Response to Editor’s and Reviewers’ Comments

We sincerely thank the editor and reviewers for their effort in reviewing our manuscript and for the valuable comments. In this document, we respond to each of the comments given by the editor and reviewers.

Editor:

Q. “Why only 2 women were included in the study?”

We recruited the subjects by advertising through mailing lists and Twitter. As we have never contacted each candidate directly in person, we do not know how many men and women were advertised to participate in the experiment. All we can say is that two female subjects participated in the experiment.

Q. Where are the subjects located? A country break down is needed.

We have added the location information as a table (Table 2) to the section “subjects”, which shows the number of subjects with respect to each country. The location is determined based on the Twitter handles of the subjects. We manually checked each handle and identified the countries.

Reviewer 1:

Q. It would be better to bring the argument regarding researchers’ private interests page 9 (lines 322 - 327) in the introduction (introduction only discusses tweets on research interests).”

We have edited the introduction and mention the researcher’s private interests in the section “introduction” (lines 50-53).

Q. “The main drawback of the study is the way the experiment was designed. The authors asked users to indicate interesting articles and artificially calculated serendipity (eq. 5) with the recommendation list generated by the primitive recommender system (p. 7, line 275). Meanwhile, the authors could directly ask users if they find a particular article serendipitous. Please indicate explicitly in introduction and abstract that serendipity has been calculated artificially (based on primitive recommender system) and not by asking users..”

We describe how we compute the serendipity in the abstract (lines 20-22) and introduction (lines 71-73) and note that the unexpectedness of recommendations are calculated by using a primitive strategy. In addition, we add the reason why we use the serendipity score (SRDP) and why we did not directly ask users to evaluate the unexpectedness of recommendations

(see the end of the section “evaluation metric”, lines 304-311) as:
We did not directly ask subjects to evaluate the unexpectedness of recommendations, because this is not the scenario in which the recommender system is used. Rather, we were aiming to detect indirectly from the subjects’ responses, if the serendipity feature had an influence on the dependent variables. Furthermore, we wanted to keep the online evaluation as simple as possible. Asking for ‘how surprising’ a recommendation is, increases the complexity of the experiment. Subjects needed to know what a non-surprising recommendation is (in comparison). In addition, the cognitive efforts required to conduct a direct evaluation of unexpectedness is much higher and it is in general difficult for subjects to share the concept of the unexpectedness.

Q. “The way serendipity was calculated in equation 5 (p. 7) does not correspond to the definition of serendipity given in the abstract, as this equation does not take into account novelty of the article (users were not asked about novelty of articles either). Please either change the definition of serendipity used throughout the article or change the way it is calculated.”

We make the language clearer in terms of the use of the term “serendipity”. As we did not evaluate the novelty of recommendations in the article, we do not use the term “novelty”. In the abstract (lines 20-22) and introduction (lines 71-73), we explain that the unexpectedness of recommendations is inferred by the difference from the primitive strategy (i.e., baseline). In particular, Equation 5 takes unexpectedness into account be the |UE|, i.e., the “set of unexpected items that are recommended to a user”.

Q. “The discussion section should include comparison of the obtained results with the results obtained by other studies.”

We add the comparison of the obtained results in the section “discussion”.
Regarding the use of Tweets (lines 353-356), we compare the results with previous studies that use Tweets for user modeling. Previous works where webpages were recommended to users (Chen et al., 2010) and Tweets were recommended based on a user’s own Tweets (Lu et al., 2012) achieved a precision of over 0.5. In contrast, we observed only a small fraction of recommendations based on a user’s Tweets that were rated as interesting by the respective subject.
Regarding the use of diversification (lines 368-373), we confirm the existing literature that showed IA-Select works well to bring unexpected and relevant recommendations in the domains of movie and music.

Q. RQ2 is missing an important reference, which discusses the same research question in the movie domain:
Kotkov, D., Veijalainen, J., & Wang, S. (2018). How does serendipity affect diversity in recommender systems? A serendipity-oriented greedy algorithm. Computing, 1-19."
We add the suggested article to the section "related work" and discuss the proposed algorithm in the context of our work (lines 128-132).

Reviewer 2:
Q. "It lacks quite a bit of a literature review in the introduction section. For example, there are papers that present content-based recommendations such as Science Concierge (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0158423) and User Profile Based Paper Recommendation (https://ijisae.org/IJISAE/article/view/752)."
We add the suggested articles to the introduction (lines 27-29) and related works (lines 86-87). As our focus is on the serendipity of research paper recommender systems rather than content-based recommender systems, we do not mention them too much.

Q. "Adding more details in the experimental section can improve the paper. For example, I would like the author to add one a section on how do they present the recommendation result sequentially to the subject i.e. which strategy from 1 to 12 gets presented first to last."
Please see line 242-246, where we write: "The subjects were informed about the randomised order of the strategies and items on the evaluation page". Thus, the order of the strategies was blind to the subjects and in random order. This corresponds to earlier experimental setups such as our recommender system for economics (Nishioka and Scherp, 2016) and other studies like (Chen et al., 2010).

Q. "My main criticism of the paper is the usage of twitter profiles in the experiment. It can be the case that the users tweet about the publication with relatively short text with and extra of publication's URL. I assume that if all tweet text and content from the tweet URL are included in the recommendation engine, it might give a better result for twitter. This is a little hard to re-experiment but it can be noted in the discussion section."
We count the number of URL per tweet. On average tweets contain 0.52 URLs (SD: 0.59). We do not use contents from URLs in tweets, because it takes computation of user profiles longer. However, we would like to consider this information in the future. We add these descriptions in the section "discussion" (lines 396-399).

Q. "The subjects are well recruited. However, I think that the subject size is a little bit too

small (n = 22) in the experiment given that the task can be finished relatively quickly in less than 10 minutes. The current number of subjects is quite small. If possible, by increasing the number of subjects, it will make the paper much stronger."

We added the paragraph "threats to validity" in the section "discussion" (lines 400-405). We mention the results might be influenced by the number of subjects we recruited. As the review pointed out, it is very difficult to increase the number of subjects. Finding significances with few subjects is harder than with many subjects. However, there are some significances and we measured the effect sizes. We assume that adding more subjects would bring almost no additional insights (lines 402-405).

Q. "The dataset could include concepts extracted from twitter and the author can also visualize using the histogram for the length of the text from Twitter. It is a bit hard to know the distribution of tweets length in all the subjects seeing just average length and its standard deviation."

In the subsection "subjects", we add the histogram regarding the number of words in tweets per subject and the number of words in research papers per subject (lines 283-293). As shown in the histograms (Figures 2 and 3), subjects are divided into those with a small total number of words in their tweets and those with a large total number of words in their tweets. On the other hand, with regard to their research papers, there are a few subjects with a large total number of words. Most subjects have a small total number of words in their research papers because they published only few research papers so far.

Q. "The authors also only explore the publications in the computer science field. I think it would be good if they can clarify in the discussion section that it has to explore more on different fields such as biomedical science (Pubmed/MEDLINE), social science, or economics."

We add a sentence to mention a necessity of experiments in other domains in the section "discussion" (lines 400-402).