How Much Novelty is Relevant? It Depends on Your Curiosity

Pengfei Zhao, Dik Lun Lee
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{ericzhao, dlee}@cse.ust.hk

Traditional recommendation systems (*RS*s) aim to recommend items that are relevant to the user's interest. Unfortunately, the recommended items will soon become too familiar to the user and hence fail to arouse her interest. Discovery-oriented recommendation systems (*DORS*s) complement accuracy with *discover utilities* (*DU*s) such as novelty and diversity and optimize the tradeoff between the *DU*s and accuracy of the recommendations. Unfortunately, *DORS*s ignore an important fact that different users have different appetites for *DU*s. That is, highly curious users can accept highly novel and diversified recommendations whereas conservative users would behave in the opposite manner. In this paper, we propose a curiosity-based recommendation system (*CBRS*) framework which generates recommendations with a personalized amount of *DU*s to fit the user's curiosity level. The major contribution of this paper is a computational model of user curiosity, called Probabilistic Curiosity Model (*PCM*), which is based on the curiosity arousal theory and Wundt curve in psychology research. In *PCM*, we model a user's curiosity with a curiosity distribution function learnt from the user's access history and compute a curiousness score for each item representing how curious the user is about the item. *CBRS* then selects items which are both relevant and have high curiousness score, bounded by the constraint that the amount of *DU*s fits the user's *DU* appetite. We use joint optimization and co-factorization approaches to incorporate the curiosity signal into the recommendations. Extensive experiments have been performed to evaluate the performance of *CBRS* against the baselines using a music dataset from last.fm. The results show that compared to the baselines *CBRS* not only provides more personalized recommendations that adapt to the user's curiosity level but also improves the recommendation accuracy.

Keywords

Recommendation, curiosity, psychology, personalization

1. INTRODUCTION

Traditional recommendation systems (*RS*s) based on content similarity and collaborative filtering aim to achieve high accuracy by recommending items that are relevant to the user's interest. The problem with this approach is that the recommended items are very similar to the user's interest as well as between themselves. Thus, the user will quickly find the recommended items too familiar and uninteresting for exploration. We call this the "accuracy overloading problem." To prevent accuracy from domainating the recommendations, Discovery-Oriented Recommendation Systems (*DORS*s) introduce metrics called Discovery Utilities (*DU*s) as additional dimensions besides relevance for ranking the candidate items. The ranking problem is often modeled as a multi-objective optimization problem, which seeks an optimal tradeoff between the different dimensions [16]. Many *DU*s, including novelty and diversity, have been studied in the literature [3]. Although *DORS*s can help to alleviate the accuracy overloading problem, they neglect an important fact that different users have different curiosity levels, which lead to different levels of desire to discover new things. Specifically, highly curious users would find recommendations with high *DU*s interesting but those with low *DU*s boring. The reverse is true for conservative users. We refer to this curiosity-driven, personal demand of *DU*s as *DU* appetite. Without considering the *DU* appetite of each user, *DORS*s would favor items with high *DU*s (balanced by relevance) for every user. Consequently, while highly curious users are excited about high *DU* items, conservative users would find them too overwhelming. We call this "curiosity mismatch problem."

In this paper, we present a framework for Curiosity-Based Recommendation Systems (*CBRS*s) to solve the curiosity mismatch problem. It consists of the Probabilistic Curiosity Model (*PCM*), which models a user's curiosity with a curiosity distribution function learnt from the user's access history. Thus, *each user has her own curiosity model* estimated from her access behavior. It allows us to compute a *curiousness score* for each item representing how curious the user is about the item. *CBRS* then selects items which have both high relevance and curiousness scores, bounded by the constraint that the items' *DU*s should fit the user's *DU* appetite. We note that the *CBRS* framework is general enough for incorporating different *DU*s. However, since novelty is by far the most studied *DU* in *RS* research, this paper focuses on novelty, leaving the details of modeling the other *DU*s as future research.

The Oxford dictionary defines curiosity as "a strong desire to know or learn something." According to psychology research [4], a user's appetite for novelty is influenced by her *curiosity* in that the higher/lower a user's curiosity, the bigger/smaller is her appetite for novelty. For a user with a particular curiosity level, recommendations with too much novelty will cause anxiety while too little will cause boredom.[1] To exploit this psychological phenomenon, a *RS*

[1]This is an interesting finding in psychology. In most existing *RS* research, novelty is assumed to be the higher the better. However, we can learn from many real-life situations that this is not true. For example, *most* people going to a theme park for roller coaster rides

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2911488

should select recommendations with novelty commensurate with the user's curiosity, instead of blindly maximize the novelty *for everybody*.[2]

In this paper, we address several challenges facing *CBRS*. First, we adopt the famous Wundt curve in psychology to model human curiosity [4]. The Wundt curve depicts the "inverse U-shape" relationship between an item's stimulation degree (*sd*) (i.e., stimulation generated by, say, the novelty of an item upon a user) and the user's response to the item (see Fig. 2). Briefly, given a user, as the novelty of an item increases, the user will become more and more interested in it until the novelty reaches a turning point where the interest is at a maximum; beyond this point the user will become less and less interested and will ultimately become disinterested in the item. This theory is adopted in *CBRS*. Unfortunately, Wundt curve is only a qualitative description of the relationship. In *CBRS*, we model the "inverse U-shape" with a Beta distribution and learn the distribution's parameters from the user's accessed history. Second, we use the curiosity model to compute for each candidate item a curiousness score representing how curious the user is about the item and select candidate items matching the user's curiosity based on their curiousness scores. The input to the model is the stimulation degree of the item. In the context of this paper, it is the novelty of the item. To measure an item's novelty, we propose three features, namely, the user's access frequency to the item, the recency of the accesses, and the user-specified tags of the item. These features have been confirmed to be effective by psychology research [4]. Finally, we need to integrate the curiosity and relevance aspects of the items to produce the final recommendations. We study two strategies, namely, the joint optimization of relevance and curiousness with the constraint that the novelty of the items should match the user's novelty appetite, and co-factorization of the relevance and curiousness signals associated with the item.

Throughout this paper, we use music recommendation to illustrate the ideas of *CBRS*. This is because music recommendation has been widely studied in recent years [8], and research has shown that music listening behavior is highly discovery oriented, highly personal and heavily driven by curiosity [2]. The contributions of this paper are summarized as follows.

- We propose a novel framework for Curiosity-Based Recommendation Systems (*CBRS*s) to combine relevance and curiosity in the recommendation process. Although we focus on novelty in this paper, the framework is general enough to embrace other *DU*s.
- We develop a computational model called *Probabilistic Curiosity Model* (*PCM*) to model a user's curiosity based on Wundt curve developed in psychology. On the one hand, Wundt curve forms the theoretical foundation of *PCM*. On the other hand, our work provides large-scale experimental evidence to validate the Wundt curve theory. In our opinion, this is a major and interesting contribution of our work to the psychology field.
- We propose two strategies to combine curiousness and relevance in selecting recommendations, considering the constraints that the novelty of the recommended items should match the user's curiosity level.
- We use various performance metrics, including inter-user similarity, novelty fitness and precision, to evaluate the effectiveness of *CBRS*. Experimental results show that *CBRS* not only provides personalized recommendations adapted to the user's unique curiosity but, surprisingly, also improves the recommendation precision. Further, we study the impact of including previously accessed items (i.e., non-novel items) in the recommendations and show that *CBRS* can recommend the right mix of non-novel and novel items to optimize the performance. This is an important finding because in many applications (e.g., music recommendation) users may repeatedly access some items (e.g., favorite songs).

The rest of this paper is organized as follows. We review related work in Section 2. Curiosity modeling will be introduced in Section 3. Section 4 describes the recommendation strategies, and Section 5 presents performance evaluation. Section 6 concludes our findings.

2. RELATED WORK

2.1 Discovery-Oriented RSs (DORS)

Discovery-Oriented Recommendation Systems (*DORS*s) introduce various *DU*s to complement accuracy in solving the accuracy overloading problem. Common *DU*s include novelty [7, 15, 17], diversity [18, 21] and serendipity [13, 19]. Although *DORS*s were designed to produce recommendations that are more novel, diversified, or surprising, they ignore the fact that different users would accept different amount of *DU*s, which we call the *DU* appetite, because users have different curiosity levels, causing what we call the curiosity mismatch problem described in Sec. 1. Recently, methods that personalize the *DU*s of the recommendations for individual users, termed Personalized *DORS*s or *PDORS*, have been proposed to resolve the curiosity mismatch problem.

2.2 Personalized DORSs (PDORS)

Personalization has been studied for a long time in information retrieval [6]. However, to the best of our knowledge, the only works on *PDORS*[3] were reported in [17] and [7]. [17] assumes that each user has a binary "novelty-seeking status" indicating whether the user would seek completely novel (i.e., new) items or non-novel items (previously accessed items). However, this binary assumption is too strict in many situations, e.g., user may want to receive both completely novel and non-novel items in the same recommendation list, as in Youtube which recommends both new and old items. To extend this binary assumption, [7] learns a real-value novelty preference score for each user through logistic regression. The score is then used as a parameter for balancing similarity and novelty in the top-K ranking task. We call this method personalized parameter balancing *RS* or *PPBRS*. The personalized novelty preference parameter of *PPBRS linearly* scales the novelty of all items in the final ranking process. The implication is that if a user is judged to accept novel items (large novelty preference score), then all novel items benefit equally in the ranking task. As discussed throughout this paper, linear scaling of novelty does not work for human curiosity: even for users having large novelty preference score, it does not mean that they can accept items with extremely high novelty (see Sec. 1 and Sec. 3). In our proposed *CBRS*, we model a user's novelty preference as a probability distribution rather than a single real value and verify that the distribution resembles the non-linear "inverse U-shape" Wundt curve in psychology research.

would prefer rides that are not too exciting (too thrilling) or too easy (too boring).

[2]Not explicitly mentioned here, the curiosity aspect must be balanced with the relevance of the items.

[3]Since we only consider novelty in this paper, from now on, *PDORS*s refer to delivering personalized amount of "novelty" according to the user's novelty appetite.

There are several differences between *CBRS* and existing *DORS*s and *PPBRS*. Comparing to *DORS*s, (1) curiosity in *CBRS* is not a new dimension of *DU*s; instead, it governs each user's acceptance level of *DU*s; (2) *CBRS* does not use *DU*s as ranking utility directly since a given amount of *DU*s might be attractive to one user but may turn other users away. Instead, *CBRS* transforms *DU*s into a curiousness score representing a user's curiousness about an item using the user's curiosity model and the curiousness score is then used in ranking; (3) curiosity is a human trait born with a person, but *DU*s are based on an item's property. Comparing to *PPBRS*, (1) user's novelty preference in *CBRS* is not modeled as a real value but a curiosity distribution modeling the shape of Wundt curve as a probability density function *pdf*; (2) rather than assuming that every item's *DU*s is uniformly scaled by the user's novelty preference score, *CBRS* influences the discovery ranking utility with a curiousness score that is derived for each user and each item based on the user's curiosity model.

3. CURIOSITY MODEL

3.1 Preliminaries

Due to the interdisciplinary nature of our research, it is necessary to introduce terminologies widely used in the psychological curiosity field. In psychology, **curiosity** is a human trait born with a person, driving her cognitive development through life. Since different people have different levels of curiosity, we introduce **curiousness** (cur_u^i), which is a real value, to quantify a user u's curiosity to explore an item i. *CBRS* adopts the **Curiosity Arousing Model** (*CAM*) developed in psychology research [4]. In *CAM*, a user receives stimuli and would only respond to stimuli which can arouse her curiosity. Since *CAM* describes how a user selectively responds to the stimuli, it is also referred to as the **Stimulus Selection Process** (*SSP*). For recommendation systems, each recommended item presents a **stimulus** to the user. The strength of a stimulus is quantified by the Stimulus Degree, which is a real value denoted as *sd*. Note that the same item can produce different *sd*s to different users.

The *sd* of a stimulus is defined by a number of factors called **Collative Variables** (*CV*s). For this paper, CVs are the same as features, which are extracted from some measurable properties of a stimulus. The values of the *CV*s are called **Collative Variable Values**. As discussed, different users have different responses even if the stimulus is the same because of their difference in curiosity. We propose the **Probabilistic Curiosity Model** (*PCM*), which is a probabilistic view of *CAM*. It models a user's selected *sd*'s as a random variable, and curiosity as the probability distribution of the random variable, called **Curiosity Distribution** (C_u). In this way, a user's stimulus selection process (*SSP*) can be interpreted as drawing samples (stimuli) from the curiosity distribution under the guidance of the user's curiosity.

3.2 Curiosity Arousal Model

Berlyne interprets curiosity as the driving factor for *SSP*, that is, "when several conspicuous stimulus are introduced at once, to which stimulus will human respond" [4]. Following Berlyne's interpretation, we can take an *interior view* of curiosity, which is to consider curiosity as an internal factor of a person influencing her selection of stimulus. We can also take an *exterior view* of curiosity, that is, a person's *SSP* can reveal her curiosity inside. *CAM* points out the possibility that the interior curiosity trait can be estimated by the exterior *SSP*. To achieve this goal, we need to (1) quantify the stimulus and (2) model the curiosity. Sections 3.2 and 3.3 will discuss (1) and the remaining subsections will discuss (2).

Berlyne [4] discovered a principle set of features, named as *Collative Variables* (*CV*s), that can arouse curiosity. Four *CV*s were identified: *novelty*, *uncertainty*, *conflict* and *complexity*. Since *CV*s are dependent on user u, item i, time t, and user's access history H_u^t before time t, we use $sd_{u,i}^t(H_u^t)$ to denote item i's *sd* with respect to u at time t given u's access history H_u^t. Similarly, we use $cvv_{u,i}^t$ to denote the value of a *CV* associated with i at time t for u. Then, a stimulus can be quantified by Equation 1, where ψ is a scoring function. Since a stimulus is made up of CVs, it can be modeled as a (weighted) sum of *cvv*'s of the *CV*s. In this paper, we focus on novelty only, so $sd_{u,i}^t$ in Equation 1 is simplified to $Nov_{u,i}^t$, which denotes item i's novelty to user u at time t.

$$sd_{u,i}^t = \psi(u, i, t|H_u^t) = \sum_{cv} cvv_{u,i}^t \quad (1)$$

Our dataset is about the users' music listening history, in which an item corresponds to a music track. When u clicks an item i to listen to it, we say "u *accesses* i," "u accesses an artist" if u accesses at least one track performed by the artist, and u accesses a tag if at least one accessed track has the tag. From the recommendation point of view, the action that u accesses i is viewed as u's provision of a positive *feedback* on i.[4] From the curiosity point of view, u's access to i can be viewed as u *responding* to the stimulus generated by i. Thus, in this paper, we use "feedback", "access", and "response" interchangeably depending on the context.

3.3 Modeling of Novelty

In the *RS* domain, the novelty of an item reflects how much the item differs from the user's previously accessed items. In the psychology field, Berlyne suggested that novelty is inversely related to three factors [4]: (1) how often the stimulus has been experienced by the user, (2) how recent the stimulus has been experienced by the user before, and (3) how dissimilar the stimulus is to the user's previous experience. Based on the three criteria, we formally define the novelty in Equation 2,

$$Nov_{u,i}^t = \frac{1}{3} \cdot (SF_{u,i}^t + SR_{u,i}^t + Dissim_{u,i}^t) \quad (2)$$

where $SF_{u,i}^t$ denotes the scaled frequency of the user's accesses to item i before t; high frequency indicates the user's familiarity of the item, making it less novel to the user. $SR_{u,i}^t$ denotes the scaled recency of the user's access on item i w.r.t the current time t; the more recent the user's access to the item is, the less novel the item is to the user. $Dissim_{u,i}^t$ denotes the dissimilarity between item i and u's historically accessed items; a large dissimilarity indicates i is more novel to the user. Note that $SF_{u,i}$, $SR_{u,i}$ and $Dissim_{u,i}$ correspond, respectively, to Berlyne's three criteria above. In music recommendation, a user's decision to listen to a track may depend on the performing artist. For example, after picking her favorite artist, the user may play all of the tracks sequentially in the album. In this case, the reason for the user to play the tracks in the album is not the novelty of the tracks but the performing artist. Thus, we include the artist in computing a track's stimulus.

We define $SF_{u,i}^t$ formally in Equation 3:

$$SF_{u,i}^t = \frac{e^{-\rho_a \cdot |A_{u,i}^t|} + e^{-\rho_i \cdot |I_{u,i}^t|}}{2} \quad (3)$$

where $A_{u,i}^t$ is the set of u's accesses to items in the user's history H_u^t (before time t) having the same artist as i's artist, $I_{u,i}^t$ is the set of u's accesses to items i in the user's history H_u^t, $|.|$ denotes the

[4] As with most web-based applications, we only consider implicit feedback from users. A user's listening to a track indicates her preference on the track, which is a positive feedback; no negative feedback is available.

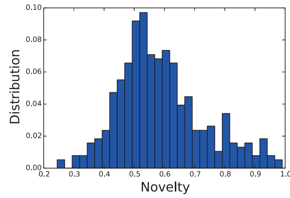


Figure 1: The global users novelty appetites distribution

cardinality of a set, and ρ_a and ρ_i are positive exponential scaling coefficients. The reason to model $SF^t_{u,i}$ based on both the frequencies of the track itself (item i) and tracks having the same artist as i's artist is that both tracks and artists are stimuli. It is clear that when $|I^t_{u,i}|$ is high, the user has accessed item i many times so i's novelty is low, and when $|A^t_{u,i}|$ is high, the user is familiar with i's artist, and so its contribution to novelty is low. Overall, a small $SF^t_{u,i}$ value means i has low novelty.

$SR^t_{u,i}$ is defined in Equation 4:

$$SR^t_{u,i} = \frac{e^{\rho_t \cdot (t - t(A^{-1}_{u,i}))} + e^{\rho_t \cdot (t - t(I^{-1}_{u,i}))}}{2} \tag{4}$$

where $t(I^{-1}_{u,i})$ and $t(A^{-1}_{u,i})$ denote, respectively, the timestamps of u's latest access in $I^t_{u,i}$ and $A^t_{u,i}$. Here, we use "day" as the time unit. ρ_t is the forgetting coefficient. Equation 3 and 4 both use the decay function with exponential forgetting rate [10] to scale frequency and recency to the 0-1 range. Note that small SR value indicates less novelty.

To model $Dissim^t_{u,i}$, for each track, we extract six tags[5] labeling the genres of the music (e.g., "pop" and "jazz") using the LastFM API. The dissimilarity is calculated based on the number of common tags between the track and the historically accessed tags, and is formally defined in Equation 5, where $Tags(i)$ denotes the set of tags associated with track i, and $I_{u,tag}$ denotes the set of accessed items in H^t_u labeled with tag. ρ_{tag} is the coefficient for tag frequency. Note that small *Dissim* value means low novelty.

$$Dissim^t_{u,i} = \frac{1}{|2 \cdot Tags(i)|} \sum_{tag}^{Tags(i)} (e^{-\rho_{tag} \cdot |I_{u,tag}|} + e^{\rho_t \cdot (t - t(I^{-1}_{u,tag}))}) \tag{5}$$

Several observations can be made from Equations 2 to 5. First, items with small *SF* (i.e., frequently accessed), small *SR* (i.e., recently accessed), and small *Dissim* (genre has been frequently and recently accessed) have small *Nov* values and thus are less novel. Second, if the user listens to a new track performed by an artist whom she has never listened to before, $SF^t_{u,i}$ and $SR^t_{u,i}$ will become 1. However, $Dissim^t_{u,i}$ may not necessarily be 1. Third, we assume that the factors in each equation are equally weighted (e.g. in Equation 2, SF, SR, $Dissim$ contribute equally to Nov). Due to space limitation, we cannot show the results for all weighting combinations and the optimal balance is application dependent anyway. Fourth, $Nov^t_{u,i}$ is defined on (u, i, t) triples and is a real value representing item i's novelty to u at time t. If we use Nov^T_u to denote the vector of $Nov^t_{u,i}$, within the time period T, when $T \to \infty$ and we omit the superscript T, then Nov_u can be viewed as a random variable representing u's inclination in accessing novel items, and each user's novelty appetite nov_u is defined as the expectation of the random variable Nov_u. Figure 1 illustrates the distribution of all users' novelty expectation. We can see that the novelty appetites of different users vary a lot. Finally, although $SF^t_{u,i}$, $SR^t_{u,i}$ and $Dissim^t_{u,i}$ are defined based on the music recommendation dataset, it is not difficult to extend them to other domains such as restaurant recommendation and movie recommendation, since user behaviors in these applications can be formulated as (u, i, t) triples.

[5] For tracks which have less than six tags, we extract all of the tags available.

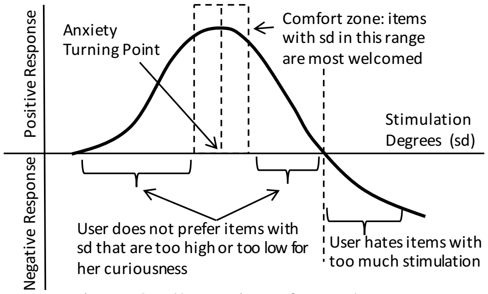


Figure 2: Illustration of Wundt curve

3.4 Wundt Curve and Probabilistic Curiosity Model

Curiosity modeling has been studied in psychology for a long time. In 1870s, Wundt introduced the theory of "optimal level of stimulation" and postulated an inverted "U-shape" relationship between stimulation level and hedonic response caused by the stimulus, which is referred to as the "Wundt curve." Figure 2 is an illustration of Wundt curve, where the x-axis denotes stimulus degrees *sd*, and the y-axis denotes the user's hedonic response. Berlyne formed the "intermediate arousal potential" (*IAP*) theory, which states that too little stimulation results in boredom while too much stimulation results in anxiety. From Fig. 2, we can see that the user's positive hedonic response increases first as *sd* increases. However, after reaching a certain threshold, the positive hedonic response will drop with further increases of *sd*. We name the turning point as user u's *anxiety turning point* (ATP_u). It means that beyond the threshold the user will become anxious due to overwhelming *sd*. An important note about applying Wundt curve to *RS* is that since we assume each of the user's interaction with an item reflects her positive feedback on the item, there is no negative feedback in our our music recommendation application, leading to a Wundt curve that is entirely above the x-axis. Since a user's curiosity can be revealed from her *SSP* and generally behaves like Wundt curve, the central task now is how to model Wundt curve.

From the probability point of view, we model each user's selected *sd*'s as a random variable and model the user's curiosity as a probability distribution of the random variable, which determines the user's *SSP*. The distribution is named *Curiosity Distribution*, denoted by C_u. In this way, *SSP* is viewed as a sampling process from the user's personal curiosity distribution, with which curiosity is able to guide *SSP*. Given a large amount of user-item interaction data and u's accessed *sd*'s, C_u can be estimated. We name our model *Probabilistic Curiosity Model* (*PCM*), which can be considered the probabilistic view of *CAM*. There are two main reasons we use *PCM* to model Wundt curve and curiosity. First, human's *SSP* behaves in a probabilistic manner. For example, a curious user may also select small *sd*'s, although the chance is small compared with a conservative user. Second, although web data lacks explicit user preference data, abundant implicit feedback data are available. By modeling curiosity probabilistically, we can utilize web-scale user interaction data for the estimation of curiosity distribution.

Figure 3 illustrates a user's *SSP* on five items with different sd values. According to the interior view of *CAM*, the user's *SSP* (either acceptance or rejection of an item) is guided by her curiosity. *PCM* views a user's *SSP* as continuously drawing samples from her curiosity distribution in such a way that stimuli whose *sd*'s best meet the user's curiosity have a high chance to be drawn. This ensures that the stimuli with *sd*'s that are too large or small, reflecting

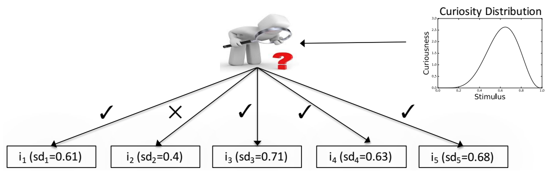


Figure 3: Illustration of the probabilistic curiosity model

the fact that the items are too novel or boring for the user, have a lower chance to be selected by the users. In the example, the user selected four items and ignored one item and the average sd of the selected items is about 0.65.

3.5 Estimation of Curiosity Distribution

Section 3.4 shows how curiosity guides a user's *SSP* from the psychology and probability points of view. In this section, we will introduce how to estimate the curiosity distribution.

Since both *CAM* from psychology and our *PCM* try to model a user's curiosity, it is natural to expect that the probability density function (*pdf*) of C_u exhibits the shape of Wundt curve. Figure 4 shows the histograms of 5 users' accessed *sd*'s, simulating the curiosity distribution. Several important findings can be obtained. First, the distribution generally fulfills the "inverted-U" shape of Wundt curve, which shows that *PCM* is suitable for modeling Wundt curve. Second, the curiosity distributions of different users are different. For example, the means for the five users are about 0.2, 0.4, 0.5 and 0.6, 0.7, respectively, showing that the users tend to show different levels of curiousness when they interact with the recommendation system. u_1 is relative more conservative compared with u_4 and u_5. Besides, the variance of the users' distributions are different. u_1 and u_2 have relative small variance while u_3 to u_5 have relative large variance. Small variance means that curiosity is stable, which means that users' curiosity tends not to change with topic or time, while large variance shows that the user's curiosity may vary with topic and time. Third, generally, there is an optimal *sd* for each curiosity distribution, indicating the *sd* has large chance to be accessed.

According to the above findings, we propose to use the Beta distribution, which has a flexible *pdf* and well studied parameter estimation techniques, to model the curiosity distribution C_u. Wundt curve's inverse U-shape can be estimated with Beta parameters α and β larger than 1. We apply the method of moments for the estimation of α and β. The red curves in Figure 4(a) to 4(e) illustrate the estimated *pdf*'s of five users' *Nov* random variable.

Once the curiosity distribution is estimated, we can obtain the likelihood that the user is curious about an item with *sd*, i.e., the user's *curiousness* on item i given its *sd*, denoted by $cur_u^i = pdf_u(sd)$, where *pdf* is the probability density function of C_u. cur_u^i can be viewed as a curiousness score mapped from an item's stimulus on the curiosity distribution. According to the psychology of interest [14], the pleasant feeling obtained from the process of exploring novel and surprising items is an important component of human interest. Thus, in addition to relevancy, the fact that an item's *sd* satisfies the user's curiosity is also a reason for the user to access an item. In the following section, we will introduce how to incorporate a user's curiousness about the items into the ranking utility of an *RS*.

4. CURIOSITY BASED RECOMMENDATION SYSTEM

4.1 Joint Optimization of Relevancy and Curiousness

We model the recommendation problem as a top-K problem, which selects the top K items to recommend based on the relevancy of the items to the user and the user's curiosity on the items. Let *I* be the item set of size $|I|$, and U be the user set of size $|U|$. The relevancy matrix *R* with dimension $|U| \times |I|$ is calculated using existing accuracy-based recommendation techniques. Each element $r_{u,i}$ in *R* represents the relevancy of item i to user u. We use *CR* with dimension $|U| \times |I|$ to represent the curiousness matrix, with each element $c_{u,i}$ in *CR* recording u's curiousness on i (see Section 3.5). We use *SD* with dimension $|U| \times |I|$ to represent the stimulus degree matrix, where each element $sd_{u,i}$ denotes i's stimulus degree to u. The vectors $\mathbf{R}_u$, $\mathbf{CR}_u$ and $\mathbf{SD}_u$ represent, respectively, *I*'s relevancy to u, u's curiousness over *I* and *I*'s stimulus degree to u, and correspond to one row of *R*, *CR* and *SD*, respectively. It is useful to represent the recommendation list $Rec \subseteq I$ using an $|I|$ dimensional indicator vector $\mathbf{y}$, such that $y(i) = 1$ if $i \in Rec$ and $y(i) = 0$ otherwise. $\mathbf{y}_u$ represents u's indicator vector, denoting the items to be recommended to u. In the *CBRS* framework, we try to recommend items which are highly relevant to the user and stimulative to her curiosity. This is bounded by the constraint that the items in the recommend list should not exceed her anxiety turning point. With these expressions, it is possible to express the trade-offs between relevancy and curiosity as constrained optimization problems.

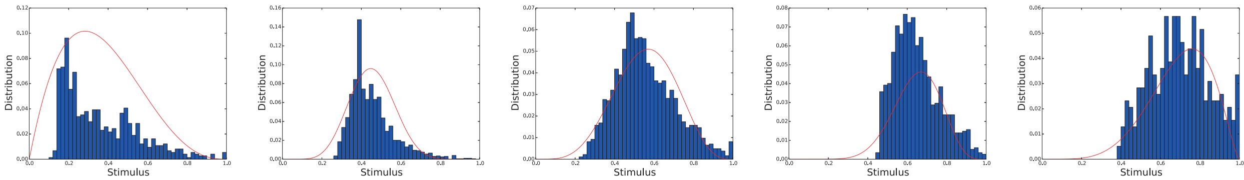
Given a fixed parameter $\theta \in$ [0,1], find the vector $\mathbf{y}^*$ such that

$$\begin{aligned} \max_{\mathbf{y}^*} \quad & (1-\theta)\alpha \mathbf{R}_u^T \cdot \mathbf{y} + \theta\beta \mathbf{CR}_u^T \cdot \mathbf{y} \\ \text{s.t.} \quad & \mathbf{SD}_u^T \cdot \mathbf{y} \leq t_{tol} \\ & \mathbf{1}^T \cdot \mathbf{y} = K \\ & y_i \in 0, 1 \forall i \in 1, ..., |I| \end{aligned} \tag{6}$$

In this optimization problem, we seek to jointly optimize relevancy and curiosity, controlled by the parameter θ. Optimal θ can be tuned with a validation set. The final two constraints specify that $\mathbf{y}$ is a binary vector with K non-zero values. Recall that the anxiety turning point introduced in Section 3.4 is an item's optimal stimulation. If the item's stimulus exceeds this optimal stimulation, the user will feel anxious. In the constraints above, t_{tol} is the aggregate tolerance threshold of the K items. Here, we define $t_{tol} = c \cdot k \cdot ATP_u$, where c is a coefficient within [0,1], K is the number of items to be selected in the recommendation list. The constraint $\mathbf{SD}_u^T \cdot \mathbf{y} \leq t_{tol}$ is referred to as the "*ATP* constraint". The *ATP* constraint tends to be more loose when c is close to 1 and more strict when c is close to 0.

4.2 Curiosity in Matrix Factorization

Matrix-factorization-based collaborative filtering (*MFCF*) has become popular in recent years due to its high accuracy [9]. From the psychology of interest [14], we know that curious emotion is an important component of interest. Thus, we believe a user's curiosity affects her choice of items and try to incorporate this additional signal into *MFCF*. There are several approaches for the incorporation, e.g. co-factorization, ensemble and regularization. Due to the space limit, we only apply the co-factorization-based method, which jointly predict the missing preferences and curiousness of the items. Other approaches will be left for future work. Specifically, *MFCF* maps both users and items to a latent space, denoted as $R \approx U^T V$, where $U \in \mathbb{R}^{l \times m}$ and $V \in \mathbb{R}^{l \times n}$ with $l < min(m, n)$, represent the users' and items' mapping to the latent space, respectively. In order to incorporate the curiosity information, we create a user-item curiousness matrix C with the same size as R, and each entry $c_{u,i}$ denotes u's curiousness about item i. Then, learning of latent factors is done by minimizing the



(a) User 1 (b) User 2 (c) User 3 (d) User 4 (e) User 5

Figure 4: Illustration of 5 users' stimulus degree distribution

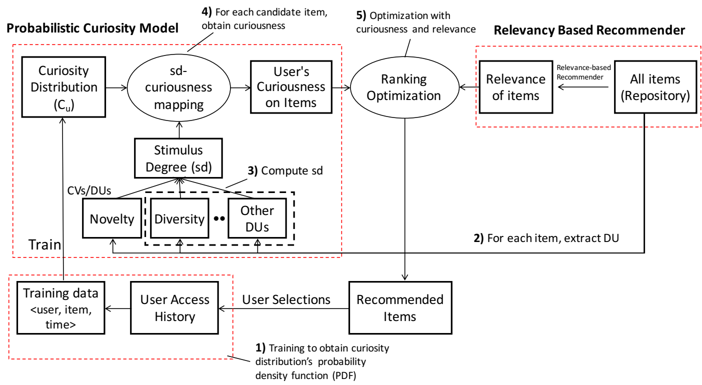


Figure 5: Illustration of CBRS framework

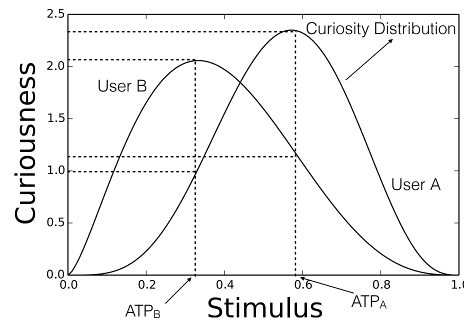


Figure 6: Illustration of two users' curiosity distribution

following sum-of-squared-errors objective functions with quadratic regularization terms:

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}^{R}(R_{ij} - U_i^T V_j)^2 + \frac{\lambda_C}{2}\sum_{i=1}^{m}\sum_{k=1}^{n} I_{ik}^{C}(C_{ik} - U_i^T Z_k)^2$$
$$+ \frac{\lambda_U}{2}||U||_F^2 + \frac{\lambda_V}{2}||V||_F^2 + \frac{\lambda_Z}{2}||Z||_F^2, \tag{7}$$

where $\mathcal{L}$ denotes the joint loss function, I^R is the preference indicator matrix where $I_{i,j}^R = 1$ if U_i has an access on V_j in R and 0 otherwise, I_*^C is defined in the same way as the curiousness matrix C, Z denotes an item's mapping to the curiosity latent space, and $\lambda_C, \lambda_U, \lambda_V and \lambda_Z$ denote the preset coefficients. Note that this minimization task is equivalent to maximizing the log-posterior distribution over U, V, Z if Gaussian priors are assumed [11]. Since we do not have users' explicit ratings in our dataset, we use u's access frequency on i to approximate the rating $r_{i,j}$ in R. This setting is commonly used in music recommendation systems [5]. Since $C_{i,j}$ is within the range [0,1], in order to normalize R and C into the same [0,1] scale, we apply logistic function $\frac{1}{1+e^{-x}}$ to each entry of R. A local minimum of the objective function given by Eq. 7 can be found by performing gradient descent in U_i, V_j and Z_k.

$$\frac{\partial \mathcal{L}}{\partial U_i} = \sum_{j=1}^{n} I_{ij}^R (U_i^T V_j - R_{ij})V_j + \lambda_C \sum_{k=1}^{n} I_{ik}^C (U_i^T Z_k - C_{ik})Z_k + \lambda_U U_i$$
$$\frac{\partial \mathcal{L}}{\partial V_j} = \sum_{i=1}^{m} I_{ij}^R (U_i^T V_j - R_{ij})U_i + \lambda_V V_j$$
$$\frac{\partial \mathcal{L}}{\partial Z_k} = \lambda_C \sum_{i=1}^{m} I_{ik}^C (U_i^T Z_k - C_{ik})U_i + \lambda_Z Z_k \tag{8}$$

4.3 Framework

The *CBRS* framework is illustrated in Figure 5. In the training phase (Step 1), by recording a user's access history on the recommendation list, we can collect the user's responses to the stimuli (i.e., the (u, i, t) triples). The stimulus degree (sd) associated with a triple is computed as described in Section 3.3. The user's curiosity distribution is then estimated based on the selected *sd*'s. In Steps 2 and 3, for each candidate item in the item repository, we compute the *DU*s based on Equation 1.[6] In Step 4, the user's curiousness on the item is obtained by mapping the item's *sd* to the *pdf* corresponding to the user's curiosity distribution C_u. In Step 5, the recommender then integrates the user's curiousness and the relevancy (calculated by traditional accuracy-based recommendation methods) of the item by solving the optimization problem depicted in Equation 6, or formulating the co-factorization problem denoted in Equation 7. The top K items which are both relevant and with high curiousness are selected into the recommendation list.

Figure 6 shows the curiosity distributions of two users *A* and *B* to illustrate how two users with different curiosity models respond differently to the same item. According to their curiosity distributions, *A* is more curious than *B*. Now, a candidate item is recommended to both *A* and *B*, and its *sd* is high (at 0.6, denoted by ATP_A in the figure). According to the mapping, *A* will have a higher curiousness score than *B*. If the item's relevance is more or less the same to both users, the the item's high curiousness score for *A* would push the item into *A*'s recommendation but its low curiousness score for *B* may not be able to push itself into *B*'s recommendation. However, if the item has small *sd* (at 0.3, denoted by ATP_B in the figure), the situation will be reversed. *B*'s curiousness score will be higher than that of *A*, and the item will be recommended to *B* instead of *A*.

5. EXPERIMENTAL RESULTS

In this section, we first describe the dataset used in the experiment. We then apply various metrics and vary the parameters to evaluate the performances of *CBRS* and the baselines.

5.1 Dataset and Parameter Setting

We use the public dataset "Last.fm Dataset - 1K users" [1] in the experiments. The whole dataset is 2.53GB in size. It contains 19,150,868 chronologically ordered listening records of 992

[6] In this paper, only novelty is considered. Other *DU*s such as diversity and serendipity can be plugged into the framework by developing the corresponding formula to compute the *sd* of each *DU*.

Table 1: Statistics of the Dataset used in the Experiments

No. of users	937
No. of artists	83,908
No. of tracks	960,415
No. of records	16,955,328
Avg history span	762 days

unique users over several years till May 5, 2009. Each record has the format "user id, time stamp, artist id, artist name, track id, track name". The primary key is "(user id, track id, time stamp)", and we use (u, i, t) to denote the record of user u having accessed track i at time t. We remove all records without track id or artist id (every record in the dataset has a time stamp) and users with too few records for learning their curiosity distributions. The statistics of the cleaned dataset is given in Table 1.

To conduct an experiment, we need a training dataset, a historical dataset and a test dataset derived from the original dataset. The training dataset is used to learn the curiosity function for each user. The historical dataset is needed for computing the stimulation degree *sd* for each record in the training dataset, because *sd* is determined by novelty and novelty in turn depends on a user's historical accesses to the items (see below for further details). When a method is evaluated, we use the historical and training datasets to produce the recommendations and the test dataset as ground truth to evaluate the performance metrics of the recommendations.

Since the original dataset contains a large number of records, we divide it into 10 consecutive windows, $w_0, ..., w_9$, each of which contains one-tenth of the records in the dataset (denoted as $|w|$). We use three consecutive windows w_i, w_{i+1}, w_{i+2} to form, respectively, the training, historical and test datasets for one experimental run. That is, we use w_0, w_1 and w_2 in the first run, w_1, w_2 and w_3 for the second, ..., and w_7, w_8 and w_9 for the eighth run. Thus, we can perform eight experimental runs and average the results to obtain the value of a performance metric.

In the training phase, we need to compute the stimulation degree $sd^t_{u,i}$ for each (u, i, t) record in the training dataset. Given a record $r = (u, i, t)$ in the training dataset, we take the $|w|$ records preceding r as r's historical window[7] $H^t_{u,i}$, and compute r's $sd^t_{u,i}$ using Equations 1 to 5. The procedure is repeated for all records in the training dataset. At the end, each (u, i, t) record in the training dataset is associated with a $sd^t_{u,i}$ value. For each user u, we can obtain a series of sd values indicating u's accessed *sd*'s in the training dataset and estimate u's curiosity distribution as described in Sectionrefsubsec:est.

The parameter setting in the experiments is as follows: the frequency scaling coefficients corresponding to ρ_a and ρ_i in Equation 3, and ρ_{tag} in Equation 5 are set to 0.1, the time scaling coefficient corresponding to ρ_t in Equations 4 and 5 is set to 0.01. The goal of these settings is to ensure that different novelty components ($SF^t_{u,i}$, $SR^t_{u,i}$, $sim^t_{u,i}$) are within the (0,1) scale. We set $\lambda_C, \lambda_U, \lambda_V and \lambda_Z$ in Equation 8 to 0.02. K denotes the number of items in the recommendation list, which is set to 10 by default and varied in Table 2 to study its impact on performance.

5.2 Performance Metrics

This subsection introduces the performance metrics used in the experiments.

[7]The historical window for each record r contains the same number of accessed items to avoid bias because *sd* depends on the number of accessed items in the historical window H^t_u (see Equations 3 to 5).

5.2.1 Novelty Fitness

As discussed in Section 1, a major difference between *DORS*s and *CBRS* is that *CBRS* delivers recommendations with the proper amount of novelty to suit a user's curiosity. We introduce *novelty fitness* (NF_u) to measure the fitness between the novelty of the recommended items and the user's novelty appetite nov_u (defined in Section 3.3). As defined in Equation 9, NF_u is the root-mean-square error (RMSE), where K denotes the top K recommended items and $nov_{u,i}$ denotes the novelty of the recommended item i to u. A smaller *NF* value means that the novelty of the top K recommended items has a better fit to the user's novelty appetite.

$$NF_u = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(nov_{u,i} - nov_u)^2} \tag{9}$$

5.2.2 Recommendation Precision

We use precision to measure how accurate the recommendations produced by a recommendation method predict the user's future accessed items. Formally, we define precision as $\frac{1}{|U|}\sum_{u=1}^{|U|}\frac{|R_u \cap T_u|}{|R_u|}$, where U denotes the user set, R_u denotes the recommendation list for user u, T_u denotes the set of tracks that u has accessed in the test dataset (i.e., the ground truth). Note that we do not use recall as an evaluation metric, since the size of the recommendation list in Top-K recommendation is fixed.

5.2.3 Inter User Similarity

Since *CBRS* makes recommendations adapted to a user's personal curiosity, we expect its recommendations to have larger differences across users comparing to traditional *RS*s. To test the validity of this expectation, we use inter-user similarity (*IUS*) proposed in [20] to evaluate the system-wide personalization effect of a recommender. Equation 10 defines $IUS_{i,j}$ as the proportion of overlap between recommendation lists L_i and L_j received by users i and j,

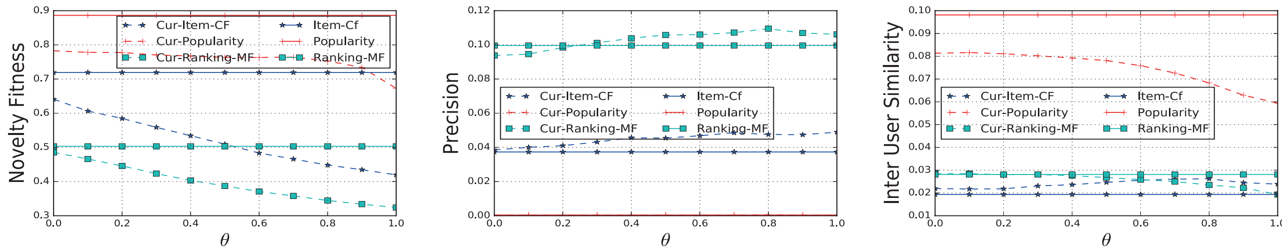
$$IUS_{i,j} = \frac{|L_i \cap L_j|}{K} \tag{10}$$

The *IUS* of a *RS* is the average of $IUS_{i,j}$ over all pairs of users. A large $IUS_{i,j}$ means a high similarity between recommendation lists received by users i and j, and a large *IUS* indicates a weak personalization effect of the *RS*.

5.3 Influence of Relevancy-Curiousness Balance Weight

According to Equations 6, the selection criteria of *CBRS* is to pick the items which are both relevant and stimulative to the user's curiosity, bounded by the *ATP* constraint. The relative weight of an item's relevancy and the user's curiousness on the item is controlled by θ. A large θ promotes items with high curiousness, while a small θ favors items with high relevance. In this subsection, we investigate the effect of θ on the performance of a recommendation system in terms of the three metrics introduced in the preceding subsections. Since the optimization task for all items is very time consuming, we take the top 50 items as candidate items for each baseline recommender. Then, the top-K recommendation task is to select K items from the candidates.

To facilitate comparison, we pick two sets of recommenders, non-curiosity-aware (*non-CBRS*) and curiosity-aware (*CBRS*). For *non-CBRS*, we pick three baseline recommenders, namely, Popularity, Item-based collaborative filtering (Item-CF) and ranking ma-



(a) Novelty Fitness, small better (b) Precision, large better (c) Inter User Similarity, small better

Figure 7: Variation of θ in joint optimization (Equation 6)

trix factorization (Ranking-MF).[8] For *CBRS*, we incorporate curiosity into the three baselines by solving the joint optimization problem in Equation 6. The *CBRS* recommenders are denoted as *Cur-Popularity*, *Cur-Item-CF* and *Cur-Ranking-MF*.

Figure 7 illustrates the influence of θ on the performance metrics. A general observation for all three metrics is that as θ becomes larger (i.e., curiousness plays a more important role in recommendation), the performance of the recommenders generally increases. In Fig. 7(a), a small *NF* means a better fit between the novelty of the recommendations and the users' novelty appetite. We can observe that *CBRS* has lower *NF* (i.e., better) than *non-CBRS*, because *non-CBRS* does not consider curiosity at all and hence cannot optimize the recommendations for novelty. This also explains why the *NF* of *non-CBRS* is flat w.r.t θ. Further, the *NF* of *CBRS* improves as θ increases, because as θ increases *CBRS* will bias towards curiosity, thus producing recommendations with better novelty fitness. Finally, for both *CBRS* and *non-CBRS*, Ranking MF is better than Item-CF, which in turn is better than Popularity in terms of *NF*. It is interesting to note that even for *non-CBRS*, which does not consider curiosity, Popularity has the worst *NF*, because popular items have less novelty.

In Fig. 7(b), we can observe the same relative performance as in Fig. 7(a), that is, Ranking-MF has higher precision than item-CF, which in turn has higher precision than Popularity for both *CBRS* and *non-CBRS*. It is interesting to note that although *CBRS* is designed to balance relevance with curiousness, its precision not only has not suffered but is better than *non-CBRS*. This can be attributed to the fact that *CBRS* optimizes novelty according to *each user*'s curiosity distribution. For example, if a user has low curiosity, *CBRS* would favor items that are less novel (i.e., more familiar) to the user. Since conservative users tend to listen to familiar tracks, precision is improved. In Fig. 7(c), a small *IUS* indicates a large difference and hence higher personalization effect between the recommendation lists received by the users. Again, Popularity (both *CBRS* and *non-CBRS*) performs the worst because it favors hot items and thus its recommendations to different users tend to be more similar. We can also see that Item-CF has the lowest *IUS* among all recommenders because it does not recommend non-novel items. Further, its *IUS* is also lower than Cur-Item-CF because the latter would consider non-novel items (which could be hot items) if the user is judged to be conservative.

From the previous experiments, we can see that optimal performance is reached when $\theta = 1.0$. This is equivalent to re-ranking the 50 candidate items purely based on the user's curiousness on the items (see Equation 6) with the *ATP* constraint. This is because the candidate items produced by the recommenders already have very high relevance so further optimizing the items' relevance has much smaller effect than re-ranking the items based on their curiousness scores. Thus, in the remainder of this paper, when we perform recommendation with the joint optimization method discussed in Section 4.1, we use $\theta = 1.0$. In summary, take the best performer *Ranking-MF* as an example, *CBRS* produces 35.6%, 6.6%, 31.3% improvement in *NF*, precision, and *IUS*, respectively, compared to *non-CBRS* when $\theta = 1.0$. It demonstrates the superior effectiveness of *CBRS*.

5.4 Comparing to DORS

From Section 2, we know that existing *DORS*s have proposed various *DU*s to complement accuracy. However, they suffer from the curiosity mismatch problem. In contrast, *CBRS* believes a user's need of novelty is non-linear and depends on her curiosity. In this section, we compare the performance of *CBRS* and *DORS*s. Figure 8 illustrates the result.

To setup the experiment, we pick the best performers *Ranking-MF* and *Cur-Ranking-MF* in Section 5.3 as baselines. Our proposed method is labeled as *Pure* and *Rel-Cur-ATP* in Figure 8, representing, respectively, pure application of *Ranking-MF* and joint optimization according to Equation 6, which considers an item's relevancy, curiousness and the *ATP* constraint. *Rel-Nov* performs joint optimization between relevancy and novelty as specified in Equation 6 without the *ATP* constraint, reflecting the implementation of traditional *DORS*s. *Rel-Nov-ATP* is the same as *Rel-Nov* except the addition of the *ATP* constraint. From the comparison between *Pure* and *Rel-Nov*, we can find that with novelty complementing relevancy, *IUS* decreases, indicating the users' recommendation lists become less similar to each other. This is an advantage of *DORS*s. However, recommendation precision is sacrificed and the recommendations deviate more from the user's real novelty appetite (i.e., novelty fitness score increases). This trade-off is a common phenomenon in many *DORS*s. From the comparison between *Rel-Nov* and *Rel-Nov-ATP*, we can observe that with the *ATP* constraint, *Rel-Nov-ATP* can retain the benefit of small *IUS* in *DORS*s without suffering from decreased precision. The novelty fitness is even better compared to *Pure*. This could be attributed to the incorporation of the *ATP* constraint, which models individual user's personalized novelty appetite. From the comparison between *Rel-Nov-ATP* and *Rel-Cur-ATP*, we can observe the importance of modeling a user's personalized non-linear novelty appetite, since novelty fitness and precision are improved about 33% and 29%, respectively at $\theta = 1.0$, while retraining the improvement of *IUS*.

5.5 Incorporate Previously Accessed Items

Most existing *RS*s aim to recommend completely novel (new)

[8]Popularity and Item-CF denote methods based on item popularity and similarity, respectively. Ranking-MF denotes the factorization machine approach [12]. An overview and implementation details of the three recommenders can be obtained from Dato Graphlab (https://dato.com/) API.

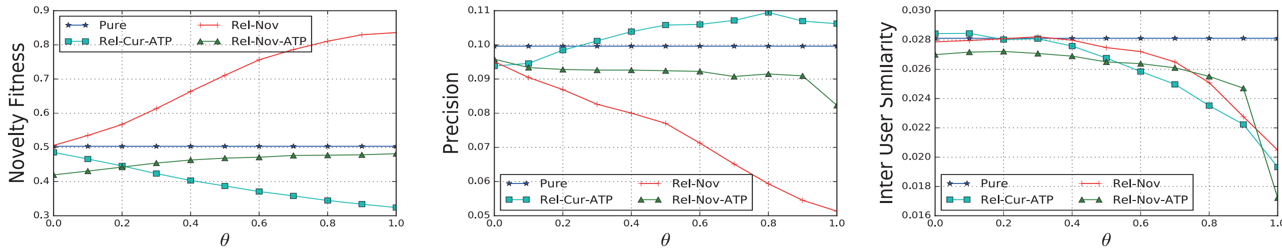


Figure 8: Compare *CBRS* to *DORS*, proving the importance of modeling user's personalized needs of novelty

items, because they assume that users will not be interested in items they have accessed before. This assumption is reasonable in some applications such as movie recommendation, but in some applications such as music listening users may repeatedly access previously accessed items. It is important to note that novelty in *CBRS* is not binary. Instead, novelty is a continuous value affected by many factors. For example, a music track that the user has listened to can still have high novelty to the user if she listened to the track or similar tracks only occasionally and a long time ago. In this subsection, we will study how the inclusion/exclusion of previously accessed items in recommendations affects performance. We use the same experiment setting as in Section 5.3, except that in addition to the 50 candidate items we randomly add N items that the user has previously accessed (*old items*). Now the recommendation problem is to pick the top K items for each user from these $N+50$ candidates. We set $K = 10$, and vary N from 2 to 20.

The result is shown in Figure 9, from which several observations can be made. First, with the additional N old items, both *non-CBRS* and *CBRS* have improved performance in all of the three performance metrics. This can be attributed to the users' habit of listening to their favorite songs frequently, and adding old items in the recommendations will capture this user behavior. Second, although the performance of both *non-CBRS* and *CBRS* is improved, *CBRS* produces much larger improvement than *non-CBRS*. For example, compared with *Ranking-MF*, the best performer *Cur-Ranking-MF* achieves 85.5%, 166.7%, and 77.9% improvement on *NF*, precision, and *IUS*, respectively, when $N = 20$. This is attributed to *CBRS*'s ability to capture a user's unique novelty appetite. That is, if a user likes to listen to old songs, *CBRS* will model her as a conservative listener and recommend old songs to her and the contrary is also true. Thus, *CBRS* can better utilize the N old items to satisfy the user's preference on old songs. This is a significant contribution of this paper since very few works, if any, have discussed the importance of non-novel items to the users and how to incorporate them into a recommendation list. Third, even when $N = 0$, i.e., only new items are recommended, *CBRS* still outperforms *non-CBRS* even though the improvement is not as much as when old items are included. This can be attributed to the fact that new items still have different degrees of novelty to the user because of their properties (e.g. a new track may have low novelty to a user when the user is familiar with the genre or performer of the track). Finally, with N becomes larger, the performance difference among the *CBRS* recommenders narrows down. As discussed, users would listen to their favorite tracks repeatedly. Thus, as N increases, *CBRS* recommenders would recommend more from the N old items. This would greatly improve their performance and thus narrow down their performance gap. This is indeed an advantage of *CBRS*, since curiosity can be incorporated into any specific recommender to make it curiosity-aware and improve its performance.

5.6 Overall Performance Comparison

From Sections 5.3 to 5.5, all recommendations are based on Equation 6, which performs joint optimization between relevancy and curiousness. In this subsection, we will compare *CBRS* to other state-of-the-art *PDORS* methods. As introduced in Section 2, the only two works on *PDORS* are [17] and [7]. We choose the *PPBRS* method proposed in [7] as the baseline, because it is more recent and closer to our work. *PPBRS* learns a personalized novelty preference score for each user by logistic regression. For comparison, we use *Ranking-MF* as the reference recommender, instead of the Item-based collaborative filtering in [7]. We also choose the standalone *Ranking-MF* as another baseline, because it performs best among the traditional accuracy-based recommenders (shown in Figure 7) and does not consider curiosity. As introduced in Sections 4.1 and 4.2, we propose two ways to incorporate curiosity information into an *RS*. Thus, our proposed *CBRS* methods are Cur-Ranking-MF and Cur-MFCF, corresponding to Equations 6 and 7, respectively. The results are shown in Table 2.

Comparing the two baselines *PPBRS* and *Ranking-MF*, we can see that although *PPBRS* can provide more personalized novelty to the users (small *NF* score), it is limited in enhancing precision. This is because although *PPBRS* learns a personalized "relevancy-novelty" balance parameter, it applies the parameter to *all items* in the same way. That is, if the parameter favors novelty, then all items' novelty scores will be emphasized in the same way. This may result in some recommended items overly novel to the user because they may already have high novelty scores and amplifying them further will make them too novel for the user's curiosity level. For the same reason, some recommended items could become too boring to the user. Comparing *PPBRS* and *CBRS* (*Cur-Ranking-MF* and *Cur-MFCF*), we can see that *CBRS* performs better in both novelty fitness and precision (e.g. *Cur-Ranking-MF* achieves 32.1%, 7.8%, 18.8% improvement over *PPBRS* on *NF*, precision, *IUS*, respectively, when $K = 10$) because it is able to capture a user's non-linear novelty appetite. Between the two *CBRS* recommenders, *Cur-Ranking-MF* performs better than *Cur-MFCF*. This might be attributed to the ability of explicit joint optimization to combine an item's relevancy and the user's novelty appetite more optimally. Detailed study on how to better incorporate the curiosity signal into the traditional recommendation framework is beyond the scope of this paper and will be treated in future work.

6. CONCLUSION

Various Discovery-Oriented Recommendation Systems (*DORS*s) have been proposed to address the accuracy overloading problem of traditional recommendation systems. They introduce Discovery Utilities (*DU*s) such as novelty, diversity and serendipity as additional ranking dimensions to ensure that recommendations are not dominated by accuracy. However, *DORS*s do not consider the fact that different users have different appetite for *DU*s and suffer

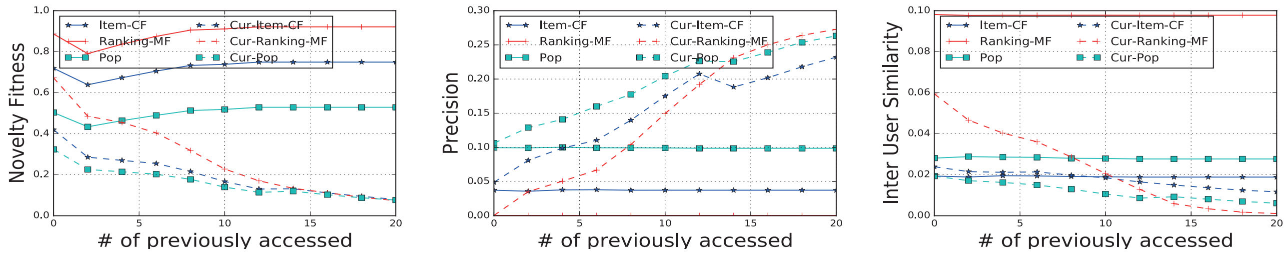


Figure 9: Performance when previously accessed items are included

Table 2: Performance comparison between *CBRS* and *PPBRS*

Model	NF			Precision			IUS		
	k=5	k=10	k=20	k=5	k=10	k=20	k=5	k=10	k=20
Ranking-MF	0.455	0.503	0.463	0.105	0.0973	0.0857	0.0220	0.0281	0.0392
PPBRS	0.417	0.477	0.449	0.0995	0.0983	0.0961	0.0209	0.0238	0.0262
Cur-MFCF	0.324	0.368	0.361	0.112	0.106	0.101	0.0195	0.0208	0.0223
Cur-Ranking-MF	0.278	0.324	0.268	0.120	0.106	0.113	0.0170	0.0193	0.0244

from the curiosity mismatch problem. In this paper, we present the Curiosity-based Recommendation System (*CBRS*) framework for integrating the curiosity and relevance aspects of recommendations. The Probabilistic Curiosity Model (*PCM*) models a user's unique appetite for novelty. *PCM* is formulated based on the curiosity arousal theory and Wundt curve in psychology [4]. It uses the Beta Distribution to model Wundt curve representing a user's non-linear desire of novelty. With *PCM*, for each user we can obtain a curiousness score for each candidate item. Joint optimization and matrix factorization methods are used to incorporate the curiosity signal into *RS*s. Experimental results show that *CBRS* significantly outperforms the baselines in various performance metrics. An important finding is that *CBRS* can provide personalized recommendations adapted to an individual user's curiosity and at the same time improve the recommendation accuracy. We also study the impact of allowing previously access items (i.e., non-novel items) in the recommendations and show that *CBRS* can recommend the right mix of non-novel and novel items to optimize the performance. This is an important finding because in many applications (e.g., music recommendation) users may repeatedly access some items (e.g., favorite songs).

For future work, we plan to study other *DU*s such as diversity and surprisal and integrate curiosity and social relationship into the *CBRS* framework. Finally, we recognize the mutual benefit between our work and psychology research and plan to apply web-scale data in the study of human's psychological curiosity behavior.

Acknowledgement

Research reported in this paper was supported by the Research Grants Council, HKSAR, GRF No. 615113.

7. REFERENCES

[1] Lastfm homepage, 2015.
[2] A. Anglade, O. Celma, B. Fields, P. Lamere, and B. McFee. Womrad: 2nd workshop on music recommendation and discovery. 2011.
[3] I. Avazpour, T. Pitakrat, L. Grunske, and J. Grundy. Dimensions and metrics for evaluating recommendation systems. Springer Berlin Heidelberg, 2014.
[4] D. Berlyne. *Conflict, Arousal and Curiosity*. McGraw-Hill, 1960.
[5] E. Bernhardsson. Collaborative filtering at spotify, 2013.
[6] W. Fan, M. D. Gordon, and P. Pathak. Personalization of search engine services for effective retrieval and knowledge management. ICIS '00.
[7] K. Kapoor, J. A. Kumar, and P. Schrater. "i like to explore sometimes": Adapting to dynamic user novelty preferences. RecSys '15.
[8] P. Knees and M. Schedl. Music retrieval and recommendation: A tutorial overview. SIGIR '15.
[9] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 2009.
[10] X. Li and W. B. Croft. Time-based language models. CIKM '03.
[11] H. Ma, Yang, and I. King. Sorec: Social recommendation using probabilistic matrix factorization. CIKM '08.
[12] S. Rendle. Factorization machines. ICDM '10.
[13] G. Shani and A. Gunawardana. Evaluating recommendation systems. *Recommender Systems Handbook*, 2011.
[14] P. J. Silvia. Interest the curious emotion. *Current Directions in Psychological Science*, 2008.
[15] P. Vargas, Castells. Rank and relevance in novelty and diversity metrics for recommender systems. RecSys '11.
[16] S. Vargas. Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. SIGIR '14.
[17] F. Zhang, K. Zheng, Yuan, and X. Zhou. A novelty-seeking based dining recommender system. WWW '15.
[18] M. Zhang and N. Hurley. Avoiding monotony: Improving the diversity of recommendation lists. RecSys '08.
[19] Y. C. Zhang and T. Jambor. Auralist: Introducing serendipity into music recommendation. WSDM '12.
[20] Y. Zhou, T.and Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 2010.
[21] C.-N. Ziegler and G. Lausen. Improving recommendation lists through topic diversification. WWW '05.