Inferring Searcher Attention by Jointly Modeling User Interactions and Content Salience

Dmitry Lagun*
Google
dlagun@google.com

Eugene Agichtein†
Emory University and Yahoo Labs
eugene@mathcs.emory.edu

ABSTRACT

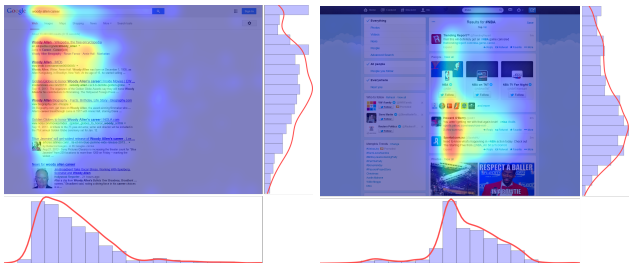
Modeling and predicting user attention is crucial for interpreting search behavior. Current applications include quantifying web search satisfaction, estimating search quality, and measuring and predicting online user engagement. The most direct way to measure attention is through eye gaze tracking, which is not yet widely available. While prior research has demonstrated the value of mouse cursor data and other interactions as a rough proxy of user attention, accurately predicting where a user is looking on a page remains a challenge. This problem is exacerbated when moving beyond the traditional search result pages to other domains, where high diversity of content and visual presentation often affect how users examine a page. We posit that in order to accurately model user attention online, interaction signals should be grounded to the underlying content. To this end, we introduce a principled model to connect interaction signals with page content features, which we call Mixture of Interactions and Content Salience (MICS). To our knowledge, our model is the first to effectively combine user interaction data with visual prominence, or salience, of the page content elements. Extensive experiments on multiple popular types of Web content demonstrate that our model significantly outperforms previous approaches to searcher gaze prediction, which use only the interaction information. Grounding the observed interactions to the underlying page content provides a general and robust approach to user attention modeling, which could enable more powerful tools for search behavior interpretation, and ultimately search quality improvements.

Keywords

searcher attention; search behavior; content salience

1. INTRODUCTION

Inferring searcher attention in Web search, and more general online settings, has been recognized as a key aspect of relevance evaluation, search quality, and user interface design. More generally, inferring and measuring user attention is key for diverse areas such as online advertising, education, and crowdsourcing. The availability of accurate user attention data at scale could potentially enable vast opportunities for search quality evaluation and richer models of user interaction with generic Web page content. The challenge is how to infer attention robustly and for diverse kinds of content - the problem we aim to address in this paper.



(a) (b)

Figure 1: User attention on a Web search result page (a) vs. on a social network Web site (b). The color indicates the time spent viewing a region of the page, ranging from red (high concentration of attention) to blue (low attention). The marginal distributions are projected onto the horizontal and vertical axes.

Inferred attention has already been used in a variety of applications ranging from improvements of web site usability [24, 28], to search relevance estimation[15, 16], and automatic generation of attention-biased summaries [1]. User attention data gains even greater importance as Web search shifts towards addressing user information needs directly on the search result page, which does not require a click on the result, making click-based evaluation of search quality more challenging (e.g., [13, 19, 32, 26]).

Previous work on attention modeling from cursor interactions on the Web has mostly focused on Web search and E-learning settings, where a user's eye gaze, and mouse cursor positions, are somewhat coordinated[34, 14, 18].

When extending the attention prediction task to other Web page types, prediction becomes more challenging. For example, as search over social media gains popularity, attention models must be adapted to take into account more complicated page layouts, with some content static (not query-specific), whereas other, related content in fact might answer the search intent directly. Figure 1 (b), based on the data reported by [25], illustrates searcher attention heatmap for a popular social media site. Note that attention is shifted to

*Work done as student at Emory University.
†Work done at Emory University.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767745.

the right and towards the bottom of the page – contrasting it with the more well known "golden triangle" examination pattern for a traditional Web search result page, illustrated in Figure 1 (a). As search spans increasingly diverse domains, such differences abound. Furthermore, as search engines increasingly incorporate images and other visually attractive elements into the search result, models of searcher attention have to be revisited accordingly.

Our aim is to develop a robust, yet principled model of searcher attention that combines both content and interaction signals. For this, we adapt techniques and ideas from computational neuroscience of *visual saliency* to develop a Mixture of Interaction and Content Salience (*MICS*) model, that is able to integrate content-based static signals, together with the user's interaction data, in order to predict where on a page a searcher is paying attention. We show that our model achieves significantly lower error compared to previously reported state-of-the-art techniques using interaction-only signals. Specifically, our contributions are:

- A novel model of combining content and behavioral signals to predict searcher attention (Section 3).
- Significant accuracy improvements in predicting eye gaze position for a given user, in multiple popular search domains (Section 6).

Next, we review previous work on predicting searcher attention. Then, in Section 3 we present out general *MICS* model for inferring user attention on Web pages, and describe the specific reference implementation for popular search domains (Section 4). In Sections 5 and 6 we present empirical results on using our *MICS* model to predict user attention. Finally, we outline the implications and directions for future work in Section 7, which concludes the paper.

2. BACKGROUND AND RELATED WORK

Our work bridges three main areas of research: computational modeling of visual attention, primarily developed in the fields of computational neuroscience and computer vision, user engagement and attention modeling on the Web, and searcher interaction modeling, primarily focused on the web search domain, from the fields of human computer interaction and information retrieval.

2.1 Computational Visual Salience

There has been extensive research on automatically identifying the most important, or *salient* regions in a given image, where a person examining the image is likely to attend. As we build on the ideas explored in computational visual salience research for Web search tasks, we briefly introduce the underlying ideas and techniques.

Different formulations of salience have been proposed, e.g., focusing on identifying the image regions which initially attracts attention, or the aggregate attention distribution after a longer period of examination. In order to model salience computationally, three major factors were identified that affect human attention during visual examination of an image (or a Web page): (i) the visual importance, or salience of areas in the scene, (ii) memory and expectations about where to find the information, and (iii) the task and information need at hand. Depending on the modeling choices, the resulting models can be either task-agnostic (i.e., only consider the salience of the image regions based on content alone) or task-driven (i.e., that consider salient regions for a given task). These models are typically categorized as either bottom-up models or top-down models.

In bottom-up models the salience of image areas is typically computed based on low-level image characteristics, particularly contrast, color, intensity, edge density, and edge orientation (see [20]). One well known (*bottom-up*) salience model was introduced by Itti [20]. As do other *bottom-up* approaches, this framework attempts to simulate human attention as a feed-forward neural network. That is, it takes various features of the stimulus, such as the color contrast, gradient and motion maps, as input, and produces a single salience map that highlights the locations where the human gaze is mostly likely to attend. According to at least some neuroscience theories[20], a representation similar to the described salience map may be used by the human brain to control the human oculomotor system and to direct eye gaze to explore the stimulus, such as an image or a web page.

While it is thought that the human eye is initially attracted by the most salient regions of the image, theories about subsequent examination differ. For example, some have argued that the subsequent examination points are planned in order to maximize the resulting *information gain* [20, 21, 39]. While low level visual salience may direct the first gaze position, or fixation, it is believed that ultimately memory and expectations (e.g., about what is shown in the image and where to find specific objects) begin to also play important roles in subsequent examination positions. To take advantage of this insight, *top-down* (task specific) models which account for these effects were introduced [20, 30, 31].

2.2 User Attention and Engagement on the Web

Modeling searcher attention on the Web introduces additional challenges as content can be more varied semantically (e.g., combining both images and text), dynamic, and interactive. Stone and Dennis [38] used latent semantic analysis of topics to predict eye movement fixations on a generic Web page given position of text elements and hyperlinks. Our work is different in that we incorporate much richer information about elements semantics. In addition, we aim to combine Web page content salience with user interactions on the page. More recently, reference[35] predicted aggregated salience of Web pages based on visual content alone. As in most visual salience approaches, the authors used pixel based information to construct variety of salience maps (based on pixel level contrast, color, orientation) that are combined to produce a single map that approximates distribution of eye gaze fixations on a Web page. Our method builds on some of the ideas in these papers, but takes a different approach by modeling user attention with a parametric distribution derived on a relatively small number of prominent Web page elements. It allows us to avoid dealing with pixel level salience (and high computational complexity associated with it) and develop a unified model of user attention on Web pages that takes advantage of Web page element importance and user interaction with the page.

More closely related to our work, Buscher el al. [4] analyzed the factors determining fixation time, i.e., the time a user spends carefully examining, or fixating on, individual elements of a Web page. They found that the size of an HTML element and its proximity to the top left corner of the Web page play a major role in amount of attention this element receives. Instead of implicitly modeling scrolling activity like in [4], our work focuses on modeling user attention on a visible portion of the page (i.e., *viewport*).

In the Web search domain, studies of user attention using eye tracking provided numerous insights about typical content examination strategies, such as top to bottom scanning of Web search results [11, 29]. Other studies have investigated the effect of caption length and quality on searcher attention, identifying relevant page regions, and for many other tasks. Perhaps the most well known outcome of this line of work is the "golden triangle" of attention on Web search results, often present in the traditional search result layouts. A more comprehensive overview of classical models of searcher attention is available in [17]. More recently, a line of research on user engagement with online content has emerged[27]. For example, the users' engagement with news examination is often influenced by affect, sentiment, and visual clues[2] – which in turn can be used to better predict attention.

2.3 Searcher Attention Prediction

Perhaps the most practical application of attention modeling is to predict searcher attention – which could be applied for tasks such as search evaluation, improving search quality, and search advertising. Due to the high cost and the necessarily small scale of eye tracking studies (due to requiring specialized eye tracking equipment), a considerable amount of research has been devoted to finding more scalable methods of attention measurement. In particular, mouse cursor tracking was proposed as a cheap alternative to eye tracking by inferring searcher gaze position from mouse cursor position. The relationship between cursor and gaze has been studied empirically[8, 34, 14, 18, 32]. Chen et al. [8] was one of the first to study coordination patterns between mouse cursor and gaze. They classified mouse cursor movement into five classes: "Stay Nowhere", "Go Nowhere", "Stay the Same Region" and "Go to New Region". They found that the distance between mouse cursor and gaze position was smallest when the user moved mouse cursor in order to perform an action, e.g., pointing or clicking. However, when the cursor remained inactive, the reported accuracy of attention measurement degraded.

In the context of Web search, the coordination between mouse cursor and eye movements was first reported by Rodden et al. [34]. They reported the alignment between the user's eye movements and mouse movements when scanning a web search results page, and identified three patterns of active mouse usage: *following the eye position vertically*, *following the eye position horizontally*, and *using the mouse to mark a relevant result*. Guo and Agichtein [14] proposed a natural extension Rodden's work - to predict eye-mouse coordination (i.e., whether the mouse cursor is in close proximity to eye gaze at any given point in time). These works were further extended by Huang et al. [18] to directly predict the gaze position from mouse cursor movement, showing that the cursor and eye gaze are best aligned when the user is performing click action, and have the largest average distance in periods of cursor inactivity. Navalpakkam et al. [32] studied the coordination of cursor and gaze on non linear search result page layouts, e.g., in the presence of rich information panel on the right side. They showed that a non-linear regression model offers more accurate predictions of gaze position, and outperforms previous approaches. Also related to our work is the research of Diaz et al.[9], which proposed a two-dimensional attention transition model based on mouse cursor movement over grid layouts.

In this work, we build upon these ideas[18, 32, 9] and instead of predicting gaze position in isolation (without spatial constraints) we develop the first model that combines the evidence provided by interaction data with the information about Web page elements visible to the user through the browser's visible part of the page, or viewport. Furthermore, unlike in previous work in gaze prediction, we extend our model to predict attention on pages in other domains, moving beyond web search result pages.

3. MICS: COMBINING CONTENT AND INTERACTION SIGNALS

In this section we present our *MICS* model that allows us to more effectively infer user attention on Web pages by combining content and interaction signals. This is an even more challenging problem than predicting attention in images, as is done in computational visual salience research: Web pages contain extensive layout structure and multiple layers of meaning encoded in the text, layout, and metadata about a page.

To address this challenge, we exploit the observation that web pages, more so than image-only stimuli, can be effectively annotated with areas of interests (i.e., potential targets in the *top-down* models terminology), that can enable more accurate modeling of user gaze during web browsing or information seeking activities. Such annotations can be based on set of rules or rely on an automatic classifier to segment page elements that take part in the model. While the ultimate accuracy of the model is likely to depend on the quality of the page segmentation, for now let us assume that for popular types of Web pages (e.g., Search Engine Result pages or social media news feeds) such segmentation is available. The details of the particular page segmentation algorithm used in this work are provided in Section 4.1. Assuming a page segmentation is given to us, we can now define our Mixture of Interactions and Content Salience (MICS) model.

3.1 MICS: Definition

Our approach to modeling the allocation of user attention on a page is derived from the general idea of the *mixture of experts* model in machine learning [22]. Our goal is estimating the task-specific (top-down) element importance on the page, and then *refine* the prediction based on how long each element on a page was displayed to the user, and where it was in the viewport[1] and what interactions the user performed.

MICS operates by sub-dividing the visual space into regions - each corresponding to a particular Web page element. While the distribution of gaze positions within each element is determined only by the features of the element, the probability of attending a page element depends on relative attractiveness of *all the elements* displayed in the visible portion of the Web page. Intuitively, in our model each element "competes" for user attention against other visible elements on the page. Unlike previous approaches, which mainly use the visual stimuli information on pixel level (i.e., visual salience) to predict attention, our model takes advantage of the information about page element rendering (how elements are displayed by an Web browser to a user) and constructs compact, yet expressive, distribution of user attention in the browser viewport.

More formally, our model defines a probability distribution of gaze position over the visual space (browser viewport)

[1]We use the term *viewport* to denote the portion of a Web page visible to the user at given point of time.

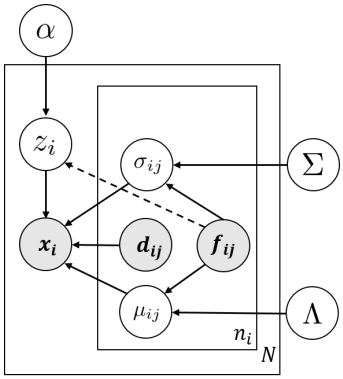


Figure 2: The *MICS* model for search attention modeling.

Variable	Description
N	number of gaze data points
n_i	number of Web page elements at i-th viewport
$\mathbf{x_i} \in \mathbb{R}^2$	i-th gaze position
$z_i \in \{1, .., n_i\}$	index of the Web page element being viewed at i-th viewport
$\mu_{ij} \in \mathbb{R}^2$	mean of the j-th element Normal distribution
$\sigma_{ij} \in \mathbb{R}^2$	variance of the j-th element Normal distribution
$d_{ij} \in \mathbb{R}^2$	position of the j-th element
p	dimensionality of element's feature space
$\alpha \in \mathbb{R}^p$	feature weights for the element importance distribution (z_i)
$\Lambda \in \mathbb{R}^{(p \times 2)}$	feature weights for the element means μ_{ij}
$\Sigma \in \mathbb{R}^{(p \times 2)}$	feature weights for the element variances (σ_{ij})
$\mathbf{f_j} \in \mathbb{R}^p$	feature vector of j-th page element.

Table 1: Summary of the notation used in the *MICS* model.

which is represented as a mixture of distributions - each corresponding to a particular web page element. This can be viewed as a particular type of a mixture of experts model (MoE, [22]), where each expert corresponds to a distribution representing the individual Web page elements. The reason MoE formulation is particularly well-suited to this setting is that it naturally manages uncertainty about the "attractiveness" of each element, which can be refined using additional features of the content element itself, or, later on, with interaction data.

MICS can also be viewed as a generative model. Figure 2 presents the *MICS* model diagram in plate notation. Table 1 defines the notation used in Figure 2. In the diagram, i stands for each data point, which consists of the set of elements and their locations on a page, visible at that time on the page, and the corresponding gaze position coordinates x_i. *MICS* states that the $i-th$ gaze position is generated from the observed element positions d_{ij} with their corresponding features $\mathbf{f_{ij}}$. The element's μ_{ij} parameters are defined as:

$$\mu_{ij}^{(x)} = d_{ij}^{(x)} + width_{ij} \cdot sigmoid(\Lambda^{(x)} \cdot \mathbf{f_{ij}})$$

where $\mu_{ij}^{(x)}$ is the horizontal component of element's Normal distribution mean parameter, $d_{ij}^{(x)}$ is the element's top left coordinate x, $width_{ij}$ is the width of the element, Λ is a free parameter estimated during training, and $\mathbf{f_{ij}}$ is vector of element's features. The element's variance parameter σ_{ij} is computed as:

$$\sigma_{ij} = exp(\Sigma \cdot \mathbf{f_{ij}})$$

where Σ is free parameter estimated during training. The probabilities of viewing an element are parametrized using the softmax function with the free parameter α:

$$P(z_i = j|\alpha, \mathbf{f}) = \frac{exp(\alpha \cdot \mathbf{f_{ij}})}{\sum_{j=1}^{n_i} exp(\alpha \cdot \mathbf{f_{ij}})}$$

3.2 MICS Training

To make the model training more tractable, we make a simplifying assumption that all gaze positions are generated independently from each other. This allows us to derive an efficient inference and learning algorithm. Our algorithm learns the element importance weights α for the *MICS* model as follows. Let the dataset $D = \{\mathbf{x_i}\}_{i=1}^{N_k}$ collection of N_k gaze positions for k-th page view. Note that depending on the scroll position of the browser, there could be a different number of elements visible in the viewport, we denote this number as n_i. We assume that information about position of page elements (d_{ij}) and their features f_{ij} is available.

In order to find plausible values for model parameters $\Theta = \{\alpha, \Sigma, \Lambda\}$ we perform maximum likelihood estimation That is we optimize log-likelihood of gaze observations given the model parameters:

$$\mathcal{L}(\mathbf{x_i}|\Sigma, \Delta, \alpha) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{n_i} P(z_i = j) \log \mathcal{N}(\mathbf{x_i}|d_{ij} + \mu_{ij}, \sigma_{ij}^2)$$

In order to optimize the log-likelihood we use Stochastic Gradient Ascent (SGA) method with learning rate annealing. The model is implemented using symbolic differentiation tool Theano[3] that automatically generates code for gradient computation.

3.3 MICS Inference

Once the *MICS* model is trained, gaze prediction distribution is computed as:

$$P(\mathbf{x_i}|\Sigma, \Delta, \alpha) = \sum_{j=1}^{n_i} P(z_i = j|\alpha, \mathbf{f_{ij}}) \mathcal{N}(d_{ij} + \mu_{ij}, \sigma_{ij}^2)$$

Note that $P(z_{ij} = j)$ gives us an importance weight (from 0 to 1) for each of page element d_{ij}. Thus, we could view it as a mixture distribution of n_i Normal distributions associated with the attractiveness and uncertainty predicted for each element, respectively. Computing the density of this distribution over a fixed grid of 2-dimensional points is tractable, as we demonstrate in the experiments section. Given the predicted density, the expected gaze position can be obtained by computing the maximum likelihood estimate of the x and y values under the predicted density distribution.

Since *MICS* is a generative model, for completeness we describe the generative process of how gaze positions could be generated by our trained model. The following generative process can be used to generate an i-th sample from our model:

1. Generate z_i with probability $P(z_i = j|\alpha, \mathbf{f_{ij}})$
2. Generate gaze position $\mathbf{x_i} \sim \mathcal{N}(d_{ij} + \mu_{ij}, \sigma_{ij}^2)$, where d_{ij} is position of j-th element on the screen,

In practice, in order to avoid computationally costly sampling procedure, we could obtain estimates of gaze positions by numerically computing expectation over the predictive density:

$$\bar{\mathbf{x}}_\mathbf{i} = \sum_{l=1}^{L}\sum_{m=1}^{M} \mathbf{x_{lm}} P(\mathbf{x_{lm}}|\Sigma, \Delta, \alpha, \mathbf{f_{ij}}, \mathbf{d_{ij}})$$

where L and M are number if integration points in $\mathbf{x_{lm}}$ are nodes in two dimensional grid used for computing the expectation.

Having defined the general *MICS* model, we now turn to the specific implementations of *MICS* to be validated on two increasingly difficult tasks, as described next.

4. MODEL IMPLEMENTATION

We now describe the specific implementation of the *MICS* model including Web page segmentation and content features that were used in this work.

4.1 Extracting Prominent Web Page Elements

Identifying most prominent Web page elements is not always a trivial task. Often, Web pages contain thousands of HTML elements, many of which are not even displayed to the user. As our goal is to model attention in presence of significant (visible and important) page elements, it is desirable to eliminate page elements that are unlikely to attract user attention, thus, considerably simplifying modeling complexity. To this end, our web page content analysis consists of first segmenting a web page into HTML DOM elements, then selecting a subset of the elements to consider, and finally extracting content features just from that subset. To take advantage of all Web pages in our dataset we employ both rule based segmentation, applied for frequent page types, and classifier based segmentation, applied for less frequent page types in our dataset. We would like to emphasize that this is just one of many ways to implement content element segmentation and other variations could be explored in future work to further improve performance.

For web pages that occur relatively frequently in our data, such Google search result pages or Twitter pages, we implement manually engineered segmentation. This is a common approach taken in previous work, and is applicable to a large and important subset of web pages which tend to share the same layout and page template.

For less frequent pages, we apply a supervised automatic classifier that for each web page *layout* element outputs a binary decision - whether this needs to be segmented or not. This makes our approach potentially applicable to a wider range of web pages. To perform this classification we use Gradient Boosting Decision Tree classifier (GBDT) [10]. The classifier uses page element's features to determine if element needs to be included or not. In order to train the this classifier we manually annotated page segmentation for 20 pages. Table 2 shows features used by our classifier. We utilized several types of information including the element's DOM Tree features (e.g. amount of links), the element's position information and size (e.g., width and height), as rendered by the browser at the time of page visit, and the element's style (e.g. visibility and text font size).

Figure 3 shows example of the page segmentation output for a Google search result page. While the granularity of the segmented elements varies for different page types, we see that the elements carrying most important content information are captured. The fact that such segmentation only eliminates page elements that are not displayed in the browser or used only for layout or formatting, simplifies the salience modeling in a sense that we do not need to account for thousands of elements in our model.

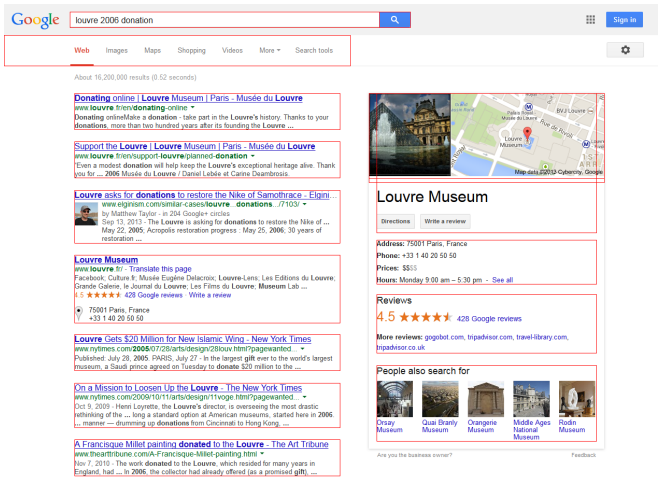


Figure 3: Example of page segmentation for a search result page (bottom of screen shot is cropped to fit).

4.2 Content and Interaction Features

We re-use content features employed by page segmentation algorithm (shown in Table 2, Content feature group). Our features encode information about element size, position on the page, style and font size, and simple information content measures such as number of words normalized by area. As discussed, additional more sophisticated content representation features could be invented, but in this reference implementation we opted for simplicity and generality. Despite the simplistic representation, the *MICS* model is able to use these features effectively, as shown in the experiments below.

MICS naturally allows to enrich the previously proposed regression models by allowing the features to be *element-specific*. For example, how *MICS* can exploit the information on how close is the mouse cursor to the particular element or whether a mouse cursor will hover over the element in the next few seconds. Such features allow the *MICS* model to learn cursor gaze coordination patterns not only on the overall behavior level, but on the element level as well. For example, if the mouse cursor hovers over the search box element, it is very likely that the user is going to reformulate the query terms, which implies the attention is focused on the search box. In contrast, if the cursor hovers over the elements located on the right side the search result page, it is less likely that the user's attention is following the cursor. Thus, to capture user interaction with the given element we include features that encode relative position of the cursor the element, cursor velocity, binary features indicating whether cursor is currently hovering the element or user is clicking on the element. Table 2 lists both Content and Interaction features used in our model. To account for a potential lag between interaction and eye gaze movement we concatenate features in the Interaction group at adjacent time d steps. The offsets for the adjacent time steps are $\{\pm 1, \pm 2, \pm 4, \pm 8, \pm 16\}$.

We train the *MICS* model using Stochastic Gradient Ascent algorithm with minibatch size = 100 and learning rate 0.001. To improve convergence speed we randomly shuffle

training examples before start of the training.

To obtain the predicted gaze position using the *MICS* model the we use expected $\bar{\mathbf{x}}$ under the predicted attention distribution as described in Section 3.3 with number of integration steps $L = M = 100$.

5. EXPERIMENTAL SETUP

In order to compare the effectiveness of different approaches for attention modeling, we performed a realistic user study, with eye tracking to collect the eye gaze data as ground truth. In the rest of the section we provide the details on data collection, the baseline models that were proposed in prior work, and the evaluation metrics used for the experiments in the next section.

5.1 User Study

In order to investigate the effects of domain and task on searcher attention, we systematically varied the *scope* of the search task, and the search *domain*. The tasks were modeled on the studies in [12, 33] to be representative of the common search tasks in common search domains. Specifically, the task (*scope*) was designated as either *Focused* or *Broad*. The *Focused* information need required the users to find specific information, e.g., "How many megapixels does Nexus 5 camera have?" in *Web Search* domain, while *Broad* information tasks had no specific answer, but rather asked the users to learn about a particular topic, e.g., "Learn what people on Twitter are saying about gay marriage" in the *Social Network* domain. Two focused and two broad tasks were performed by each user in each of the five common search domains: *Web Search* (Google), *Shopping* (Amazon), *Social Network* (Twitter), *News* (CNN) and *Wikipedia*, for a total of 20 tasks per user. We randomized the presentation order of the tasks to eliminate possible learning effects. To reduce the biases in the training data, we balanced the study design by ensuring that the same amount of data was collected for each (*domain*, *scope*) pair.

For the user study, we recruited 20 undergraduate and graduate students (11 of them males) from a major university. Each user was asked to perform four "warm-up" practice tasks to become familiar with the study flow, followed by the 20 tasks that we use in our analysis. All user actions, including query input, page navigation, clicks and mouse cursor movements were recorded using a custom extension to the Firefox internet browser. To capture the user's eye movements we used the Tobii T60 eye tracker system built into a 17" monitor with 1280×1024 screen resolution, recording eye gaze positions with frequency of 60 Hz. The Tobii system is head-free, where the participant could sit and interact with the computer naturally without being locked into a specific head position or body posture, making the collected interaction data more realistic. The eye movement data was pre-processed using Tobii Studio software to segment the data points into eye fixations (i.e., times of slow and detailed examination) and saccades (i.e., times of fast movement when the eye jumps to examine a new position).

5.2 Data

Overall, the data includes eye movement and interactions data for 2,890 page views, with 673 page views corresponding to the search pages. The eye gaze data contains 93,290 fixations. As in prior work [18, 32] we interpolated the gaze and cursor data every 100ms using nearest neighbor interpolation method, which resulted in 233,225 aligned eye gaze and cursor data points. Discretizing gaze data with fixed sample rate greatly simplifies prediction task by eliminating need to infer fixation duration [18, 32]. The dataset size is comparable to previously reported studies, and was primarily limited by the effort required to recruit and supervise the user study participants.

5.3 Baseline Models

Several works have attempted to infer the user's gaze position from cursor interactions. Most of the prior approaches trained a regression model to estimate the gaze position from cursor interaction features. We describe the two recent well known models that we use as state-of-the-art baselines for the subsequent experiments.

Linear Regression (LR). Huang et al. [18] proposed to directly predict a searcher's eye gaze position from mouse cursor movements on search result pages. They used a linear regression model to learn the relationship between eye gaze and cursor movement features. Their model can be formulated as:

$$\mathbf{x_i} = \langle \mathbf{w}, \mathbf{v_i} \rangle$$

where $\mathbf{w}$ is the vector of feature weights, $\mathbf{v_i}$ is vector of features for the i-th data point. During training the model computes an optimal vector of weights $\mathbf{w}$, such that the discrepancy between model predictions and the actual gaze positions is minimized. Specifically, the optimization minimizes the *squared error* between the actual eye gaze positions and the predicted positions.

Non-Linear Regression with Kernels (KR). Recently, a more sophisticated model of attention prediction from was introduced by Navalpakkam et al. [32]. Unlike the LR model, which assumes a linear relationship between cursor features and gaze position, the KR model is able to capture non-linearities of the data, but adding an additional transformation ϕ over the feature vector. The KR model is defined as:

$$f(\mathbf{v_i}) = \langle \mathbf{w}, \phi(\mathbf{v_i}) \rangle$$

where $\mathbf{w}$ is the vector of feature weights, $\mathbf{v_i}$ is vector of features for the i-th datapoint, and $\phi(\mathbf{v_i})$ is the Nystrom approximation of the Gaussian Radial Basis Function kernel matrix [36] with $n = 500$ basis vectors. This transformation is known as a "kernel trick" that helps capture non-linear dependencies between cursor features and gaze, while maintaining relatively easy training procedure. During model training we find the optimal set of vectors $\mathbf{w}$ so that it minimizes the discrepancy between the model predictions and gaze positions in the training data.

Baseline Interaction Features In our experiments we use the same basic cursor movement features for all models for a fair comparison. This set of features include and extend the published features used in prior work [18, 32]. These feature vectors are computed to represent each data point (mouse cursor position):

- Mouse cursor x and y positions
- The time since the page load
- The absolute values of cursor speed (vertical and horizontal), and movement direction (angle)
- The cursor distance traveled in the page up to this point
- The time and position of the most recent click on the page, if any

Group	Feature Name	Description
Content	Num{Links, Images,P }	Number of {a, img, p} tags in the given element
	IsTagName	Collection of binary features which equal 1 if the element's tag matches particular type (otherwise feature value is zero). The tags are $\langle a \rangle$, $\langle img \rangle$, $\langle p \rangle$, $\langle div \rangle$, $\langle span \rangle$, $\langle h1 - h3 \rangle$, $\langle em \rangle$, $\langle b \rangle$, $\langle li \rangle$, $\langle ol \rangle$, $\langle ul \rangle$
	Left, Top, Width, Height	Position and size information (in pixels)
	TimeOnPage	Time since the page load
	NumChildren	Number of child elements
	{Text, Image}Area	Total area of all {Text, Image} elements inside of the given element
	FontSize	Font size of the element's text
	TextToAreaRatio	Number of words (tokenized by white spaces) in the element divided by the element's area
Interaction	Cursor{X, Y}	Cursor position in pixels
	Speed{X, Y, Abs}	Horizontal, vertical and absolute speed of cursor movement
	Cursor{On, L, R, T, B}	Binary features indicating cursor position with respect to the element position (OnElement, Left, Right, Top, Bottom)
	CursorSame{Vert, Horiz}	Binary features indicating whether cursor position overlaps with the element vertically or horizontally
	DistX, DistY, DistEuclidean	Distance from the cursor to the element's center
	ClickOn	Non-zero if cursor click occurs on the element at given time step
	ClickDistX, ClickDistY, ClickDistEuclidean	Distance from the click position to the element center (zero if there is no click)
	TimeToScroll, TimeSinceScroll	Time since last scroll and time to the next scroll.
	OffsetFromScreenCenter{X, Y}	Vertical and horizontal offset of the element with respect to center of the viewport

Table 2: Content and interaction features used by MICS.

- The vertical scroll position, in pixels
- The time since the last scroll event, if any.

To account for longer range dependencies between the gaze and cursor movement for each time step we include features from previous time steps, logarithmically spaced, following the approach of [32]. The time step offsets were chosen as $\{\pm 1, \pm 2, \pm 4, \pm 8, \pm 16\}$, capturing the 100ms to 1.6 second "history" of the mouse movements. While this is a minor extension compared to previously proposed approaches, all the compared methods benefited from this additional contextual information.

5.4 Evaluation Metrics

For comparing the performance of the *MICS* model against the baseline LR and KR models, we use the root mean squared error (RMSE) and mean absolute error (MAE) metrics, used in prior work for this task.

More formally, given a sequence of true and predicted gaze positions $\mathbf{x}^{(\mathbf{i})}_{\mathbf{gaze}}$ and $\mathbf{x}^{(\mathbf{i})}_{\mathbf{pred}}$, RMSE is computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\mathbf{x}^{(\mathbf{i})}_{\mathbf{gaze}} - \mathbf{x}^{(\mathbf{i})}_{\mathbf{pred}}|^2}$$

where N is the number of gaze data points, $\mathbf{x}^{(\mathbf{i})}_{\mathbf{gaze}}$ is the actual gaze position at step i, and $\mathbf{x}^{(\mathbf{i})}_{\mathbf{pred}}$ is the predicted position, and the difference is the square of the Eucledian distance between the two. While RMSE is convenient from the optimization perspective (both LR and KR minimize the mean squared error, or MSE, on the training data), it dis-proportionally weights large errors. Therefore, we also consider mean absolute error, also used in prior work, which does not introduce this bias. The MAE is computed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\mathbf{x}^{(\mathbf{i})}_{\mathbf{gaze}} - \mathbf{x}^{(\mathbf{i})}_{\mathbf{pred}}|$$

where the sum is over the Euclidean distance between the actual and the predicted gaze positions.

To achieve more robust estimates of models' performance, all the experiments were performed with 3-fold cross validation (CV). Each of the metrics is computed as the average across the hold-out (test) folds.

6. RESULTS AND DISCUSSION

Table 3 summarizes prediction performance for the baseline models LR and KR and our *MICS* model, averaged across the hold-out samples, in the cross validation setting. *MICS* performs significantly better than LR and KR in all of the domains ($p < 0.001$, two tailed t-test). Reduction in error varies from 7% in *Shopping* domain to 35% in *Social Network* domain $RMSE$=237.8 px. The lowest prediction error was obtained in the *Social Network* domain ($RMSE$=237.8px, MAE=206.3px), while the *Shopping* domain appeared to be the most difficult to predict resulting in the highest error ($RMSE$=335.7 px, MAE=298.6px). We believe that the reason for the large performance improvements lie in the additional power available to the *MICS* model. Both LR and KR models make strong assumptions about the relationship between gaze and cursor interactions, relying on a constant bias term independent of the actual content shown to the user. Since user attention distribution heavily depends on what is shown the screen (e.g, see Figure 1), a constant bias that works for different types of pages may not exist. In contrast, *MICS* , by design, follows the content, and is able to supply a multi-modal predictive distribution dictated by the Web page elements visible to the user.

Interestingly, on the Web search domain, *MICS* also exhibits substantial reduction in error on the horizontal dimension (RMSE_x and MAE_x), making it even more appealing for evaluation – when search results may be shown to the right of the organic search results [32]. Such results attempt to provide users with direct answers to their information needs without requiring users to click. Previously, it has been proposed [32, 26] to utilize user attention for evaluation, providing a natural application of *MICS* for this task.

Our results demonstrate that it is possible to learn Web page element salience or attractiveness that is generalizable across different page types. This is even more encouraging

Web Page Domain	Method	$RMSE_x$	$RMSE_y$	RMSE	MAE_x	MAE_y	MAE
Web Search	LR	234.0	236.6	332.7 (N/A)	181.9	194.1	294.8 (N/A)
	KR	207.4	220.6	302.8 (-8%)	172.3	181.3	273.4 (-7%)
	MICS	156.1	202.6	**255.8 (-23%)**	128.8	160.1	**225.7 (-23%)**
News	LR	262.2	279.1	383.0 (N/A)	209.7	229.5	340.6 (N/A)
	KR	176.4	247.8	304.2 (-21%)	144.5	204.7	272.4 (-20%)
	MICS	174.5	208.3	**271.7 (-29%)**	138.1	167.0	**237.3 (-30%)**
Wikipedia	LR	219.7	242.0	326.8 (N/A)	173.6	195.5	288.1 (N/A)
	KR	290.0	272.5	398.0 (+22%)	242.5	223.9	360.8 (25%)
	MICS	87.2	277.4	**290.8 (-11%)**	70.2	210.9	**235.7 (-18%)**
Shopping	LR	249.1	259.2	359.5 (N/A)	196.3	211.2	319.5 (N/A)
	KR	281.5	285.9	401.2 (+12%)	225.6	231.1	359.3 (+12%)
	MICS	257.6	215.3	**335.7 (-7%)**	201.4	179.6	**298.6 (-7%)**
Social Network	LR	260.4	256.2	365.3 (N/A)	206.5	207.0	322.1 (N/A)
	KR	205.6	263.0	333.9 (-9%)	162.2	210.0	293.3 (-9%)
	MICS	146.9	187.0	**237.8 (-35%)**	113.3	146.7	**206.3 (-36%)**

Table 3: Predictions results for LR, KR and *MICS* , for different Web page domains. The *MICS* model consistently outperforms prior methods in all domains (differences in RMSE and MAE are significant p<0.001 with two tailed t-test).

since the Web search engines are constantly experimenting with various ways to improve user interface of search results and maintaining an attention model that can only work for a certain page configuration would severely impact its use cases. While *MICS* outperforms prior approaches in the gaze prediction task, it provides a general and principled way to integrate page content information into the attention model. The behavioral features allow MICS to make more sensible, time dependent predictions and capturing cursor-gaze coordination patterns.

6.1 Discussion

We have shown that *MICS* is able MICS to learn salience of Web page elements and to combine it with information about user interaction. In this section we highlight implications of this work, provide more intuition on why *MICS* is able to outperform previous models and discuss potential limitations of our approach.

We first analyze the contributions of the content-based vs. interaction-based signals to better understand the performance improvements of the *MICS* model. To this end, we performed a feature ablation experiment, where we compare the model variants while removing the corresponding feature groups (Table 4). We find that certain page types benefit from *Content* and *Interaction* to different extent. In *Web Search* and *Wikipedia* domains we find that both Content and Interaction feature groups are not particularly helpful independently of each other. *Interaction* features appear to be more important for *News*, *Shopping* and *Social Network* domains, where model performance drops substantially, compared to the model with ablated *Content* features. In all cases, the combination of both content and interaction features performs better than either signal alone.

To further understand *which* Web page elements *MICS* finds important we examined element importance weights given by $P(z_i)$. Figure 4a shows an example prediction for a part of a Social Network Web page. Red boxes boxes indicate the segmented Web page elements. Line width of the bounding boxes reflects the element's relative importance weight given by the *MICS* model. Hence, only a few elements stand out as important in Figure 4a. We see that the Twitter message displayed in the center of the page draws most of the attention. The next most prominent element on the page is another message, displayed towards the bottom of the page. In contrast, the "recommended users" feature elements receive much smaller weights - an indication that *MICS* is able to find elements which are likely to attract user attention. Another example is shown on Figure 4b. It visualizes the relative importance weights for a Wikipedia page. We see that text paragraphs near the center of the screen (on the right column) are the most prominent. Menu items and contents navigation blocks receive smaller weights. This is in agreement with prior work of Buscher et al. [5] that analyzed aggregated pattern of user attention on the screen. These examples illustrate that *MICS* identifies task-specific, salient page elements across different page layouts. These elements (as described in Section 3) are then used to construct a mixture distribution that helps predict user attention on any given viewport.

Weight	Feature
-0.821	DistEuclidean(dt=+16)
0.763	Height
0.720	CursorOn(dt=+8)
0.615	IsH3
0.483	NumP
-0.414	SpeedAbs(dt=0)
0.405	CursorSameVert(dt=+2)
0.402	CursorOn(dt=0)
0.383	SpeedY(dt=-4)
0.372	IsP

Table 5: Features with largest weights learned by the *MICS* model.

To further understand which element and interaction features *MICS* considers important with respect to "attracting" attention, we report the weights α assigned to the element features (Table 5). For interaction features we include the time step offset modifier. The DistEuclidean feature has largest negative weight, which reduces element importance when cursor moves further away (DistEuclidean increases) from the element in next 16 time steps which in our data translates to 1.6 seconds. Height is positively related with element importance and has weight of 0.763. Interestingly, CursorOn(dt+=8) has almost two times higher weight than 0.402 CursorOn(dt=0) – which capture element hover interaction. This is consistent with findings of Huang et al. [18] who reported a positive lag of 700ms between gaze and cursor positions. *MICS* assigns relatively high weights

MICS Features	Web Search		News		Wikipedia		Shopping		Social Network	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
All	**290.8**	**235.7**	**271.7**	**237.3**	**290.8**	**235.7**	**335.7**	**298.6**	**237.8**	**206.3**
– ContentSalience	331.8	284.2	275.3	244.0	331.8	284.2	338.6	302.8	240.8	209.9
– Interactions	333.3	283.8	305.8	269.2	333.3	283.8	363.1	327.4	267.8	236.0

Table 4: Regression performance of the *MICS* model when ablating *Content* and *Interaction* feature groups in MICS.

to features related to the element's tag (IsH3–header and IsP–paragraph). Regardless of the element importance, the weights decrease when the cursor starts moving (SpeedAbs has negative weight -0.414), indicating the attention is no longer on the element. Finally, *MICS* finds CursorSameVert to be positively related with the element importance.

This work can be potentially improved or extended in an number of ways. In particular, web page segmentation approaches (rule based and classifier based) used in the current implementation can be improved. That is, rule-based segmentation may only capture a limited set of pages with pre-defined HTML template or layout. Alternatives include other popular Web page segmentation approaches, such as [7], or using semantic relationship between Web page blocks [37], or a combination of visual, text and link information [6]. These more powerful segmentation mechanisms can be naturally incorporated into the *MICS* implementation with operates over whatever page elements are provided by the segmentation step. Another area of improvement is efficiency of the implementation. More accurate prediction performance comes at a price of higher computational complexity during training, compared to the baseline models. While LR and KR models enjoy the benefits of convex optimization and even allow closed form solutions (in a matrix form), optimization problem that comes with *MICS* model is inherently non-convex and requires application of iterative methods. In our experiments, *MICS* converged in about 10-20 iterations, which translate to 1-2 hours of run time using Theano [3] generated code on Tesla K20 GPU. At inference time, *MICS* incurs the additional cost of computing the expected gaze position through a numerical integration described in Section 3.3. Our results demonstrate that our optimization method (Stochastic Gradient Ascent) is able to successfully train a model that performs well on the evaluation metrics. As another promising future direction, Navalpakkam et al.[32] showed that gaze prediction error may be further reduced by personalization – i.e., by additionally tuning the set of vectors $\mathbf{w}$ for each user. While personalization is orthogonal to the ideas proposed in this paper, further personalizing the *MICS* model can be explored in future work.

7. CONCLUSIONS

We have introduced *MICS* , a robust and principled method for connecting interaction data with the underlying page content for predicting user attention. Results validated against eye gaze tracking data show that *MICS* is more accurate than previous state of the art models that consider interactions alone. We have shown that the cursor interaction features allow *MICS* to make more sensible predictions that capture cursor-gaze coordination patterns on specific Web page content. Importantly, *MICS* forces the most likely gaze position to be within, or in close proximity to, the prominent web page elements. This feature could potentially offer better gaze prediction performance for users who do not use the mouse pointer actively, or only to perform necessary actions. This could be particularly important for attention prediction in mobile phones and tablets [16, 26]. The *MICS* implementation as well as the user study data, including the eye gaze ground truth, will be made available to the research community[2].

Our work can be expanded in multiple directions. First, other complementary content features could be designed, capturing visual attractiveness, semantic concepts embedded in the text, or readability. Similarly, more complex behavior or mouse cursor movement features could be incorporated into the model, including cursor patterns discovered automatically, e.g., as in [23]. These richer features can be naturally incorporated into the *MICS* model to further improve prediction performance. In turn, accurately inferring search attention prediction from observable user behavior data, obtained at Web scale, could enable numerous improvements to the user experience in search and other online settings.

ACKNOWLEDGMENTS: This work was supported by the National Science Foundation grant IIS-1018321, the National Institutes of Health grant R01EB014266, the DARPA grant D11AP00269, and the Yahoo FREP program.

8. REFERENCES

[1] Mikhail Ageev, Dmitry Lagun, and Eugene Agichtein. Improving search result summaries by using searcher behavior data. In *Proc. of SIGIR*, 2013.

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

[3] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a cpu and gpu math expression compiler. In *Proceedings of SciPy*, volume 4, 2010.

[4] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proc. of CHI*, pages 21–30, 2009.

[5] Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye movements as implicit relevance feedback. In *CHI Extended Abstracts*, pages 2991–2996, 2008.

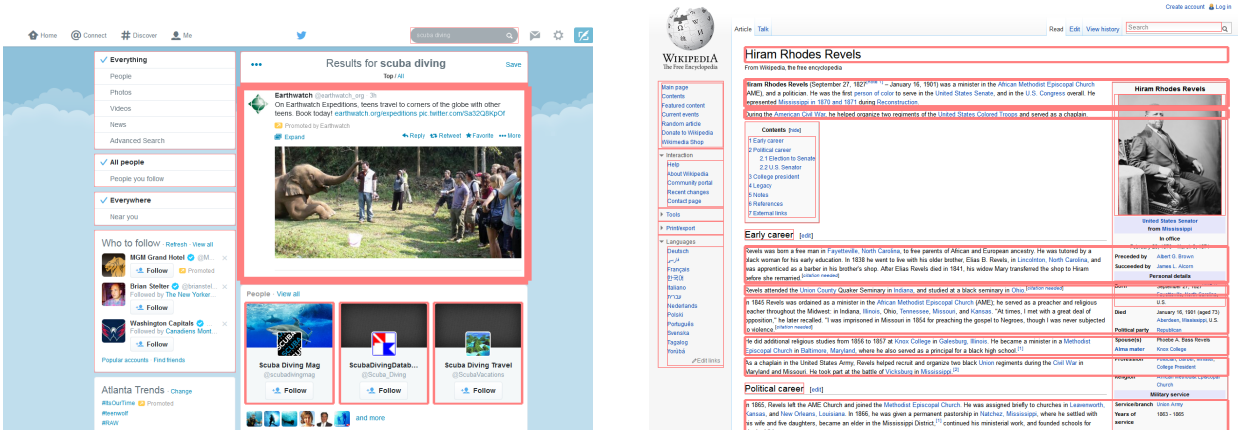
[6] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the ACM Int. Conf. on Multimedia*, pages 952–959. ACM, 2004.

[7] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Vips: A vision-based page segmentation algorithm. Technical report, Microsoft technical report, MSR-TR-2003-79, 2003.

[8] Mon Chu Chen, John R Anderson, and Myeong Ho Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI extended abstracts*, pages 281–282, 2001.

[9] Fernando Diaz, Ryen White, Georg Buscher, and Dan Liebling. Robust models of mouse movement on dynamic

[2]See `http://ir.mathcs.emory.edu/software-data/`



(a) Social Network Web Page

(b) Wikipedia Web Page

Figure 4: Example of element importance for user examination of different Web pages. Line thickness indicates the element's relative importance compared to all other elements within the viewport.

web search results pages. In *Proceedings of CIKM*, pages 1451–1460, 2013.

[10] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[11] Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proc. of SIGIR*, pages 478–479, 2004.

[12] Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proc. of SIGIR*, 2004.

[13] Qi Guo and Eugene Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *Proceedings of SIGIR*, pages 130–137, 2010.

[14] Qi Guo and Eugene Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI Extended Abstracts*, pages 3601–3606, 2010.

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

[16] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proc. of SIGIR*, pages 153–162, 2013.

[17] Marti Hearst. *Search user interfaces*. Cambridge University Press, 2009.

[18] Jeff Huang, Ryen White, and Georg Buscher. User see, user point: gaze and cursor alignment in web search. In *Proc. of CHI*, pages 1341–1350, 2012.

[19] Jeff Huang, Ryen W White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proc. of CHI*, pages 1225–1234, 2011.

[20] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[21] Laurent Itti, G Rees, and J Tsotsos. Models of bottom-up attention and saliency. *Neurobiology of Attention*, 582, 2005.

[22] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[23] Dmitry Lagun, Mikhail Ageev, Qi Guo, and Eugene Agichtein. Discovering common motifs in cursor movement data for improving web search. In *Proc. of WSDM*, 2014.

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

[25] Dmitry Lagun and Eugene Agichtein. Effects of task and domain on searcher attention. In *Proc. of SIGIR*, 2014.

[26] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proc. of SIGIR*, pages 113–122, 2014.

[27] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. Models of user engagement. In *User Modeling, Adaptation, and Personalization*, pages 164–175. Springer, 2012.

[28] Luis Leiva. Automatic web design refinements based on collective user behavior. In *CHI Extended Abstracts*, pages 1607–1612, 2012.

[29] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Proc. of JASIST*, 59(7):1041–1052, 2008.

[30] Vidhya Navalpakkam and Laurent Itti. A goal oriented attention guidance model. In *Biologically Motivated Computer Vision*, pages 453–461. Springer, 2002.

[31] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

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

[33] Bing Pan, Helene A Hembrooke, Geri K Gay, Laura A Granka, Matthew K Feusner, and Jill K Newman. The determinants of web page viewing behavior: an eye-tracking study. In *Procc. of ETRA*, pages 147–154. ACM, 2004.

[34] Kerry Rodden, Xin Fu, Anne Aula, and Ian Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI Extended Abstracts*, pages 2997–3002, 2008.

[35] Chengyao Shen and Qi Zhao. Webpage saliency. In *ECCV 2014*, pages 33–46. Springer, 2014.

[36] Alex J Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. 2000.

[37] Ruihua Song, Haifeng Liu, Ji-Rong Wen, and Wei-Ying Ma. Learning block importance models for web pages. In *Proc. of WWW*, pages 203–211, 2004.

[38] Benjamin Stone and Simon Dennis. Using lsa semantic fields to predict eye movement on web pages. In *Proc. CSS*, pages 665–670, 2007.

[39] Yuan-Chi Tseng and Andrew Howes. The adaptation of visual search strategy to expected information gain. In *Proc. of CHI*, pages 1075–1084, 2008.