What you have read and what you need to read: understansing daily learning process

Lei Zhang
Georgia State University
Atlanta,Georgia,U.S.A
lzhang14@student.gsu.edu

Ying Zhu
Georgia State University
Atlanta,Georgia,U.S.A
yzhu@cs.gsu.edu

ABSTRACT
In this paper, we present our research on detecting habitual reading region on the screen when a user is doing his daily online reading. Instead of traditional eye-tracking based solution, we developed an unintrusive user attention tracking(UUAT) solution. In UUAT, we developed a Reading-Clue technology, which helps to pinpoint user reading attention when necessary. At the same time, user reading behavioral data(mouse click and mouse scroll) is collected in background. With these behavioral data, UUAT calculates a user's preferred reading range. Our experiment indicates that the accuracy of UUAT attention tracking is as high as 93.2%.

Author Keywords
Attention Tracking; Reading; Mouse movements; Scrolling;

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

See: `http://www.acm.org/about/class/1998/` for more information and the full list of ACM classifiers and descriptors. Optional section to be included in your final version, but strongly encouraged. On the submission page only the classifiers letter-number combination will need to be entered.

INTRODUCTION
Personal Data Mining[8](PDM) has been proposed as a complement to commercial/business data mining, which aims to get useful knowledge to improve business opportunity or products popularity. Unlike the concept of commercial data mining, PDM is proposed for a users individual good. In PDM, user behavioral data is collected by a neutral utility. The user can choose third-party tools(local) or service(cloud) to analyze his personal data, aiming to improve individual productivity or quality of life.

As the first step of PDM, Self-Tracking[9] was proposed to log personal data, so that future analysis can be conducted for a users good in health. Self-Tracking has been mainly involved in personal physical data, such as, weight, heartbeat, diet and sports. Moreover, the way of logging data has been mainly referred to manual logging, e.g. a user input his weight before going to bed everyday. In this paper, we argue that the reign of personal data is more than physical data and the ways to collect personal data can be extended to an automatic and stealthy way.

In this research, we consider a users daily reading as a source of personal data. An immediate reason is that to understand a users daily reading helps to build a users personal knowledge demography, let alone For many self-trackers, the goal is unknown[10]. We do not consider what a user have read as static data, instead, we consider it as dynamic. For example, if a user spend more unit reading time on paragraph 1 than average, this might indicate that the user might have pondered on the information in paragraph 1 or have reading struggles on paragraph 1. In either case, it might indicate that the user has encountered new knowledge and he takes time to digest this new information. From this point of view, it might be a mileage for personal knowledge acquisition, e.g., today is the day I first come to know the concept of shale gas, although I have read the same phrase years ago. To log this type of personal data, it is unlikely to adopt the same user-aware way, a more intelligent and stealthy method is required to obtain personal reading data.

In this research, we focus on how to collect objective and detailed reading behavioral data, so that the PDM can be conducted on an accurate and correct basis. To be specific, we focus on a users daily online reading activity, our goal is to accurately identify what texts a users read that day and to what extent he spent on specific parts of an article. To realize this goal, we developed a browser-based solution to obtain a users preferred reading scope on a screen, by analyzing the users mouse/keyboard input, we provide details on users dwell time on each information piece(currently our granularity is paragraph-based). By doing this, we enable a more extensive PDM applications which we will discuss in future work of this paper. Our contributions are as follows: 1.We developed an almost UnIntrusive User Attention Tracking utility, which helps to get a users preferred reading scope in a stealthy way. 2. We developed a solution to get a users daily reading data which can be used as input for deep personal data mining.

RELATED WORK
The idea of quantified self-tracking is well known proposed in [10]. In [10], the author advocates collecting individual quantified data. He presents cases where individual data collection helped to solve person-specific problems. For example, Barbier used her personal data to find a way to cure her insom-

nia[10], Seth Roberts, a Caltech professor, analyzed his personal data and find an optimum diet(flaxseed oil) to improve his math performance[10]. In this paper, the author advocates a diversified personal data collection, as long as the data is quantified, albeit you might not know the goal of your data collection immediately.

Since the current personal data mainly includes personal physical data, such as weight, glucose, heartbeat, sports, food/medication consumption. The pervasive methods of personal data collection are mainly by a user himself, such as the methods adopted by PatientsLikeMe[1].

Our proposed research is very similar to the work in the field of computer human interaction(HCI). To get a users preferred reading scope on a screen, there are mainly two categories of solutions, method of eye-track based and method of user input tracking. Although having the advantages of highly goal-revealing and providing more details, eye-tracking[5] methods are also tagged as intrusive and expensive. It is impractical to popularize an eye-tracking based application to a large scale, in fact, the eye-tracking method are mainly applied in the software usability test.

Compared to eye-tracking method, our proposed research is closer to the method of user input tracking [7, 4, 6, 2, **?**]. The authors of [7] collected user mouse data, at the same time they collected users eye gaze data, they proved that the user mouse data can indicate user behavior, e.g. which part of the screen is more attractive, the user distraction behavior can be detected and the user experience can be predicted by mouse analysis in an accuracy of 80By literature review we distinguish our research from existing works as follows: the data collection in our research is designed to be more scalable, so our data collection method can be tagged as ”unintrusive”, unlike the eye-gaze technology adopted in [7, 6]. Meanwhile our goal is personal data collection instead of usability test, we proposed a new CHI technology to accurately infer users preferred reading range on the screen, so that we can collect high quality data that can describe a users daily reading behavior objectively.

WHERE DO YOU READ?

To understand online reading in an inattentive way, it is necessary to determine a user’s Reading Range(RR). A Reading Range is defined as the screen area where a user’s reading attention covers. Only those texts in the RR are possibly read and learnt by a user. In this section we will first model RR and introduce our method to compute behavioral RR and realtime RR.

Model of Reading Range

The previous research on user attention tracking [3] have disclosed some facts when a user is doing daily reading on articles or document papers. Vertically, in most of the time, the reading progresses towards the end of the screen in a line-by-line manner. However, the reading does not start from the first line on the screen until the last line of the screen. Generally, a user has a preferred reading range on screen, which in this paper, we name it as behavioral RR(BRR), as indicated in figure 1. A BRR can be identified by two parameters Δ_0 and w. Δ_0 indicates the offset from the head of the screen and w indicates the vertical span of the whole region. If a user wants to read the contents outside the RR, a scroll action is expected. In this research, we take aforementioned assumption and model the user attention range as displayed in figure 1.

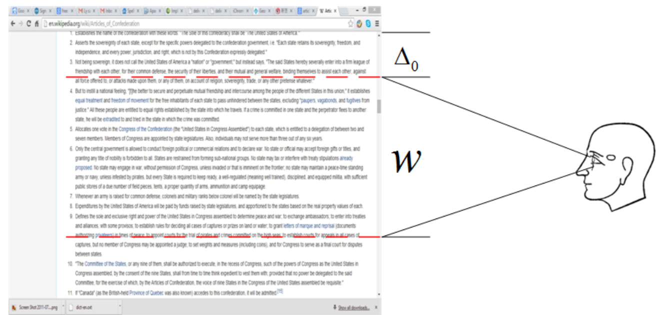

Figure 1. The preferred reading region on the screen

Taking this model into consideration, our next research question is: What behavioral data should be collected and analyzed to compute RR?

Eye tracking data can be easily mentioned to answer the question, especially when a ”visual” reading range is the goal. However, we argue that the eye tracking data has twofold major drawbacks: 1. The eye tracking collection requires extra hardware, which means the eye tracking based solution is not scalable and very costly. 2. What is even worse, eye tracking is ”intrusive”. A subject person is aware that he is wearing a headset device and he is under a camera coverage, his behavior is very likely to be different from how he behaves when he is doing reading alone. For these two reasons, we can not use eye tracking data. To meet the requirement of inattentive but objective, the only option left is the user interaction data collected when the reading is ongoing. In the next subsection we will discuss how to collect and analyze user interaction data to get the user preferred region.

Calculating BRR by Interaction Data

When a user is doing daily reading, since the keyboard input rarely happens during the reading process, we are more interested in the mouse-generated data. There are two categories of mouse-generated data, mouse move(mouse click) and scroll. We consider the mouse-generated data to be informational behavioral data. To be specific, a mouse click indicates the position of user’s reading attention at that time instant, a page scroll indicates how much contents has been moved out of reading region. In this subsection, we answer the question of how to infer BRR by mouse-generated data.

Now we examine a typical online reading event, during which a user finishes reading of a whole article. We suppose the article is relatively long, which requires many scroll action. The whole reading process can be illustrated by figure 2

In figure 2, a user conducts a scroll action when he finishes reading contents of current reading region. At ith scroll action, a Δ_i of page height will be moved out of region. So when the article reading is finished, we have equation (1). In

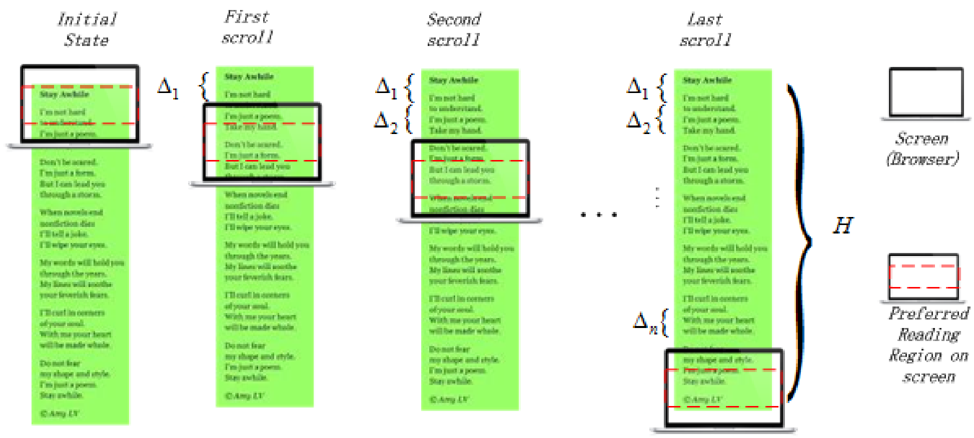

Figure 2. Reading of a long article in a browser

equation (1) , Δ_0 and w are defined in figure 1. Ideally, equation (2) should hold and the parameter w can be calculated in equation (3). In practice, it is unlikely for equation (2) to strictly hold, however, equation (1) can still be applied.

$$H = \sum_{i=1}^{n} \Delta_i + \Delta_0 + w \tag{1}$$

$$\Delta_1 = \Delta_2 = ... = \Delta_n = w \tag{2}$$

$$w = \frac{\sum_{i=1}^{n} \Delta_i}{n} \tag{3}$$

Since equation (2) does not always hold, if the variance of Δ_i is relatively large(actually it is), we need to consider the question how to infer user reading region on each single scroll. Figure 3 displays a the events happened during reading between two scroll actions. Here red dots are collected mouse click data with the position on the corresponding place on the screen. Since we take the assumptions that the reading process progresses line by line and the mouse click is an indication of instantaneous user attention, we can calculate a minimum reading range after $i_t h$ scroll action, $w_i^{\min}$, which can be identified by highest click and lowest click(as indicated in figure 3).

However, $w_i^{\min}$ is a minimum approximate of w_i, we can have a more accurate approximate if we take the neighboring scroll actions into consideration. In figure 4, we put the click data both before ith scroll(purple dots) and after $i+1$th(yellow dots). We can see there is a gap between two neighbor minimum approximates($g1$ and $g2$ in figure**??**). So we can have a moderate estimate of w_i:

$$w_i = w_i^{\min} + (g_1 + g_2)/2 \tag{4}$$

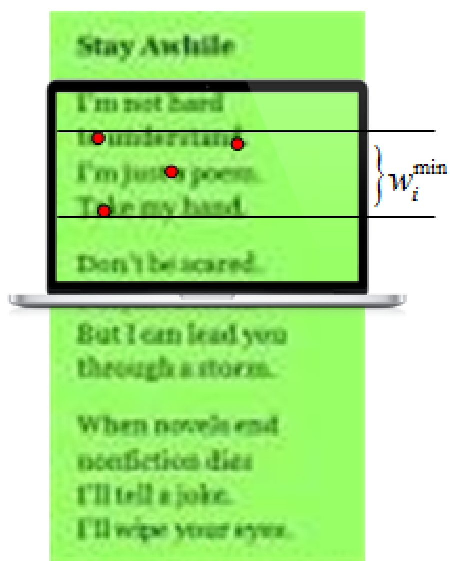

Figure 3. Mouse click and reading region

WHAT AND TO WHAT EXTENT YOU HAVE READ

Once we compute the BRR and RRR, we can conduct reading contents collection. In order to formulate our solutions here, we introduce the following notations. Given two points $p_1(x_1, y_1)$ $p_2(x_2, y_2)$ on a screen at time instant t, the context identified by the two points can be notated as : $C(t, p_1, p_2)$, which is indicated as yellow background texts in figure 5

Intuitively,

REFERENCES

1. Patients Like Me, howpublished = `http://www.patientslikeme.com/`, note = Accessed: 2010-09-30.

2. Atterer, R., and Schmidt, A. Tracking the interaction of users with ajax applications for usability testing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2007), 1347–1350.

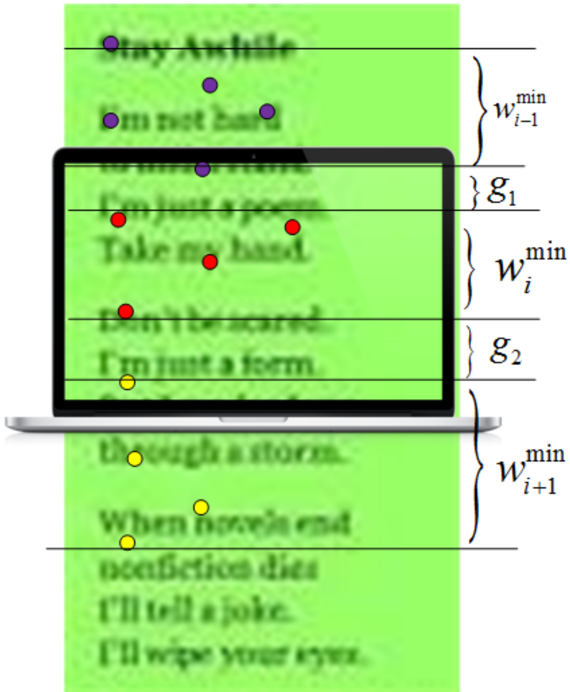

Figure 4. Mouse click and reading region

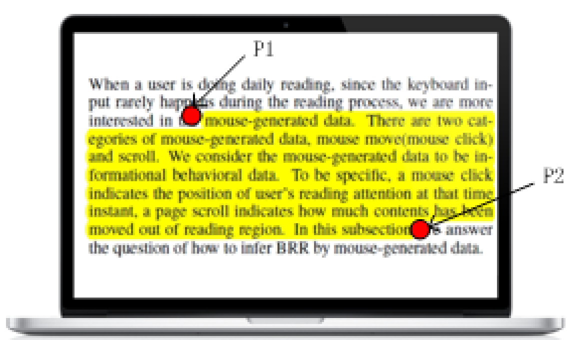

Figure 5. Contexts between points

3. Buscher, G., Biedert, R., Heinesch, D., and Dengel, A. Eye tracking analysis of preferred reading regions on the screen. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, ACM (2010), 3307–3312.

4. Huang, J., White, R. W., and Dumais, S. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2011), 1225–1234.

5. Jacob, R. J., and Karn, K. S. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind 2*, 3 (2003), 4.

6. Lagun, D., and Agichtein, E. Viewser: Enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, ACM (2011), 365–374.

7. Navalpakkam, V., and Churchill, E. Mouse tracking: measuring and predicting users' experience of web-based content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 2963–2972.

8. Ozzie, R. E., Gates III, W. H., Flake, G. W., Bergstraesser, T. F., Blinn, A. N., Brumme, C. W., Cheng, L., Connolly, M., Dani, N. V., Glasgow, D. A., et al. Personal data mining, Apr. 19 2011. US Patent 7,930,197.

9. Swan, M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *International journal of environmental research and public health 6*, 2 (2009), 492–525.

10. Wolf, G. The data-driven life. *The New York Times 28* (2010), 2010.