How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation

Li Chen
Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

Yonghua Yang
Alibaba Group
China
huazai.yyh@alibaba-inc.com

Ningxia Wang
Hong Kong Baptist University
Hong Kong, China
nxwang@comp.hkbu.edu.hk

Keping Yang
Alibaba Group
China
shaoyao@alibaba-inc.com

Quan Yuan
inspirAI Co. Ltd.
China
quanyuan007@gmail.com

ABSTRACT

Recommendation serendipity is being increasingly recognized as being equally important as the other beyond-accuracy objectives (such as novelty and diversity), in eliminating the "filter bubble" phenomenon of the traditional recommender systems. However, little work has empirically verified the effects of serendipity on increasing user satisfaction and behavioral intention. In this paper, we report the results of a large-scale user survey (involving over 3,000 users) conducted in an industrial mobile e-commerce setting. The study has identified the significant causal relationships from novelty, unexpectedness, relevance, and timeliness to serendipity, and from serendipity to user satisfaction and purchase intention. Moreover, our findings reveal that user curiosity plays a moderating role in strengthening the relationships from novelty to serendipity and from serendipity to satisfaction. Our third contribution lies in the comparison of several recommender algorithms, which demonstrates the significant improvements of the serendipity-oriented algorithm over the relevance- and novelty-oriented approaches in terms of user perceptions. We finally discuss the implications of this experiment, which include the feasibility of developing a more precise metric for measuring recommendation serendipity, and the potential benefit of a curiosity-based personalized serendipity strategy for recommender systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; *User models*; *User studies*; • **Information systems** → *Recommender systems*; *Personalization.*

KEYWORDS

Recommender systems; serendipity; curiosity; user satisfaction; large-scale user evaluation

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313469

ACM Reference Format:
Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313469

1 INTRODUCTION

Recommender systems have been popularly used in many Web applications to eliminate users' information overload, with recommended items/information being tailored to the users' individual interests. Traditionally, accuracy is the primary metric for judging a recommender system's effectiveness [15, 28, 40]. However, it may cause the "filter bubble" phenomenon in which users are trapped in a subspace of options that are too similar to their profile, and hence lose the opportunity to explore outside alternatives that may potentially match their preferences [26, 28, 34].

Therefore, some beyond-accuracy objectives, especially *diversity*, *novelty*, and *serendipity*, have been emphasized in the recent literature, because they are targeted to allow users to discover new and different items to broaden their horizons [12, 15, 26]. By definition, *diversity* refers to the difference between the current recommendation and the user's profile (e.g., the user's previously purchased items) or the system's prior recommendations [21][1]. *Novelty* stresses whether the item is unknown to the user. *Serendipity* emphasizes whether the user feels surprised when s/he sees a relevant recommendation, which implies that the item should not only be relevant to the user's interests, but also be unexpected (i.e., not intentionally looked for by the user). Therefore, different from diversity and novelty that may compromise accuracy to a certain degree, serendipity aims to preserve both accuracy and positive emotional response evoked by the recommendation [19, 20, 24].

However, because the "surprising" nature of serendipity is difficult to measure and simulate [6, 15, 29], the actual benefit of serendipity for recommender systems is still unclear. It is not conclusive whether the increased serendipity of recommendation would

[1] Another definition of diversity refers to the intra-list diversity within a set of recommended items [48], which is beyond the scope of this paper.

necessarily lead to users' higher satisfaction with the recommendation [19], and even their higher intention to purchase the recommended item if in an e-commerce domain.

In this study, we conducted a large-scale user survey (involving over 3,000 participants) to measure users' perceptions of recommendations, in relation to serendipity and its components (unexpectedness and relevance), novelty, diversity, user satisfaction, and purchase intention. The results are thus likely to help identify the causal relationships among these factors. For example, *are diversity and novelty significantly related to serendipity?* And *does serendipity significantly improve user satisfaction and purchase intention?* To design the questionnaire, we selected the popularly used questions from related literatures and translated them into Chinese (because our experimental system was plugged into a Chinese industrial mobile e-commerce app called *Mobile Taobao*). It is worth noting that the Chinese translation of the term "serendipity" is "惊喜" that contains two characters: '惊' (surprising) and '喜' (relevant), so it would be intuitive for Chinese users to easily interpret the meaning of "serendipity" when they answer the related questions.

In addition to acquiring users' perceptions of recommendations, we included a curiosity quiz to measure whether they have a strong desire for new knowledge or experiences in general. Indeed, curiosity has been widely regarded as an important antecedent of users' appetite for novelty in the field of psychology [2, 23, 39]. It can greatly affect the level of pleasure a user experiences when s/he explores new and surprising things. For instance, if an item under or over matches her/his curiosity, s/he may feel bored or overwhelmed, rather than pleasant [29, 47]. In this sense, the degree in which a recommendation is perceived as serendipitous might be more or less dependent on the user's curiosity. To verify this point, we integrated curiosity as a moderating factor into our path model to test its effects.

Our third contribution lies in the comparison of four recommender algorithms in relation to user perceptions, which include a non-personalized popularity-based approach, and three variants of the collaborative filtering (CF) based method that are respectively relevance-, novelty-, and serendipity-oriented. To the best of our knowledge, most of the existing algorithm evaluations for e-commerce products have relied on offline metrics [1, 12, 15, 30]. Given there is often a gap between offline measurements and user perceptions (especially regarding the beyond-accuracy objectives) [37], we believe it should be meaningful to conduct an online user study to evaluate whether users would be practically more satisfied with an algorithm's recommendation if it is perceived of higher serendipity. The results could also help improve the existing offline metrics (e.g., more precisely measuring serendipity).

In short, this study makes four major contributions. First, we conducted a large-scale online user survey in an industrial mobile e-commerce setting to collect user feedback on various beyond-accuracy assessments. Second, we validated a hypothesized path model that reveals the causal relationships from novelty, unexpectedness, relevance, and timeliness to serendipity, and from serendipity to user satisfaction and purchase intention. Third, we identified the moderating effects of user curiosity on the relationships from novelty to serendipity and from serendipity to satisfaction. Fourth, we experimentally compared four algorithms for recommending e-commerce products, and found a serendipity-oriented approach is significantly more effective in terms of enhancing user perceptions of the recommendation.

The remainder is organized as follows. We first introduce the related works on serendipity in recommender systems and those considering user curiosity to improve recommendations (Section 2). We then present our user survey, including the experimental procedure, data collection, questionnaire design, and hypothesized model (Section 3). The results are then analyzed through model validation, moderation test, and algorithm comparison (Section 4). We discuss the practical implications of our findings in Section 5, and draw the final conclusions in Section 6.

2 RELATED WORK

2.1 Serendipity in Recommender Systems

The original definition of serendipity is "[...] *making discoveries, by accidents and sagacity, of things which they were not in quest for* [...]" [3, 24, 42]. [12] is one of earlier papers to list serendipity as one of the beyond-accuracy objectives for recommender systems. Since then, serendipity has been used to measure how surprising and relevant a recommendation is [6, 20, 28]. In other words, this factor is comprised of two core components: *surprise* in the sense of unexpected (i.e., different from the user's expectations) and *relevance* (i.e., useful to the user).

However, the assessment of *surprise* is not straightforward because it involves the user's emotional response to a recommendation. In previous work, surprise (or called unexpectedness) was defined as the deviation from a primitive prediction method that produces expected recommendations [10, 30]. For example, [10] proposed a formulation that combines unexpectedness with item relevance, for which the set of unexpected recommendations is acquired by subtracting from those items generated by a primitive prediction model, and the relevance of a recommendation is approximated offline based on users' ratings. Considering this approach is sensitive to the choice of the primitive prediction model, [1] regarded items rated by the user and those similar to the rated ones to be the expected recommendations.

As stated in [15], evaluations of recommendation serendipity that do not involve user feedback can be unreliable, because it is not evident whether the recommended item would be perceived to be serendipitous by the user, even though some objective metric indicates it is. So far, few studies have empirically measured recommendation serendipity and its actual benefits from users' perspective. For instance, [46] conducted a small-scale user study (with 21 participants) to evaluate user perceptions of a serendipity-enhancing system that recommends music artists. The results showed most of the users overall preferred the serendipity-oriented system to the accuracy-oriented baseline especially in terms of serendipity and novelty, but gave lower ratings to serendipity-based recommendations. The authors concluded that the users are willing to compromise accuracy for serendipity, which nevertheless is contrary to the original definition of serendipity as comprising both relevance and unexpectedness. [33] designed a fusion-based recommender interface that allows users to experience extrinsic and intrinsic accidents for discovering serendipitous books. They recruited 9 users to evaluate the interface, which demonstrated its advantage over the standard Amazon interface as for serendipity.

In a more recent work, [6] combined the questionnaire and facial expression detection approaches in a preliminary user experiment (involving 40 subjects) performed on their proposed graph-based random walk algorithm. Two questions were asked to respectively assess a user's perceived relevance of a recommended movie and its unexpectedness. The user's facial expression was detected simultaneously to implicitly infer her/his degree of surprised when seeing the item. The experimental results showed their algorithm was perceived to be more relevant and serendipitous than a random recommendation method. They also found a moderate agreement between the questionnaire answers and facial expression detection results.

Given that there are different interpretations of serendipity in the related literature, [19] conducted a user survey (participated by 475 users) to acquire users' responses to eight different serendipity definitions in a movie recommender. The survey also aimed to verify whether it is meaningful to recommend serendipitous items. Specifically, by running a regression model on the users' answers to various serendipity assessments with the two dependent variables *preference broadening* and *user satisfaction*, the authors found that most definitions of serendipity helped broaden user preferences, but none of them affected user satisfaction.

The novelty of our work lies in revealing the relationships among the three beyond-accuracy factors, *novelty*, *diversity*, and *serendipity*, and their respective impacts on *user satisfaction* and *purchase intention*. From the algorithm's perspective, we compared a serendipity-oriented approach with three related methods, focusing on identifying their differences in terms of user perceptions. In addition, differing from the small-scale user studies in the related work, our study, to the best of our knowledge, is the first one performed on an industrial mobile e-commerce platform for a large-scale user experiment.

2.2 Recommendation based on User Curiosity

Curiosity, as a particular psychological trait that can affect a user's desire for exploration, has been found to likely influence the user's reaction to recommendations [29]. For example, it was shown that more curious people prefer recommendations containing some unexpected items, but less curious individuals prefer recommendations that are mostly similar to what they experienced before [29, 47].

Thus, some studies have attempted to incorporate curiosity into their recommendation methods. For example, [47] developed a curiosity-based music recommender, in which the recommendations can be personalized to the individual user's curiosity level. They mainly implemented a probabilistic curiosity model learnt for each user based on her/his access history. With this model, a curiousness score is computed for each item to indicate how curious the target user will be about it. Those items that are both relevant and suited to the user's curiosity level are then recommended. A major finding of their offline experiment is that the proposed method can improve the recommendation personalization and accuracy at the same time. In [29], a personalized recommendation architecture was presented in a tourism context, which aimed to adapt the degree of surprise of a recommended item to the user's curiosity. For this purpose, they proposed to predict a user's curiosity from her/his data shared on social network websites like Facebook, and then generate a recommendation list with the degree of unexpectedness tailored to the user's curiosity value. [24] also considered users' curiosity, with the objective of enhancing a TV program recommender. They adopted a psychological curiosity theory that defines two appraisals of curious emotion: *novelty check* and *coping potential check*. They then proposed to estimate the first appraisal based on an item's dissimilarity to those in the user profile and the second one based on the diversity of items within the user profile. A user study involving 165 participants showed that using curiosity to guide recommendations can be promising in terms of balancing serendipity and precision.

However, the above-described studies mostly inferred users' curiosity from their behavioral data and then incorporated the inferred curiosity into the process of generating recommendations. Little ground work has been done to verify the exact effect of curiosity on users' perceived recommendation serendipity. Therefore, we particularly investigated this effect in our user survey, which may be constructive for developing more personalized serendipity strategy based on user curiosity.

3 USER SURVEY

3.1 Setup and Measurement

We conducted a user survey on a popular mobile e-commerce platform in China (*Mobile Taobao*) from Dec. 21, 2017 to Jan. 11, 2018. The users were able to access the survey's link after they logged in the system. If a user volunteered to take part in, s/he first received a recommended product (that was generated by one of our tested algorithms, and was from one of various product domains such as "clothes," "toys," "home appliances," "foods," etc.) including its name, image, short description, and price. The user then completed a questionnaire that assessed her/his immediate feedback on this recommendation. S/he was also asked to answer several questions about her/his personal background (e.g., age, gender), and fill out a psychological curiosity quiz. As the incentive, all of the participants were placed in a lottery draw with customized presents as awards given to the winners.

3,039 users joined in our survey. We carefully checked their responses in order to filter out invalid answers. For example, if a user did not answer all of the questions, or gave the same rating to all the questions (some were asked in the reverse way), her/his response was deleted. We also removed redundant responses (by the same user) and only kept her/his first response. As a result, 2,401 users remained (1,651 females). An analysis of the users' historical behaviors over the past three months (from Oct. to Dec., 2017) showed that all of them had clicked at least one item before taking the survey (98.5% had more than 100 clicks).

In the following, we introduce the variables assessed in the user survey.

3.1.1 User Perceptions of Recommendation. Because the survey was completed on the user's mobile device, s/he would have been less patient in responding to a lengthy questionnaire consisting of many questions [41]. Therefore, to reduce the survey duration, we adopted a short version of ResQue (a widely used user-centric evaluation framework for recommender systems [36]), which, as

claimed by the authors, can provide a fast and reliable way to assess user perceptions of recommendations. Specifically, we used the suggested question for each of the four variables (see Table 1): *recommendation relevance* (Q1), *recommendation novelty* (Q2), *satisfaction* (Q8), and *purchase intention* (Q9). For recommendation diversity, because we emphasize the difference of the currently recommended item from those the user previously experienced, we included two specific questions: one is about the difference from the user's previously purchased product types (Q3: *pur_diversity*), and the other is about the difference from the system's prior recommendations (Q4: *rec_diversity*). In addition, we asked two questions particularly related to recommendation serendipity: *unexpectedness* (Q5) and *serendipity* (Q6). As mentioned before, there is a popularly used Chinese word "惊喜" that can well convey the meaning of serendipity to Chinese users, so we asked about the user's perceived serendipity directly, instead of choosing indirect statements designed to address the interpretation difficulty of the English word "serendipity" [19, 24, 37]. We also added a question about the recommendation's context compatibility [36], i.e., whether the item is recommended at the right time (Q7: *timeliness*).

Overall, nine questions were included in our survey to assess a user's perceptions of the recommendation (each responded on a 5-point Likert scale from "strongly disagree" to "strongly agree"). In Section 3.2, we present a hypothesized path model to depict their potential causal relationships.

3.1.2 Curiosity Instrument. According to the psychological theory [2, 23, 39], curiosity is an intrinsic human trait, triggered when there is a gap between the person's current knowledge level and the desired level. Because different people may have different desires for new knowledge, in recommender systems, some users may feel excited when they see an unexpected recommendation, while some may be reluctant to explore unfamiliar items [29, 47]. Thus, in our survey, we asked the participants to respond to a popular curiosity instrument, Curiosity and Exploration Inventory-II (CEI-II) [16] (an improved version of the original Curiosity and Exploration Inventory (CEI) [17]), because it was proven capable of assessing individual differences in the "general tendency to embrace novel or uncertain situations and to seek out new experiences." Specifically, it is a 10-item self-report scale embodying two factors: *Stretching* ("motivation to seek out knowledge and new experiences") and *Embracing* ("willingness to embrace the novel, uncertain, and unpredictable nature of everyday life"). Each item is rated on a 5-point Likert scale from "very slightly or not at all" to "extremely." This instrument has been shown to have acceptable internal reliability and stable validity across time. It is also short enough to be possibly completed within two minutes [5]. Another favorable consideration is that it was validated as having good psychometric properties in a Chinese context [45].

3.2 Hypothesized Path Model

Figure 1 shows our hypothesized path model covering all of the observed variables. We first linked the two major components, *unexpectedness* and *relevance*, to *serendipity*, because it has been claimed that a serendipitous recommendation should be not only unexpected (surprising), but also relevant to the user's preferences

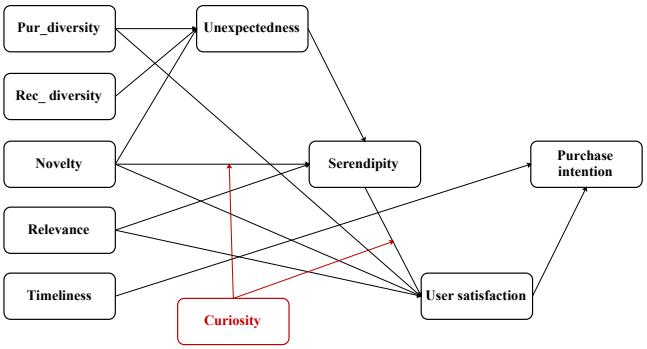

Figure 1: Hypothesized path model.

[15, 19, 24]. We were hence interested in verifying the two components' respective contributions to serendipity. In some related work, *novelty* has also been closely related to serendipity as one key component [12, 19, 28]. For example, [12] stated that a serendipitous item should be both novel (unknown to the user) and surprising. Therefore, we also linked novelty to serendipity, so as to identify to what degree a user may feel a recommendation is serendipitous when s/he perceives it as novel.

In addition, we postulated that *novelty* and *diversity* lead to *unexpectedness* [15]. That is, a novel item, or an item different from the user's previously purchased items (*pur_diversity*) or from the system's prior recommendations (*rec_diversity*), would make the user feel that it is unexpected. In this sense, unexpectedness may act as a mediator in passing some of the effect from novelty to serendipity, and also build a bridge between diversity and serendipity to reflect the diversity's indirect effect.

More importantly, we aimed to identify the factors that empirically influence *user satisfaction* with the recommendation. Although it has been recognized that accuracy alone does not always result in user satisfaction [12], few studies have verified whether and how the other beyond-accuracy factors, especially *serendipity*, have a positive effect in this regard. For instance, in a recent work [19], the effect of serendipity on satisfaction was not found to be significant. As for *diversity* and *novelty*, some user studies showed that diversity positively affects user satisfaction, whereas novelty has a negative effect [7]. However, in another work [36], novelty was demonstrated to have positively higher impact than diversity on satisfaction. We therefore expected that by collecting a larger amount of user feedback in a realistic recommending environment (e.g., mobile e-commerce in our experiment), we could empirically identify these factors' exact roles.

Finally, we postulated that the context compatibility and user satisfaction lead to higher *purchase intention* [36]. In particular, context compatibility refers to the timely compatibility (i.e., *timeliness*) of the current recommendation in our experiment. For example, although a user generally likes digital cameras, s/he may not want to buy a new one at present because s/he has already owned one. It was hence assumed that if one user is satisfied with a recommendation that also meets her/his current requirement, s/he would be more likely to purchase it.

Table 1: Assessment of user perceptions of the recommendation and descriptive statistics (of 2,348 responses after filtering out invalid records and outliers)

Subjective variable and assessment question	Mean	Std.	Median	Kolmogorov-Smirnov test	Skewness	Kurtosis
Relevance: *Q1. "The item recommended to me matches my interests."*	3.32	1.410	4.00	0.255^{***}	-0.419	-1.192
Novelty: *Q2. "The item recommended to me is novel."*	3.06	1.424	3.00	0.235^{***}	-0.146	-1.391
Pur_diversity: *Q3. "The item recommended to me is different from the types of products I bought before."*	3.39	1.215	4.00	0.221^{***}	-0.400	-0.813
Rec_diversity: *Q4. "The item recommended to me is similar to the system's prior recommendations."*(reversed)	2.93	1.302	3.00	0.206^{***}	0.214	-1.109
Unexpectedness: *Q5. "The item recommended to me is unexpected."*	3.16	1.437	3.00	0.207^{***}	-0.199	-1.337
Serendipity: *Q6. "The item recommended to me is a pleasant surprise."*	2.73	1.456	2.50	0.193^{***}	0.195	-1.400
Timeliness: *Q7. "The item recommended to me is very timely."*	3.00	1.484	3.00	0.207^{***}	-0.074	-1.450
User satisfaction: *Q8. "I am satisfied with this recommendation."*	3.21	1.140	3.00	0.210^{***}	-0.286	-0.466
Purchase intention: *Q9. "I would buy the item recommended, given the opportunity."*	2.83	1.456	3.00	0.191^{***}	0.003	-1.418
Curiosity: Curiosity and Exploration Inventory-II (CEI-II) with a 10-item self-report scale [16]	3.13	0.831	3.10	0.043^{***}	0.088	-0.402

Note: $^{***}p < 0.001$; all of the questions were accompanied by Chinese translations.

Furthermore, considering that *curiosity* is a personal trait, we incorporated it as a moderating factor into our path model. By definition, "moderation" means that this factor will moderate the effect of one variable on another variable [22]. For instance, highly curious users may be more inclined to perceive a novel item as serendipitous, because they tend to embrace uncertain situations (i.e., curiosity moderates the causal relationship *from novelty to serendipity*). Moreover, they may be more satisfied with a serendipitous item, compared to less curious users (i.e., curiosity moderates the relationship *from serendipity to satisfaction*).

3.3 Algorithms

Another objective of our user survey was to compare several recommender algorithms in terms of the users' perceived serendipity and the other subjective variables, which may help explain whether and why some particular algorithm would be preferred over others. Specifically, we implemented a popularity based approach as the baseline, and three variants of the collaborative filtering (CF) based method that are respectively tailored to highlight relevance, novelty, and serendipity of the recommendation.

(1) The baseline recommends a product with the most clicks to users (referred to as *HOT* henceforth), so it is purely popularity based without considering the target users's preferences.

(2) For the second method, we intended to strengthen a recommendation's relevance to the user's preferences. For this purpose, we revised the standard user-based CF by calculating a domain-specific similarity score $sim_d(u, v)$ between two users u and v (see Equation (1)), so that if they have often clicked the same item within the same domain (e.g., "clothes," "toys," "home appliances") and the clicking time is close, their overall similarity will be enhanced. This approach is shortened as *Rel-CF*.

$$sim_d(u,v) = \frac{\sum_{i \in I_{du} \cap I_{dv}} W_i^2 / (1 + \alpha \cdot |t_{ui} - t_{vi}|)}{\sqrt{\sum_{i \in I_{du}} W_i^2} \sqrt{\sum_{j \in I_{dv}} W_j^2}} \quad (1)$$

where I_{du} is the item set that the user u has clicked in domain d, and t_{ui} is the time stamp when u clicked i (so $|t_{ui} - t_{vi}|$ gives the time interval between u's and v's clicks on the same item i). W_i is the weight of an item i, computed via $W_i = \frac{1}{\log_2(3+q_i)}$, so it will be higher if the item has been less frequently clicked (q_i is the clicking frequency of i). The score for an unknown item for user u is predicted as:

$$score(u,i) = \sum_{d \in D} \sum_{\substack{v \in (U_i \\ \cap S_d(u))}} sim_d(u,v) * r(v,i) \quad (2)$$

where D is the set of all domains (36 domains in our experiment), $S_d(u)$ contains the top n users who are most similar to user u w.r.t. domain d according to $sim_d(u, v)$, U_i is the set of users who have clicked item i, and $r(v, i)$ is the rating score that v gave to i (that is the implicit clicking behavior in our case, i.e., 1 "clicked" and 0 "not clicked").

(3) The third method aims to be novelty oriented, by recommending an item from a category/domain outside of the user's profile (so unlikely known by the user). We adopted the item-based CF, but only calculated the similarity between two items if they belong to two different categories of a domain or two different domains (one domain can consist of multiple categories, e.g., the domain "clothes" includes the categories "T-shirt," "skirt," "sweater," etc.). This approach is shortened as *Nov-CF*.

$$sim(i,j) = \frac{\sum_{(c_i \neq c_j) \& (u \in U_i \cap U_j)} W_u^2 / (1 + \alpha \cdot |t_{ui} - t_{uj}|)}{\sqrt{\sum_{u \in U_i} W_u^2} \sqrt{\sum_{v \in U_j} W_v^2}} \quad (3)$$

where c_i is the category of i, and W_u is the weight of user u, computed via $W_u = \frac{1}{\log_2(3+q_u)}$, so if a user has clicked fewer items (i.e., q_u), her/his weight will be higher. The prediction score for an unknown item for user u is hence computed as:

$$score(u,i) = \sum_{j \in (S(i) \cap I_u)} sim(i,j) * r(u,j) \quad (4)$$

where $S(i)$ is the set of the top n items that are most similar to item i according to $sim(i, j)$, and I_u is the set of all items that have been clicked by user u.

(4) The fourth is more serendipity oriented (shortened as *Ser-CF*), because its recommendation's relevance to the target users's preferences is more strengthened on top of *Nov-CF*. Specifically, it considers the time sequence of two items clicked by a user as well as the other items clicked between them in the sequence:

$$sim_{seq}(i,j) = \sum_{(u \in U_i \cap U_j) \& (p_{uj} < p_{ui})} W_{uij} \cdot s_u(p_{ui}, p_{uj}) \quad (5)$$

where $s_u()$ gives the sum of similarities between any two adjacent items positioned from p_{uj} to p_{ui} (p_{uj} is the position of item j within the time sequence of items user u has clicked, and $p_{uj} < p_{ui}$ indicating that j was clicked before i). $W_{uij} = \frac{\sum 1/|I_u \cap I_v|}{|U_i \cap U_j| - 1}$ denoting the distance of u from other users who also clicked both items i and j (i.e., $v \in (U_i \cap U_j) \backslash \{u\}$), so that if the whole set of items s/he clicked (I_u) is largely different from other users', her/his weight will be higher.

$$s_u(p_{ui}, p_{uj}) = \sum_{\substack{(p_{uj} \le p_{um}, p_{un} \le p_{ui}) \\ \&(p_{un} = p_{um}+1)}} sim(m, n) \tag{6}$$

where $sim(m, n)$ is computed via Equation (3). The prediction score is thus computed as:

$$score(u, i) = \sum_{j \in (S(i) \cap I_u)} sim_{seq}(i, j) * r(u, j) \tag{7}$$

Each participant was randomly assigned one of the four algorithms' recommendations, which basically adheres to the between-subjects experimental design for simulating real-world experiences [11]. The numbers of users assigned to the four algorithms are respectively: 570 for HOT, 596 for Rel-CF, 589 for Nov-CF, and 593 for Ser-CF (out of 2,348 users after filtering out invalid records and outliers).

4 RESULTS ANALYSIS

We chose IBM SPSS Amos to run the path analysis [4], because it allows us to test the causal relations among the nine observed variables (see Table 1). We first checked whether our data meet its assumption of multivariate normality. The results show the multivariate kurtosis value is 7.749 (Mardia's coefficient), with a critical ratio (c.r.) of 5.798, which is greater than the desired level 1.96 [9]. Therefore, we deleted some obvious outliers with very large Malanobis d-squared distances (i.e., improbably far from the centroid). In consequence, there were 2,348 users' responses left (1,621 females), which were found to be normal by a Mardia's coefficient of 1.440 (c.r. = 1.904). We then conducted normality testing for each variable. The p values of the Kolmogorov-Smirnov test (suitable for sample size greater than 2,000 [25]) are all less than 0.001 (see Table 1), showing that the null hypothesis of normal distribution is rejected for each variable (including curiosity). Therefore, we chose non-parametric tests that do not assume normality for the following correlation and comparison analyses. But we still used Amos to test our path model, because the multivariate distribution of the observed variables is normal, and each variable's univariate skewnesses and kurtosis are in the acceptable ranges of -1.0 to 1.0 and -2.0 to 0 respectively (so not likely to inflate the Chi-squared statistics of path analysis according to [31]).

In addition, we checked whether there are high inter-correlations among those variables, because such multicollinearity phenomenon may cause disturbance and unreliability in the regression analysis [8]. For this purpose, we calculated the Spearman's correlation coefficients[2], which are all below 0.80 suggesting that there is no serious problem with multicollinearity (we later performed collinearity diagnostics for each outcome variable for further confirmation).

[2]Spearman's correlation is a non-parametric test used to measure the degree of association between two variables.

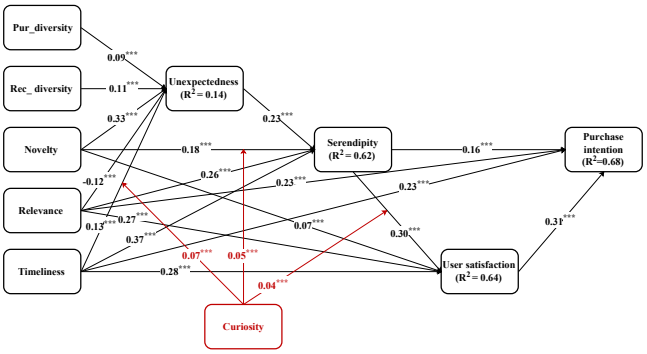

Figure 2: Path model fit. The value associated with each path is the standardized regression coefficient via bootstrap estimates, and the interaction term involving curiosity is standardized for moderation analysis (*$p < 0.001$, **$p < 0.01$, and *$p < 0.05$).**

4.1 Path Model Validation

We then tested our hypothesized path model (see Figure 1). To obtain unbiased estimates given univariate non-normalities, we performed bootstrapping (on 500 bootstrap samples at 95% biased-corrected confidence intervals) [4]. Figure 2 shows the model after the suggested modifications[3], which obtains superior goodness of fit: $\chi^2 = 8.608$ ($df = 7$, $p = 0.282$), $CFI = 1.000$, $AGFI = 0.995$, $RMSEA = 0.010$, all surpassing the suggested values of these model fit indices [13][4]. The measurement portion of the model is also good, with all of the R^2 estimates being larger than 0.10, and thus appropriate and informative to examine the significances of the paths associated with the outcome variables [36]. More notably, except that the R^2 estimate for *unexpectedness* is 0.14, the other R^2 values all exceed 0.60, indicating that the associated predictors can account for a large proportion of the variance in the corresponding outcome variable.

In more detail, we found there are significant causal relationships from *novelty*, *pur_diversity*, *rec_diversity*, *relevance*, and *timeliness* to *unexpectedness*, which implies that if a user perceives the recommended item as novel ($\beta = 0.327$ [5], $p = 0.002$), different from her/his previously purchased product types ($\beta = 0.095$, $p = 0.006$), different from the system's prior recommendations ($\beta = 0.108$, $p = 0.008$), not relevant to her/his preferences ($\beta = -0.119$, $p = 0.003$), or timely ($\beta = 0.134$, $p = 0.004$), s/he will be likely to perceive the item as unexpected. As a whole, all of these predictors account for 14% of the variance in unexpectedness.

Regarding *serendipity*, there are four significant predictors (in the descending order of influence): *timeliness* ($\beta = 0.369$, $p = 0.009$), *relevance* ($\beta = 0.264$, $p = 0.004$), *unexpectedness* ($\beta = 0.234$, $p = 0.004$), and *novelty* ($\beta = 0.181$, $p = 0.001$), which altogether

[3]We started with a full model and then went through model trimming by deleting one non-significant path at a time.

[4]χ^2, i.e., the chi-square value, should better be with probability $p \ge 0.05$ to indicate absolute fit; the cutoff values of the other fit indices CFI (Comparative Fit Index), AGFI (Adjusted Goodness of Fit), and RMSEA (Root Mean Square Error of Approximation) are respectively ≥ 0.90, ≥ 0.90, and ≤ 0.05 [13].

[5]This is the bootstrap estimate of the standardized regression weight (the same to the other β values).

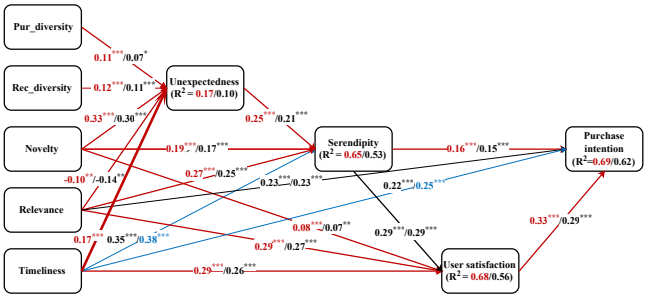

Figure 3: Multi-group path model analysis between *high* and *low* curiosity groups. Red highlights higher β coefficients and R^2 estimates in the *high* curiosity group, and blue highlights higher values in the *low* curiosity group.

account for 62% of the variance in serendipity, inferring that if a recommendation is timely, relevant, unexpected, and novel, the user will very likely perceive it as serendipitous. Because there are also direct paths from novelty, relevance, and timeliness to unexpectedness, it suggests that some of their effects on serendipity are partially mediated by unexpectedness.

Serendipity, allied with *timeliness*, *relevance*, and *novelty*, can be further used to predict *user satisfaction*, with standardized regression weights of $\beta = 0.305$ ($p = 0.006$), $\beta = 0.275$ ($p = 0.003$), $\beta = 0.273$ ($p = 0.003$), and $\beta = 0.075$ ($p = 0.003$) respectively. The R^2 value associated with user satisfaction is 0.65, indicating that 65% of its variance can be accounted for by the set of these four variables, among which the effects of serendipity, timeliness, and relevance are much higher than that of novelty. Here we also see the mediated paths passing through serendipity, suggesting that serendipity has a certain mediating effect on the causal relationships from timeliness, relevance, and novelty to user satisfaction.

The last outcome variable is *purchase intention*, which is significantly affected by *user satisfaction* ($\beta = 0.314$, $p = 0.004$), *timeliness* ($\beta = 0.232$, $p = 0.002$), *relevance* ($\beta = 0.228$, $p = 0.002$), and *serendipity* ($\beta = 0.161$, $p = 0.005$), with 68% of the variance explained by them. In combination with the results above, this suggests that timeliness and relevance can directly affect users' purchase intention, while serendipity and satisfaction mainly play mediating roles.

Collinearity diagnostics were then conducted for all of the four outcome variables (i.e., *unexpectedness*, *Serendipity*, *user satisfaction*, and *purchase intention*). Because the values of Tolerance and VIF ($Tolerance > 0.34$ and $VIF < 3$) all meet the desired standards[6], we can confirm there is no multicollinearity issue in our path analysis.

4.2 Moderating Effect of User Curiosity

Next, we conducted an in-depth investigation on the role of *curiosity* in affecting user perceptions of the recommendation, for which we first checked the internal reliability of our used curiosity instrument (10-item scale). Cronbach's alpha value is 0.876 and the item-total correlations are all between 0.381 and 0.701, which suggest that the 10 items all reliably measure the same variable [35]. We also performed principal components analysis (PCA), by which two factors were extracted: questions 1-3, 5 of higher loadings on one factor, and questions 4, 6-10 of higher loadings on the other factor (all exceeding the acceptable level of 0.5 [35]). However, because this 2-factor structure does not fully replicate the original Stretching/Embracing structure[7] [16], we used the 1-factor structure in our analysis (i.e., taking the 10 items together to represent *curiosity*).

We then performed two types of analysis. First, multi-group analysis enabled us to compare observations across two population groups (i.e., low curiosity vs. high curiosity). To divide all users into the two groups, we used the median split method [14], resulting in 1,132 users in the low curiosity group (< median 3.1 that is the cutoff point) and 1,216 in the high curiosity group (>= 3.1). Then, by imposing structural weights constraint on the path model, we found the Chi-square difference statistic is significant ($\chi^2 = 45.542$ with 17 *df*, $p < 0.001$), which implies significant inequalities of path coefficients across the two groups[8] (see Figure 3).

In particular, we detected three major differences between the two groups: 1. For the low curiosity group, there is no significant path relationship from timeliness to unexpectedness; 2. the regression weights of paths originating from *novelty*, *unexpectedness*, and *serendipity* are mostly higher in the high curiosity group; 3. the R^2 estimates for the outcome variables are all higher in the high curiosity group (for example, 0.65 for *serendipity* in the high curiosity group, vs. 0.53 in the low curiosity group), indicating that the associated predictor variables can be more accurate to estimate the corresponding outcome variable.

Second, we performed the moderation analysis, aiming to identify whether curiosity has actual moderating effects on the causal relationships in our path model[9]. The results show that it primarily acts as a moderator for three relationships (see Figure 2): *from timeliness to unexpectedness* (the standardized regression weight (s.r.w.) of the interaction term *curiosity×timeliness* is 0.073, $p = 0.006$), *from novelty to serendipity* (the s.r.w. of *curiosity×novelty* is 0.048, $p = 0.004$), and *from serendipity to satisfaction* (the s.r.w. of *curiosity×serendipity* is 0.041, $p = 0.005$). In other words, it suggests that more curious users are more likely to perceive a timely recommendation as unexpected, a novel recommendation as serendipitous, and be satisfied with the serendipitous item, while these relationships are not so strong for less curious people. The results hence validate the observations of the above multi-group analysis, implying the prominent role of curiosity in influencing users' perceived novelty, serendipity, and satisfaction with the recommendation.

[6] To be a problem, Tolerance has to approach 0 and VIF has to approach 10 [32].

[7] Similar phenomenon was found in [45] that was also tested in a Chinese context, so more studies might be needed to verify this 2-factor structure under different cultural backgrounds.

[8] We also conducted multi-group analyses with respect to age (young vs. old) and gender (female vs. male), but no significance was found ($\chi^2 = 17.059$ with 17 *df*, $p = 0.450$ w.r.t. age; and $\chi^2 = 16.943$ with 17 *df*, $p = 0.458$ w.r.t. gender).

[9] The path model still fits well after being incorporated with the curiosity and interaction terms ($\chi^2 = 25.466$ ($df = 18$, $p = 0.113$), $CFI = 0.999$, $AGFI = 0.992$, $RMSEA = 0.013$).

Table 2: Results of algorithm comparison by Kruskal-Wallis 1-way ANOVA test (the superscript indicates that the corresponding algorithm significantly outperforms the numbered one, with $p < 0.05$ adjusted by the Bonferroni correction)

Subjective variable	Mean rank				Sig.
	HOT[1]	Rel-CF[2]	Nov-CF[3]	Ser-CF[4]	
Relevance	1047.61	1168.76[1]	1130.39	**1346.05**[1,2,3]	0.000
Novelty	1079.22	1148.04	1182.39[1]	**1284.84**[1,2,3]	0.000
Pur_diversity	1201.83	1164.10	1182.21	1151.03	0.578
Rec_diversity	**1391.74**[2,3,4]	1203.96[3,4]	1092.20	1017.82	0.000
Unexpectedness	**1300.69**[2,3,4]	1178.18	1101.88	1121.64	0.000
Serendipity	1062.11	1142.97	1182.73[1]	**1306.05**[1,2,3]	0.000
Timeliness	1002.74	1110.57[1]	1182.43[1]	**1395.98**[1,2,3]	0.000
User satisfaction	1013.12	1154.61[1]	1184.08[1]	**1340.10**[1,2,3]	0.000
Purchase intention	1033.38	1142.73[1]	1162.42[1]	**1354.07**[1,2,3]	0.000

4.3 Algorithm Comparison

The third objective of this experiment was to compare the four recommender algorithms (see Section 3.3) in terms of user perceptions. We used the Kruskal-Wallis 1-way ANOVA test[10] to handle the non-normally distributed dependent variables. As above mentioned, there were 570, 596, 589, and 593 users respectively assigned to the HOT, Rel-CF, Nov-CF, and Ser-CF algorithms. Table 2 reports the results. It can be seen that there is very strong evidence of differences among the four algorithms, in terms of *novelty*, *relevance*, *rec_diversity*, *timeliness*, *unexpectedness*, *serendipity*, *satisfaction*, and *purchase intention*. Moreover, the mean rank of Ser-CF seems to perform the best compared to the other three, except unexpectedness and rec_diversity. It is interesting to see that the HOT algorithm was perceived the most unexpected and the most different from the system's prior recommendations, but it was the worst as for most of the other variables including user satisfaction and purchase intention. Because this method only returns the most popular item without taking into account the user's preferences, it implies that unexpectedness can reflect unpleasant surprise in this condition.

We further ran the Mann Whitney Wilcoxon test[11] for the pairwise comparisons. The results show strong evidence ($p < 0.05$ adjusted by the Bonferroni correction for multiple tests) of the differences between Ser-CF and the other three algorithms. Specifically, Ser-CF is significantly better than HOT, Rel-CF, and Nov-CF, in respect of *novelty*, *relevance*, *timeliness*, *serendipity*, *satisfaction*, and *purchase intention*. The comparison between Nov-CF and HOT shows that the former obtains significantly higher mean ranks regarding *novelty*, timeliness, serendipity, satisfaction, and purchase intention; while the comparison between Rel-CF and HOT reveals the former is particularly better at *relevance* and timeliness, which also results in increased satisfaction and purchase intention. However, there are no significant differences between Rel-CF and Nov-CF as for most of the variables. One more interesting observation is that unexpectedness and rec_diversity are significantly higher for HOT relative to the other three algorithms.

[10] A nonparametric test used to compare the mean ranks of more than two independent groups, when the dependent variable is ordinal or not normally distributed [27].
[11] It is also a nonparametric test for comparing two independent groups.

All of the results suggest that the Ser-CF algorithm is capable of accommodating both *novelty* and *relevance* to more effectively increase users' perceived recommendation *serendipity* and *satisfaction*, compared to Rel-CF that is more relevance focused, and Nov-CF that is more novelty focused. In addition, we found that people with a high curiosity level were more satisfied with Ser-CF, because its differences from Nov-CF and Rel-CF are significant in the high curiosity group regarding *user satisfaction* (Ser-CF vs. Nov-CF: 686.59 mean rank vs. 611.69, $p = 0.032$; Ser-CF vs. Rel-CF: 686.59 vs. 595.45, $p = 0.004$), but not significant in the low curiosity group (Ser-CF vs. Nov-CF: 630.68 vs. 580.93, $p = 0.365$; Ser-CF vs. Rel-CF: 630.68 vs. 565.39, $p = 0.082$). These results imply the effect of curiosity on users' evaluation of a particular recommender algorithm.

We also recorded all participants' post-survey behavior to see whether they revisited the recommended item after the survey[12]. A Chi-square test reveals a significant association between the algorithms and users' revisiting behavior ($\chi^2 = 188.027, p < 0.001$). The values of Phi and Cramer's V again indicate that the strength of this association is strong (both are 0.283, $p < 0.001$). More specifically, more Ser-CF users (over 17%) clicked the recommended item after the survey, whereas fewer users who used algorithms Nov-CF (6.1%), Rel-CF (1.0%), and HOT (0%) revisited their recommendations. However, because the revisiting behavior did not occur immediately after our experiment (e.g., one user revisited the item 6 hours after taking the survey), we cannot claim it was necessarily caused by the recommendation. Therefore, in the future, we will conduct more studies to measure the correlation between users' perceptions of a recommendation and their actual behavior (e.g., clicking and purchasing).

5 DISCUSSION

In this section, we summarize the major findings of our experiment and their practical implications, and discuss the limitations of our work.

5.1 Serendipity and User Satisfaction

One major finding is that serendipity significantly positively affects user satisfaction and purchase intention. Its contribution is largely higher than that of novelty, more direct relative to diversity's, and comparable to those of relevance and timeliness. As a result, more than 60% of the variance in the two outcome variables (satisfaction and purchase intention) can be explained by the associated predictors.

Moreover, as relevance, timeliness, and novelty can directly affect serendipity, it discloses serendipity's mediating role in transmitting part of their effects on user satisfaction. However, two diversity related variables (i.e., pur_diversity and rec_diversity) are not directly related to serendipity, but have a positive influence on unexpectedness. This suggests that if the recommendation is different from the user's previously purchased product types or the system's prior recommendations, s/he may first feel surprised, which will then combine with the other factors (i.e., relevance, timeliness, novelty) to induce serendipity (altogether accounting for over 60% of the variance in serendipity).

[12] We measured a user's revisiting behavior during the day s/he took part in the survey.

Compared with the related work that primarily takes unexpectedness, novelty, and relevance as principal components of serendipity [6, 20, 28], we have two new observations. First, ***timeliness*** can be more important than unexpectedness, novelty, and relevance, in terms of affecting serendipity. However, this factor is often neglected in the related literature. For example, in [36], this variable was eliminated due to low correlations with other variables. In our study, we find it has a significantly positive impact on multiple perception variables, such as serendipity, satisfaction, and purchase intention. One possible explanation is that our experiment was conducted in a mobile e-commerce environment, so whether the recommendation was given at the best possible time can be regarded as critically important by users.

Second, we find that ***unexpectedness*** and ***novelty*** take different roles in influencing users' perceived serendipity. Novelty acts more positively as it has a significantly direct relationship with serendipity and user satisfaction, whereas unexpectedness is only related to serendipity. In addition, unexpectedness may reflect unpleasant surprise in some conditions (e.g., when the recommendation is not relevant to the user's preferences). Another finding is that although unexpectedness is affected by several variables, only 14% of its variance is explained by them all, suggesting that there would be other unexplored variables greatly impacting it.

For novelty, our results are basically consistent with those of [36] that also indicated a positive effect of novelty on user satisfaction, but contradict the findings of [7] that showed its negative influence. As discussed in [15], this may be caused by the studied product domain (e-commerce products vs. movies in [7]), or the formulation of novelty-related question (positive tone such as "*The item recommended to me is novel*" vs. negative tone such as "*Which list has more movies you would not have thought to consider?*" in [7]). More studies are thus needed to verify these confounding effects.

Implication. Because the experiment shows there are four variables, namely *timeliness*, *relevance*, *unexpectedness*, and *novelty*, are significantly related to serendipity, we may improve the existing offline metrics of measuring serendipity [1, 10], in a more precise way. For instance, we could add the component *timeliness*, as well as more clearly distinguishing the roles of *unexpectedness* and *novelty* in the formulation. It would also be meaningful to introduce some indirectly related factors (such as *pur_diversity*), and other potential factors that may lead to unexpectedness.

5.2 The Role of User Curiosity

The second finding is about the exact role of user curiosity in moderating the effect of one perception variable on another. The results of moderation analysis show that curiosity can not only strengthen the positive effect of *novelty on serendipity*, but also that of *serendipity on satisfaction*. It hence implies that a more curious person will be more likely to perceive a novel recommendation as serendipitous, and be more satisfied with the serendipitous item.

Furthermore, we find that the path models are significantly different between the high and low curiosity groups of users through the multi-group analysis. In particular, *novelty*, *unexpectedness*, and *serendipity* behave more actively in positively affecting the other variables in the high curiosity group. The accuracy of predicting serendipity (and furthermore user satisfaction and purchase intention) can also be likely higher in the high curiosity group. Overall, these results verify our hypothesis about the effect of curiosity on users' appetite for serendipitous recommendations.

Implication. Recently, some studies have attempted to incorporate curiosity into recommendation generation [24, 29, 47], but their methods were mainly based on some assumption without empirical validation of the curiosity's actual effect. The results of our study could be constructive for the related work to further strengthen the role of curiosity in their recommender algorithms. For instance, given that curiosity significantly moderates the causal relationships from novelty to serendipity and from serendipity to satisfaction, a personalized recommendation strategy might be developed to dynamically adjust the serendipity degree according to the target user's curiosity value, instead of being simply maximized for everyone.

5.3 User Evaluation of Recommender Algorithms

This experiment also explains why a particular recommender approach is preferred to others. We compared four algorithms: HOT, Rel-CF, Nov-CF, and Ser-CF, which are respectively popularity, relevance, novelty, and serendipity oriented. The results show that Ser-CF is the best in terms of novelty, relevance, timeliness, serendipity, user satisfaction, and purchase intention; Nov-CF and Rel-CF lie in the middle; and HOT provides the most unexpected recommendations, but performs worst regarding most of the other variables. It thus infers that a serendipity-oriented algorithm might be more satisfying and stimulating users' intention to purchase its recommended items. The observations related to HOT suggest that unexpectedness can reflect users' unpleasant surprise if the recommendation is purely popularity based without taking into account the user's preferences.

Implication. The way we evaluated recommender algorithms might be referential to related researchers for assessing user perceptions of their algorithms. On the other hand, since the advantage of the serendipity-oriented algorithm (Ser-CF) is not very obvious among less curious users, it would be meaningful to have a personalized serendipity approach as we discussed above. Recently, a personality-based framework was proposed to adapt recommendations' diversity to individual users' intrinsic needs [44]. We could extend it to accommodate serendipity, so as to achieve optimal user satisfaction with the recommender system.

5.4 Limitations of Our Work

Our work has several limitations. First, the users who took part in our experiment mainly represent the population that has used mobile apps to buy products [43]. The results might have been different if it was conducted in another setting (e.g., traditional e-commerce) or even other domains (such as social media and tourism). Therefore, we are interested in performing a cross-platform validation in the future, so as to identify any platform-specific features. Second, a cross-cultural study would also be necessary, because our survey mainly involved Chinese users. It will be interesting to see whether similar findings would occur among users in other countries, or there would be any cultural differences. Third, we used the short

version of ResQue questionnaire [36] for the sake of mobile survey, which limited our statistical approach to path analysis that does not contain latent variables to account for measurement error [38]. It would therefore be better to verify the findings by using the complete version of ResQue [36], and the other user-centric evaluation frameworks [18] (given that different formulations of the survey questions may engender different user responses [15]). Fourth, we did not test the state-of-the-art algorithms, but emphasized on tailoring the classical CF methods to strengthen the recommendation's relevance, novelty, and serendipity respectively. For the future work, it will be useful to identify whether the results could be generalizable to other algorithms, especially those that are serendipity oriented [24, 33, 46].

6 CONCLUSIONS

Different from related work on recommendation serendipity that either focused on algorithm development or conducted small-scale user evaluations, we have been engaged in empirically revealing the causal relationships between serendipity and various user perceptions via a large scale online study performed on an industrial mobile e-commerce platform. The results show that there are four major components of serendipity, namely *timeliness*, *relevance*, *unexpectedness*, and *novelty*. Furthermore, *serendipity* behaves more effectively than the other two beyond-accuracy objectives (novelty and diversity) in affecting *user satisfaction* and *purchase intention*.

Additionally, this study verifies the moderating effects of *user curiosity* on the causal relationships from novelty to serendipity and from serendipity to satisfaction. This implies that more curious users will be more likely to perceive a novel item as serendipitous and be satisfied with a serendipitous recommendation. Last but not least, the user evaluation of four recommender algorithms indicates that a serendipity-oriented method (*Ser-CF*) performs the best. Inspired by those findings, we believe that a personalized serendipity approach based on curiosity would be potentially able to optimize user satisfaction with the recommender system.

ACKNOWLEDGMENTS

This work was partially supported by Hong Kong Research Grants Council (RGC) (project RGC/HKBU12200415).

REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *ACM Trans. Intell. Syst. Technol.* 5, 4, Article 54 (Dec. 2014), 32 pages. https://doi.org/10.1145/2559952

[2] Daniel Ellis Berlyne. 1960. *Conflict, Arousal and Curiosity.* McGraw-Hill.

[3] Toine Bogers and Lennart Björneborn. 2013. Micro-serendipity: Meaningful Coincidences in Everyday Life Shared on Twitter. In *Proceedings of the iConference 2013.* 196–208.

[4] Barbara M. Byrne. 2016. *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming* (3rd ed.). Routledge.

[5] Curiosity and Exploration Inventory (CEI-II) [n. d.]. http://www.midss.org/content/curiosity-and-exploration-inventory-cei-ii.

[6] Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Cataldo Musto. 2015. An Investigation on the Serendipity Problem in Recommender Systems. *Inf. Process. Manage.* 51, 5 (Sept. 2015), 695–717. https://doi.org/10.1016/j.ipm.2015.06.008

[7] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14).* ACM, New York, NY, USA, 161–168. https://doi.org/10.1145/2645710.2645737

[8] Donald E. Farrar and Robert R. Glauber. 1967. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics* 49, 1 (1967), 92–107.

[9] Shengyi Gao, Patricia L. Mokhtarian, and Robert A. Johnston. 2008. Nonnormality of Data in Structural Equation Models. *Transportation Research Record* 2082, 1 (2008), 116–124. https://doi.org/10.3141/2082-14

[10] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10).* ACM, New York, NY, USA, 257–260. https://doi.org/10.1145/1864708.1864761

[11] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommendation Systems. In *Recommender Systems Handbook.* Springer US, 265–308. https://doi.org/10.1007/978-1-4899-7637-6_8

[12] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* 22, 1 (Jan. 2004), 5–53. https://doi.org/10.1145/963770.963772

[13] Daire Hooper, Joseph Coughlan, and Michael Mullen. 2008. Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods* 6, 1 (2008), 53–60.

[14] Dawn Iacobucci, Steven S. Posavac, Frank R. Kardes, Matthew J. Schneider, and Deidre L. Popovich. 2015. The Median Split: Robust, Refined, and Revived. *Journal of Consumer Psychology* 25, 4 (2015), 690–704. https://doi.org/10.1016/j.jcps.2015.06.014

[15] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1, Article 2 (Dec. 2016), 42 pages. https://doi.org/10.1145/2926720

[16] Todd B. Kashdan, Matthew W. Gallagher, Paul J. Silvia, Beate P. Winterstein, William E. Breen, Daniel Terhar, and Michael F. Steger. 2009. The Curiosity and Exploration Inventory-II: Development, Factor Structure, and Psychometrics. *Journal of Research in Personality* 43, 6 (2009), 987–998. https://doi.org/10.1016/j.jrp.2009.04.011

[17] Todd B. Kashdan, Paul Rose, and Frank D. Fincham. 2004. Curiosity and Exploration: Facilitating Positive Subjective Experiences and Personal Growth Opportunities. *Journal of Personality Assessment* 82, 3 (2004), 291–305. https://doi.org/10.1207/s15327752jpa8203_05

[18] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (Oct. 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[19] Denis Kotkov, Joseph A. Konstan, Qian Zhao, and Jari Veijalainen. 2018. Investigating Serendipity in Recommender Systems Based on Real User Feedback. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC '18).* ACM, New York, NY, USA, 1341–1350. https://doi.org/10.1145/3167132.3167276

[20] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A Survey of Serendipity in Recommender Systems. *Knowledge-Based Systems* 111 (Nov. 2016), 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

[21] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal Diversity in Recommender Systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10).* ACM, New York, NY, USA, 210–217. https://doi.org/10.1145/1835449.1835486

[22] Todd D. Little, Noel A. Card, James A. Bovaird, Kristopher J. Preacher, and Christian S. Crandall. 2012. Structural Equation Modeling of Mediation and Moderation with Contextual Factors. In *Modeling Contextual Effects in Longitudinal Studies.* Taylor and Francis, 207–230. https://doi.org/10.4324/9780203936825

[23] George Loewenstein. 1994. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological Bulletin* (1994), 75–98.

[24] Valentina Maccatrozzo, Manon Terstall, Lora Aroyo, and Guus Schreiber. 2017. SIRUP: Serendipity In Recommendations via User Perceptions. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17).* ACM, New York, NY, USA, 35–44. https://doi.org/10.1145/3025171.3025185

[25] Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (1951), 68–78. https://doi.org/10.1080/01621459.1951.10500769

[26] Christian Matt, Thomas Hess, Alexander Benlian, and Christian Weiß. 2014. Escaping from the Filter Bubble? The Effects of Novelty and Serendipity on Users' Evaluations of Online Recommendations. (2014). https://EconPapers.repec.org/RePEc:dar:wpaper:66193

[27] John H. McDonald. 2014. *Handbook of Biological Statistics, Third Edition.* Sparky House Publishing.

[28] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06).* ACM, New York, NY, USA, 1097–1101. https://doi.org/10.1145/1125451.1125659

[29] Alan Menk, Laura Sebastia, and Rebeca Ferreira. 2017. CURUMIM: A Serendipitous Recommender System for Tourism Based on Human Curiosity. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI).* 788–795. https://doi.org/10.1109/ICTAI.2017.00124

[30] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. 2008. Metrics for Evaluating the Serendipity of Recommendation Lists. In *New Frontiers in Artificial Intelligence*. Springer Berlin Heidelberg, 40–46.

[31] Bengt Muthén and David Kaplan. 1985. A Comparison of Methodologies for the Factor Analysis of Non-Normal Likert Variables. *Brit. J. Math. Statist. Psych.* 38, 1 (1985), 171–189.

[32] Robert M. O'Brien. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity* 41, 5 (01 Oct 2007), 673–690. https://doi.org/10.1007/s11135-006-9018-6

[33] Kenta Oku and Fumio Hattori. 2012. User Evaluation of Fusion-Based Approach for Serendipity-Oriented Recommender System. In *Proceedings of the Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2012)*. 39–44.

[34] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin UK.

[35] Robert A. Peterson. 1994. A Meta-Analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research* 21, 2 (1994), 381–391.

[36] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962

[37] Alan Said, Ben Fields, Brijnesh J. Jain, and Sahin Albayrak. 2013. User-centric Evaluation of a K-furthest Neighbor Collaborative Filtering Recommender Algorithm. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1399–1408. https://doi.org/10.1145/2441776.2441933

[38] Randall E. Schumacker and Richard G. Lomax. 2015. *A Beginner's Guide to Structural Equation Modeling* (4th ed.). Routledge.

[39] Paul J. Silvia. 2008. Interest–The Curious Emotion. *Current Directions in Psychological Science* 17, 1 (02 2008), 57–60.

[40] Kirsten Swearingen and Rashmi Sinha. 2001. Beyond Algorithms: An HCI Perspective on Recommender Systems. In *ACM SIGIR. Workshop on Recommender Systems*, Vol. 13. Citeseer, 1–11.

[41] Paula Vicente, Elizabeth Reis, and Maria Santos. 2009. Using Mobile Phones for Survey Research: A Comparison with Fixed Phones. *International Journal of Market Research* 51, 5 (2009), 613–633. https://doi.org/10.2501/S1470785309200852

[42] Horace Walpole. 1960. To Mann, Monday 18 January 1754. In *Horace Walpole's Correspondence*. Yale University Press, 407–411.

[43] Jen-Her Wu and Shu-Ching Wang. 2005. What Drives Mobile Commerce? *Information and Management* 42, 5 (July 2005), 719–729. https://doi.org/10.1016/j.im.2004.07.001

[44] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing Recommendation Diversity Based on User Personality. *User Modeling and User-Adapted Interaction* 28, 3 (01 Aug 2018), 237–276. https://doi.org/10.1007/s11257-018-9205-x

[45] Shengquan Ye, Ting Kin Ng, Kin Hang Yim, and Jun Wang. 2015. Validation of the Curiosity and Exploration Inventory-II (CEI-II) Among Chinese University Students in Hong Kong. *Journal of Personality Assessment* 97, 4 (2015), 403–410. https://doi.org/10.1080/00223891.2015.1013546 PMID: 25774779.

[46] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/2124295.2124300

[47] 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 (SIGIR '16)*. ACM, New York, NY, USA, 315–324. https://doi.org/10.1145/2911451.2911488

[48] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists Through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web (WWW '05)*. ACM, New York, NY, USA, 22–32. https://doi.org/10.1145/1060745.1060754