One Size Does Not Fit All: Modeling Users' Personal Curiosity in Recommender Systems

Abbas, Fakhri
University of North Carolina at Charlotte
fabbas1@uncc.edu

Niu, Xi
University of North Carolina at Charlotte
xniu2@uncc.edu

ABSTRACT

Today's recommender systems are criticized for recommending items that are too obvious to arouse users' interest. That's why the recommender systems research community has advocated some "beyond accuracy" evaluation metrics such as novelty, diversity, coverage, and serendipity with the hope of promoting information discovery and sustain users' interest over a long period of time. While bringing in new perspectives, most of these evaluation metrics have not considered individual users' difference: an open-minded user may favor highly novel or diversified recommendations whereas a conservative user's appetite for novelty or diversity may not be that large. In this paper, we developed a model to approximate an individual's curiosity distribution over different levels of stimuli guided by the well-known Wundt curve in Psychology. We measured an item's surprise level to assess the stimulation level and whether it is in the range of the user's appetite for stimulus. We then proposed a recommendation system framework that considers both user preference and appetite for stimulus where the curiosity is maximally aroused. Our framework differs from a typical recommender system in that it leverages human's curiosity to promote intrinsic interest with the system. A series of evaluation experiments have been conducted to show that our framework is able to rank higher the items with not only high ratings but also high response likelihood. The recommendation list generated by our algorithm has higher potential of inspiring user curiosity compared to traditional approaches. The personalization factor for assessing the stimulus (surprise) strength further helps the recommender achieve smaller (better) inter-user similarity.

KEYWORDS

Surprise, Curiosity, Recommender System, Psychology

ACM Reference format:
Abbas, Fakhri and Niu, Xi. 2019. One Size Does Not Fit All: Modeling Users' Personal Curiosity in Recommender Systems. In *Proceedings of RecSys '19: the 13th ACM Conference on Recommender Systems, Copenhagen, Denmark, 2019 (Recsys '19),* 9 pages.
DOI: 10.1145/1122445.1122456

Recsys '19, Copenhagen, Denmark

978-1-4503-9999-9/18/06…$15.00
DOI: 10.1145/1122445.1122456

1 INTRODUCTION

Today's recommender systems have been criticized for having the problem of "information filter bubble" [26] or "echo chamber" [2] by offering people close matches with what they have seen already, but not exposing them to a broader range of information. To burst the bubble and break the chamber, the recommender systems research community has incorporated some "beyond accuracy" objectives such as novelty [38], unexpectedness [1, 23], serendipity [10, 12]. Among these "beyond accuracy" objectives, one that receives little attention is curiosity, a strong desire to know or learn something. Curiosity is central in human information seeking [17] and therefore believed important in recommender systems to promote users' intrinsic interest to continue using the system.

In this paper, we built a personal curiosity distribution curve for each user. The users' access history with the system was used to estimate their curiosity levels for different recommendation stimuli. The estimation was then used to suggest new items that were highly likely to stimulate the user's curiosity. The curiosity model has been incorporated into a traditional recommender system. The result is a new proposed recommender framework that predicts user preference, infers what they are curious about, and then synthesizes recommendations.

Specifically, our curiosity distribution curve was inspired by the probabilistic curiosity model (PCM) developed by Zhao et al. [38]. PCM was guided by the early German psychologist Wilhelm Wundt, who proposed the Wundt curve [37] that describes the relationship between the amount of stimulus and the pleasant feeling. According to the curve, as in Figure 1, too little stimulus will not be exciting whereas too much will cause anxiety. This creates a stimulus "sweet spot" where the pleasant feeling is near its peak. This "sweet spot" is highly dependent on an individual. Built on PCM, we proposed to use surprise to represent the stimulus and curiosity to represent the pleasant feeling. We developed computational approaches to quantify both concepts of surprise and curiosity, and approximated the Wundt curve in a quantitative way. Then we used an item's stimulation distance to the "sweet spot" as a criterion to re-rank the items predicted by the traditional collaborative filtering techniques. The re-ranking algorithm promotes the items that have sufficient surprise amount to be exciting but not too much to be intimidating. The evaluation experiments have demonstrated that our recommender framework has balanced relevance with curiosity in order to increase the user response likelihood.

We used a book recommendation dataset from Amazon [19] to illustrate the idea. The dataset is information rich not only because of its large size but also the abundant users' access history and rating history which date back to the year 1996. Also, book reading behavior is highly driven by personal taste and curiosity.

2 RELATED WORK

This research brings together the concept of curiosity, incorporation of curiosity in intelligent computational systems, and computational models of surprise in artificial intelligence (AI).

2.1 The Concept of Curiosity

This study was guided by the early German psychologist Wilhelm Wundt, who proposed the Wundt curve [37], as shown in Figure 1, that describes the positive response from a stimulus initially increased. As the stimulus grew more intense, the aversion or anxiety overtook it. This creates a stimulus "sweet spot", within which positive response is near its peak. This peak is highly dependent on an individual's experiences. The seeking of stimuli and experiences within this zone is known as curiosity, and as curiosity leads to new knowledge, the "sweet spot" shifts or expands, leading to renewed curiosity about newly adjacent knowledge. This iterative development cycle is the grounding for our recommender framework: the hypothesis that encouraging curiosity will promote new information discovery and sustain users' long-term interest.

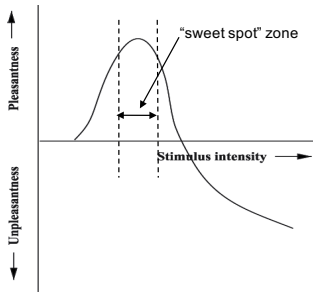


Figure 1: Illustration of the Wundt curve

2.2 Incorporation of Curiosity in Intelligent Computational Systems

In the field of Artificial Intelligence (AI) and Robotics, various computational models have been developed to simulate and stimulate curiosity. According to Wu et al. [36], most of these computational approaches model the curiosity arousal process as a two-step process: identify one or several stimulus variables and appraise the stimulus level; then based on the stimulus level, evaluate the curiosity level. In the first step, some models used a single variable to determine the stimulation value. For example, Saunders and Gero [28] developed a computational model of curiosity for intelligent design agents, focusing on the appraisal of novelty. Novelty is the key for evaluating the curiosity arousal and therefore the selection of good design patterns. Other models combined several stimulus variables to determine the stimulation level. For example, in Wu et al. [36], they used the concept of curiosity in a virtual companion to detect potentially interesting learning objects for users and help them avoid the feeling of being lost. They considered four stimulus variables: novelty, uncertainty, conflict, and complexity, and proposed a measure for each of them. These previous studies have marked milestones for applying curiosity in intelligent systems. They have inspired our motivation of applying such concept into recommender systems where human exploring and information discovery is also desirable.

As the second step in the two-step process of modeling curiosity, the level of curiosity is evaluated through a mapping from the stimulation value to the curiosity value. Some models assumed a linear relationship between stimulation and curiosity such as [36]. Other models simply used the stimulation value as the curiosity value such as [18, 25, 30]. Still other models followed the principle of "sweet spot" by explicitly simulating the Wundt curve, which represents a nonlinear mapping from stimulation to curiosity such as models in [20, 28]. These models avoided too small and too big stimuli in their stimulus selection approaches. In this study, we are informed by these studies and further used a mathematical approach to quantify the thresholds and the "sweet spot" along the Wundt curve.

2.3 Computational Models of Surprise in Artificial Intelligence

Surprise, as a potential stimulus variable, has received substantial attention in AI research these years. Studies of computational creativity find that unexpected discovery leads to reflective thinking of the current problem, which in turn leads to further unexpected discoveries [33]. According to Grace et al. [6], this reflective behavior suggests that surprise is one possible trigger for curiosity. There are three interpretations for surprise in the literature of computational curiosity. The first one interprets surprise as the difference between an expectation and the real outcome. Prediction error matches well with this interpretation and has been utilized in many curiosity models to measure the level of surprise, such as the studies in [3, 29, 31, 34]. The second interpretation describes surprise as the change of knowledge. Storck et al. [32] modeled this type of surprise using the information gain before and after an observation. The third one is using improbability of existence of an item or an event, as proposed by Macedo and Cardoso [18]. Using improbability as surprise, a series of studies by Grace and Maher [7–9] have developed a personalized curiosity engine called PQE that recommends surprising and interesting recipes to users to encourage their curiosity and help diversify their diet. Their surprise model was based on how unlikely the ingredients co-exist in a recipe. Niu et al. [24] adopted several Information Theory metrics such as entropy and mutual information to calculate how surprising a news article is to its reader. These previous studies informed this study of the basic idea of using low likelihood or rare occurrence to measure surprise. Built on but different from these studies, this study further factored a person's previous experience into surprise calculation because the same item is believed to carry different amounts of surprise and therefore has different stimulation levels for different users.

3 THE FRAMEWORK ARCHITECTURE

Our proposed recommender framework consists of three main components, as shown in Figure 2. The Preference Model, the Curiosity Model, and the Recommendation Generator. The Preference Model captures the user interest to recommend preferred items. The Curiosity Model estimates what makes the user curious using the user's previous accessed items. The Recommendation Generator

uses the knowledge from both the Preference Model and the Curiosity Model, searches for items, ranks them based on a balance between preference and curiosity, and recommends to the user.

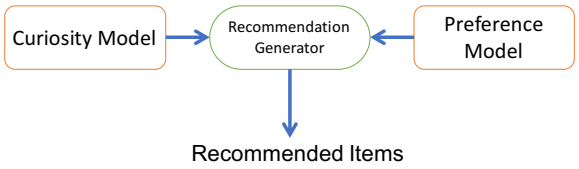


Figure 2: The architecture of the proposed recommender system framework

3.1 Preference Model

The Preference Model makes use of a user's previous ratings as the user profile and then adopts the state-of-the-art collaborative filtering (CF) recommender techniques to identify a set of items that are most preferable to the user. Collaborative filtering (CF) techniques typically have higher accuracy compared to the content-based techniques, and are more generalizable in different domains independent of the item content representation. Having this Preference Model as a separate component enables us to experiment with different off-the-shelf CF algorithms without affecting other components of the framework. This facilitates the later experimentation, evaluation, and rapid deployment.

3.2 Curiosity Model

The Curiosity Model, a core component of the recommender framework, also uses a user's access history to infer what the user will feel curious about. Inspired by Zhao et al.'s study [38] that developed a probabilistic curiosity model (PCM), this component develops a probabilistic curiosity curve (PCC) for each individual user, and informs the Recommendation Generator about where the stimulation sweet spot is to stimulate the user's curiosity with a high likelihood.

3.2.1 Preliminaries: Probabilistic Curiosity Model (PCM) by Zhao et al. [38]. According to Berlyne's curiosity arousal model [4], a user receives stimuli and would only respond to stimuli which can arouse their curiosity. The curiosity arousal model essentially describes a process of how a user selectively responds to the stimuli. For a recommender system, each recommended item presents a stimulus to the user. The strength of a stimulus (SI) could be defined by a number of factors that are extracted from some measurable properties of a stimulus. It is noteworthy that the same item (stimulus) may produce different SIs to different users because of individual difference in curiosity. In order to capture the individual difference in curiosity, Zhao et al. [38] proposed a probabilistic curiosity model (PCM), which is a probabilistic view of the Berlyne's model. It models a user's selected or responded SI as a random variable, and curiosity as the probability distribution of the random variable. In this way, a user's stimulus selection (response) process can be interpreted as drawing a sample (s stimulus) from her curiosity distribution. Adopting PCM, this component (the Curiosity Model) develops a probabilistic curiosity curve (PCC) for each individual user based on the user's past access history. The left panel in Figure 3 illustrates a PCC for a hypothetical user. The right panel in this Figure lists some example points along the curve, depicting a user's stimulus selection process under the guidance of the user's PCC. SIs around 0.6 are the level where the user's curiosity will be maximally aroused, therefore will be selected (responded) with a maximal probability. The user may also select other SIs, but the chance is smaller. The stimulus point where the curiosity is maximally aroused is called stimulus prime (SP).

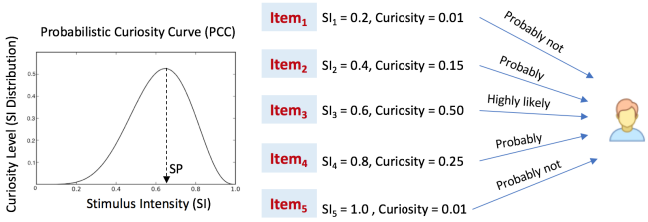


Figure 3: Illustration of the probabilistic curiosity model

Built on Zhao et al.'s work [38], our study contributes to (1) propose and use personalized surprise metrics to quantify SI, (2) fit a curve for PCC for each user using a mathematical distribution curve, and identify the stimulus prime (SP) point where curiosity is maximally inspired, (3) calculate the stimulation distance to SP as a way to assess whether the stimulus is in the range of the "sweet spot", and incorporate the distance measure into recommender algorithms, and (4) propose evaluation metrics, such as Discounted Cumulative Curiousness (DCC) to test whether proximity to such such "sweet spot" zone can arouse the actual user response likelihood (curiosity). Below, we will introduce these new contributions in more detail.

3.2.2 Quantifying SI: computational measure of personalized surprise amount. We used the amount of surprise as the curiosity stimulus, since surprise captures all the elements of stimulus factors identified by Berlyne [4], such as novelty, conflict with expectation, hard to explain, etc. We follow the definition of surprise as violation of expectation [22]. A low likelihood of the expectation would be a surprise to the user. This surprise should be personalized in that the surprise is specific to the user, but not necessarily to others or the entire society. To quantify surprise, a computational measure of surprise was proposed in this study, which consists of two steps. First we built an objective surprise measure based on the society's collective knowledge as expectation. Second, a personalization factor was incorporated to discount the objective surprise to reflect the personalized level of surprise.

In the first step, we adopted the computational model in Niu et al.'s study [24] for its proven validity. We will briefly introduce the model here. Each item was represented as a "bag" of its elements. For example, the book "The Prophet" could be represented as a bag of its topics: humanities, religion, and love poems. We then measured objective surprise as how unlikely these topics co-occur

in one book. The topic religion tends to co-occur with humanities with a high likelihood, but not as much co-occurring with love poems. Expectations of co-occurrence likelihood have been implicitly formed by our collective knowledge, and were computationally constructed using a large collection of such items or some external knowledge base. A surprise in that sense is: "Seeing the topic religion is surprising given seeing the topic love poems."

To capture the heuristics of co-occurrence likelihood, Pointwise Mutual Information (PMI) [5] was used to calculate how much more likely than expected it is that an element e_i occurs given the occurrence of another element e_j. We call this pairwise surprise score s, as in Equation 1:

$$s(e_i, e_j) = -PMI(e_i, e_j) = -log_2 \frac{p(e_i, e_j)}{p(e_i)p(e_j)} \tag{1}$$

where $p(e_i)$ and $p(e_j)$ represent the individual occurrence probabilities of the elements e_i and e_j, and $p(e_i, e_j)$ represents the joint occurrence probability of the two. In this equation, the lower part of the log fraction represents the expectation of these two elements in the collection, and the upper part represents the actual or observed likelihood for this particular combination. The ratio between the observed likelihood and the expected likelihood reflects the amount of pairwise surprise.

Since many items have more than two elements, the pairwise surprise s will be calculated for all possible pairwise combinations, and the highest of those values becomes the overall surprise score, S. This is shown in Equation 2, where E is the set of all possible pairwise combinations belonging to the item. We adopt the highest surprise on the recommendation of Maher and Grace [7], based on the idea that the peak element-level surprise dominates the item-level surprise.

$$S = max_E s(e_i, e_j) \tag{2}$$

The second step of the computational surprise measure is to calculate the personalization factor. Guided again by the study of Berlyne [4] where the stimulus intensity is believed to be influenced by how often the stimulus has been experienced by a user. The idea is that the more frequent the user has accessed the item or similar items, the less surprising the item will be. To mimic the impact of past access frequency on the current feeling, we used an exponential decay function $e^{-\lambda t}$, commonly used to describe a natural decreasing process at a rate proportional to its current value and with an exponential forgetting rate [15]. Therefore, the personalization factor is represented as in Equation 3:

$$P^t_{u,i} = e^{-\lambda F^t_{u,i}} \tag{3}$$

where λ is the forgetting rate and $F^i_{u,t}$ is the frequency that the user u has experienced the items related to the item i before time t. Note that $F^i_{u,t}$ is a variable that is user-dependent, item-dependent, and also time-dependent. Therefore $SI^t_{u,i}$, the stimulus intensity of the item i for user u at the moment t, is the multiplication of the personalization factor and the objective surprise of the item i, represented as Equation 4:

$$SI^t_{u,i} = P^t_{u,i} S_i \tag{4}$$

Although a simplified personalization model that may not capture all the factors impacting the personal feeling of surprise, this approach reasonably makes use of a user's past access frequency to approximate a person's familiarity level with an area, the most important element in forming an expectation [7]. Surprise just reflects how strongly an encounter violates such expectation.

3.2.3 Approximating the Wundt curve: fitting a curve for PCC. Since we view a stimulus selection process as drawing samples (stimuli) from a person's PCC, it is natural to expect that PCC follows the probability density function (PDF) of the random variable SI. Specifically in this study, the empirical (observational) PDF of SI is the distribution of a series of $SI^t_{u,i}$ in a user's past access history, as shown in the histogram for the hypothetical user in Figure 4.

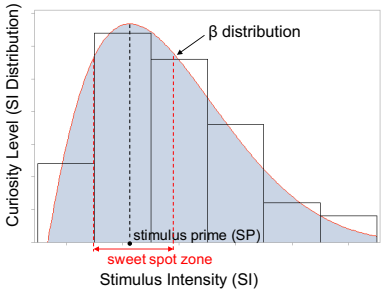


Figure 4: Illustration of SI distribution

In order to get a continuous PDF from the observational PDF histogram, we used the β distribution to fit a curve for the empirical PDF. β distribution has been applied to modeling random variables of human behavior limited to intervals of finite length in a wide variety of disciplines. It is a family of curves controlled by the parameters α and β to approximate any probability distribution. The fitted curve, as shown in the curve in Figure 4, serves as PCC, and also the approximation of the Wundt curve. Generally, the PCC generated using the β distribution has three characteristics: first, distribution generally follows the "inverted-U" shape, suggesting that probability density captures the degree of pleasantness implied by the Wundt curve; second, from the fitted distribution curve, we are able to calculate the stimulus prime (SP) and the stimulus sweet spot zone where curiosity are highly likely to be stimulated. Both the SP and the sweet spot zone are illustrated in Figure 4, and are different for different individuals; and third, the fitted distribution quantifies the Wundt curve using a probabilistic view, which reflects the natural process that humans tend to select the pleasant stimuli more frequently. A person may also respond to a less pleasant stimulus, but the chance is smaller. Overall speaking, a curious user's response zone shifts rightward compared to that of a conservative user.

3.3 Recommendation Generator

We model the recommendation problem as a top-K item ranking problem which selects the top-K items to recommend considering both user preference and curiosity inspiration potential. Specifically, the Recommendation Generator obtains top N items from

the Preference Model as a candidate pool for future recommendation. It then re-ranks the N items according to the proximity, $-dist(SI^t_{u,i}, SP_u)$, between the item's amount of surprise ($SI^t_{u,i}$) to that user's surprise prime SP_u. The re-ranking favors smaller horizontal distance between $SI^t_{u,i}$ and SP_u with the hypothesis that a stimulus closer to the surprise prime (SP) point will have a higher likelihood of stimulating curiosity. This way, the Recommender Generator considers both recommendation accuracy (represented by the Preference Model) and the potential to arouse the user's curiosity (represented by the Curiosity Model).

4 IMPLEMENTATION OF THE RECOMMENDER FRAMEWORK

In this section, we first described the dataset and then some implementation details for both the Models of our proposed recommender framework.

4.1 Dataset

We used book recommendation as our dataset to implement our recommender framework. The dataset is a subset of the Amazon books dataset [19]. The original dataset contains 8,026,324 users, 2,370,585 books, 22,507,155 user-book ratings and the rating timestamps. To supplement the original dataset with the book topic information, we utilized Amazon Product Advertising API to crawl the main topics for each book from the Amazon website. The dataset was pre-processed to exclude books that did not have the topic information available. In addition, in order to avoid the data sparsity problem, we have removed users with fewer than 10 ratings. The final dataset used in this study is summarized in Table 1.

Table 1: The Amazon book dataset used in this study

No. of users	127,627
No. of books	494,108
No. of ratings	3,668,757
Average rating history span	4.7 years

The dataset was split into a training set (80%) M and a test set T (20%). The training set is used to train the CF recommendation algorithms to predict the ratings of the items in the test dataset, as well as to plot the curiosity distribution, fit the PCC for future look-up for item's curiosity level in the test set.

4.2 Preference Calculation: User Rating Training

As mentioned in the Preference Model, we used the "off-the-shelf" collaborative filtering techniques to identify books that are preferred by the user. Specifically, we used three state-of-the-art recommender algorithms: Bayesian Personalized Ranking using Matrix Factorization (BPR-MF) [27], an algorithm that is formulated to maximize the likelihood that the user prefers one item to another. Weighted Approximate-Rank Pairwise Loss using Matrix Factorization (WARP-MF) [35], which maximizes the rank of positive examples by sampling negative examples until a rank violation occurs. And finally, a Variational Autoencoder with multinomial likelihood (Multi-VAE) [16], a deep learning model that extends variational autoencoders.

Each base recommendation algorithm (WARP-MF, BPR-MF, and Multi-VAE) was implemented to identify a set of N candidate recommendations. N has been set to be 100 in this study, a reasonably large pool of candidate items to search for curiosity-inspiring items without sacrificing recommendation accuracy too much.

4.3 Surprise and Curiosity Calculation

For calculating those probabilities for objective surprise in Equation 1, we went beyond the current book dataset, the size of which is limited for deriving accurate estimate of the society's collective expectation. We used a knowledge base - the English Wikipedia corpus with approximately 5 million articles written in English, to calculate the individual occurrence probability $p(e_i)$, $p(e_j)$, and the joint occurrence probability $p(e_i, e_j)$. We used the *search* API introduced in Wikipedia API MediaWiki [1] to obtain the number of articles mentioning e_i, and e_j respectively as well as both e_i and e_j, calculating against the total number of articles in the corpus, in order to estimate those probabilities.

We calculated personalized surprise for each book for each user in the training set. Since books are our items, e_i in Equation 1 is a main topic in a book. In order to measure the personalization factor $P^t_{u,i}$ in Equation 3 for each book and each user at each time point t, we need to calculate $F^t_{u,i}$, the frequency that the user has accessed the books related to the book i before the moment t. The related books in this study were defined as the books that shared a topic with the book i and the shared topic must be one of the two topics that featured the objective surprise level of the book i. Therefore, $F^t_{u,i}$ was calculated this way:

$$F^t_{u,i} = \frac{F^t_{u,Topic1} + F^t_{u,Topic2}}{2} \tag{5}$$

where $Topic1$ and $Topic2$ are the topic pair that features the objective surprise level of the book i as in Equation 2. $F^t_{u,Topic1}$ and $F^t_{u,Topic2}$is the number of times that the user u has accessed $Topic1$ and $Topic2$ respectively before time t. Time t is defined as the access moment of the book i, which means for each accessed book i, we have only considered the access history before this book through the timestamps information of the dataset.

All the calculations were conducted using Python's *math* and *pandas* packages. As mentioned, the distribution of SIs served as the empirical (observational) curiosity distribution. To further turn this empirical distribution to a continuous PDF distribution, we fit the distribution using Python's *stats.beta* library in the *SciPy* package. The library took observational frequency distribution as the input, and output the beta distribution parameters α, β, and the curve's lower and upper limits. These values were used later to plot PCC for each user using Python's *matplotlib* plotting package. The SP points were also calculated through the parameters α and β.

As the result of surprise calculation, the distribution of the objective surprise S as in Equation 2 for all the books in the training dataset is presented in the left panel in Figure 5. The distribution generally follows a normal distribution with the average amount

[1] https://en.wikipedia.org/w/api.php

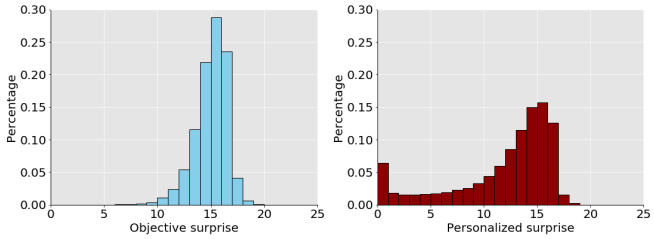


Figure 5: The distribution of the objective and personalized surprise

of objective surprise around 14 or 15. The right panel shows the distribution of the personalized surprise $SI^t_{u,i}$, which has lower bar height and spreads out to the lower end after the personalization due to the exponential decay function proposed in Equation 3. More interestingly, this $SI^t_{u,i}$ distribution could serve as an aggregate empirical curiosity distribution for all the users in the training set. Its lower bar height and spreading toward the lower end suggests that users' tastes were very different, suggesting the value of personalization.

To illustrate what empirical curiosity distribution and fitted PCC look like for different users, Figure 6 on the top of next page shows the histograms of the SIs (after normalization) and the fitted PCC (the blue curves) for five users in our training dataset. The SPs for the five users are about 0.2, 0.4, 0.5, 0.6, and 0.7, respectively, showing that the users tend to respond to different average levels of stimulus. User 1 is relative more conservative compared with User 4 and User 5. Besides, the variance of the users' distributions are different. User 1 and User 2 have relative small variance while User 3 to User 5 have relative large variance. Small variance means that curiosity level is stable, suggesting that users' curiosity tends not to change much with different levels of stimulus, while large variance shows that the user's curiosity may vary greatly. Generally, for each curiosity distribution, there is an optimal SP which has the largest chance to be responded to, and SP is different for different users.

4.4 Combining Preference and Distance to SP

As mentioned in Section 3.3, Recommender Generator obtains a top-N items from a baseline recommender algorithm as a candidate pool, and re-ranks the N items according to the proximity, $-dist(SI^t_{u,i}, SP_u)$, between the item's amount of surprise ($SI^t_{u,i}$) to that user's surprise prime SP_u. The new algorithms are labeled as BPR-MF+Cur, WARP-MF+Cur, and MultiVAE+Cur, meaning a baseline counterpart plus curiosity re-ranking.

5 EVALUATION STUDIES

In this section, we proposed and applied four performance metrics to evaluate our recommender framework. We then presented the evaluation results in terms of the four metrics.

5.1 Evaluation Metrics

We proposed four metrics to evaluate our recommender framework:

5.1.1 Recall. This work adapts the *one plus random* evaluation method [14] with some modification. It randomly splits each user's rated items into a training set M and test set T. An additional probe set P is constructed by selecting up to 10 highly rated items (e.g., those having a four- or five-star rating on a 1 to 5 scale) from the user's test set T. Then, for each user u, predictions will be computed to select the top N ($N = 100$ in this study) unrated items as the candidate pool (introduced in Section 3.3) plus all the p items in P. The set of 100 + p items is ranked according to a baseline algorithm (BPR-MF, WARP-MF, or MultiVAE), or an experimental algorithm (BPR-MF+Cur, WARP-MF+Cur, or MultiVAE+Cur). We will examine whether the experimental algorithm is able to rank the p items higher among the 100 + p items than the baseline algorithm . The underlying belief is since all the items in P represent both high ratings (relevance) and high response likelihood (curiosity), they should be ranked higher compared to the candidate set N.

Specifically, for each user u, whether the items in P is ranked higher is calculated by Recall@K, which is defined as:

$$Recall_u@K = \frac{\text{number of items in P ranked in top K}}{\text{the total number of items in P}} \tag{6}$$

The overall value of $Recall@K$ is the average of $Recall_u@K$ for all the users. $Recall@K$ is an important metric to evaluate whether a recommender algorithm is able to recommend curiosity inspiring items with higher response likelihood, as well as relevance.

5.1.2 Discounted Cumulative Curiousness(DCC). Inspired by the measure of Discounted Cumulative Gain (DCG) [11] that considers both relevance and ranking position to measure the ranking quality in terms of relevance, we propose a measure, called Discounted Cumulative Curiousness (DCC), to measure the ranking quality in terms of curiosity-inspiring potential, represented as:

$$DCC_u@K = \sum_{i=1}^{K} \frac{\text{curiousness score}}{log_2(i+1)} \tag{7}$$

where $DCC_u@K$ is the result list's DCC for user u at each position i from the first position up to the position K. How to measure curiousness is the key problem for applying this measure. Since the ranking is generated (predicted) by ordering the candidate items by horizontal distance between $SI^t_{u,i}$ and SP_u, we will evaluate curiousness in a different way than the prediction - using the "ground truth" data: observational curiousness values offered by the height of a histogram bar in a user's curiosity distribution, representing the actual response likelihood of items with that stimulation level.

The overall value of DCC@K is the average of $DCC_u@K$ for all the users. The higher the value of $DCC@K$, the more potential the recommender has to arouse users' curiosity.

5.1.3 Inter-User Similarity (IUS). Since our recommender framework quantifies a stimulus in a personalized way, we expect that its recommendations are different for different users. To test this expectation, we use inter-user similarity (IUS) proposed in [39]. The $IUS_{i,j}$ between the user i and j is the proportion of overlap between two recommendation lists L_i and L_j for the user i and j.

$$IUS_{i,j} = \frac{|L_i \bigcap L_j|}{|K|} \tag{8}$$

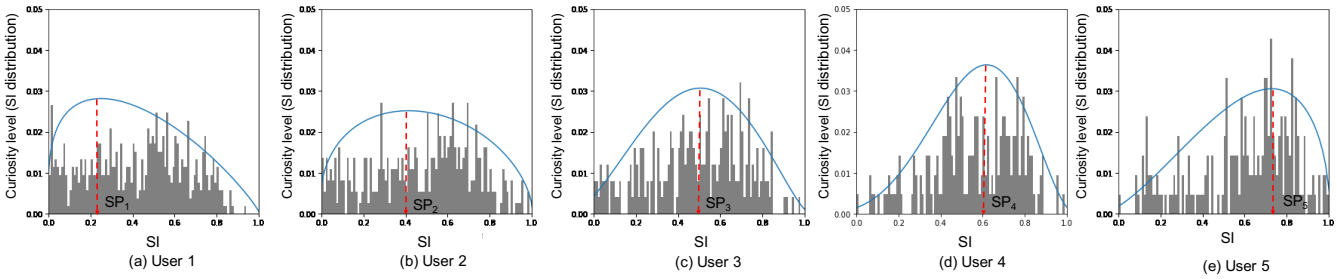


Figure 6: Examples of five users' PCC

The overall value for IUS for all the users is the average of $IUS_{i,j}$ between all pairs of users. A large value of IUS means a high similarity between users and therefore less effect of personalization.

5.1.4 Recommendation Accuracy. Re-ranking the recommendation list returned from the Preference Model means some degree of sacrifice to relevance in order to accommodate the curiosity requirement. We will use Kendall's Tau to measure the agreement between the test items in T's ranking r_1 generated by our algorithm (either baseline or experimental) and their "ground truth" ranking r_2 according to their ratings. This way, we will test how much sacrifice of relevance the system needs to make. The equation of Kendall's Tau [13] for a specific user u is given by Equation:

$$\tau(r_1, r_2)_u = \frac{(C - D)}{\sqrt{(C + D + U_1) * (C + D + U2)}} \tag{9}$$

where C is the number of concordant pairs, D is the number of discordant pairs, U_1 is the number of ties only in r_1, and U_2 is the number of ties only in r_2. If a tie occurs for the same pair in both r_1 and r_2, it is not added to either U_1 or U_2. The overall value of τ for all the users is the average of $\tau(r_1, r_2)_u$ for all the users. A large value of τ means high agreement between the predicted ranking and the "ground truth" ranking, and therefore high recommendation accuracy.

5.2 Evaluation Results

We have conducted two sets of evaluation studies for our recommender framework. The purpose of the first set is to evaluate the effectiveness of incorporating curiosity into the recommender system whereas the second set is to test the effectiveness of personalization in measuring the stimulus intensity.

5.2.1 Evaluating the Curiosity Model. In this set of evaluation, we investigated the effect of different values of K on the four metrics we proposed: Recall, DCC, IUS, and Kendall's Tau. We compared two sets of algorithms: BPR-MF, WARP-MF, and MultiVAE without considering the Curiosity Model, as three baseline algorithms; and BPR-MF+Cur, WARP-MF+Cur, and MultiVAE+Cur as our experimental algorithms.

Figure 7(a) shows the recall levels at different Ks for each recommender algorithms. All the three experimental algorithms outperformed their baseline counterparts at varying K values. This confirms our hypothesis that re-ranking the candidate items by the proximity to the user's appetite will result in a higher chance of hitting an item with high response likelihood as well as a high rating. The performance curves behave as expected since as K increases the chance of hitting is larger. Among the three experimental algorithms, the performance curves of BPR-MF+Cur and WARP-MF+Cur are about the same, both better than MultiVAE+Cur.

As in Figure 7(b), BPR-MF+Cur and WARP-MF+Cur generally have higher DCC values than their baseline counterpart algorithms, especially when K is larger, backing up our hypothesis again that re-ranking by the closeness to a person's comfort zone of response will generate a list of recommendations with higher potential of curiosity. In contrast, the Multi-VAE+Cur algorithm's DCC values are lower at the beginning compared to its baseline algorithm. As K increases to 30 and beyond, the performance is catching up and going above the baseline.

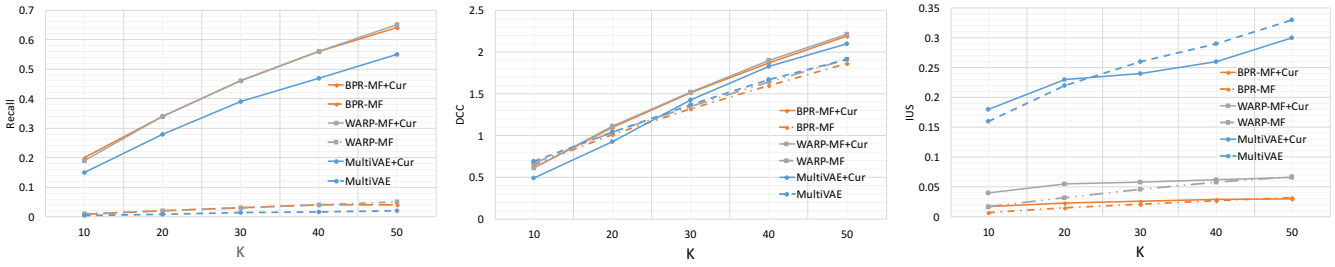
Figure 7(c) presents the IUS curves for the six recommender algorithms. A small value of IUS indicates large effect of personalization factor, which is therefore desired. Unexpectedly, compared to the baselines, BPR-MF+Cur and WARP-MF+Cur have slightly larger IUS levels, probably because in the current Amazon books dataset, there is a small set of popular books which have been highly rated by many users. In order to increase the response likelihood, the experimental algorithms tend to recommend some books from this set, which slightly lower IUS. The result reflects the well-known phenomenon of "the rich get richer" [21] in the dataset we used in this study. Comparing the three experimental algorithms , both BPR-MF+Cur and WARP-MF+Cur have outperformed MultiVAE+Cur in terms of IUS.

Table 2 shows the results of the six algorithms' Kendall's Tau, representing recommendation accuracy based on the user ratings. The lower τs of the experimental algorithms suggests the sacrifice that the experimental algorithms need to make in order to accommodate the curiosity need. This confirms the trade-off relationship between accuracy and curiosity.

Table 2: Kendall's Tau

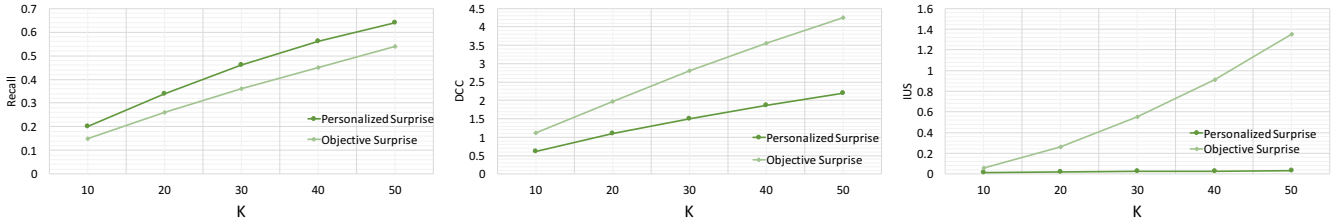
Baseline Algorithm	τ	Experimental Algorithm	τ
BPR-MF	0.11	BPR-MF+Cur	0.008
WARP-MF	0.13	WARP-MF+Cur	0.008
MultiVAE	0.07	MultiVAE+Cur	0.006

5.2.2 Evaluating personalized surprise vs. objective surprise. To follow up with the phenomenon of "the rich get richer" in the Amazon book dataset we used, we want to conduct analysis on the effect of personalization and whether personalization helped



(a) Recall, large better (b) DCC, large better (c) IUS, small better

Figure 7: The first set of evaluation results with varying K



(a) Recall, large better (b) DCC, large better (c) IUS, small better

Figure 8: The second set of evaluation results with varying K

mitigate such a problem. In this study, we have calculated the personalized surprise for each user based on Equation 4 with the expectation that the same item may contain different amounts of surprise to different individuals. This second set of evaluation studies is to evaluate whether using personalized surprise to assess stimulus level brings value in finding curiosity-inspiring books as well as inter user similarity, compared to if we just use the objective surprise as stimulus level: the same item carries the same amount of stimulus for everyone.

We selected one algorithm, BPR-MF+Cur from the last evaluation because of its better performance compared to the other experimental algorithms. We applied this algorithm into two settings: using objective surprise as $SI'^{t}_{u,i}$ or using personalized surprise as $SI^{t}_{u,i}$, and compared its performance in these two settings.

Figure 8 illustrates the evaluation results. In terms of Recall, the personalized approach outperforms the objective approach as shown in Figure 8(a). This confirmed our hypothesis that personalized surprise better reflects the stimulus intensity specific to a user and therefore results in a higher chance of hitting of curiosity-inspiring and relevant items. In Figure 8(b), the personalized approach has lower DCC values, probably because it diversifies the items, deviating from the popular set by adding a personalization factor. Figure 8(c) presents that the personalized approach has constantly achieved a smaller IUS across different values of K, suggesting the effectiveness of personalization. This observation supports our belief that using personalized surprise has alleviated the problem of convergence to some books in the popular set.

6 CONCLUSION AND FUTURE WORK

This paper presents a recommender framework that considers both user preference and curiosity inspiring potential. The Probabilistic Curiosity Curve (PCC) is constructed for each individual user to model their unique appetite for stimulus. To quantify stimulus, we proposed to use surprise as the stimulus factor and developed a measure for evaluating personalized amount of surprise an item contains. Moreover, we have quantified the classic "sweet spot" concept by finding a surprise prime point from the fitted curve and measured the distance between an item's stimulus level to such a prime point. A book recommendation dataset from Amazon has been adopted as the use case to illustrate our idea. In the evaluation studies, we have shown than our algorithms are able to rank higher those items with not only high ratings but also high response likelihood. The personalization factor for assessing the stimulus (surprise) amount helps the recommender achieve smaller inter-user similarity.

For the near future, we plan to apply the framework into other domains, like a recipe recommender system to arouse people's curiosity to different food. We will also extend the framework to generate a sequence of recommendations that are able to transport user from the borders of their current comfort zone (around SP) to "as-yet-too-alien" items that the system might persuade them to appreciate. Finally, since our idea relies on the availability of the user access and rating history with a recommender system, how to apply the framework in a "cold-start" mode without relying much on user history is our future research questions.

REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2015. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2015), 54.

[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[3] Andrew G Barto, Satinder Singh, and Nuttapong Chentanez. 2004. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*. 112–19.

[4] Daniel E Berlyne. 1966. Curiosity and exploration. *Science* 153, 3731 (1966), 25–33.

[5] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* (2009), 31–40.

[6] Kazjon Grace and Mary Lou Maher. 2015. Surprise and reformulation as metacognitive processes in creative design. In *Proceedings of the Third Annual Conference on Advances in Cognitive Systems ACS*. 8.

[7] Kazjon Grace, Mary Lou Maher, David Wilson, and Nadia Najjar. 2017. Personalised specific curiosity for computational design systems. In *Design Computing and Cognition'16*. Springer, 593–610.

[8] Kazjon Grace, Mary Lou Maher, David C Wilson, and Nadia A Najjar. 2016. Combining CBR and deep learning to generate surprising recipe designs. In *International Conference on Case-Based Reasoning*. Springer, 154–169.

[9] Kazjon Grace, Mary Lou Maher2 Maryam Mohseni, and Rafael Pérez y Pérez. 2017. Encouraging p-creative behaviour with computational curiosity. In *Proceedings of the 8th International Conference on Computational Creativity. Association for Computational Creativity*.

[10] Leo Iaquinta, Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, and Piero Molino. 2008. Introducing serendipity in a content-based recommender system. In *Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on*. IEEE, 168–173.

[11] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[12] Marius Kaminskas and Derek Bridge. 2017. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2017), 2.

[13] Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika* 33, 3 (1945), 239–251.

[14] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 426–434.

[15] Xiaoyan Li and W Bruce Croft. 2003. Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*. ACM, 469–475.

[16] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. *arXiv preprint arXiv:1802.05814* (2018).

[17] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.

[18] Luís Macedo and Amílcar Cardoso. 2001. Modeling forms of surprise in an artificial agent. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 23.

[19] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based Recommendations on Styles and Substitutes. *CoRR* abs/1506.04757 (2015). arXiv:1506.04757 http://arxiv.org/abs/1506.04757

[20] Kathryn Merrick and Rob Saunders Mary Lou Maher. 2008. Achieving adaptable behaviour in intelligent rooms using curious supervised learning agents. (2008).

[21] Robert K Merton. 1968. The Matthew effect in science: The reward and communication systems of science are considered. *Science* 159, 3810 (1968), 56–63.

[22] Wulf-Uwe Meyer, Rainer Reisenzein, and Achim Schützwohl. 1997. Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion* 21, 3 (1997), 251–274.

[23] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. 2007. Metrics for evaluating the serendipity of recommendation lists. In *Annual Conference of the Japanese Society for Artificial Intelligence*. Springer, 40–46.

[24] Xi Niu, Fakhri Abbas, Mary Lou Maher, and Kazjon Grace. 2018. Surprise Me If You Can: Serendipity in Health Information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 23.

[25] Pierre-Yves Oudeyer. 2004. Intelligent adaptive curiosity: a source of self-development. (2004).

[26] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

[27] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.

[28] R Saunders and John S Gero. 2001. A curious design agent. In *CAADRIA*, Vol. 1. 345–350.

[29] Jürgen Schmidhuber. 1991. Adaptive confidence and adaptive curiosity. In *Institut fur Informatik, Technische Universitat Munchen, Arcisstr. 21, 800 Munchen 2*. Citeseer.

[30] Jürgen Schmidhuber. 1991. Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*. IEEE, 1458–1463.

[31] Jürgen Schmidhuber. 1999. Artificial curiosity based on discovering novel algorithmic predictability through coevolution. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, Vol. 3. IEEE, 1612–1618.

[32] Jan Storck, Sepp Hochreiter, and Jürgen Schmidhuber. 1995. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, Vol. 2. Citeseer, 159–164.

[33] Masaki Suwa, JS Gero, and Terry Purcell. 1999. Unexpected discoveries and s-inventions of design requirements: A key to creative designs. *Computational Models of Creative Design IV, Key Centre of Design Computing and Cognition, University of Sydney, Sydney, Australia* (1999), 297–320.

[34] Emre Ugur, Mehmet R Dogar, Maya Cakmak, and Erol Sahin. 2007. Curiosity-driven learning of traversability affordance on a mobile robot. In *Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on*. IEEE, 13–18.

[35] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, Vol. 11. 2764–2770.

[36] Qiong Wu, Chunyan Miao, and Zhiqi Shen. 2012. A curious learning companion in virtual learning environment. In *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*. IEEE, 1–8.

[37] Wilhelm Max Wundt. 1874. *Grundzüge de physiologischen Psychologie*. Vol. 1. W. Engelman.

[38] Pengfei Zhao and Dik Lun Lee. 2016. How much novelty is relevant?: It depends on your curiosity. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 315–324.

[39] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.