Browsing Regularities in Hedonic Content Systems

Ping Luo,[1] **Ganbin Zhou,**[1,2] **Jiaxi Tang,**[3*] **Rui Chen,**[4*] **Zhongjie Yu,**[5*] **Qing He**[1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. {luop@ict.ac.cn}
[2]University of Chinese Academy of Sciences, Beijing, China.
[3]School of Computing Science, Simon Fraser University, Cannada.
[4]School of Information, University of Michigan, USA.
[5]Deparment of Statistics and Finance, University of Science and Technology of China, China.

Abstract

Various hedonic content systems (e.g. mobile apps for video, music, news, jokes, pictures, social networks etc.) increasingly dominates people's daily spare life. This paper studies common regularities of browsing behaviors in these systems, based on a large data set of user logs. We found that despite differences in visit time and user types, the distribution over browsing length for a visit can be described by the *inverse Gaussian* form with a very high precision. It indicates that the choice threshold model of decision making on continuing browsing or leave does exist. Also, We found that the stimulus intensity, in terms of the amount of recent enjoyed items, affects the probability of continuing browsing in a curve of inverted-U shape. We discuss the possible origin of this curve based on a proposed *Award-Aversion Contest* model. This hypothesis is supported by the empirical study, which shows that the proposed model can successfully recover the original inverse Gaussian distribution for the browsing length. These browsing regularities can be used to develop better organization of hedonic content, which helps to attract more user dwell time in these systems.

1 Introduction

Recent years have witnessed a fast increase of the usage of mobile apps to consume various hedonic content (e.g. video, music, news, jokes, pictures, social networks etc.) for fun. These systems allow inexpensive and fast access to hedonic content, which has dominated our spare time. For example, according to the report[1] from Tencent, more than 55.2% of the 549 million active users visit Wechat for more than 10 times per day, summing up to more than 40-min usage. Due to the explosive usage of these systems, more research into the user behaviors in them is needed.

In these *hedonic content systems* (HCS), the user behaviors are often casual and task-less. It means that users might not have a concrete task at all for the use of these systems, except spending time and having leisure. Hence, users usually quit the HCSs after a few minutes, while sometimes users become couch potatoes even there is nothing worth reading. In this study, we aim to discover common regularities, which drive users' browsing behaviors in HCSs, through extensive empirical studies.

Specifically, we mainly study the behaviors of the decisions on whether a user continues browsing or leave the HCS. To support the easy browsing of hedonic content with overwhelming quantity and diversity, content items are usually organized in a sequence ranked by recency, popularity, or relevance [Lerman and Hogg, 2014]. Although there are many *exogenous* factors, such as content quality, network speed, external context, visit time etc., which may affect the decision to proceed or quit, there might be some *endogenous* factors from psychology and behavior to drive this decision.

Some recent research works in psychology focus on the decision model for two-choice decision tasks [Ratcliff and McKoon, 2008; Wagenmakers, 2009]. In their test of two alternative forced-choice (2-AFC) task, the participants are required to choose an answer out of two choices (e.g., yes or no). The psychologists proposed a diffusion model to simulate this decision process, in which people collect evidence for decision making. It assumes that a person has an action bound in making choices and would not make decision until the evidence of one choice exceeds the bound.

Motivated by this *choice threshold* model, we propose a browsing model in HSC. It assumes that users continue browsing until the evidence for quit exceeds the bound. With this model, the first passage time to the action bound is given by the two-parameter inverse Gaussian distribution [Seshardri, 1993]. To test the validity of this model, we analyzed the data collected from a typical hedonic mobile app, Wallpapers Plus for iOS 8[2]. It shows a strong fit of the empirical data to the theoretical distribution, and this strong fit always occurs despite differences in visit time and user types. This inverse Gaussian distribution indicates that there exists a page

[*]This work was done when Jiaxi, Rui and Zhongjie were visiting Institute of Computing Technology, CAS, China. At that time, Jiaxi and Rui were the undergraduate students at Wuhan University, China.

[1]http://tech.qq.com/a/20150127/018482.htm#p=1

[2]https://itunes.apple.com/en/app/bi-zhi+-zhu-ti-for-ios8/id557074482

position (the mode of the distribution, around the 6th page in the considered wallpaper application) at which the probability to leave is maximized.

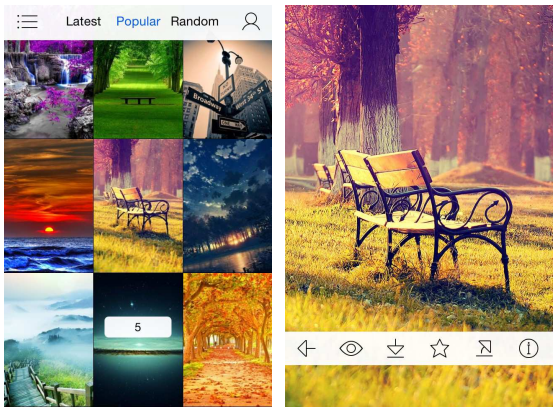
Furthermore, we noticed that the number of recent enjoyed items, as the stimulus intensity, might affect the decision on continuing or leaving. Here, this stimulus intensity can be measured by the number of clicks, votes or favorites on the content items, which occur during the browsing process. To this end, we explore how the probability of continuing browsing change over this stimulus intensity. For the first time, we observe that this curve does not monotonically increase, but shows inverted-U shape. It suggests that very low or very high levels of stimulus intensity will drive users to leave. Meanwhile, there is an intensity level of moderate degree at which the probability of continuing browsing is maximal. This observation is actually an empirical counterpart to the old adage recommending everything in moderation.

Finally, we discuss the possible origin of this inverted-U curve based on a proposed model of *Award-Aversion Contest*. This hypothesis is supported in the sense that the fitting with this model can successfully recover the original inverse Gaussian distribution for the number of visit pages.

We argue that all these browsing regularities are helpful to generate better content organization, which drives more dwell time of users in HCSs. Specifically, the inverse Gaussian distribution suggests that the content provider should pay more attention on the content, shown at the mode position of the distribution. Since the probability to leave is maximal at this position, more attractive content should be arranged here. After the mode position, since the probability to leave decreases gradually some advertisements could be inserted. Additionally, the inverted-U curve suggests that putting the high-quality or low-quality items into a small page range, which might attract very high or very low intensity of items being clicked and saved, might not be a good choice for content organization. The high and low quality items should be interleaved in the content sequence, leading to a moderate level of item clicks and saves. Only with this moderate level of stimulus intensity, the probability of continuing might be maximal.

2 Data Background and User Logs

In this study we analyze the user logs from a typical hedonic mobile app, Wallpapers Plus for iOS 8[3]. This app ranked in the top 20 list in the China market of Apple free apps and has more than fifty million users. Here, we briefly describe its function as follows. This app provides plenty of beautiful wallpapers, presented in a sequence, to users for viewing and downloading. After entering this app, as shown in Figure 1(a), users see one-page of wallpapers, where 9 pictures are shown in the 3×3 grid. This is the thumbnail mode for wallpapers. In this mode, users can swipe left on the screen to fetch the next pages of wallpapers. Also, if they like one of the 9 pictures, users can click it to see its full-screen version, as shown in Figure 1(b). In this full-screen mode, users can download, share, or bookmark this picture (by clicking the corresponding buttons at the bottom of the screen), or they can click the return button (the leftmost one) to get back to the thumbnail mode.



(a) The thumbnail mode (b) The full-screen mode

Figure 1: The screenshot of the app of Wallpapers Plus for iOS 8

In this study we only consider the following two kinds of user actions in this application: 1) turning to the next-page of wallpapers; 2) clicking certain picture to see its full-screen version. Table 1 summarizes all the notations used, with which we detail the user logs for this study as follows.

Table 1: The notations

Symbol	Meaning
$\vec{S}_i$	the i-th sequence
t_i	the happening time of the i-th sequence
u_i	the user ID of the i-th sequence
A_{ik}	the actions on the k-th page of the i-th sequence
C_{ik}	the set of clicked items on the k-th page of the i-th sequence
U_{ik}	the set of un-clicked items on the k-th page of the i-th sequence
$\mathbb{S}$	the set of all the sequences
$\mathbb{S}_{\geq k}$	the set of all the sequences whose size is not smaller than k
$\mathbb{S}_{>k}$	the set of all the sequences whose size is bigger than k
$\mathbb{S}_{=k}$	the set of all the sequences whose size is equal to k
e_k	the event that a user turns to the k-th page

Specifically, each browsing sequence $\vec{S}_i$ can be denoted as

$$\vec{S}_i = (A_{i1}, \cdots, A_{ik}, \cdots, A_{im_i}),$$

where A_{ik} records the actions on the k-th page of this sequence, and $|\vec{S}_i| = m_i$ denotes the size of this sequence, meaning that the user only visited m_i pages in this sequence. Note that each sequence $\vec{S}_i$ begins from the first page. Further, A_{ij} can be represented as a tuple

$$A_{ik} = (C_{ik}, U_{ik}),$$

[3]https://itunes.apple.com/en/app/bi-zhi+-zhu-ti-for-ios8/id557074482

where C_{ik}, U_{ik} denote the sets of clicked and un-clicked items in this page, respectively. $|C_{ik}|, |U_{ik}|$ denote the numbers of items in these sets, respectively. Altogether, we have the sequence set $\mathbb{S}$, denoted by

$$\mathbb{S} = \{(\vec{S}_i, t_i, u_i)|i = 1, \cdots, n\},$$

where $|\mathbb{S}| = n$ is the number of sequences in this set, t_i is the happening time of the i-th sequence, and u_i is the user ID of the i-th sequence. Additionally, we define $\mathbb{S}_{\geq k}$ as

$$\mathbb{S}_{\geq k} = \{(\vec{S}_i, t_i, u_i) \mid |\vec{S}_i| \geq k, \vec{S}_i \in \mathbb{S}\},$$

denoting the set of all the sequences whose size is not smaller than k. Similarly, we define $\mathbb{S}_{>k}$ and $\mathbb{S}_{=k}$ as the set of all the sequence whose size is bigger than k and equal to k, respectively.

Note that users may have different intentions to use the wallpaper app. Some may have the clear intention for changing wallpapers on the phones, while others may just browse for leisure and fun. Also, in this app users can search for the pictures they want by text queries. Since in this study we focuses on the "pleasure-seeking" behaviors without the clear information needs, rather than the "information-seeking" behaviors with the clear information requirements, we deliberately removed the logs from the search function for the following analysis.

Totally, we have the user logs for the 41 days from Nov. 4 to Dec. 14, 2014. After the data preprocessing, we have 1,545,950 sequences for the whole analysis in the following. Note that since the pictures are the content items in this application, the two terms, picture and item, will be used exchangeably in the rest of this paper.

3 Continue or Leave

Here, we describe the strong regularity of browsing patterns in HCSs through extensive empirical studies. This regularity can be described by a law of browsing, which determines the probability distribution of the depth, namely the number of pages a user browses within a visit of HCSs.

3.1 Stochastic Process for Browsing

We start by deriving the probability $p(K)$ of the number of pages K that a user browses in a visit. This can be done by assuming that there is action bound α for leaving. The browsing process is actually a procedure of evidence accumulation, and it continues until the evidence value X for leave reaches the action bound α. We assume that at the beginning $X_0 = 0$, and X_k grows as

$$X_k = vk + \sigma W_k, \tag{1}$$

where $v > 0$ is a constant for drift, σ is a parameter, and W_t is a standard Brownian motion [Shreve, 2008]. It means that X_k continuously increases at the speed of v plus some random process σW_t. With this stochastic process, a particular sequence of browsing is one of its realizations.

In this process, once the accumulated evidence X_k for leaving reaches the action bound α, the process ends. The number of pages a users follows before X_k first reaches its bound is a random variable K. Then, for the random walk in Eq. (1), the probability distribution of this first passage time to α is given by the two parameter inverse Gaussian distribution [Seshardri, 1993],

$$p(K) = \sqrt{\frac{\lambda}{2\pi K^3}} \exp[\frac{-\lambda(K-\mu)^2}{2\mu^2 K}], \tag{2}$$

with mean $E(K) = \mu$, variance $Var(K) = \mu^3/\lambda$, and mode equaling to $\mu[(1 + \frac{9\mu^2}{4\lambda^2})^{\frac{1}{2}} - \frac{3\mu}{2\lambda}]$ (λ is a scale parameter).

3.2 Empirical Studies on the Distribution of K

To test the validity of Eq (2), we analyze the data described in Section 2. Specifically, for a certain size $k \in \mathbb{N}$ we calculate the value,

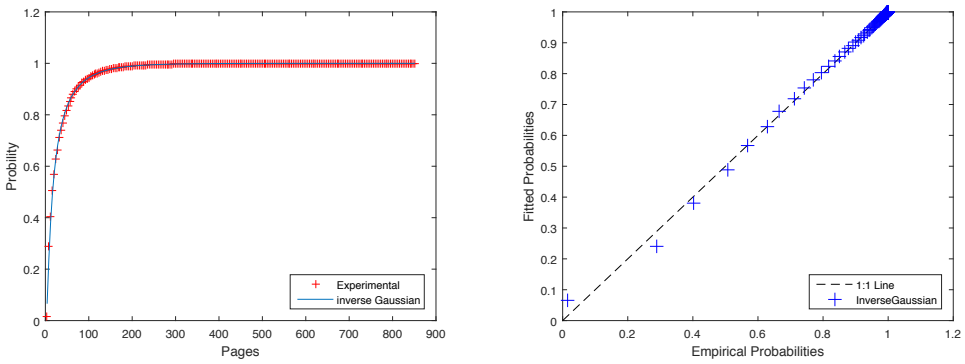
$$p(k) = \frac{|\mathbb{S}_{=k}|}{|\mathbb{S}|}$$

where $p(k)$ is the measured probability that a user browses exactly k pages. The measured cumulative distribution function (CDF) of the depth K for all the data is shown in Figure 2(a). Then, we fit this measured line as the inverse Gaussian distribution by the simplex search method [Lagarias *et al.*, 1998] with the objective of $L1$-norm. We test the quality of the fit by analyzing a quantile-quantile against the fitted distribution. As shown in Figure 2(b), it shows a strong fit with the significance level $P < 0.05$.

This inverse Gaussian distribution has two aspects worth stressing. First, it has a very long tail, indicating that some sequences are extended to a much farther length, which deviates a lot from the average number of visited pages. Second, since the distribution mode and mean are different, the typical behavior of users will not be the same as their average behavior. For the fitted distribution in Figure 2(a), its mode is 6.43. It means that it is at the positions of page 6 or 7 that users are most likely to leave. However, its mean is 23.03, meaning that the average number of pages users visit in this HCS is around 23. This distribution mean is greatly increased by the extremely long sequences occurred. These characteristics on the inverse Gaussian distribution support the observations that most frequently users leave the HCSs in a short time while sometimes users stay in the HCSs for a long time.

We also divide the data into subgroups with different visit time. We consider two dimensions of visit time. The first is weekday or weekend while the second one is morning (8am to 12am), afternoon (2pm to 6pm), or night (8pm to 12pm). The data from all these subgroups show the same strength of fit to the inverse Gaussian distribution with nearly the same parameters. For example, the mean and mode for weekday visit is 23.11 and 6.33 respectively while the mean and mode for weekend visit is 22.72 and 6.38 respectively.

Furthermore, we divide the data into subgroups with different types of users, i.e. new or old ones. We aim to check whether there exists some difference between the patterns for new and old users. Specifically, we use the following process to judge whether a user is new or old. Originally, we have the whole 41-day data. For each sequence occurred in the last 26 days, we check whether the user of this sequence visited the HCS in the first 15 days or not. If yes (or no), this sequence belongs to the subgroup for old (new) users. Again, these two



(a) The CDF of the sequences as a function of sequence length.

(b) The quantile-quantile plot against the fitted distribution.

Figure 2: The distribution fitting. The fitted inverse Gaussian distribution has $\mu = 23.03$, $\lambda = 20.91$, and the mode equals to 6.43.

subgroups show the strong strength of fit to the inverse Gaussian distribution, however, their parameters are significantly different. The mean and mode for old users are 25.65 and 8.73 respectively, while the mean and mode for new users are 22.06 and 5.83. It shows that old users visit more pages than new users in terms of both distribution mean and mode.

Finally, we analyze the browsing sequences, in which there are only picture clicks, but no saves. We believe that the behaviors in these sequences are for "pleasure-seeking" to more degree. Based on these data, we actually observe the similar pattern of Inverse Gaussian. In short, we argue that this strong regularity does exist in the "pleasure-seeking" behaviors.

4 Stimulus Intensity in Moderation

Here, we consider how the stimulus intensity in terms of the degree that users have enjoyed in HCSs affect their decisions on continuing or leaving. In HCSs, this stimulus intensity can be measured by the number of clicks, votes or favorites on the content items, which occur in the browsing process. Before this study, we have the conjecture that the more a user enjoys the more likely that she continues browsing in HCSs. However, the collected data tell that this conjecture of "the more the merrier" is not true. On the contrary, only the intensity level of moderate degree drives the probability of continuing browsing to be maximal.

4.1 Empirical Studies on Stimulus-Action

Here, the number of items a user clicked or saved recently can be used as the measure for the items she enjoyed. Due to the space limitation, we only show the patterns from the measure for clicked items, and the similar patterns still exist with the measure for saved items.

For a specific page position k in a sequence, the range for counting the clicked items can be any page interval before k. For example, the recent l pages is denoted by $R = [k - l + 1, k]$. Then, we have

$$\mathbb{S}^{w,R}_{>k} = \{\vec{S}_i \mid \sum_{j \in R} |C_{ij}| = w, \vec{S}_i \in \mathbb{S}_{>k}\},$$

denoting all the sequences whose size is bigger than k and there are exactly w clicks in the page range of R. Similarly, we define $\mathbb{S}^{w,R}_{\geq k}$, which is different from $\mathbb{S}^{w,R}_{>k}$ only in the sense that the size of these sequences is not smaller than k. With these two notations, we can compute $g_R(k, w)$ as

$$g_R(k, w) = \frac{|\mathbb{S}^{w,R}_{>k}|}{|\mathbb{S}^{w,R}_{\geq k}|},$$

measuring the empirical probability that a user turns to the $(k+1)$-th page after she clicks exactly w pictures in the page range of R. The sub-index R can be omitted if the position range R is given clearly in the context.

Considering the effect of *memory decay* [Das Sarma *et al.*, 2012], we first consider the range R of the latest 8 pages to the position k, namely $R = [k - 7, k]$. Figure 3 shows the values of $g(k, w)$, where each solid curve stands for the values of $g(k, w)$ with a fixed k. It shows that when k is fixed, $g(k, w)$ first monotonically increases along w, reaches its maximum around $w = 5$, and then drops monotonically to a low value. Additionally, this pattern exists for different values of k. Thus, the inverted-U curve always exists for this stimulus-action relationship.

It is also worth mentioning that the similar patterns in Figure 3 still exist if we change the position range to the latest 4 pages to position k, namely $R = [k - 3, k]$. We also change R to an intermediate range far away from k, namely $R = [k - l_1, k - l_2]$ (where $8 \leq l_1 < l_2$). However, this time $g_R(k, w)$ becomes very flat along the increase of w. It indicates that the behaviors in a far-away range do not affect the decision at the current position. It agrees with the effect of *memory decay*.

Additionally, we see that in Figure 3 the curves corresponding to bigger values of k are always above the ones with smaller k. It indicates that $g(k, w)$ increases when $k \geq 8$. It is consistent with the observation on inverse Gaussian distribution that after the page of distribution mode (around 6) the probability of continuing browsing increases monotonically.

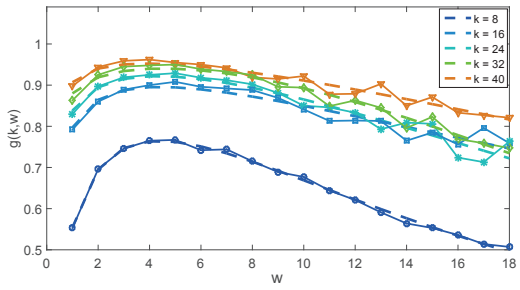


Figure 3: The curves of $g_R(k, w)$ with $R = [k - 7, k]$ (solid lines) and its fitting (dotted lines). Only the curves for $k = 8, 16, 24, 32, 40$ are shown for clear presentation. Better view with color.

4.2 Modeling the Inverted-U Curve

Next, we will propose a possible model to explain the observation of inverted-U curve. We argue that during the browsing process two kinds of values, namely *reward* and *aversion*, develop simultaneously on users. On one hand, high-quality contents fulfil users with reward, which attracts users to stay for more time. On the other hand, continuous consumption on these contents produces to users aversion, which impels users to leave. Then, to stay or leave the HCSs is jointly determined by reward and aversion.

Modeling Reward and Aversion

In this study, we model reward and aversion respectively as follows:

$$\begin{aligned} reward(w) &= a_1 \cdot w^{\alpha} + b_1 \quad (0 < \alpha < 1) \\ aversion(w) &= a_2 \cdot w^{\beta} + b_2 \quad (\beta > 1) \end{aligned} \tag{3}$$

As shown in Figure 4, the stimulus intensity w is the number of recent clicked items. We control the exponent values in range of $0 < \alpha < 1$ and $\beta > 1$ such that the values of reward and aversion increase in return-diminishing and return-increasing manner, respectively. These settings agree with the intuitive understandings on them that: with the increase of enjoyed items, reward goes up fast first, and then its increase tendency slows down. Nevertheless, aversion goes up slowly at the early time, but its increase speed becomes large gradually.

Reward-Aversion Contest

Then, with the modeling on reward and aversion we propose the following probabilistic process to determine the action of stay or leave:

1) After a user clicks w pictures, two values, $reward(w)$ and $aversion(w)$ in Equ. (3), are evoked by these clicked pictures.

2) Then, we sample a value v from the Beta distribution $Beta(reward(w), aversion(w))$. Namely, we have

$$v \backsim Beta(reward(w), aversion(w))$$

Here, $reward(w)$ and $aversion(w)$ are used as the two parameters of the Beta distribution.

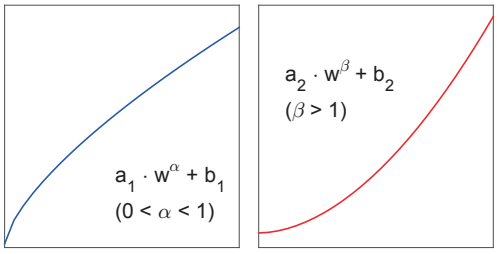
3) Finally, the action of continuing browsing, denoted by Y, is the random variable of the Bernoulli distribution with the success probability of v. Namely, we have

$$Y \backsim B(v).$$

With this stochastic process we can get the expectation of Y as follows;

$$E(Y) = E(E(Y|v)) = E(v) = \frac{reward(w)}{reward(w) + aversion(w)}$$

The above equation intuitively shows that the probability of continuing browsing is controlled by the contest between reward and aversion. The bigger this fraction of $\frac{reward(w)}{reward(w)+aversion(w)}$ is, the more likely a user will continue browsing.



(a) The curve for reward. (b) The curve for aversion.

Figure 4: Modeling on reward and aversion

4.3 Fitting with Reward-Aversion Contest Model

With the proposed reward-aversion contest model, for a fixed k we fit the curve $g(k, w)$ by the function of

$$\begin{aligned} \hat{g}(k, w) &= \frac{reward(w)}{reward(w) + aversion(w)} \\ &= \frac{(a_1 \cdot w^{\alpha)} + b_1}{(a_1 \cdot w^{\alpha} + b_1) + (a_2 \cdot w^{\beta} + b_2)} \end{aligned}$$

such that $0 < \alpha < 1$ and $\beta > 1$. The fitted curves are the dotted ones in Figure 3, which shows the strong fit with the significance level $P < 0.001$.

For further confirmation of the model, we use the fitted values of $\hat{g}(k, w)$ to recover the inverse Gaussian distribution on K. Specifically, the probability that a user continues browsing at page k can be computed as

$$g(k) = \sum_{w} g(k, w) f(k, w), \tag{4}$$

where $f(k, w)$ is the probability that a user click w items in the page range of $[k - 7, k]$. It can be empirically estimated from the data. Then, we have

$$p(k) = (1 - \sum_{i=1}^{k-1} p(i))(1 - g(k)), \tag{5}$$

where $(1 - \sum_{i=1}^{k-1} p(i))$ is the probability that a user browses at least k pages. With Eq. (4) and (5), we can recover the

values of $\hat{p}(k)$ from the fitted values of $\hat{g}(k,w)$. Specifically, we have

$$\hat{p}(k) = (1 - \sum_{i=1}^{k-1} \hat{p}(i))(1 - \hat{g}(k))$$
$$= (1 - \sum_{i=1}^{k-1} \hat{p}(i))(1 - \sum_{w} \hat{g}(k,w) f(k,w))$$

Then, based on the values of $\hat{p}(k)$, we also get a strong fit of these values to an inverse Gaussian distribution with $\mu = 23.17$ and $\lambda = 21.17$. And its mode is 6.50. Compared with the true inverse Gaussian distribution in Figure 2(a) based on the empirical values of $p(k)$, we cannot reject the null hypothesis that these two distributions are the same with the significance level $P < 0.001$. It indicates again that $\hat{g}(k,w)$ is a strong fit to $g(k,w)$.

5 Related Work

Since HCSs increasingly dominate our spare time, more research into these casual and task-less scenarios is needed. Towards this end, this study exposes the inverse Gaussian distribution on the depth of the browse sequences. Similar regularity was discovered by Huberman et. al. [Huberman *et al.*, 1998] for the behavior patterns in world wide web surfing. Surfing in world wide web is usually task-oriented, while browsing in HCSs is often task-less except for spending time and having leisure. This study actually confirms this regularity in these more casual scenarios. Additionally, for the first time we study how the stimulus intensity in terms of the number of recently clicked items affects the probability of continuing browsing. We observe the pattern of inverted-U, and give a possible explanation for this observation by proposing the reward-aversion contest model, which is based on the research in psychophysics [Murray, 1993]. We show a strong fit on this pattern in the sense that the fitted curves by the reward-aversion contest model can successfully recover the original inverse Gaussian distribution on the depth of the browse sequences.

The research on HCSs mostly focuses on assessing the *appeal* of content items in order to identify interesting ones in a timely manner for users [Lerman and Hogg, 2014]. Due to the overwhelming quantity and diversity of hedonic content (e.g. pictures, news, videos generated every day), these studies help to answer the questions, such as which of the thousands of daily news on Digg and Reddit are worth reading, and which of the many videos uploaded every day are worth watching. Usually, HCSs provide a voting function to allow users to express their opinions, and then rank the content items based on these opinions. The concrete ranking strategies can be *sort by voting numbers* (e.g. in Dig and Reddit) or *sort by voting recency* (e.g. latest retweet items appears at the top of a follower's stream). However, these collective judgements often produces "rich-get-richer" and "irrational herding" phenomenon [Szabo and Huberman, 2010; Yin *et al.*, 2012], in which the inequality and unpredictability of high-quality items will be increased by position bias [Lerman and Hogg, 2014] and social influence [Lorenz *et al.*, 2011; Salganik *et al.*, 2006]. Thus, these ranking schemes are often biased and inconsistent, with the similar items ending up with totally different numbers of votes.

There are also some other studies, focusing on user browsing models in a search engine as a typical task-oriented system [Chapelle and Zhang, 2009; Craswell *et al.*, 2008; Joachims, 2002; Joachims *et al.*, 2005; Srikant *et al.*, 2010; Zhang and Jones, 2007]. These studies aimed to leverage the click logs to improve the search engine performance. The key issue in this area is to infer the true relevance for query-document pairs while removing the effect from document positions. Along this line, Joachims [Joachims, 2002; Joachims *et al.*, 2005] conveyed the click-through data as relative relevance judgements instead of absolute relevance judgements, and proposed the ranking SVM algorithm to learn a better ranking function. This is the indirect method to interpret the click logs from a search engine. On the other hand, there are also some studies to infer the document relevance directly [Chapelle and Zhang, 2009; Craswell *et al.*, 2008; Srikant *et al.*, 2010; Zhang and Jones, 2007]. Their basic idea is to model the sequence of user browsing and clicking search results as a probabilistic process with the true relevance as the latent variables. This probabilistic process is usually based on the assumption that: the probability of being clicked decays by position, and users examine the results sequentially and terminate once a relevant document is found. Chapelle and Zhang [Chapelle and Zhang, 2009] further model the *perceived relevance* (the probability that an item is clicked) and *actual relevance* (the probability that the user is satisfied given that the item is clicked) respectively, and the true relevance is defined as the product of these two terms. Srikant et al. [Srikant *et al.*, 2010] also proposed the model where examination depends on how the prior results are clicked.

Compared with task-oriented systems, user behaviors in taskless HCS are more casual. High-quality content brings more reward to users and thus prompt them to explore more. Meanwhile, aversion also increases after users consume more content and then impel users to leave. Therefore, the dilemma on exploring more or not is a contest between reward and aversion.

6 Conclusion

In this paper, we study the mechanism behind the "pleasure-seeking" behaviors in HCSs. We found that the distribution over browsing length for a visit can be described by the inverse Gaussian form with a very high precision. Also, we found that the amount of recent enjoyed items affects the probability of continuing browsing in a curve of inverted-U shape. Furthermore, we propose the possible models to explain why these patterns always exist. Clearly, all these regularities help to guide the content arrangement in the HCSs, which may attract more user dwell time. In the future, based on all these regularities we will formally formulate the problem of content organization optimization, which aims to maximize the average user dwell time. Also, we plan to extend this study into micro-perspective level for the relationship between the characteristics of different user communities and the model parameters for these regularities.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.61473274), National High-tech R&D Program of China (863 Program) (No.2014AA015105).

References

[Chapelle and Zhang, 2009] Olivier Chapelle and Ya Zhang. A dynamic bayesian network click model for web search ranking. In *WWW*, 2009.

[Craswell *et al.*, 2008] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *WSDM*, 2008.

[Das Sarma *et al.*, 2012] Anish Das Sarma, Sreenivas Gollapudi, Rina Panigrahy, and Li Zhang. Understanding cyclic trends in social choices. In *WSDM*, 2012.

[Huberman *et al.*, 1998] Bernardo A. Huberman, Peter L. T. Pirolli andJames E. Pitkow, and Rajan M. Lukose. Strong regularities in world wide web surfing. *Science*, 1998.

[Joachims *et al.*, 2005] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, 2005.

[Joachims, 2002] Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

[Lagarias *et al.*, 1998] J.C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 1998.

[Lerman and Hogg, 2014] Kristina Lerman and Tad Hogg. Leveraging position bias to improve peer recommendation. *PloS one*, 2014.

[Lorenz *et al.*, 2011] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 2011.

[Murray, 1993] David J. Murray. A perspective for viewing the history of psychophysics. *Behavioral and Brain Sciences*, 1993.

[Ratcliff and McKoon, 2008] R. Ratcliff and G. McKoon. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 2008.

[Salganik *et al.*, 2006] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 2006.

[Seshardri, 1993] V. Seshardri. *The Inverse Gaussian Distribution*. Clarendon, Oxford, 1993.

[Shreve, 2008] Steven E Shreve. *Stochastic Calculus for Finance II: Continuous Time Models*. Springer, 2008.

[Srikant *et al.*, 2010] Ramakrishnan Srikant, Sugato Basu, Ni Wang, and Daryl Pregibon. User browsing models: relevance versus examination. In *KDD*, 2010.

[Szabo and Huberman, 2010] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 2010.

[Wagenmakers, 2009] E.-J. Wagenmakers. Methodological and empirical developments for the ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, 2009.

[Yin *et al.*, 2012] Peifeng Yin, Ping Luo, Min Wang, and Wang-Chien Lee. A straw shows which way the wind blows: Ranking potentially popular items from early votes. In *WSDM*, 2012.

[Zhang and Jones, 2007] Wei Vivian Zhang and Rosie Jones. Comparing click logs and editorial labels for training query rewriting. In *WWW*, 2007.