The Impact of Semantic Context Cues on the User Acceptance of Tag Recommendations: An Online Study

Dominik Kowald*, Paul Seitlinger**, Tobias Ley**, Elisabeth Lex*

*Know-Center Graz & Graz University of Technology (Austria), **Tallinn University (Estonia)

dkowald@know-center.at,paul.seitlinger@tlu.ee,tley@tlu.ee,elisabeth.lex@tugraz.at

ABSTRACT

In this paper, we present the results of an online study with the aim to shed light on the impact that semantic context cues have on the user acceptance of tag recommendations. Therefore, we conducted a work-integrated social bookmarking scenario with 17 university employees in order to compare the user acceptance of a context-aware tag recommendation algorithm called *3Layers* with the user acceptance of a simple popularity-based baseline. In this scenario, we validated and verified the hypothesis that semantic context cues have a higher impact on the user acceptance of tag recommendations in a *collaborative tagging setting* than in an *individual tagging setting*. With this paper, we contribute to the sparse line of research presenting online recommendation studies.

Keywords. Tag Recommendations; Online Evaluation; 3Layers

1 INTRODUCTION

Tag recommendation algorithms support users in finding descriptive tags for their bookmarked resources. Related research has proposed various ways to calculate tag recommendations (see e.g., [5]) and in our previous work [4], we have shown that the factors of tag usage frequency, recency and semantic context influence the accuracy of tag recommendations. While frequency and recency were found to be important in both individual tagging settings (i.e., narrow folksonomies) and collaborative tagging settings (i.e, broad folksonomies), the semantic context only increased the tag recommendation accuracy in the collaborative tagging setting.

Therefore, it is the aim of this paper to validate these offline evaluation results also in an online tag recommendation study, which enables us to test the following hypothesis: "Semantic context cues have a higher impact on the user acceptance of tag recommendations in a *collaborative tagging setting* than in an *individual tagging setting*". Thus, the contributions of our paper are two-fold: (i) we shed light on the impact that semantic context cues have on the user acceptance of tag recommendations, and (ii) with this paper, we contribute to the sparse line of research presenting results of online tag recommendation studies (see e.g., [2]).

2 METHOD

We constructed a work-integrated bookmarking scenario, in which 17 university employees from the areas of computer science and

Tagging setting	$|T|$	$|Y|$	$|B|$	$|MostPop|$	$|3Layers|$
Individual (narrow)	119	191	53	29	24
Collaborative (broad)	127	262	62	29	33
Full dataset	213	453	115	58	57

Table 1: The data collected by the 17 participants during the four weeks of our online bookmarking study: $|T|$ is the number of distinct tags, $|Y|$ is the number of tag assignments, $|B|$ is the number of bookmarks, $|MostPop|$ is the number of bookmarks supported by the *MostPop* algorithm and $|3Layers|$ is the number of bookmarks supported by the *3Layers* algorithm.

psychology have explored a given topic (i.e., "designing workplaces that inspire people") for a period of four weeks. Each participant was assigned the task to collect topic-relevant resources (i.e., three Web links or documents per week) in the DropBox-like tool *KnowBrain*[1]. Furthermore, they had to annotate these resources using a tagging interface.

In this tagging interface, the participants were not only asked to provide a set of tags but also to select from a predefined set of categories (i.e., "Gamification & Playfulness", "Inspiration sources & techniques", "Collaboration technologies", "Personalization services", "Augmented reality", "Interior design", "Wellbeing & health", and "Socializing"). These category selections were then used as semantic context cues for the calculation of tag recommendations.

To test our hypothesis, the participants were split into two groups, where the participants in the first group conducted the task individually and the participants of the second group conducted the task collaboratively. The collected data for both groups as well as for the full dataset is shown in Table 1. For details on the study design, please see [6].

Tag recommendations. The tagging interface supported the annotation process by showing 7 tag recommendations to the users, which were either calculated via a baseline approach (i.e., *MostPop*) or an algorithm, which incorporates semantic context cues (i.e., *3Layers*). For each bookmark, the recommendation algorithm was chosen by the system at random.

MostPop. The *MostPop* algorithm is a simple tag recommendation approach, which ranks the tags by their overall frequency. Thus, *MostPop* recommends the 7 most frequently used tags.

3Layers. 3Layers builds on MINERVA2, which is a parameter-free formalization of how people make use of context cues (e.g., semantic categories) to (i) search memory for contextually similar episodes (e.g., of other bookmarks with similar category combinations), and (ii) access relevant items, such as tags that have frequently co-varied with these episodes (e.g., popular tags of similar bookmarks) [1]. For implementing this context cue-dependent

---

[1]https://github.com/learning-layers/KnowBrain

(a) Individual tagging setting (narrow)   **(b) Collaborative tagging setting (broad)**   **(c) Full dataset (narrow + broad)**
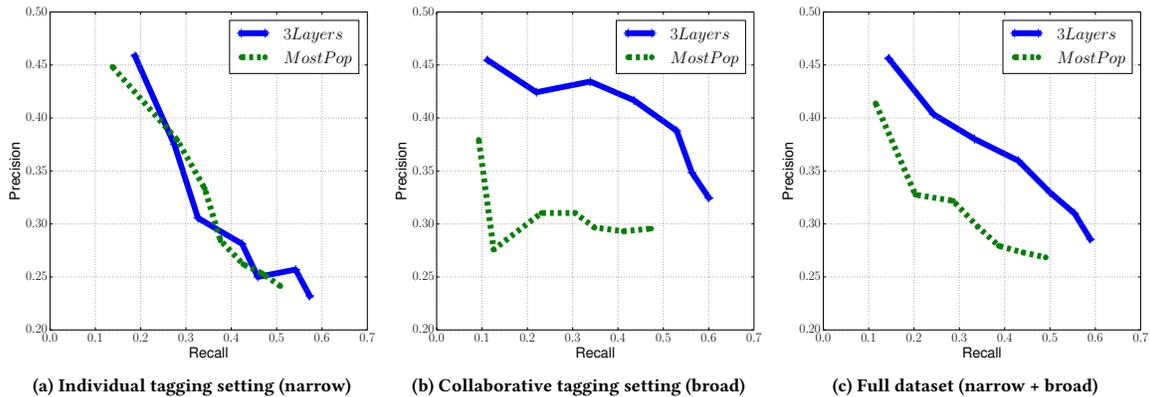
Figure 1: User acceptance rates for the two tagging settings (a and b), and the full dataset (c) measured by Precision and Recall. The results show that semantic context cues have a higher impact on these metrics in the collaborative setting than in the individual one.

search process, we realized a simple ranking principle that weights a tag's usage frequency by the extent to which it has co-occurred with categories that apply to the current resource (see [5]).

User acceptance. We measured the user acceptance by means of Precision and Recall for k = 1 - 7 recommended tags. Hence, for each bookmark, we compared the set of recommended tags with the set of tags the user has actually used in the bookmark. Additionally, we report the F1-score for k = 5 (i.e., $F1@5$) as also used in [3].

3 RESULTS

In Figure 1, we present the results of our online study related to the two tagging settings. Additionally, we present the results for the full dataset in order to relate our results to previous offline studies.

Individual tagging setting. In the individual setting (plot (a) of Figure 1), we see overlapping Precision / Recall plot-lines and thus, cannot observe a significant difference between the user acceptance of *MostPop* and *3Layers*. This is in line with the F1-score for k = 5 since $F1@5$ = .309 for *3Layers* and $F1@5$ = .302 for *MostPop* (p > .5)[2]. These results show that semantic context cues have only a small impact in an individual (i.e., narrow) tagging setting.

Collaborative tagging setting. In the collaborative setting (plot (b) of Figure 1), we identify a significant advantage of *3Layers* with $F1@5$ = .418 over *MostPop* with $F1@5$ = .302 (p < .05). We attribute this to the fact that in this collaborative setting, *3Layers* is better able to exploit the additional information of the semantic context than in the individual tagging setting. These results mirror our previous work since in [4], we showed that the semantic context has a low influence in narrow folksonomies (i.e, in an individual setting) and a high influence in broad folksonomies (i.e., in a collaborative setting).

Full dataset. In the full dataset (plot (c) of Figure 1), *3Layers* with $F1@5$ = .372 significantly outperforms *MostPop* with $F1@5$ = .302 (p < .05). This validates our previous results presented in [5], where we showed that *3Layers* outperforms a context-unaware algorithm (i.e., Latent Dirichlet Allocation – LDA) in an offline study setting, now also in an online study setting.

4 CONCLUSION

In this paper, we presented the results of an online study to identify the impact of semantic context cues on the user acceptance of tag recommendations. Based on our stated hypothesis, our two main findings showed (i) that the impact of semantic context cues greatly depends on the given tagging setting (i.e, individual vs. collaborative), and (ii) that in a collaborative setting, a tag recommendation algorithm, which incorporates semantic context cues, reaches a higher user acceptance than a solely frequency-based algorithm. These findings are also well in line with the results of previous offline tag recommendation evaluation studies [4, 5], which are now verified in an online setting. For future work, we plan to extend our study design by a third tag recommendation algorithm, which incorporates the temporal context of tag assignments (i.e., BLL_{AC} from [4]). We hypothesize, that this time-aware approach positively influences the user acceptance of recommendation in individual tagging settings.

REFERENCES

[1] Douglas L Hintzman. 1984. MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers* 16, 2 (1984), 96–101.

[2] Robert Jäschke, Folke Eisterlehner, Andreas Hotho, and Gerd Stumme. 2009. Testing and Evaluating Tag Recommenders in a Live System. In *Proc. of RecSys'09*. ACM, New York, NY, USA, 369–372.

[3] Dominik Kowald and Elisabeth Lex. 2015. Evaluating tag recommender algorithms in real-world folksonomies: A comparative study. In *Proc. of RecSys'15*. ACM, 265–268.

[4] Dominik Kowald and Elisabeth Lex. 2016. The Influence of Frequency, Recency and Semantic Context on the Reuse of Tags in Social Tagging Systems. In *Proc. of HT'16*. ACM, 237–242.

[5] Paul Seitlinger, Dominik Kowald, Christoph Trattner, and Tobias Ley. 2013. Recommending tags with a model of human categorization. In *Proc. of CIKM'13*. ACM, 2381–2386.

[6] Paul Seitlinger, Tobias Ley, Dominik Kowald, Dieter Theiler, Ilire Hasani-Mavriqi, Sebastian Dennerlein, Elisabeth Lex, and Dietrich Albert. 2017. Balancing the Fluency-Consistency Tradeoff in Collaborative Information Search with a Recommender Approach. *International Journal of Human–Computer Interaction* (2017), 1–19.

---

[2]According to a Wilcoxon rank-sum significance test.