LuckyFind: Leveraging Surprise to Improve User Satisfaction and Inspire Curiosity in a Recommender System

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

Ahmad Al-Doulat
adoulat@uncc.edu
University of North Carolina at Charlotte
Charlotte, NC, USA

ABSTRACT

The growing amount of online information today has increased opportunity to discover interesting and useful information. Various recommender systems have been designed to help people discover such information. No matter how accurately the recommender algorithms perform, users' engagement with recommended results has been complained being less than ideal. In this study, we touched on two human-centered objectives for recommender systems: user satisfaction and curiosity, both of which are believed to play roles in maintaining user engagement and sustain such engagement in the long run. Specifically, we leveraged the concept of surprise and used an existing computational model of surprise to identify relevantly surprising health articles aiming at improving user satisfaction and inspiring their curiosity. We designed a user study to first test the validity of the surprise model in a health news recommender system, called LuckyFind. Then user satisfaction and curiosity were evaluated. We find that the computational surprise model helped identify surprising recommendations at little cost of user satisfaction. Users gave higher ratings on interestingness than usefulness for those surprising recommendations. Curiosity was inspired more for those individuals who have a larger capacity to experience curiosity. Over half of the users have changed their preferences after using LuckyFind, either discovering new areas, reinforcing their existing interests, or stopping following those they did not want anymore. The insights of the research will make researchers and practitioners rethink the objectives of today's recommender systems as being more human-centered beyond algorithmic accuracy.

CCS CONCEPTS

• **Information systems** → **Users and interactive retrieval**; **Recommender systems**; • **Human-centered computing** → **User studies**.

KEYWORDS

surprise; user satisfaction; curiosity; recommender systems

CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia

ACM ISBN 978-1-4503-8055-3/21/03…$15.00
https://doi.org/10.1145/3406522.3446017

ACM Reference Format:
Xi Niu and Ahmad Al-Doulat. 2021. LuckyFind: Leveraging Surprise to Improve User Satisfaction and Inspire Curiosity in a Recommender System. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3406522.3446017

1 INTRODUCTION

Understanding user satisfaction and curiosity is a complicated research problem in many information discovery systems, such as search engines and recommender systems. Today's recommender systems have been complained to provide close matches to a user's previous access history rather than promoting richer information discovery that satisfies users and inspires their curiosity to sustain long-term engagement. Previous studies have shown that surprising or unexpected discovery may attract user attention and arouse pleasant feelings, such as interest, like, and curiosity (e.g., [13, 27]). In this study we will "engineer" some surprise into a health news recommender and evaluate whether that would result in user satisfaction and curiosity.

One type of recommender systems is the knowledge-based that presents relevant items based on a user's self-reported interests. When a user accesses the recommender system for the first time, they are asked to select their preferred topics from a list. The selected topics constitute the user's profile and are called profile topics (PT); in contrast, the topics that are not selected by the user are called non-profile topics (NPT). However, a problem arises due to users' lack of awareness or due to their inability to articulate their full range of interests. Therefore, the PT topics that the systems use may be only a partial set of topics that a user is actually interested in. The state-of-the-art matching algorithms, either through conventional machine learning or deep learning techniques, are likely to "entrap" the user in a narrow scope. With the purpose of helping users learn their full range of interests, our study implements the concept of "surprise" by delivering richer information that is outside users' expectation, but favorable and inspiring. We hope such recommendations could incrementally reinforce, expand, or shift user interests they may otherwise have not been able to recognize, and finally sustain long-term user engagement with the system. Using existing research on computational surprise models in artificial intelligence, for example the studies of [4, 9, 11–14, 18, 26, 34, 35, 42–44, 46], this study applies one of them into a recommender system for online health news. It first tests the validity of the surprise model, and then evaluates the impact of surprising health news on user satisfaction and curiosity.

The major contributions are summarized as: 1) although the computational surprise model is from an existing study, we are the first to design a user study to test the surprise model's validity in a recommender system, i.e., whether pieces of surprising news identified by the model align with what users feel; 2) this study innovatively evaluates the subtle relationship between surprise, user satisfaction, and curiosity, which is believed to have profound impacts on the development of recommender systems; and 3) this study is the first to report the changes in users' preferences before and after being exposed to recommendations.

2 RELATED WORK

This research brings together three concepts: surprise, user satisfaction, and curiosity in the field of Information Retrieval.

2.1 The Concept of Surprise and Its Relationship with Serendipity

In cognitive science, surprise has been described as the events that are different from one's expectations [32], or are difficult to explain [10]. In neuroscience, seeking surprise has been a well-identified human trait. It has been suggested that only the surprising signal at one stage is transmitted to the next stage [39]. Hence, human sensory cortex may have adapted to predict and downplay the expected regularities of the world [8, 37], focusing instead on events that are unpredictable or surprising. Therefore, human attention greatly reduces with repeated or prolonged exposure to an initially surprising stimulus. These descriptions in cognitive and neurosciences suggest that surprise attracts human attention, which is the reason why we believe the concept of surprise is valuable in recommender systems.

There is a concept related to surprise called serendipity, which has attracted wide attention in these years. First coined by Harold Walpole in 1754 [31], the word "serendipity" means the process of making unexpected discoveries by accident. Recent studies on serendipity involve two groups of researchers: one in social sciences who have made attempts to define and characterize serendipity. Examples of this stream include [2, 3, 28–30]. The other group involves computer scientists trying to use machine learning or deep learning techniques to predict what things are serendipitous, and then recommend those to users. Examples of this stream are [1, 19, 21, 38]. Although there is some disagreement as to the precise nature of serendipity, most descriptions agree that the following two aspects are central: an unexpected encounter and a valuable discovery. This unexpected encounter, in our opinion, is surprise. We believe surprise is a critical element to trigger serendipity.

In this paper, we will study this important trigger of serendipity - surprise, and whether surprise will be valued by users in recommender systems. In particular, we will apply one existing model of surprise [35] for text-based items, and evaluate whether recommending surprising items will improve user satisfaction and inspire their curiosity.

2.2 User Satisfaction

In Human-Computer Interaction (HCI), user satisfaction of a system is based on subjective, affective, or emotional aspects of a user assessment, as a "subjective sum of the interactive experience" described by [22]. Recognition of the affective aspect of user satisfaction stems from the observation that the ultimate usability of a system is invariably determined by subjective, user-specific functions. This makes user satisfaction fundamentally distinct from other evaluation metrics, such as efficiency and effectiveness. A system may be evaluated favorably on both efficiency and effectiveness, but may not be used very much because of low user satisfaction with the system [7].

The HCI community generally agrees that user satisfaction is affected by the perceived usability and aesthetics of an interface [6, 15, 16, 20, 45], and the extent to which user needs are met [17]. Satisfaction was found to be complex, intensity of which varied with the nature of the experience. Experts believe that the amount of user interaction with a product affects users' overall satisfaction. People with long-term use experience tended to rate satisfaction higher than those with a shorter period of experience [41]. That suggests the cyclic relationship between user satisfaction and continued use.

The definition of user satisfaction in this study was inspired by the study [24] that develops two constructs for the subjective value of information: Willingness-to-Pay (WTP) and Experienced Utility (EU). WTP is suggested to reflect the instrumental-rational value that an information object has in problem-solving tasks, while EU reflects the aesthetic-emotional value that an information object has in its own right as the user engages with the object. Our user satisfaction definition reflected both aspects by collecting users' opinions on whether the news article is *useful* and whether it is *interesting*.

2.3 Curiosity

In the classic study by Berlyne in Psychology in 1966 [5], curiosity has been described as both a trait (C-Trait) and a state (C-State). The C-Trait of curiosity refers to individual differences in the capacity to experience curiosity, while C-State means the same individual's difference in response to a particular stimulus. While most HCI studies have focused on C-State, this study, however, measures both C-State and C-Trait with the hypothesis that surprising news will increase C-State in general; and users possessing a high level of C-Trait will experience a wider range of situations as curiosity-arousing than do users possessing a low level of C-Trait.

Prior research has suggested that curiosity is a manifest of user engagement [36]. Curiosity has a strong association with attention, intrinsic interest, and motivation. According to Berlyne [5], curiosity is human response to external stimulus. He further identified a set of stimulus factors that can arouse curiosity, such as novelty, uncertainty, conflict and complexity. His study inspired many research studies in artificial intelligence (AI) that use a single or several stimulus factors to arouse curiosity. For example, Saunders and Gero [40] focused on appraisal of novelty of architecture design patterns, as the key for evaluating the curiosity arousal and therefore the selection of good design patterns. In Wu et al. [47], they considered four stimulus variables: novelty, uncertainty, conflict, and complexity, to encourage curiosity in a virtual learning environment. Macedo and Cardoso [25, 27] incorporated novelty, surprise, and uncertainty into their curious intelligent agents to simulate human-like exploratory behavior.

Among these stimulus factors for curiosity, *surprise* has been chosen in this study to inspire curiosity, since we believe surprise is complex and reflects other stimulus factors, such as novelty, uncertainty, and conflict mentioned by Berlyne [5].

3 RESEARCH QUESTIONS

We will develop a recommender, called LuckyFind, which uses an existing computational model of surprise to identify surprising health news articles. Specifically, two variations of LuckyFind will be implemented: Knowledge-Based (KB) and Adaptive Knowledge-Based (AKB). For the KB variation, the news articles will be recommended to users based on their levels of surprise scores. However, the limitation of the KB variation lies in the possibility that an NPT (non-profile topics) may be over-represented in the recommendation list. If a user is not interested in this NPT, repeated presentation of it will reduce the chance for other NPTs to be seen. To solve this problem, an Adaptive Knowledge-Based variation will be implemented to incorporate users' real-time feedback. This adaptive method excluded the presentation of certain NPT topics that users have low ratings. For the evaluation purpose, we will also implement a baseline system that recommends articles randomly selected from the PTs (profile topics). We will compare the articles from the KB or AKB variation with the baseline systems in terms of user surprise, satisfaction, and curiosity ratings. Specifically, four research questions are put forward:

- RQ 1: Is the computational surprise model able to find surprising contents? Is there any difference in KB and AKB in terms of finding surprising contents?
- RQ 2: Does presenting surprising contents increase user satisfaction? Is there any difference in KB and AKB in terms of user satisfaction?
- RQ 3: Does presenting surprising contents inspire user curiosity? Is there any difference in KB and AKB in terms of user curiosity?
- RQ 4: What are the impacts of using LuckyFund on users' preference change?

4 RESEARCH METHOD

4.1 Computational Surprise Model

The computational surprise model in this study is from the study [35]. We will briefly introduce it here. We represented each news article as a "bag" of its topics. For example, the article "Optometrists: Got the Flu? Take Out Your Contacts" could be represented as a bag of its topics: *flu*, *infectious disease*, and *eye health*. We then measured the article surprise as how unlikely a topic combination is in one article. For example, *flu* tends to co-occur with *infectious disease* with a high likelihood, but not as much co-occurring with *eye health*. Expectations of co-occurrence likelihood have been implicitly formed by our collective knowledge as the expectation, and were computationally constructed using a large collection of news articles. A surprise in that sense is: seeing the topic *flu* is rare (surprising) given seeing the topic *eye health* in the same article. To capture the heuristics of co-occurrence likelihood, as in [35], Pointwise Mutual Information (PMI) was used to calculate how much more likely than expected it is that a topic t_i occurs given the occurrence of another topic t_j. We call this pairwise surprise score s, as in Equation 1:

$$s(t_i, t_j) = -PMI(t_i, t_j) = -log_2 \frac{p(t_i, t_j)}{p(t_i)p(t_j)} \quad (1)$$

where $p(t_i)$ and $p(t_j)$ represent the individual occurrence probabilities of the topics t_i and t_j, and $p(t_i, e_j)$ represents the joint occurrence probability of the two. In this equation, the lower part of the log fraction represents the occurrence expectation of these two topics 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 divergence from expectation or surprise. A smaller ratio indicates a rare combination, and therefore a higher pairwise surprise score.

Since many items have more than two topics, the pairwise surprise s is calculated for all possible pairwise combinations and the highest of those values becomes the overall surprise score S, based on the idea that the peak element-level surprise dominates the item-level surprise. This is shown in Equation 2, where E is the set of all possible pairwise combinations belonging to the same article.

$$S = max_E s(e_i, e_j) \quad (2)$$

4.2 Health News Corpus

The health news collection in this study is a corpus we scraped from the Medical News Today (MNT) website since its launch in 2003 to the present. MNT is one of the leading websites to provide quality and up-to-date health news for average readers in the U.S. This corpus includes 268,850 articles, labeled with 135 different health topics by health professionals working with MNT. These topics include a wide range of health-related topics, such as *cancer*, *depression*, *men's* health, *women's health*, and *anxiety*. For our recommender system study, the corpus was further divided into 135 sub-corpora with each sub-corpus involving one topic. It is noteworthy that these sub-corpora may have large overlap articles because many articles are labeled with several topics. Since our computational surprise measure relies on topic co-occurrence, we have removed articles that only contain one topic. The final corpus includes 123 topics with 181,102 articles.

To illustrate the result of the computational measure of surprise, the distribution of S for all the articles used in this study is presented in Figure 1. It shows that S ranges from -7.5 to 6.5. The distribution roughly follows a normal distribution with the majority articles centered in the middle. There are very few highly surprising or highly non-surprising articles. Table 1 presents the top five topics that have the highest average pairwise s with the other topics, and the bottom five topics that have the lowest average pairwise s. The average s with other topics indicates the "incompatibility" with other topics. From this Figure, we understand that those top five topics such as *fibromyalgia* and *medical innovation* are very "incompatible" with other topics, therefore are relatively more surprising while co-occurring with other topics.

Table 2 lists some examples of the most and least surprising articles based on S. As we can see in the topic labels column of this Table, the most surprising articles (highlighted in gray) have

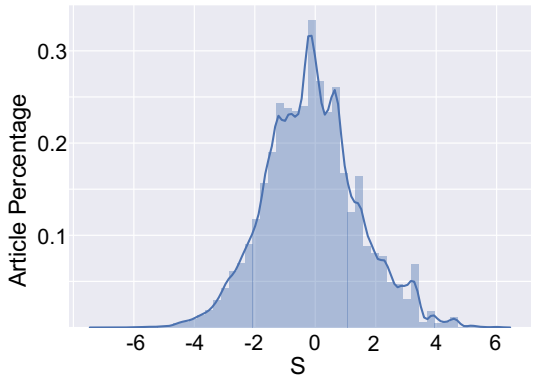

Figure 1: Distribution of S in the entire corpus of MNT

Table 1: The top and bottom five topics' average pairwise s with another topic

Topic	Average s
Fibromyalgia	0.08
Medical Innovation	-0.03
Irritable-Bowel Syndrome	-0.10
Gout	-0.13
Compliance	-0.17
Aid/Disasters	-1.68
Water-Air Quality/Agriculture	-1.70
Nursing/Midwifery	-1.71
Flu/Cold/SARS	-2.01
HIV/AIDS	-2.05

Notes: The shaded area indicates the top five topics' average pairwise s with another topic, while the white area indicates the bottom five topics

those rare topic combinations, such as a combination of *pregnancy* and *lung cancer*; whereas the least surprising articles have common combinations like *statins* and *cholesterol*.

4.3 Two Variations of LuckyFind: KB and AKB

LuckyFind used the computational surprise model to find health news articles based on their S scores. The system then recommended those health news articles to users on a session basis. Each session was a list of five to ten articles depending on how many topics a user had selected as the PTs before any recommendation. Since the study did not evaluate the impact of the specific ranking of articles in a list, we adopted a random position approach for each topic in a session. How to pick articles in a session depended on whether it is the KB or AKB variation.

4.3.1 Knowledge-Based Approach (KB). For KB, all the news articles in one PT sub-corpus were grouped into five percentile range segments as in Table 3. The articles in Segment 1 with the lowest 10^{th} to 20^{th} percentile of S were recommended in the first and the second session, followed by the articles in Segment 2 in the third and the fourth session, and so on so forth. The last two sessions had the recommendations with the highest (Segment 5) S scores (90^{th} to 100^{th}). The reason for doing these segmented sessions was

Table 2: Example articles based on S

Example Article Title	Topic Labels	S
Fetal Exposure to Carcinogens Leading to Cancer Depends on Dose, Timing	Pregnancy, Ovarian Cancer, Lung Cancer, Public Health	6.08
Optometrists: Got the Flu? Take Out Your Contacts	Flu, Infectious Diseases, Eye Health	6.07
Sneezing in Times of A Flu Pandemic	Swine Flu, Public Health, Psychology, Immune System	5.80
Statins: Uses, Side Effects, and Risks	Statins, Cholesterol	-6.04
General Osteopathic Council Backs Awareness Week	Back Pain, Body Aches	-5.88
What Is the Difference Between Food Allergy and Food Intolerance?	Allergy, Food Intolerance	-5.61

Notes: * The shaded area indicates the surprising articles whereas the white area indicates the non-surprising ones
** Topic labels in red are the pair combination with the highest s.

to investigate the validity and sensitivity of that computational surprise model in this health news recommender, i.e., testing whether the higher S really matched what users thought as more surprising. For each segment, two consecutive sessions were offered for the measurement reliability.

Table 3: S percentile ranges in a PT sub-corpus

Segment	S Percentile Range	Session
Segment 1	10^{th} - 20^{th} (the lowest S)	1^{st} and 2^{nd}
Segment 2	30^{th} - 40^{th}	3^{rd} and 4^{th}
Segment 3	50^{th} - 60^{th}	5^{th} and 6^{th}
Segment 4	70^{th} - 80^{th}	7^{th} and 8^{th}
Segment 5	90^{th} - 100^{th} (the highest S)	9^{th} and 10^{th}

In order to better explain how LuckyFind recommends session-based articles, let us assume one participant chooses six preferred topics: *anxiety*, *diabetes*, *depression*, *sleep disorder*, *breast cancer*, and *hypertension*. For this user, each recommended session contains six articles and there are ten consecutive sessions in total, as shown in Figure 2 below.

4.3.2 Adaptive Knowledge-Based Surprise (AKB). If a recommendation does not receive high user satisfaction but similar recommendations are presented repeatedly, LuckyFind may lose the user's attention and have a high opportunity cost. Based on the KB approach, the AKB approach was developed to solve the problem of over-presentation of non-satisfactory recommendations. The AKB variation is similar to KB but it incorporated users' real-time ratings. Specifically, one NPT was penalized (removed) if a recommended article with such a NPT had received a negative user satisfaction rating (less than 6 out of 10 on the sum of the interestingness and usefulness ratings).

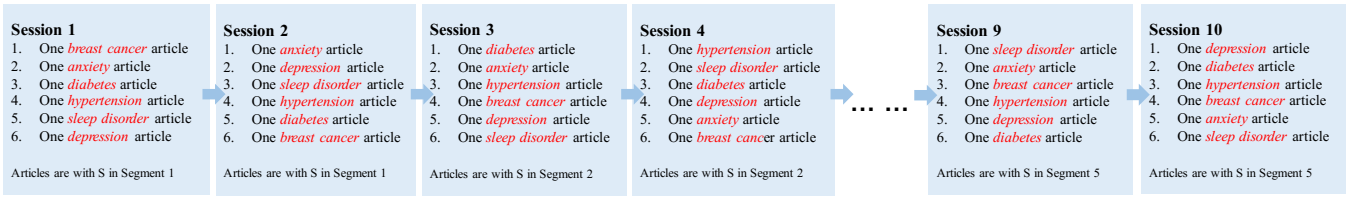

Figure 2: Examples of recommended sessions

4.4 User Satisfaction and Curiosity Measurement

For user satisfaction, we used an interactive process to collect the satisfaction information. Before any recommendation session, the user was asked to select 5 - 10 topics as their PTs. These selected PTs made the scope of articles for applying the computational model of surprise, ensuring that all surprise would be relevant to the user's selections. During the recommended sessions, we asked participants to rate on two 5-point Likert-scales to indicate how useful and how interesting they thought that article was. Usefulness is to represent the instrumental aspect while interestingness is to represent the emotional aspect, as in [24]. The sum of the two ratings served as the final user satisfaction rating.

As for curiosity, we have adopted the C-Trait and C-State instruments developed by Naylor [33] with proven validity and reliability. We have adapted them into our recommender system context. Specifically, since our C-Trait and C-State were more about epistemic curiosity (drive to learn knowledge) [23] rather than perceptual curiosity (visual, auditory, or tactile curiosity) [23], we have removed the questions related to the perceptual curiosity such as "I enjoy exploring new food", and focused on the epistemic curiosity questions such as "I think learning about things is interesting and exciting". As the result, there were 20 statements in the C-Trait questionnaire, which was applied before the user using LuckyFind; and 20 statements in the C-State questionnaire, which was applied afterwards. For both questionnaires, we asked participants to rate on 4-point Likert-scales (to avoid neutral answers) to indicate how much they agree with each statement. Both questionnaires' statements are listed in Table 4.

Table 4: C-Trait and C-State questionnaires

C-Trait Statements
1. I think learning "about things" is interesting and exciting.
2. I am curious about things.
3. I enjoy taking things apart to "see what makes them tick".
4. I feel involved in what I do .
5. My spare time is filled with interesting activities.
6. I like to try to solve problems that puzzle me.
7. I enjoy exploring new places.
8. I feel active.
9. New situations capture my attention.
10. I feel inquisitive.
11. I feel like asking questions about what is happening.
12. The prospect of learning new things excites me.
13. I feel like searching for answers.
14. I like speculating about things.
15. I like to experience new sensations.
16. I feel interested in things.
17. I like to enquire about things I don't understand.
18. I feel like seeking things out.
19. I want to probe deeply into things.
20. I feel absorbed in things I do.

C-State Statements
1. I want to know more.
2. I feel curious about what is happening.
3. I am feeling puzzled.
4. I want things to make sense.
5. I am intrigued by what is happening.
6. I want to probe deeply into things.
7. I am speculating about what is happening.
8. My curiosity is aroused.
9. I feel interested in things.
10. I feel inquisitive.
11. I feel like asking questions about what is happening.
12. Things feel incomplete.
13. I feel like seeking things out.
14. I feel like searching for answers.
15. I feel absorbed in what I am doing.
16. I want to explore possibilities.
17. My interest has been captured.
19. I want more information.
20. I want to enquire further.

5 EVALUATION STUDY

After the implementation of the two variations of LuckyFind, a user study was conducted to evaluate whether the computational surprise model identified recommendations that were surprising, satisfactory, and curiosity-inspiring. The study adopted a one-way design. The independent variable is the variations: KB, AKB, or a baseline system that recommends articles randomly selected from the profile topics. The comparison between KB and AKB is a between-subject design to reduce the learning effect between the two variations as well as alleviate the burden on each user. Each subject was randomly assigned to one of the two groups. The comparison between KB and the baseline approach or AKB and the baseline approach was within-subject for its stronger statistical power. The placement of the two baseline sessions in the KB or the AKB group was counterbalanced to remove the order effect: they were inserted into one of the six possible points (as shown in Figure 3) along the ten KB and AKB sessions, making it twelve sessions in total. Several screenshots of a session are presented in Figure 4.

Thirty undergraduate students without any formal medical education were recruited. After the introduction and collection of

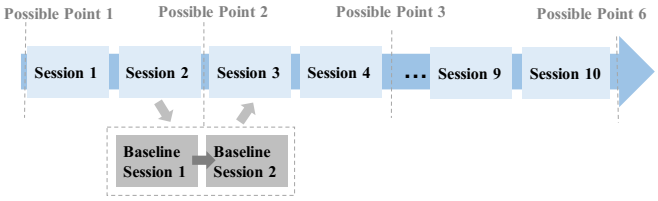

Figure 3: Placement of the baseline sessions

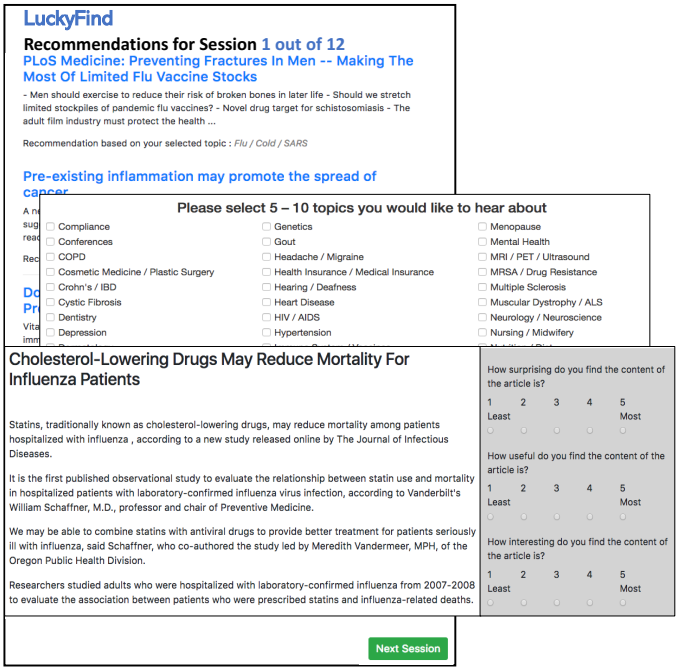

Figure 4: Screenshots of a recommendation session

consent, a series of entry questionnaires were administered to collect demographic information, their preferred health topics, as well as their curiosity levels. Each participant then experienced twelve recommended sessions, of which ten of them were either KB or AKB, and the other two were baseline. In each session, the articles were recommended according to their S percentile segments as in Table 3. Each session contains a list of recommended articles corresponding to the user's PTs. In each session, participants were encouraged to click on whatever articles they would like to read. For each clicked article, they were required to provide their ratings on three 5-point Likert-scales: whether the article content was surprising, useful, and interesting.

After the recommended sessions, each user was asked to complete a topic preference questionnaire similar to the pre-study one in order to test whether there was topic preference change after using LuckyFind. They also needed to complete the C-State questionnaire. In addition, users were interviewed about their experience and perception.

6 EVALUATION RESULTS

All of the thirty subjects have completed the study. The average age of them is 23. On average they have 11.6 years of experience seeking information online and 5.9 years seeking health information online. 23 subjects mentioned Google and 12 subjects mentioned WebMD as their primary sources of health information. Mayo Clinic, PubMed, Reddit, Twitter, and Wikipedia were also mentioned a few times. Subjects have selected 70 health topics as their preferred topics, which is a reasonable coverage of the 123 health topics we have. The most frequent selected health topics were *mental health*, *anxiety/stress*, *depression*, *sleep/sleep disorder/insomnia*. When asked about their general satisfaction with their past experience with online health information, most people gave moderate ratings, 3.4 out of 5 on average.

6.1 Surprise Ratings

In order to answer RQ1, Figure 5 shows the average surprise rating for each session broken down by each variation. Although the two baseline sessions were not necessarily the 11^{th} and the 12^{th} sessions during the user study, for visualization purpose, we have put them in Session 11 and Session 12 for easy comparison with KB and AKB.

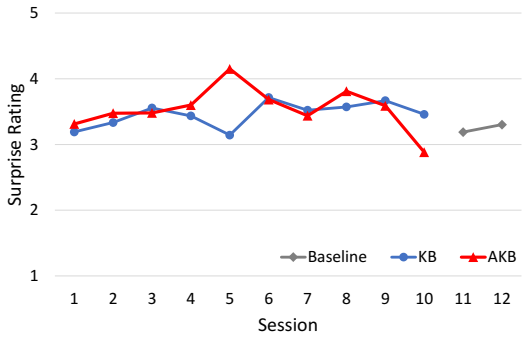

Figure 5: The average surprise rating in each session by different variations

Overall speaking, both the KB and AKB's surprise ratings are higher than the baseline system. We have conducted two repeated measure t-tests for KB vs baseline and AKB vs. baseline respectively. Both t-test results are significant ($t(56) = 2.1206$, $p = 0.0192$; $t(38) = 1.7568$, $p = 0.0435$), suggesting that the computational surprise model is able to identify significantly surprising articles than the baseline system. However, the surprise ratings between KB and AKB are comparable to each other, as the t-test result is not significant ($t(423.77) = 0.0291$, $p = 0.5116$). This is expected because both KB and AKB used the same computational model of surprise; their difference was only the incorporation of users' real time feedback on user satisfaction.

For KB, there is a slightly increasing trend for the surprise ratings along with the sessions, which is expected since we have allocated the articles with the higher S in those later sessions. However, for the AKB approach, such an increasing trend is not that obvious. The reason is probably that after incorporating the real-time user

feedback, we may have removed those "bold" articles that were very surprising but failed to satisfy users.

6.2 User Satisfaction Ratings

User satisfaction comprises user-perceived usefulness and interestingness. The usefulness ratings for each session and each variation are shown in Figure 6. Although there is some fluctuation, the repeated measure t-tests show that neither the comparisons between KB and the baseline or AKB and the baseline approach is significant ($t(56) = -0.3391$, $p = 0.3679$; $t(38) = 1.0690$, $p = 0.8541$), meaning that the user-perceived usefulness is not lower for KB or AKB than the baseline approach. Both KB and AKB are able to maintain a comparable usefulness level with the baseline approach, while searching for surprising contents. The comparison between KB and AKB is not significant either ($t(423.77) = -0.4215$, $p = 0.6632$).

The variation among the sessions demonstrates certain interesting patterns. AKB has seen a higher usefulness level in later sessions despite some fluctuation, suggesting the effectiveness of incorporating real-time user feedback in improving the usefulness ratings in later sessions. Such improvements are believed to be stronger if more sessions were offered as the system collected more and more user feedback.

As for the interestingness ratings, the repeated measure t-tests show that there is no significant difference between KB and the baseline group ($t(56) = -0.7560$, $p = 0.7736$) or between AKB and the baseline group ($t(38) = 0.1390$, $p = 0.4451$). That means the surprising articles identified by LuckyFind are on a par with those random PT articles in terms of interestingness. The comparison between KB and AKB is not significant either ($t(423.77) = -0.7535$, $p = 0.7742$). The fluctuation of the interestingness ratings along the sessions is not as large as that of the usefulness ratings. For the AKB approach, a slight improvement has been seen in the later sessions. It is noteworthy that compared to the usefulness ratings, interestingness ratings are higher in general, which was backed up by the follow-up interview finding that most users thought those surprising news more interesting than being useful.

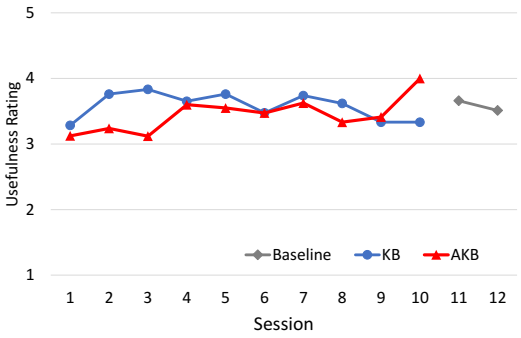

Figure 6: The average usefulness rating in each session by different variations

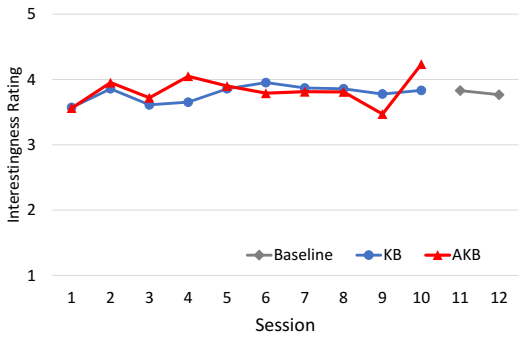

Figure 7: The average interestingness rating in each session by different variations

6.3 Which Approach Better Stimulates User Curiosity

RQ3 is about user curiosity. As mentioned in Section 2 Related Work, users' curiosity state has also been affected by their C-Trait [5]. Our hypothesis is that users possessing more C-Trait will experience a wider range of situations as curiosity-arousing than do users possessing less C-Trait. To test this hypothesis, C-Trait information was collected using the C-Trait questionnaire with 20 questions before a user using LuckyFind. Each question was presented on a 4-point Likert scale, and therefore the total C-Trait score for each individual is 80 (4 x 20). C-State information was collected after the user experienced LuckyFind via another 20 questions, each also on a 4-point Likert scale.

As the result, the average of the C-Trait scores is 68.0 out of 80. We have selected ten users with the highest C-Trait scores (ranging from 73 to 75) and ten users with the lowest C-Trait scores (ranging from 51 to 61) and compared the two groups in terms of their average C-State scores on each question. The comparison was also broken down by the KB and AKB variations. The result is presented in Figure 8. As shown in this figure, AKB has higher C-State levels than KB no matter which C-Trait group the user belonged to, and the difference is significant ($F(1,16) = 4.8908$, $p = 0.0419$). The finding means converging the surprise to the satisfying surprise via the adaptive approach helps stimulating a high level of curiosity. As we hypothesized, the high C-Trait group has obtained significantly higher C-State ratings than the low C-Trait group ($F(1,16) = 5.1150$, $p = 0.038$).

6.4 User Preference Change

RQ 4 concerns whether users' topic preference has changed, hopefully broadened, after using LuckyFind. The preference change was inferred by comparing the user profiles before and after the recommendation sessions. As the result, 19 out of the 30 users have added topics to their profiles, suggesting a substantial percentage of users who have broadened their preferences. Out of the 19 users, 10 were from the KB group and the other 9 from the AKB group. In total, the 19 users added 49 topics, of which 26 were after using the KB approach and 23 after the AKB approach. To check the potential reason for these added topics, the ratings on the articles with

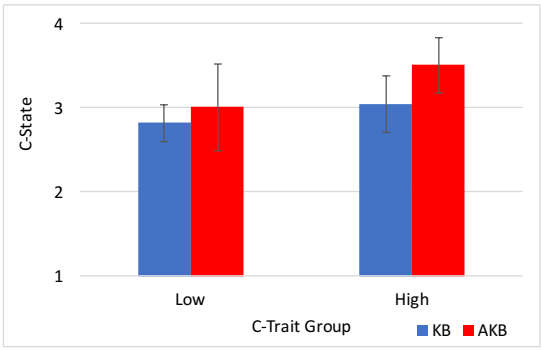

Figure 8: C-State by different approaches and different C-Trait groups

those added topics were analyzed. It was found that 69% ratings on surprise were positive (either 4 or 5 out of 5); 78% ratings on usefulness and 87% ratings on interestingness were positive.

For those topics that stayed in both pre- and post-study profiles, users were also asked to update their preference levels after using LuckyFind. As the result, 41 topics have received increased levels of preference. That means other than helping people discover new interesting areas, LuckyFind also helps them strengthen the interest levels of existing profile topics.

We have also found 4 users who have removed topics from their profile and 4 users who reduced the level of preference for the existing PTs. Digging into those topics that were removed or reduced the preference level, their ratings on surprise, usefulness, and interestinness were all significantly lower (all p-values smaller than .05) than the ratings for the topics that remained the same preference level before and after using LuckyFind. The reasons could be found in the follow-up interviews: people may realize the topic they selected were actually not what they expected after reading articles from that topic.

6.5 Follow-Up Interview: What Users Say

Upon finishing using LuckyFind, participants were asked about whether they encountered surprising articles, whether they were satisfied with those surprising pieces, and whether, why or why not they thought LuckyFind was curiosity-inspiring.

Whether the users encountered surprising contents: All participants indicated that they came across some surprising articles at varying degrees. When asked to give examples of surprising articles, one participant cited a piece of news about training three dogs to sniff out the odorants that indicate a woman has ovarian cancer. The news was "shocking" to him. One participant stated "I read an article about the link between Type 1 diabetes and celiac disease. They found certain DNA that links these two diseases, and I didn't think that the two diseases have any connection." Another participant stated that she read a suggestion on personalizing medication dosage based on weight. For example, for a overweight person, the antibiotics dosage should be more, "which was surprising to me because all antibiotics on market are the same dosage as long as you are an adult, but it (the news) makes sense." Still another participant cited that an article that talked about a type of medicine for schizophrenia could actually reduces the chances of having cancer. The majority of these examples suggested some kind of rare or special co-occurrence or connection between two medical topics, like Type 1 diabetes and the celiac disease, dogs and ovarian cancer, that made the users feel surprising. These examples support the effectiveness of the computational surprise model. However, not every example is so, like the article about the dosage of antibiotics. According to that user, it is contrary to common practice and therefore it was felt surprising, not because of a hidden connection between two things. One user also mentioned that he was surprised to see many articles on swine flu after his topic selection was flu: "Surprised. I realized it is different than what I wanted". The interesting comment made us rethink about our definition of surprise, which could be content-based and intention-based. Apparently, the computational surprise model focuses on the content-based surprise. On the other hand, the intention-based surprise tend to be unpleasant to users, which will try to avoid anyway.

Were they satisfied with the surprising contents: Participants mentioned interestingness more than usefulness when describing their feeling of the surprising contents, as stated by one participant "not that much useful as being interesting". One example is that a participant found an article claimed that over-weight women tend to have stronger sex drive. He said that he thought it should be the opposite, and stated that this was interesting but not that "practically useful" for him. Occasionally, a few (4 out of 30) participants mentioned they found some articles somewhat useful. Those articles were mostly close to their daily life. For example, one participant mentioned an article about eye health. In this article they claimed that men should not wear tight neckties since it might increase the chances of developing glaucoma, and some other serious eye diseases. The participant believed the article informed what he should avoid in future and therefore was useful. Also, another participant cited an article about pollution causing problems in people's brain, and that was useful to him since it would make him more aware of the problem of air pollution and try to stay away from the polluted areas.

Two Participant mentioned that they did not trust some seemingly surprising news, like the article on air pollution in China caused by salt fuel. This article claimed that 20 to 30 millions deaths have been caused by salt fuel pollution in China. The participant thought this number might exaggerate the fact and he questioned the source. Another participant mentioned that an article on a type of antioxidant was in fact advertising for a cosmetic product. Although the content was surprising, she had doubts on the credibility of the content.

Was LuckyFind curiosity-inspiring: Half (15 out of 30) of the participants said yes to the question that whether they feel the content of the news has inspired their curiosity. Their answers include: "I will read more of those articles instead of watching Netflix"; "It encourages me to read more to get more information."; "In general I feel like I need to be more educated on those topics than just looking at people's opinions in some of these articles"; and "There were a lot of stuff that I definitely want to go online and look into the source and see more about." Also, one participant cited an article about ADHD which he used to have when he was young. He said that the article has peaked his curiosity to know more especially when it attributed ADHD partly to genetics. He

was also more interested in the topic of genetics than before to learn more about its association with ADHD. Therefore he added *genetics* to his profile after using LuckyFind.

The other half of the participants indicated their hesitance. One participant said: "for some research and scientific topics, I did not really need to know as much as what a simple Google search gives me... I am not much into those clinical trial articles". Another participant stated: "I would say out of the twelve articles I read, there were only maybe three that I wish I would kept reading about, but not the majority".

About users' preference change: Several participants mentioned why they have changed the topic preference after using LuckyFind. The reasons mentioned for adding a topic are that the recommendations have helped them find interesting topics they were not aware of, as in the example of the user who has added *genetics* after learning the association between *genetics* and *ADHD*.

One participant mentioned that he removed some topics because he was more interested in "a subset of the topic" and the topic "was giving a broad range." He was interested in the dietary effects of cholesterol and the recommendations were about every aspect of cholesterol.

7 CONCLUSION AND FUTURE WORK

This study leveraged computational surprise to address two user-centered objectives for a health news recommender system: user satisfaction and curiosity. Surprise has been implemented computationally using an existing model, resulting in two variations of a recommender system called LuckyFind. The finding shows that the computational surprise model is able to identify surprising contents at little cost of user satisfaction. The cost could be mitigated by adaptively incorporating user real-time feedback. As for the curiosity-inspiring objective, the AKB approach has stimulated a higher level of curiosity than KB, after adjusting for different levels of C-Trait people have before the experiment. Over half of the users have changed their preferences after using LuckyFind, either discovering new areas, reinforcing their existing interests, or stopping following those they did not want anymore.

This study offers new insights for research on recommender systems not to only focus on algorithmic accuracy, but also on those user-centered objectives via deep understanding of both users and recommended items. Practically, for recommender systems developers, this study contributes a feasible implementation that aims for getting people out of the information bubble by promoting information that is not always too obvious and a little outside their "comfort zone". Additionally, for the human-centered design researchers, this study demonstrated the value of incorporating surprise and real-time user feedback for improving user satisfaction and inspiring curiosity.

This study is the first step in investigating the potential of our model of surprise to inspire curiosity. We plan to investigate other models of surprise with the assistance of recent advancements in deep learning, word embedding, and natural language processing techniques with the hope of going deeper into the contents of text-based information beyond metadata (topics). Personalization of surprise is also our future plan. Beyond an interactive process, we will also leverage the established content-based or collaborative filtering recommender algorithms, and turn them into a computational model of relevance to be combined with the surprise model, with the hope of improving the user satisfaction and curiosity. The relative weights on how to combine the two components will be tested empirically. In addition, the adaptive approach in this study is preliminary. Future work includes more sophisticated ways to penalize the articles that failed to satisfy users.

ACKNOWLEDGMENTS

This research is supported by National Science Foundation (NSF) (Award #1910696). We would like to thank NSF to make this research possible.

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] Naresh Kumar Agarwal. 2015. Towards a definition of serendipity in information behaviour. *Information research: an international electronic journal* 20, 3 (2015), n3.
[3] Paul André, Jaime Teevan, and Susan T Dumais. 2009. From x-rays to silly putty via Uranus: serendipity and its role in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2033–2036.
[4] 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.
[5] Daniel E Berlyne. 1966. Curiosity and exploration. *Science* 153, 3731 (1966), 25–33.
[6] Simone Borsci, Jasna Kuljis, Julie Barnett, and Leandro Pecchia. 2016. Beyond the user preferences: Aligning the prototype design to the users' expectations. *Human Factors and Ergonomics in Manufacturing & Service Industries* 26, 1 (2016), 16–39.
[7] John P Chin, Virginia A Diehl, and Kent L Norman. 1988. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 213–218.
[8] Valentin Dragoi, Jitendra Sharma, Earl K Miller, and Mriganka Sur. 2002. Dynamics of neuronal sensitivity in visual cortex and local feature discrimination. *Nature neuroscience* 5, 9 (2002), 883.
[9] Xiangyu Fan and Xi Niu. 2018. Implementing and Evaluating Serendipity in Delivering Personalized Health Information. *ACM Transactions on Management Information Systems (TMIS)* 9, 2 (2018), 7.
[10] Meadhbh Foster and Mark T Keane. 2013. Surprise: Youve got some explaining to do. *arXiv preprint arXiv:1308.2236* (2013).
[11] Kazjon Grace, Mary Lou Maher, Douglas Fisher, and Katherine Brady. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3, 3-4 (2015), 125–147.
[12] Kazjon Grace, Mary Lou Maher, 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*.
[13] 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.
[14] 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.
[15] Marc Hassenzahl. 2018. The thing and I: understanding the relationship between user and product. In *Funology 2*. Springer, 301–313.
[16] Marc Hassenzahl and Noam Tractinsky. 2006. User experience-a research agenda. *Behaviour & information technology* 25, 2 (2006), 91–97.
[17] Marc Hassenzahl, Annika Wiklund-Engblom, Anette Bengs, Susanne Hägglund, and Sarah Diefenbach. 2015. Experience-oriented and product-oriented evaluation: psychological need fulfillment, positive affect, and product perception. *International journal of human-computer interaction* 31, 8 (2015), 530–544.
[18] Laurent Itti and Pierre F Baldi. 2006. Bayesian surprise attracts human attention. In *Advances in neural information processing systems*. 547–554.
[19] Marius Kaminskas and Derek Bridge. 2016. 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 (2016), 1–42.

[20] Sangwon Lee and Richard J Koubek. 2012. Users' perceptions of usability and aesthetics as criteria of pre-and post-use preferences. *European Journal of Industrial Engineering* 6, 1 (2012), 87–117.

[21] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, Guojun Wang, and Kenli Li. 2020. Directional and Explainable Serendipity Recommendation. In *Proceedings of The Web Conference 2020*. 122–132.

[22] Gitte Lindgaard and Cathy Dudek. 2003. What is this evasive beast we call user satisfaction? *Interacting with computers* 15, 3 (2003), 429–452.

[23] Jordan A Litman and Charles D Spielberger. 2003. Measuring epistemic curiosity and its diversive and specific components. *Journal of personality assessment* 80, 1 (2003), 75–86.

[24] Irene Lopatovska and Hartmut B Mokros. 2008. Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts. *Information processing & management* 44, 1 (2008), 92–104.

[25] Luís Macedo and Amílcar Cardoso. 1999. Towards artificial forms of surprise and curiosity. In *Proceedings of the European Conference on Cognitive Science, S. Bagnara, Ed.* Citeseer, 139–144.

[26] 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.

[27] Luís Macedo and Amílcar Cardoso. 2005. The role of surprise, curiosity and hunger on exploration of unknown environments populated with entities. In *Proceedings of the Portuguese Conference on Artificial Intelligence*. 47–53.

[28] Stephann Makri, Elaine Toms, Lori McCay-Peet, and Ann Blandford. 2011. Encouraging serendipity in interactive systems.

[29] Lori McCay-Peet and Elaine Toms. 2011. Measuring the dimensions of serendipity in digital environments. *Information Research: An International Electronic Journal* 16, 3 (2011), n3.

[30] Dana McKay, Stephann Makri, Shanton Chang, and George Buchanan. 2020. On Birthing Dancing Stars: The Need for Bounded Chaos in Information Interaction. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 292–302.

[31] Robert K Merton and Elinor Barber. 2011. *The travels and adventures of serendipity: A study in sociological semantics and the sociology of science*. Princeton University Press.

[32] 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.

[33] Frank D Naylor. 1981. A state-trait curiosity inventory. *Australian Psychologist* 16, 2 (1981), 172–183.

[34] Xi Niu. 2018. An Adaptive Recommender System for Computational Serendipity. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 215–218.

[35] 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.

[36] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.

[37] Bruno A Olshausen and David J Field. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 6583 (1996), 607.

[38] Gaurav Pandey, Denis Kotkov, and Alexander Semenov. 2018. Recommending serendipitous items using transfer learning. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1771–1774.

[39] Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* 2, 1 (1999), 79.

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

[41] J Sauro. 2011. Does prior experience affect perceptions of usability. *Retrieved December* 20 (2011), I1.

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

[43] 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.

[44] 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.

[45] Noam Tractinsky. 1997. Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM, 115–122.

[46] 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.

[47] 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.