nature of engagement. Foremost, when examining gaze behaviour, we observed significant differences between the news articles of different interestingness that spanned across several eye metrics, like time to first fixation, duration of fixation, fixation count, and total visit duration. These differences were also noticed in cursor interactions and were found to be correlated to the eye metrics. The message here is that engagement manifests in different forms such as the gaze behaviour of users that develop an emotional, cognitive, and behavioural connection with a digital resource (e.g., an interesting news article), but also as observable and distinct mouse cursor patterns. Although this may not be a new finding, considering previous research on mouse-gaze interactions [18, 24, 25], it provides evidence that connects our methodological approach to analysing cursor behaviour with gaze and demonstrates the utility of mouse tracking as a scalable, cost-effective alternative to eye tracking.

Furthermore, we identified several significant correlations between cursor behaviour and focused attention, preNAS, prePAS, postNAS, postPAS, and affect. More specifically, we noticed that certain types of mouse gestures are negatively correlated with focused attention, with negative affect at pre- and post-task, and at a lesser extent with positive affect at pre- and post-task. This basically translates to negative emotions being more influential on cursor behaviour than positive ones. This observation is consistent with previous work [24] demonstrating that mouse-related signals are sensitive to frustrating and unpleasant experiences. Considering the challenge of explaining user behaviour using cursor pattern analysis, but also identifying to what extent these patterns are good indicators of affect, this makes it a novel and noteworthy finding. Additionally, we observe a more profound correlation between cursor behaviour and our user engagement measures compared to the correlation reported between gaze behaviour and user engagement measures [1]. This leads us to conclude that affect is, to some extent, "measurable" and can be anticipated to a certain degree.

Our prediction experiments also revealed that it is possible to measure content interestingness by accounting for the frequency of the cursor patterns changes and, subsequently, predict user engagement over time. Our best performing model, using the stacked generalization method with classifiers `1NN` and `SMO`, attained the accuracy of 75%, which is considerably better than the baseline. These results are further supported by the statistical analysis we performed on the frequency distribution of mouse gestures, which indi-

cated a significant difference between interesting and uninteresting news articles. Overall, this suggests an interaction effect between web content interestingness and cursor behaviour. Unlike prior research efforts that have analysed mouse tracking data, our approach does not involve manual or costly attempts (e.g., eye tracking). Also, the proposed method was designed to be as independent as possible from the page layout, the type of elements it contains, or their relative position, making it applicable to broader and more heterogeneous contexts.

Finally, our work comes with certain limitations. One of them is the relatively small sample of the population we studied. To some extent, this has affected our ability to generalise our findings to the population as a whole. However, this is a very common caveat in user studies and, in our case, a necessary trade-off, given the controlled and time-demanding nature of our experiment. Also, there is room for improvement, both with respect to the accuracy of the classifiers and the feature selection. Our modelling approach can benefit from experimenting with additional machine learning techniques, and a more thorough evaluation of our feature set, to improve its discriminative power.

7. FUTURE WORK

We plan to develop a larger collection of mouse tracking data. One challenge would be to develop sufficient ground truth under real-life conditions and at large scale, with the help of other testing methods like bucket testing and query log analysis. Comparing the mouse gestures identified by the proposed method against this kind of ground truth will let us validate whether they have the required physical characteristics and are well-founded, and further ensure that our approach is grounded in more than intuitive plausibility. Finally, we will investigate the sequential nature of mouse gestures using stochastic models that can describe the process of how the data is being generated and account for the transitions between the mouse gestures.

8. REFERENCES

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

[2] J. D. Brutlag, H. Hutchinson, and M. Stone. User preference and search engine latency. *ASA*, 2008.

[3] G. Buscher, R. Biedert, D. Heinesch, and A. Dengel. Eye tracking analysis of preferred reading regions on the screen. *SIGCHI*, 2010.

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

[5] M. Chen and V. Lim. Eye gaze and mouse cursor relationship in a debugging task. In C. Stephanidis, editor, *HCI Int'l - Posters, Extended Abstracts*, 2013.

[6] S. Counts and K. Fisher. Taking it all in? visual attention in microblog consumption. *ICWSM*, 2011.

[7] F. Diaz, R. W. White, D. Liebling, and G. Buscher. Robust models of mouse movement on dynamic web search results pages. *CIKM*, 2013.

[8] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. *WWW*, 2001.

[9] A. Edmonds, R. W. White, D. Morris, and S. M. Drucker. Instrumenting the dynamic web. *Journal of Web*, 6(3), 2007.

[10] M. H. Fischer. An investigation of attention allocation during sequential eye movement tasks. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 52(3), 1999.

[11] J. Freeman, R. Dale, and T. Farmer. Hand in motion reveals mind in motion. *Frontiers in Psychology*, 2(59), 2011.

[12] Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. *SIGIR*, 2008.

[13] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. *SIGCHI*, 2010.

[14] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. *WWW*, 2012.

[15] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. *CIKM*, 2012.

[16] Q. Guo, D. Lagun, D. Savenkov, and Q. Liu. Improving relevance prediction by addressing biases and sparsity in web search click data. In *Workshop on Web Search Click Data*, 2012.

[17] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3), 2001.

[18] J. Huang, R. White, and G. Buscher. User see, user point: Gaze and cursor alignment in web search. *SIGCHI*, 2012.

[19] J. Huang, R. W. White, G. Buscher, and K. Wang. Improving searcher models using mouse cursor activity. *SIGIR*, 2012.

[20] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. *SIGCHI*, 2011.

[21] D. Lagun, M. Ageev, Q. Guo and E. Agichtein. Discovering common motifs in cursor movement data for improving web search. *WSDM*, 2014.

[22] L. A. Leiva and R. Vivó. Interactive hypervideo visualization for browsing behavior analysis. *WWW*, 2012.

[23] F. Mueller and A. Lockerd. Cheese: tracking mouse movement activity on websites, a tool for user modeling. *SIGCHI Extended Abstracts*, 2001.

[24] V. Navalpakkam and E. Churchill. Mouse tracking: measuring and predicting users' experience of web-based content. *SIGCHI*, 2012.

[25] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. J. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. *WWW*, 2013.

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14*, 2001.

[27] J. Nielsen and K. Pernice. *Eyetracking Web Usability*. New Riders, 2009.

[28] H. L. O'Brien. Exploring user engagement in online news interactions. *JASIST*, 48(1), 2011.

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

[30] H. L. O'Brien and M. Lebow. Mixed-methods approach to measuring user experience in online news interactions. *JASIST*, 64(8), 2013.

[31] A. Olinsky, S. Chen, and L. Harlow. The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, 151(1), 2003.

[32] V. Pihur, S. Datta, and S. Datta. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13), 2007.

[33] S. S. Ravi, D. J. Rosenkrantz, and G. K. Tavyi. Heuristic and special case algorithms for dispersion problems. *Operations Research*, 42(2), 1994.

[34] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. *SIGCHI Extended Abstracts*, 2008.

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

[36] M.D. Smucker, X. Sunny Guo, and A. Toulis Mouse Movement During Relevance Judging: Implications for Determining User Attention. *SIGIR*, 2014.

[37] D. Watson, L. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, 54(6), 1988.

[38] R. W. White, and G. Buscher. Text Selections As Implicit Relevance Feedback. *SIGIR*, 2012.

[39] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.