Gaze Prediction for Recommender Systems

Qian Zhao, Shuo Chang, F. Maxwell Harper, Joseph A. Konstan
GroupLens Research
Dept. of Computer Science, University of Minnesota
Minneapolis, United States
{qian, schang, harper, konstan}@cs.umn.edu

ABSTRACT

As users browse a recommender system, they systematically consider or skip over much of the displayed content. It seems obvious that these eye gaze patterns contain a rich signal concerning these users' preferences. However, because eye tracking data is not available to most recommender systems, these signals are not widely incorporated into personalization models. In this work, we show that it is possible to predict gaze by combining easily-collected user browsing data with eye tracking data from a small number of users in a grid-based recommender interface. Our technique is able to leverage a small amount of eye tracking data to infer gaze patterns for other users. We evaluate our prediction models in MovieLens – an online movie recommender system. Our results show that incorporating eye tracking data from a small number of users significantly boosts accuracy as compared with only using browsing data, even though the eye-tracked users are different from the testing users (e.g. AUC=0.823 vs. 0.693 in predicting whether a user will fixate on an item). We also demonstrate that Hidden Markov Models (HMMs) can be applied in this setting; they are better than linear models in predicting fixation probability and capturing the interface regularity through Bayesian inference (AUC=0.823 vs. 0.757).

Keywords

eye tracking; Hidden Markov Models; grid-based interface

1. INTRODUCTION

Recommender systems research has experienced a transition from modeling user preferences based on explicit feedback [36, 38], e.g. **what users are rating** to preference modeling based on implicit feedback [25, 33], e.g. **what users are clicking**. Nowadays, it has been recognized that successful recommendations also need to take into account user perceptions of recommendation properties such as diversity and serendipity [23, 31], user short-term information needs, user context, and mood, i.e. **what users are thinking**.

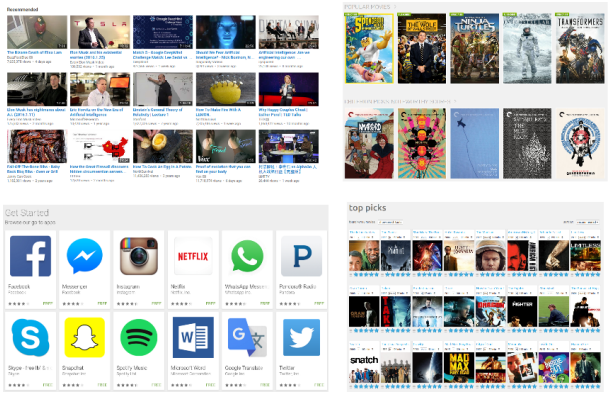

Figure 1: In this work, we predict user gaze in grid-based user interfaces. Above are four such layouts – YouTube (top-left), Hulu (top-right), Google Apps (bottom-left) and MovieLens (bottom-right).

Understanding how users look at a recommender system's content will enable further improvements to how systems model and react to user needs. As users browse a recommender system, they systematically consider or skip over much of the displayed content. It seems likely that these eye gaze patterns contain a rich signal concerning these users' state of mind. Indeed, early studies have shown the potential for improving recommender systems by incorporating eye tracking data [43, 34].

The ability to incorporate eye tracking data into a recommender system enables a variety of potential improvements. For example, recommender systems currently do not know which items are looked at and ignored versus simply not looked at. But this is a critical distinction – if the user looks and does not act, that *inaction* provides a signal that can be used to influence whether and when to display that item in the future, though interpreting whether such gazes represent interest or lack thereof may require context and further analysis.

Also, since recommender systems essentially provide decision support for users, having user gaze enables more nuanced studies on user high-level decision-making processes as demonstrated by researchers who study human decision theory through eye tracking [17].

The biggest challenge to incorporating user gaze data into recommender systems is a technical one: it requires eye

RecSys '16, September 15-19, 2016, Boston , MA, USA

DOI: http://dx.doi.org/10.1145/2959100.2959150

tracking technology, which is generally not available outside of specialized labs. It is possible that future systems will make gaze detection a common feature available to system builders, due to the ubiquitous presence of high-resolution user-facing cameras. It is also possible that eye tracking data will never be commonly used due to the privacy concerns that such low-level tracking raises.

To address these challenges, in this work we show that it is possible to model and predict user gaze without requiring the deployment of ubiquitous eye tracking technology. We predict gaze by combining easily-collected user browsing data with eye tracking data from a small number of users. Our technique is therefore able to leverage a small amount of eye tracking data to infer gaze patterns for other users.

In this research, we model gaze in the context of a grid-based user interface layout, which has become one of the most common user interface layouts in recommender systems. For example, this is the layout used in YouTube, Hulu, Google Apps, and MovieLens, as shown in Figure 1. We address the following three research questions:

- *RQ1: How accurately can we predict gaze on items in a grid-based interface?*
  - based on models trained with only **user browsing data**. (Specifically, item position in the interface, user dwell time on the page, and actions such as clicks, ratings and wishlistings on items.)
  - based on models trained with collected **eye tracking data** for a small number of users in addition to user browsing data.
- *RQ2: How is gaze distributed on different positions in a grid-based interface?*
- *RQ3: How does gaze prediction accuracy vary for different tasks or modes of usage?*

We make the following contributions in this paper:

- We show that incorporating eye tracking data from a small number of users significantly boosts gaze prediction accuracy as compared with only using user browsing data, even though the eye-tracked users are different from testing users;
- We demonstrate that Hidden Markov Models (HMMs) can be applied in this setting and that they are better than baseline models in predicting fixation probability and capturing the interface regularity through Bayesian inference.

2. RELATED WORK

Human attention theory. From cognitive sciences, two main mechanisms guide the selection of human attention: top-down and bottom-up (or endogenous vs. exogenous) processes [4]. We can volitionally focus our attention according to top-down task demands. On the other hand, our attention may be drawn by bottom-up salient stimuli. This dichotomy of attention is still under debate because it involves the fundamental question of seeing attention as a cause, as an effect or as a combination of both [37], which has significant implication for interactive applications such as recommender systems. Specific to visual attention, we focus on reviewing overt attention [16] here, i.e. gaze is directed to the attended location. Following Marr [30] and Itti's [27] seminal work, there has been plenty of research on modeling visual attention in a bottom-up approach, i.e. saliency prediction [29, 6]. However, researchers have started to criticize the over-emphasis on low-level saliency representation of visual input and develop new models of gaze allocation guided by top-down principles to account for complex natural vision [41]. Tatler et al. [42] showed that a model based solely on behavior biases and blind to current visual information can outperform a salience-based approach. Top-down task demands or regularities can be formalized with probability theories, especially Bayesian statistics. For example, Markov stochastic processes have been applied to model gaze transition behaviors, since it is intuitive to compare eye movements to random walkers. Ellis and Stark [14] developed a method to identify statistical dependencies in positions of eye fixations based on Markov matrices and found that there is statistical dependency among sequences of fixations independent of the physical placement of the points of interest. Further work built on this [20] modeled sequences of visual fixations as Markov processes and introduced a quantitative method to measure scanpath similarity based on character strings. Henderson et al. [22] demonstrated that it is possible to classify the task that a person is engaged in, i.e. cognitive states from their eye movements by multivariate pattern classification specifically naive Bayes. Haji-Abolhassani et al. [21] modeled eye trajectories as a noisy generative process centered on the foci of attention directed by cognitive processes using Hidden Markov Models. Our study builds on this top-down approach. We do not model the saliency of the displayed items on a page, but instead look at how high-level information of item position presented in the interface directs and regulates user gaze behavior.

Eye tracking and information retrieval. Joachims et al. [18] pioneered the investigation of user behavior in WWW search through eye tracking analysis. They found that a higher rank in search results attracts more attention and users do tend to scan the result list from top to bottom. In information retrieval, machine learning algorithms are used to learn the relevance between search queries and web URLs from implicit clicking feedback [1]. Joachims et al. [28] examined the reliability of this kind of feedback generated from clickthrough data using eye tracking and explicit relevance judgement. They concluded that clicks are informative but biased, i.e. the **position bias** because of search result presentation in a list layout. Following these findings, various user attention and browsing models are proposed to account for the bias in learning algorithms [8, 13, 9, 39] for information retrieval. Chapelle and Zhang [7] built and evaluated a dynamic Bayesian network model postulating explicitly examination or attention, action and satisfaction variables in addition to the observed clicking events. The temporal or dynamic aspect of the model lies in the assumption that users examine search results from top to bottom one by one, which is reasonable in a list layout interface and supported by previous work [18]. Because of the cost of eye tracking, researchers have worked on approximating gaze in two main approaches: gaze-contingent displays [12] or restricted focus viewing and predicting gaze with mouse positions [26, 19]. As an example, Buscher et al. [3] compared segment-level display time of search results with eye tracking and found that although it is much coarser, it works as well as eye-tracking-based feedback for re-ranking and

query expansion. Going beyond applications of information retrieval, Buscher et al. [2] worked on predicting gaze with web page location-based characteristics as input and generating a model that can be used to improve web page layout and design.

Eye tracking and recommender systems. Recommender system researchers have been using eye tracking in different ways. Since recommenders are considered an important decision support tool, Castagnos et al. [5] studied user decision making behaviors when purchasing products assisted by a recommender through eye tracking. They showed that users actively click and gaze at recommended products up to 40% of the time and consult the recommendation area more as they approach the end of the decision process. Another intuitive application based on eye tracking in recommenders is to infer preference or relevance from user gaze behavior. Xu et al. [43] proposed several algorithms to make recommendations relying on the attention time captured through commodity eye tracking as preference clues. Puolamaki et al. [34] combined eye movements and collaborative filtering [38] in proactive information retrieval tasks which is similar to a recommender and demonstrated its accuracy benefit in predicting whether a document is relevant. Recommender systems that rely on implicit feedback [25] could suffer from position bias as well, as demonstrated by Hofmann et al. [24] in simulated experiments. They examined this bias using different click models and showed how bias following these models would affect the outcome of recommender system offline evaluation based on implicit feedback data.

Two hypotheses for user gaze behavior in a grid. In a grid-based layout, it is likely no longer valid to assume examining results from top to bottom one by one any more, since there are two potential directions (horizontal and vertical) that users can direct their attention. We have two hypotheses regarding how users examine a grid: **F-pattern** [11] and **center effect** [40]. As pointed out by Tatler [40], observers have a tendency to fixate the center of the screen on computer monitors. They demonstrated the endurance of the central fixation bias irrespective of the distribution of image features, which implies that the center of a screen may be an optimal location for early information processing as learned by users. On contrary, because that a grid with rows and columns are different from an integral scene picture, the visual hierarchy might dominate the viewing behavior and users could exhibit a viewing pattern favoring the top and left sides [11], as suggested by the shape F going from top to bottom and left to right.

3. BUILDING MODELS FOR THE GAZE PREDICTION PROBLEM

We define a specific type of gaze prediction problem here – **Aggregated Fixation Prediction**. **Fixation** refers to the stationary period between saccades [16], in other words, the maintaining of visual gaze on the same location (we focus on predicting fixation here because in most human visual activities, we reply on fixations to take in visual information [16]). Consider a user browsing a page in a grid layout (examples shown in Figure 1), which has r rows and c columns and $r * c$ items in total. The problem is to predict **fixation probability**, i.e., whether the user has fixated, and **fixation time**, i.e., how long the user has fixated, on each of the $r * c$ items aggregating the entire browsing of the user on the page, given **item positions**, the user's **dwell time** on the page and the user's **actions** (e.g. rating, clicking or wishlisting) on some of the items. Note that the unit of prediction is for each displayed item in one page view.

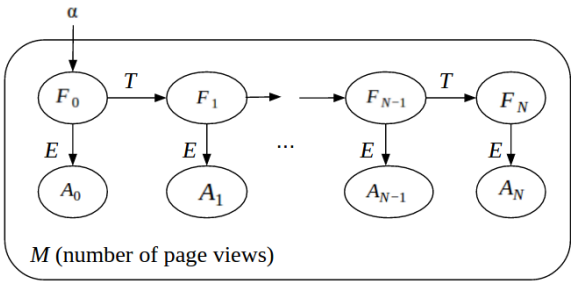

Figure 2: Graphical representation of a HMM in which F denotes fixation variable and A denotes action variable. α, T and E are parameters representing categorical conditional distributions defining the HMM. N is the length of the HMM sequence. For the gaze prediction problem, HMM gives inferred probability distributions of F when the values of A are observed.

3.1 Building Linear Models

We start with linear models to predict users' fixation. With access to a group of users' fixation data, we can build supervised machine learning models to predict future fixations for this group of users and even for other users.

For predicting fixation probability, we build a mixed-effect logistic regression model (taking into account the correlation among positions by using a random intercept for each page view) using the following three groups of features:

- *Position* features: row index, ranging from 1 to r, and column index, ranging from 1 to c;
- *Dwell time*: log transformed seconds spent on a page;
- *1/minActionDist*: the inverse of the minimum distance to actions on a page. This feature encodes the insight that users are more likely to fixate on surrounding items when they interact with an item on a page. For example, if a user acts upon (e.g. clicks) the second position, then it is likely that the user has fixated on the first or third position, which are small in terms of Euclidean distance calculated with row index and column index as the coordinates. Since there might be multiple actions, we take the minimum. We take the inverse of the minimum distance so that the coefficient of the feature is positive and the value of the feature equals zero with no action.

For predicting fixation time, we use the same set of features but a different model, a two-stage hurdle linear model [32]. This model handles zero inflation property in training data, i.e., users have not fixated on many displayed items, by modeling each data point first through a logistic regression and then through a zero-truncated negative binomial regression.

To simplify result presentation, we refer to linear models without using *1/minActionDist* feature as **linearModel**

Table 1: Different models for the gaze prediction problem, in which *bl* denotes baseline, *ub* denotes training only on user browsing data (or *training without fixation*) and *et* denotes training on eye tracking data (or *training with fixation*) as well.

Model	Model descriptions. Some depend on example action sequences a and b (nonzeros indicate the action position indices and 0 denotes no action)
	Action sequence a: 0 0 2 0 4 0 3 0 0 0 0 Action sequence b: 2 0 3 4 0 0 0 2 0 0 0
bl:simpleActionStats (simple action statistics)	Based on simple action statistics, in which page dwell time is distributed among the $r * c$ positions proportionate to the frequency of actions on that position from user action logs. They are used in predicting both fixation probability and fixation time.
ub:exactActionHmm (exact action approximation)	a does not contribute on the estimation of α, because it starts with no action. b counts once for initiating from position 2. a does not contribute on the estimation of T because there is no action transition from a non-zero position to another non-zero position. b counts once for the transition from position 3 to position 4.
ub:truncActionHmm (truncated action approximation)	After removing zeros, a will count for initiating from position 2 and transition from 2 to 4 and from 4 to 3 even though that is not exactly what have happened.
ub:RestExactActionHmm	Re-estimating using Expectation Maximization (EM, [10]) algorithm with exactActionHmm as initial values.
ub:RestTruncActionHmm	Re-estimating using EM algorithm with truncActionHmm as initial values.
et:linearModel	logistic regression model for predicting fixation probability or hurdle linear model for predicting fixation time; *1/minActionDist* feature is not used
et:linearModelActionDist	same as above except that *1/minActionDist* feature is used
et:eyeTrackingHmm	α, T and E are estimated by Maximum Likelihood Estimation.

and models using *1/minActionDist* feature as **linearModelActionDist** which are also summarized in Table 1.

3.2 Building Hidden Markov Models

We also use Hidden Markov Models, HMM for short, to predict users' fixation, as shown in Figure 2. For each page view, the dwell time is partitioned into small pieces of constant time intervals (the length of the interval is a parameter to tune). Each interval is associated with two variables: fixation F and action A. F takes $r*c$ possible values, representing the positions a user might fixate on the grid-based interface. We did not model no fixation (which could result from users' looking away or eye tracker's loss of gaze data. Time intervals with no fixation are removed from the observed F sequence) because it is less relevant for the prediction tasks. Given F, A can take $r*c+1$ possible values, i.e. one of the $r*c$ positions that a user acts upon plus the possibility of no action upon any position (we did not differentiate different types of actions). If multiple fixations are present in one time interval, we pick the one with longer fixation duration. If multiple actions are present (which rarely happens for small time intervals), we shift later actions to the following time intervals for which there is no action present or otherwise only use the first action. The parameters of this HMM are same for all users, i.e. the HMM is not personalized:

- **initiation vector** α (length of $r * c$) represents the categorical probability distribution of the initial fixation.
- **transition matrix** T (size of $r * c$ by $r * c$) represents the categorical transition probabilities from previous fixation to the next.
- **emission matrix** E (size of $r*c$ by $r*c+1$) represents the categorical action probabilities given the current fixation position.

With fixation data from some users' page views (using eye tracking), we can estimate the above parameters by maximizing the likelihood of observing both the fixations and actions. We refer to this model as **eyetrackingHmm** (summarized in Table 1 as well).

Without fixation data, we can still estimate the above parameters with observed actions as follows. Firstly, we estimate E with one assumption – users usually do not act upon items they are not fixating on. Therefore, We set small probability values (specifically, $10e-6$ in our algorithms) in E where positions of actions are different from positions of fixation. For the other case, we estimate the probability of acting upon f given that $F = f$ from frequency of actions on f in users' action logs. Secondly, we estimate α and T, by treating observed action as fixation, with the four algorithms named starting with *ub:* in Table 1.

Predicting based on HMM relies on the inferred *responsibility parameters*, i.e. posterior distribution $P(F_i|A)$, denoted by a matrix R (N by $r * c$), in which N is the length of the HMM sequence. S_i defined in Equation 1 is computed for predicting fixation probability and L_i defined in Equation 2 is computed for predicting fixation time. The rationale behind these formulas lies in that $P(F_i|A)$ tells how much *responsibility* fixating on a position takes for the observed action sequence.

$$S_i = \sum_{j=0}^{N} R_{j,i}, \text{ for } i = 1 \text{ to } r * c \quad (1)$$

$$L_i = DwellTime * \frac{S_i}{\sum_{k=1}^{r*c} S_k}, \text{ for } i = 1 \text{ to } r * c \quad (2)$$

4. METHODS

4.1 MovieLens and User Browsing Dataset

MovieLens (https://movielens.org) is a public online movie recommender service maintained by GroupLens Research at

the University of Minnesota. We focus on an interface called the **explore** page (which refers to a page with *explore* in its URL). It is a paginated three-rows-by-eight-columns grid-based layout presenting movie recommendations. Most of the explore page views are completely filled with $24(= 3 * 8)$ movie cards, as shown in Figure 1 (bottom-right). On each explore page, users can click a presented movie card, leaving the explore page and going to a movie detail page for that movie's details. Users can also rate or wishlist movies in a five-star-rating or wishlisting widget without leaving the explore page.

We use a dataset with one month of user browsing data from November 2015. For our purposes, we track page dwell time and ratings, clicks and wishlistings on movie cards. When a user leaves the explore page by clicking a movie card and directly returns, we count it as a continuing page view of that explore page, rather than a fresh new page view, so the dwell time accumulates through interruptions such as movie detail page views. In total, this data set has 102,039 page views with associated dwell times.

4.2 Eye Tracking Protocol Design and Dataset

We collected 17 subjects' gaze data using Tobii T60 Eye Tracker (0.5 degree accuracy, 60 Hz data rate, 17" screen size, 1,280x1024 resolution, roughly 65 cm viewing distance). These subjects are university students, including twelve males and five females, aging from 18 to 25 and majoring across eight disciplines. Subjects reported that they had watched five or more movies in the past two months with two exceptions: one had watched two movies and another had watched three. They had never used MovieLens before. We set up an account for each subject and asked them to perform the five tasks listed below which takes around 30 minutes after the eye tracker calibration procedure.

- *Task 1: Browsing for fun (five minutes).* This is for the subjects to get to know MovieLens features and obtain natural gaze and browsing behavior.
- *Task 2: Rate 15 movies.* Subjects were instructed to rate based on their preference on movies in a five-star rating widget.
- *Task 3: Find 10 movies you'd like to watch given a three-month holiday.* Subjects were asked to add those movies into their wishlist using the wishlisting feature.
- *Task 4: Find 5 movies you'd like to recommend to your friends.*
- *Task 5: Find 5 movies you'd like to recommend to a 12 years' old child.*

We directly used the fixation records generated by Tobii Eye Tracker (see its manual for details of the algorithms to compute fixation from raw gaze[1]). Each record has fixation duration and screen coordinates. To obtain fixations on the displayed movie cards, we recorded the movie card position coordinates, and tracked scrolling events as well to count for the position change. These positions are programmatically matched with the eye tracker's fixation records. Aggregating all 17 subjects, we collected 452 qualified page views (i.e. views of explore page completely filled with 24 movie cards.). Since the unit of prediction is with respect to each movie card display, we have $10{,}848(= 24 * 452)$ data points to use (Among them, we have 2304 for Task 1, 2760 for Task 2, 3960 for Task 3, 552 for Task 4 and 1008 for Task 5).

[1]http://www.acuity-ets.com/downloads/Tobii%20Studio%203.3%20User%20Guide.pdf

4.3 Evaluation

For each of the 17 subjects in the eye tracking study, we have their **true fixations** on each movie card in a page view, which is the **target variable** to predict. Prediction accuracy is measured with **AUC (Area Under the ROC)** for predicting fixation probability, i.e. to classify a displayed movie in a page view into being fixated or not and **MAE (Mean Absolute Error)** for predicting fixation time. The unit of MAE is in seconds.

Depending on whether using true fixation data or not to train models, the evaluation has two scenarios. In the **training-with-fixation** scenario, both fixation and browsing data from the 17 subjects in lab settings are used. We randomly pick 4 (around 20%) subjects and use their data for testing, while reserving the other subjects for training. This procedure is conducted multiple times (around 100 runs; a different set of subjects are picked each time) to compute variance of the metrics. In the **training-without-fixation** scenario, as its name suggests, only user browsing data is used for training, including both the one month dataset and user browsing data from the 17 subjects in lab settings. In order to be able to compare the accuracy between these two scenarios, testing phase of this scenario uses the exact same fixation data as the previous scenario.

5. RESULTS

RQ1: How accurately can we predict gaze on items in a grid-based interface? Figures 3 and 4 illustrate the accuracy of predicting fixation based on models trained with only user browsing data (the bottom five boxplots) and with eye tracking data (the top three boxplots). First of all, we see that there is a significant accuracy boost resulting from training on eye tracking data even though the training and testing users are different. Specifically, AUC increases from 0.693 for *exactActionHmm* to 0.823 for *eyeTrackingHmm* ($p \approx 0$) and MAE decreases from 0.466 for *simpleActionStats* to 0.332 for *linearModelActionDist* ($p \approx 0$). This result demonstrates that gaze patterns are consistent even across different users, and that our models capture these patterns very well.

In the *training-with-fixation* scenario, *eyetrackingHMM* performs significantly better than *linearModelActionDist* in predicting fixation probability ($AUC = 0.823$ vs. 0.757; $p \approx 0$). However, it performs worse in predicting fixation time ($MAE = 0.520$ vs. 0.332; $p \approx 0$). In the *training-without-fixation* scenario, *exactActionHmm* is much better than *simpleActionStats* ($AUC = 0.693$ vs. 0.580; $p \approx 0$) in predicting fixation probability (For an intuitive interpretation, Figure 5 shows the ROCs for one run of the evaluation procedure). But similarly, it has worse MAE (0.520 vs. 0.466; $p \approx 0$) in predicting fixation time. *RestExactActionHmm* and *RestTruncActionHmm* do not improve, which might be explained by overfitting to the action data set.

The above results show that HMM is more effective in capturing the interface regularity through Markov matrices and Bayesian inference in predicting binary-valued fixation vs. no-fixation, but is not very good at predicting real-valued fixation time. It might be explained by the choice of par-

tition granularity in HMM, since we have to decide on a time interval. We are using one second for all HMMs after exploring multiple choices. It might illustrate a general difficulty of a generative modeling approach such as HMMs compared with a discriminative modeling approach such as hurdle linear models, in which fewer assumptions have to be made. Actually, hurdle linear models have better accuracy than ordinary linear regression, poisson or negative binomial regression and random forest with the same set of features. Note that MAEs less than a second do not imply that predicting fixation time is an easy task. It could possibly result from the small range of the ground truth values, especially with many zeros. Instead, we found that it is hard to predict fixation time since with the best model we have, the prediction R^2(*coefficient of determination*) is 0.21. In other words, our model explains 21% of the variance in fixation time.

RQ2: How is gaze distributed on different positions in a grid-based interface? Figure 6 (drawn based on the mixed-effect logistic regression model; no significant interaction effects) illustrates user gaze behavior in a grid. It supports the *F-pattern* hypothesis, instead of *center* effect. Note that the fixation probability between either the first row and second row or the first column and second column is not significantly different. However, both the third row and third column have a significant drop ($p \approx 0$). Particularly, we omit the last column (index 7) because of data collection problem. The Tobbi eye tracker has relatively smaller screen size which leaves part of the movie card in the last column out of view. This however does not affect the conclusion for this research question. More interestingly, we found that for all positions dwell time is positively associated with fixation probability and when reaching 60 seconds, different positions on average have a very high probability (> 0.80) of being fixated.

RQ3: How does gaze prediction accuracy vary for different tasks or modes of usage? From Figure 7, we see that Task 3 – *finding ten movies for self* – has the best accuracy in predicting fixation probability ($AUC = 0.842$, $p \approx 0$). Since more data is collected for Task 3, it partially explains the accuracy advantage. Another possible explanation is that the process postulated by HMM particularly fits better to subjects' gaze behavior when engaging in this task. On contrary, as shown in Figure 8, the accuracy suffers most in

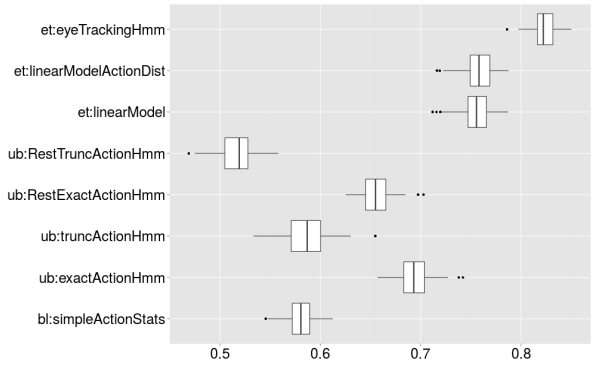

Figure 3: AUC boxplots for different models in predicting fixation probability. Higher scores are better. See Table 1 for descriptions of the models.

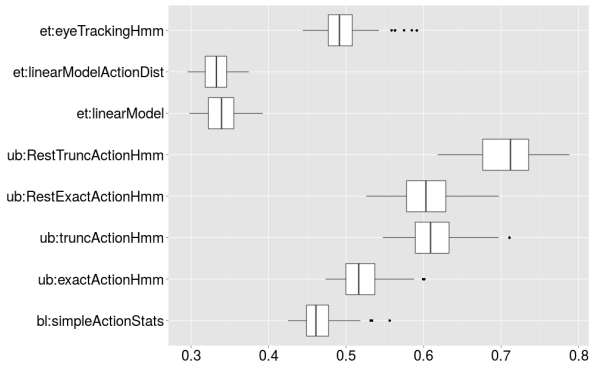

Figure 4: MAE boxplots for different models in predicting fixation time. Lower scores are better. See Table 1 for descriptions of the models.

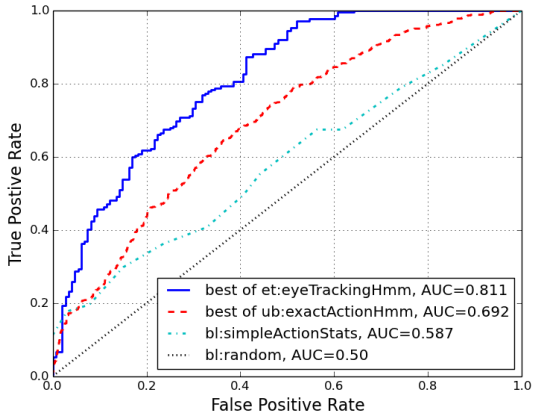

Figure 5: ROCs in classifying displayed movie cards into being fixated or not in a page view. It is from one run of the evaluation procedure.

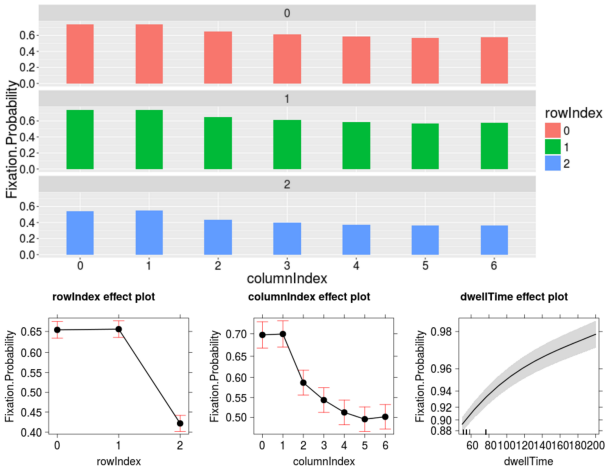

Figure 6: Fitted probabilities for different positions and the effects plot of *position* feature and *dwell time* in predicting whether a displayed movie is fixated using logistic regression. No significant interaction is found.

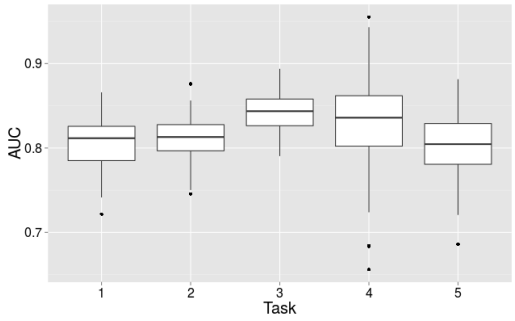

Figure 7: AUC boxplots of the best model *et:eyeTrackingHmm* in predicting fixation probability for the different tasks.

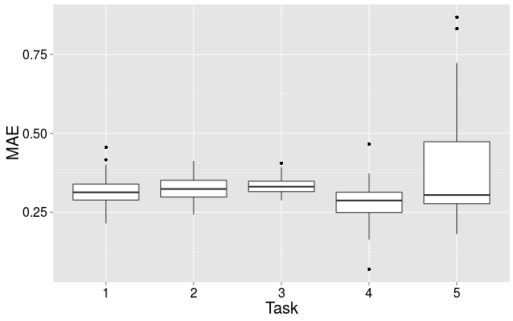

Figure 8: MAE boxplots of the best model *et:linearModelActionDist* in predicting fixation time for the different tasks.

predicting fixation time for the *finding-movies-for-children* task ($MAE = 0.340$, $p = 2.92e-08$). Subjects' gaze behavior shows substantial difference in this task from the video-recorded eye movements. Their fixations are shorter and more scattered, probably because subjects' searching strategy changed to coarser-level information scanning since most of the displayed items are not relevant anymore. Note that the better accuracy ($p \approx 0$) for Task 4 might just result from low variance in the data because we may not have enough data points for it. The general conclusion is that user gaze behavior is different in different usage modes and collecting and training on a specific task is better, especially for system designers who have knowledge about the main task that their users are engaged in.

6. FUTURE WORK

Our gaze prediction techniques imply two direct practical applications in recommender systems. First, they could be used to improve recommendation freshness. We can predict which items the user has paid attention to repeatedly without action, and replace those items with new recommendations. Second, they could be used to remove potential position bias in preference modeling with implicit feedback [24]. We have tried to directly use the predicted fixation probability to weigh the click-through observations in a matrix factorization model and achieved some accuracy improvement in predicting clicks under certain conditions. It does not always work because position bias is usually confounded with the typical relevance or preference of items shown at that position [7]. It is not straightforward to disentangle those confounding factors. A unified model on both gaze and preference may be necessary instead of simple weighting.

We envision two kinds of extensions to HMMs used in this work. First, it is possible to consider individual-level, in addition to current global modeling if a user has enough page views. Second, the fixation duration on a position is modeled implicitly through state self-transitioning, which essentially assumes that the duration follows a geometric distribution [35]. This assumption may not be valid especially when more factors are introduced such as preferences to explain fixation duration. Hidden Semi-Markov Models (HSMM) have been proposed to account for it and successfully applied in speech recognition [15]. Applying HSMM in our setting is a promising future direction.

7. CONCLUSION AND CONTRIBUTION

We conduct initial research on modeling and predicting gaze in recommender systems with a grid-based interface. We apply HMM in this setting and achieve significant accuracy improvement in predicting fixation probability. We also show that incorporating eye tracking data from a small number of users into the model training significantly boosts accuracy compared with only using normally logged user browsing data, even though the eye-tracked users are different from testing users. User gaze behavior follows an *F-pattern* rather than showing a *center effect* in a grid-based interface. In addition, we find that user gaze behavior is different in different usage modes which suggests that collecting and training on a specific task is better, especially for system designers who have knowledge about the main task that their users are engaged in.

8. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under grant IIS-1319382, and by Google under a Social Computing Focused Research Award.

9. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR'06*, pages 19–26. ACM, 2006.

[2] G. Buscher, E. Cutrell, and M. R. Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *CHI'09*, pages 21–30. ACM, 2009.

[3] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: a comparison to eye tracking. In *SIGIR'09*, pages 67–74. ACM, 2009.

[4] T. J. Buschman and E. K. Miller. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315(5820):1860–1862, 2007.

[5] S. Castagnos, N. Jones, and P. Pu. Eye-tracking product recommenders' usage. In *RecSys'10*, pages 29–36. ACM, 2010.

[6] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *NIPS'08*, pages 241–248, 2008.

[7] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *WWW'09*, pages 1–10. ACM, 2009.

[8] Y. Chen and T. W. Yan. Position-normalized click prediction in search advertising. In *KDD'12*, pages 795–803. ACM, 2012.

[9] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM'08*, pages 87–94. ACM, 2008.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[11] S. Djamasbi, M. Siegel, and T. Tullis. Visual hierarchy and viewing behavior: An eye tracking study. In *Human-Computer Interaction. Design and Development Approaches*, pages 331–340. Springer, 2011.

[12] A. T. Duchowski, N. Cournia, and H. Murphy. Gaze-contingent displays: A review. *CyberPsychology & Behavior*, 7(6):621–634, 2004.

[13] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR'08*, pages 331–338. ACM, 2008.

[14] S. R. Ellis and L. Stark. Statistical dependency in visual scanning. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 28(4):421–438, 1986.

[15] J. D. Ferguson. Variable duration models for speech. In *Proc. Symposium on the Application of HMMs to Text and Speech*, pages 143–179, 1980.

[16] J. M. Findlay. Active vision: The psychology of looking and seeing. 2014.

[17] M. G. Glaholt and E. M. Reingold. Eye movement monitoring as a process tracing methodology in decision making research. *Journal of Neuroscience, Psychology, and Economics*, 4(2):125, 2011.

[18] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR'04*, pages 478–479. ACM, 2004.

[19] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI EA'10*, pages 3601–3606. ACM, 2010.

[20] S. S. Hacisalihzade, L. W. Stark, and J. S. Allen. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(3):474–481, 1992.

[21] A. Haji-Abolhassani and J. J. Clark. A computational model for task inference in visual search. *Journal of vision*, 13(3):29–29, 2013.

[22] J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk. Predicting cognitive state from eye movements. *PloS one*, 8(5):e64937, 2013.

[23] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM TOIS*, 22(1):5–53, 2004.

[24] K. Hofmann, A. Schuth, A. Bellogin, and M. De Rijke. Effects of position bias on click-based recommender evaluation. In *ECIR'14*, pages 624–630. Springer, 2014.

[25] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *ICDM'08*, pages 263–272. Ieee, 2008.

[26] J. Huang, R. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. In *CHI'12*, pages 1341–1350. ACM, 2012.

[27] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI'98*, (11):1254–1259, 1998.

[28] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR'05*. Acm, 2005.

[29] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV'09*, pages 2106–2113. IEEE, 2009.

[30] D. Marr, T. Poggio, E. C. Hildreth, and W. E. L. Grimson. *A computational theory of human stereo vision*. Springer, 1991.

[31] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI EA'06*, pages 1097–1101. ACM, 2006.

[32] J. Mullahy. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365, 1986.

[33] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang. One-class collaborative filtering. In *ICDM'08*, pages 502–511. IEEE, 2008.

[34] K. Puolamäki, J. Salojärvi, E. Savia, J. Simola, and S. Kaski. Combining eye movements and collaborative filtering for proactive information retrieval. In *SIGIR'05*, pages 146–153. ACM, 2005.

[35] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[36] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *CSCW*, 1994.

[37] C. Roda. Human attention and its implications for human–computer interaction. *Human Attention in Digital Environments*, 1:11–62, 2011.

[38] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW'01*, pages 285–295. ACM, 2001.

[39] R. Srikant, S. Basu, N. Wang, and D. Pregibon. User browsing models: relevance versus examination. In *KDD'10*, pages 223–232. ACM, 2010.

[40] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4–4, 2007.

[41] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5, 2011.

[42] B. W. Tatler and B. T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054, 2009.

[43] S. Xu, H. Jiang, and F. Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *RecSys'08*. ACM, 2008.