Mojisola Erdt, Christoph Rensing:
Evaluating Recommender Algorithms for Learning using CrowdSourcing.
In: Prooceedings of the 14th IEEE International Conference on Advanced Learning Technologies (ICALT 2014), p. 513-517, CPS, July 2014.

Evaluating Recommender Algorithms for Learning using Crowdsourcing

Mojisola Erdt, Christoph Rensing
Multimedia Communications Lab,
Technische Universität Darmstadt,
Darmstadt, Hesse, Germany
E-mail: mojisola.erdt@kom.tu-darmstadt.de, christoph.rensing@kom.tu-darmstadt.de

***Abstract*—Keeping focused on a certain goal or topic when learning with resources found on the Web is a challenge. Creating a hierarchical learning goal structure with activities and sub-activities can help the learner to keep on track. Moreover, providing useful recommendations to such activities can further support the learner. However, recommendations need to be relevant to the specific goal or activity the learner is currently working on, as well as being novel and diverse to the learner. Such user-centric metrics like novelty and diversity are best measured by asking the users themselves. Nonetheless, conducting user experiments are notoriously time-consuming and access to an adequate amount of users is often very limited. Crowdsourcing offers a means to evaluate TEL recommender algorithms by reaching out to sufficient participants in a shorter time-frame and with less effort. In this paper, a concept for evaluating TEL recommender algorithms using crowdsourcing is presented as well as a repeated proof-of-concept evaluation experiment of a TEL graph-based recommender algorithm *AScore* that exploits hierarchical activity structures. Results from both experiments support the postulated hypotheses, thereby showing that crowdsourcing can be successfully applied to evaluate TEL recommender algorithms.**

***Keywords*—crowdsourcing; TEL; recommender systems; evaluation methods**

I. Introduction

These days, due to the vast amount of information online and the many distractions from social media, it is increasingly difficult to remain focused when learning on the Web. Setting learning goals in a hierarchical activity structure can help the learner keep focused when researching on a specific topic [1]. To further support the learner, recommender algorithms can suggest learning resources fitting the task the learner is currently working on or trying to gain knowledge about. Recommender Systems for TEL (Technology Enhanced Learning) have specific goals that go beyond simply suggesting similar things [2]. TEL recommender algorithms aim to not only provide learning resources relevant to the particular topic or activity the learner is currently focusing on but also novel and diverse learning resources [3]. There is therefore a need to focus on the user-centric evaluation of TEL recommender algorithms, especially when considering user-centric metrics such as novelty and diversity as this can really only be done by asking the users themselves. Offline evaluation methods measuring the accuracy of recommender algorithms on historical or simulated datasets can simply not fill this gap [4], [5]. But, as we all know, user studies are very time-consuming [6] and difficult to conduct multiple times. Gaining access to more than about 30 to 50 users for an experiment is also very challenging. There therefore remains a need for alternative evaluation methods for recommender systems in general [5], [7] and TEL recommender systems in particular [2]. Crowdsourcing offers a fast, repeatable alternative, giving access to sufficient participants in a less time-consuming manner [8].

In this paper, a detailed description of a crowdsourcing concept for evaluating TEL recommender algorithms is presented, as well as a repeated proof-of-concept evaluation experiment evaluating a TEL recommender algorithm *AScore* [9] - a recommender algorithm exploiting hierarchical activity structures to recommend learning resources to learners. Based on experiences gained from a user experiment performed in [10], an initial approach how crowdsourcing can be applied to evaluate TEL recommender algorithms had been investigated in [11] showing already very promising preliminary results. Final results of this experiment (Experiment Spring) are now presented and analysed in this paper and compared to the results of a repeat of the experiment (Experiment Autumn) at a larger scale.

Results from both runs of the experiment: Experiment Spring and Experiment Autumn, support the postulated hypotheses that AScore provides more relevant, novel and diverse recommendations than the state-of-the-art algorithm FolkRank. Furthermore, AScore recommends more relevant, novel and diverse resources to more specific topics in sub-activities lower in the activity hierarchy, thus benefiting learners the more detailed and precise their research on the topic becomes. Experiment Autumn supports the results of the initial Experiment Spring thereby validating the evaluation concept and affirming that neither the choice of activities nor the selected recommendations for the experiments directly influence the results obtained. Hence, these results show that crowdsourcing can be successfully applied as an evaluation method for TEL recommender algorithms.

Related work is presented in Section II where crowdsourcing is introduced as an evaluation method in research and for evaluating recommendr systems. Following this, a brief overview of existing semantic graph-based TEL recommender algorithms is given. In Section III, we present our proposed crowdsourcing evaluation concept, giving a detailed description of the preparation and execution steps. The evaluation results of the proof-of-concept experiments are finally presented in Section IV. We conclude in Section V and give an outlook on future work.

II. Related Work

A. Crowdsourcing as an Evaluation Method in Research

Crowdsourcing is an open call to users from a very large online community to solve a problem or to perform a human intelligent task. Users take part for their personal amusement, to boost their social esteem or for payments [12]. Crowdsourcing gives fast access to a large number of users, at a low cost, offering a relative high level of quality and flexibility [8]. Crowdsourcing tasks however remain sort of artificial, the setting is not very controlled and the users are unknown. Therefore the need to detect spammers or so called *gamers* remains a challenge [8]. In research, crowdsourcing has been used to solve tasks in many different domains e.g. for surveys, usability testing or classification tasks [8]. Recommender strategies have also been evaluated using crowdsourcing to determine the relevance of the recommendations made [13], [14]. In addition to relevance, user-centric metrics such as novelty, redundancy and diversity have also been measured using crowdsourcing where crowdworkers state their preference judgements for news articles to a specified information need or topic [15].

B. Semantic Graph-based TEL Recommender Systems

Graph-based recommender algorithms recommend resources based on the graph structure called *folksonomy* that results from the collaborative tagging of resources [5]. AScore is a graph-based TEL recommender algorithm based on the state-of-the-art algorithm *FolkRank* [16]. AScore extends the *folksonomy graph* [9] with activity nodes and activity hierarchy relations as well as with users belonging to or working on common activities [9]. Activities can be described as goals or tasks defined by the learner in a hierarchy [1]. When researching for a topic on the Web, a learner finds Web resources such as a news article or a blog about the topic. The learner then tags these resources with keywords and attaches them to the respective activities they are relevant to. By creating an activity hierarchy, the learner can better plan and organize his learning goals and learning resources [17]. As other learners do this as well, for example on a common collaborative learning platform such as *CROKODIL* [17], a hierarchical activity structure grows with tagged resources attached to them. Recommendations of such resources found by other learners on similar or related topics can be interesting, new and diverse when looking for learning resources to solve a particular task or activity. Another TEL recommender algorithm based on FolkRank extends the folksonomy as well but in this approach, links are added between tags by considering the taxonomy-based semantic similarity between the tags. The aim here is to increase the density of the links in the graph, thereby improving the quality of graph-based recommendations of learning resources [10]. A further extension of FolkRank has been presented as a TEL recommender algorithm for Personal Learning Environments (PLE)s, where resources are ranked according to the relevance of their tags as well as the user's tag-based attention profile which is generated from the tags the learner and his peers use most often [18].

III. Crowdsourcing Evaluation Concept

The two main steps in the proposed concept for evaluating TEL recommender systems using crowdsourcing are described in detail below, highlighting the challenges met.

A. Preparation Step

In the preparation step shown in Fig. 1, the goals of the experiment are first determined and then hypotheses are specified. In the two evaluation experiments conducted: Experiment Spring and Experiment Autumn, the aim was to evaluate the graph-based recommender algorithm AScore. One goal was to determine if the recommendations made by AScore are more relevant to a specified activity as well as being more novel and diverse to the learner when compared to recommendations made by the baseline algorithm FolkRank. A second goal was to find out if recommendations made by AScore to sub-activities lower down in the activity hierarchy (*A_Sub*) were more relevant, novel and diverse than those made by AScore to activities higher up in the hierarchy (*A_Super*). Hence, the following three hypotheses were defined:

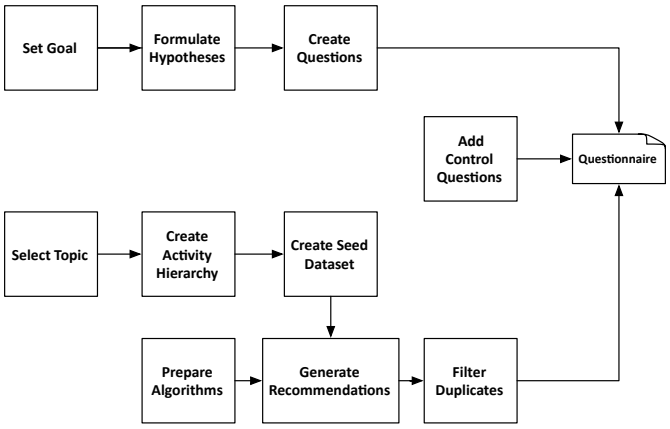


Fig. 1. Crowdsourcing Concept: Preparation Step

Hypothesis 1: Relevance - learning resources recommended by AScore are more relevant to a specified topic than learning resources recommended by FolkRank. AScore recommends more relevant learning resources to sub-activities lower down in the hierarchy (A_Sub) than to activities higher up in the hierarchy (A_Super).
Hypothesis 2: Novelty - learning resources recommended by AScore are more new or unknown to the learner than those recommended by FolkRank. AScore recommends more novel learning resources to sub-activities lower down in the hierarchy (A_Sub) than to activities higher up (A_Super).
Hypothesis 3: Diversity - AScore recommends more diverse learning resources than FolkRank. AScore recommends more diverse learning resources to sub-activities lower down in the hierarchy (A_Sub) than to activities higher up (A_Super).

From these hypotheses, questions are then formulated for the questionnaire as shown in Fig. 2. To measure each hypothesis three questions are created. Each questionnaire contains a total of 10 questions [11]: 3 questions for each hypothesis and one control question to detect gamers. Next, a topic needs to be chosen in order to create an activity structure for the initial research to create a seed dataset to generate recommendations on. The topic needs to be a currently well-known topic so most participants of the survey can understand and better judge the resources recommended to the topic. For these experiments, the topic *"Understanding Climate Change"* was chosen. An initial research on the topic was conducted where 5 experts using the platform CROKODIL [17] created a hierarchical activity

Hypothesis 1: Relevance
Q1: The given Internet resource supports me very well in my research about the topic.
Q2: If I could only use this resource, my research would still be very successful.
Q3: Without this resource just by using my own resources, my research about the given topic would still be very good.
Hypothesis 2: Novelty
Q4: The Internet resource gives me new insights and/ or information for my task.
Q5: I would have found this resource on my own/ anyway/ during my research.
Q6: There are lots of important aspects about the topic described in this resource that lack in other resources.
Hypothesis 3: Diversity
Q7: This Internet resource differs strongly from my other resources.
Q8: This resource informs me comprehensively about my topic.
Q9: This resource covers the whole spectrum of research about the given topic.
Control Questions
Q10a. How many pictures and tables that are relevant to the given research topic does the given resource contain?
Q10b. Give a short summary of the recommended resource above by giving 4 keywords describing its content.
Q10c. Describe the content of the given resource in two sentences.

Fig. 2. Questions for the Questionnaire

structure with 8 activities and tagged about 70 resources found on the Web relating to these activities. The hierarchical activity structure is shown in Fig. 3 with the activities selected for the experiments highlighted. The two recommender algorithms: AScore and FolkRank are then run on this seed dataset or *extended folksonomy* [9] comprising the users, resources, tags, and activities.

Each questionnaire contains questions to only one topic from the activity hierarchy created in the initial research mentioned above - this topic is either a sub-activity or an activity higher up in the hierarchy (a super-activity). To each topic, 5 resources were recommended either from the algorithm AScore or from FolkRank - duplicate recommendations from both algorithms are filtered out. Hence, as shown in Table I, each participant is randomly assigned to one of four treatment conditions: A_Sub where AScore recommendations to a sub-activity are in the questionnaire, A_Super where AScore recommendations are made to a super-activity, F_Sub where FolkRank recommendations to a sub-activity are made or F_Super where recommendations from FolkRank are made to a super-activity. As shown in Table I, a total of 159 participants took part in Experiment Spring. A total of 84 participants received recommendations from AScore and 75 received recommendations from FolkRank. The sub-activity chosen for Experiment Spring was *"Analyze Potential Catastrophes due to Climate Change"* with 84 participants and the super-activity was *"Understanding Climate Change"* having 75 participants (see Fig. 3). In Experiment Autumn, a total of 314 participants took part in the experiment, nearly twice as many as in Experiment Spring. 153 participants were given recommendations from AScore and 161 participants recommendations from FolkRank. In Experiment Autumn, two new activities were chosen: *"Give an overview on the history of Global Warming"* was selected as sub-activity and given to 156 participants and *"Investigate the causes for Climate Change"* as super-activity having 158 participants. Hence, the recommendations generated by AScore and FolkRank were different in both experiments, thus ensuring that the results do not depend on the activities nor the recommendations selected for the experiments.

The main challenge in the preparation step is defining suitable evaluation goals that can be broken down into small compact tasks that are solvable online by crowdworkers, who generally want to accomplish these tasks in a short period of time. It helps to pose simple, short questions to well-defined tasks that can be accomplished in about 15 - 20 minutes.

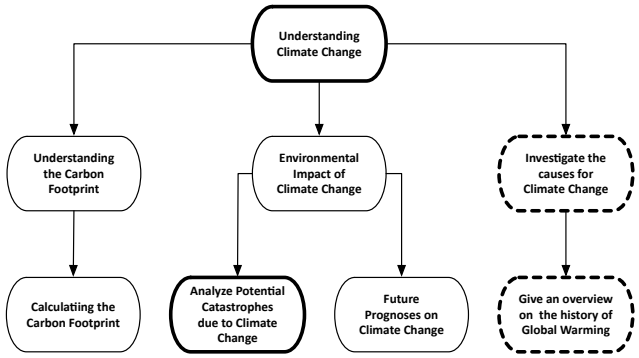


Fig. 3. Hierarchical Activity Structure

TABLE I. RANDOM ASSIGNMENT OF PARTICIPANTS ACROSS TREATMENT CONDITIONS

Experiment Spring			
	Sub-Activity	**Super-Activity**	**Total Participants**
AScore	A_Sub: 45	A_Super: 39	84
FolkRank	F_Sub: 39	F_Super: 36	75
Total Participants	84	75	159
Experiment Autumn			
	Sub-Activity	**Super-Activity**	**Total Participants**
AScore	A_Sub: 80	A_Super: 73	153
FolkRank	F_Sub: 76	F_Super: 85	161
Total	156	158	314

B. Execution Step

The execution step is shown in Fig. 4. The questionnaire is offered as a task to participants on a crowdsourcing platform. In these experiments, the platforms microWorkers[1] and CrowdFlower[2] were used. At the beginning of the experiment, after collecting general information like age, gender, level of education, country and knowledge of the topic, the participants are asked to perform a short research on the Web about the specified topic in order to be able to judge (on a 7-point Likert scale) the relevance, novelty and diversity of the recommended resources later on in the experiment. The amount

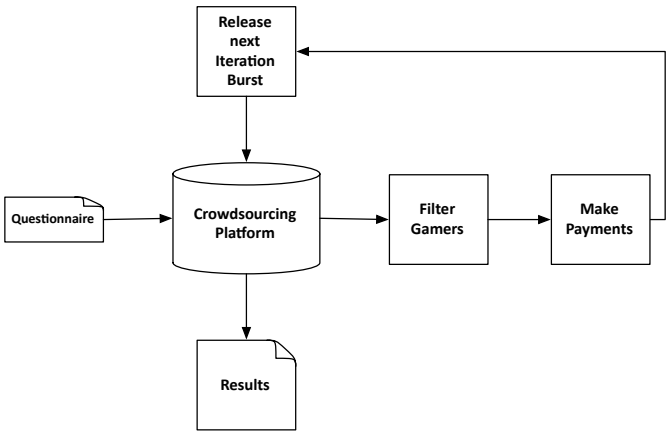


Fig. 4. Crowdsourcing Concept: Execution Step

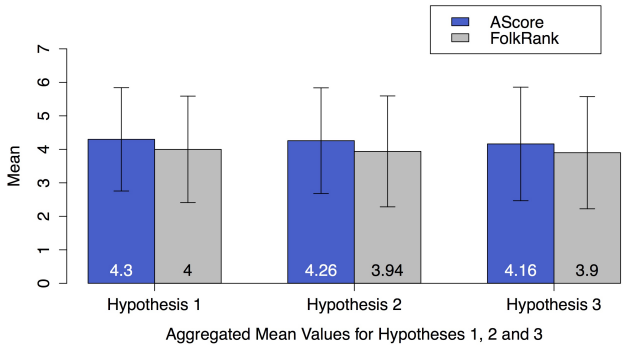
of participants taking part in the experiment in one go are

[1]http://www.microworkers.com (retrieved 27.01.2014)
[2]http://crowdflower.com (retrieved 27.01.2014)

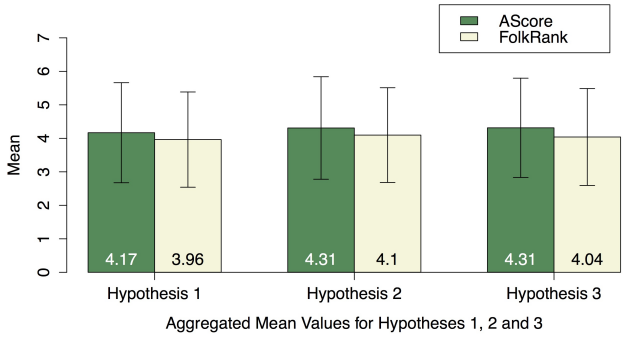
controlled by setting iteration bursts, where a limited amount, like about 50 - 100 participants are allowed to take part in the survey in one iteration. This allows for a better control of the quality of the crowdworkers as their answers need to be cross-checked, gamers identified and reported. Then the next burst of participants are released and so on. One iteration burst is usually completed by crowdworkers within a few hours of being released, however an efficient detection and filtering of gamers poses a major challenge here and posing effective control questions is crucial. After each iteration burst, the valid participants are given free for payments. For these experiments, we offered a payment of 0.75$ for each questionnaire. Finally, after all iteration bursts have been completed, the responses to the questionnaires are extracted from the crowdsourcing platform and the results analysed.

IV. Evaluation Results

The results of Experiment Spring and Experiment Autumn are shown in Fig. 5 (a) and Fig. 5 (b) where AScore is compared to FolkRank. For all three hypotheses: Hypothesis 1: Relevance, Hypothesis 2: Novelty and Hypothesis 3: Diversity, AScore receives higher mean scores than FolkRank. Additionally, for both experiments, as shown in Fig. 6 (a) for Experiment Spring and Fig. 6 (b) for Experiment Autumn, the mean scores for each hypothesis are higher for the sub-activity A_Sub than for the super-activity A_Super. Furthermore,



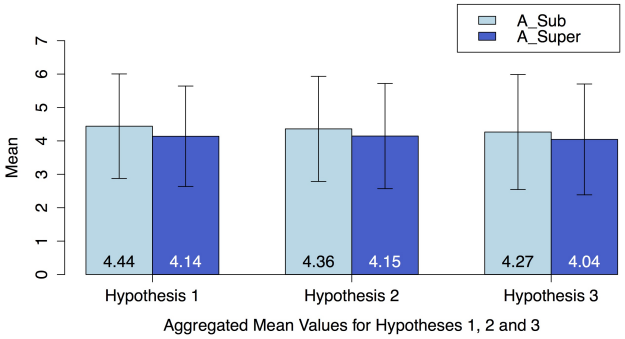
(a) Experiment Spring: AScore and FolkRank



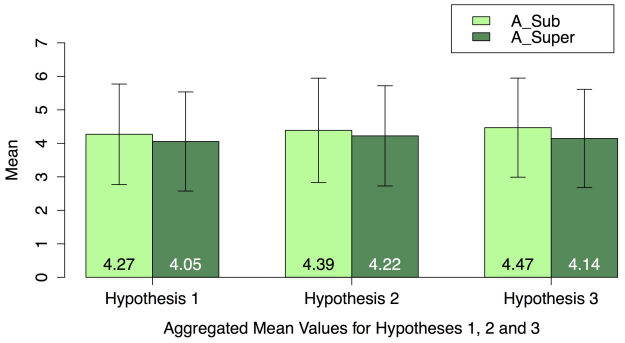
(b) Experiment Autumn: AScore and FolkRank

Fig. 5. Evaluation Results for AScore and FolkRank

independent two samples Student's t-tests [19] were conducted. The results are shown in Table II and Table III giving the mean scores (M), the standard deviation (SD), the t-values (t), the



(a) Experiment Spring: A_Sub and A_Super



(b) Experiment Autumn: A_Sub and A_Super

Fig. 6. Evaluation Results for A_Sub and A_Super

degrees of freedom (df) and the p-values for each algorithm. The threshold for p is taken at 0.05. From the results, it can be inferred that there exists a significant difference in the scores for AScore and FolkRank as $p < 0.05$ for all three hypotheses in both Experiment Spring and Experiment Autumn. These results therefore suggest that algorithm AScore recommends overall more relevant, novel and diverse resources than algorithm FolkRank. Results in Table III further show that there exists a significant difference in the scores for A_Sub and A_Super as $p < 0.05$ for all three hypotheses, for both Experiment Spring and Experiment Autumn. Therefore it can be inferred that AScore recommends more relevant, novel and diverse resources to more specific topics lower down in the activity hierarchy. In contrast and thus supporting the above inferences, the results for F_Sub and F_Super in Table IV do not show significant differences in scores except for Hypothesis 1: Relevance for Experiment Autumn.

In conclusion, the results of both Experiment Spring and Experiment Autumn support all three hypotheses: Hypothesis 1: Relevance, Hypothesis 2: Novelty and Hypothesis 3: Diversity. These repeated proof-of-concept evaluation experiments show that crowdsourcing is indeed a promising evaluation method to evaluate TEL recommender algorithms.

V. Conclusion and Future Work

In this paper, we propose an evaluation concept to evaluate TEL recommender algorithms using crowdsourcing. Results from a repeated proof-of-concept evaluation experiment shows that the algorithm AScore provides more relevant, novel and

TABLE II. EVALUATION RESULTS FOR ASCORE AND FOLKRANK

Experiment Spring						
Hypothesis	**Algo.**	**M**	**SD**	**t**	**df**	**p-value**
1: Relevance	AScore	4.30	1.54	4.65	2367	0.000003578
	FolkRank	4.00	1.59			
2: Novelty	AScore	4.26	1.58	4.82	2367	0.000001531
	FolkRank	3.94	1.66			
3: Diversity	AScore	4.16	1.69	3.78	2367	0.0001618
	FolkRank	3.90	1.67			
Experiment Autumn						
Hypothesis	**Algo.**	**M**	**SD**	**t**	**df**	**p-value**
1: Relevance	AScore	4.17	1.49	4.84	4707	0.000001362
	FolkRank	3.96	1.42			
2: Novelty	AScore	4.31	1.53	4.95	4707	0.0000007654
	FolkRank	4.10	1.41			
3: Diversity	AScore	4.31	1.48	6.42	4705	0.00000000015
	FolkRank	4.04	1.45			

TABLE III. EVALUATION RESULTS FOR A_SUB AND A_SUPER

Experiment Spring						
Hypothesis	**Algo.**	**M**	**SD**	**t**	**df**	**p-value**
1: Relevance	A_Sub	4.44	1.56	3.46	1242	0.0005654
	A_Super	4.14	1.50			
2: Novelty	A_Sub	4.36	1.57	2.40	1242	0.01666
	A_Super	4.15	1.57			
3: Diversity	A_Sub	4.27	1.72	2.30	1243	0.02176
	A_Super	4.04	1.66			
Experiment Autumn						
Hypothesis	**Algo.**	**M**	**SD**	**t**	**df**	**p-value**
1: Relevance	A_Sub	4.27	1.50	3.47	2293	0.0005306
	A_Super	4.05	1.48			
2: Novelty	A_Sub	4.39	1.56	2.58	2293	0.009999
	A_Super	4.22	1.50			
3: Diversity	A_Sub	4.47	1.48	5.26	2290	0.0000001608
	A_Super	4.14	1.47			

TABLE IV. RESULTS FOR F_SUB AND F_SUPER

Experiment Spring						
Hypothesis	**Algo.**	**M**	**SD**	**t**	**df**	**p-value**
1: Relevance	F_Sub	3.95	1.51	-1.03	1123	0.3023
	F_Super	4.05	1.67			
2: Novelty	F_Sub	3.97	1.55	0.64	1077	0.5216
	F_Super	3.91	1.76			
3: Diversity	F_Sub	3.96	1.61	1.27	1122	0.2031
	F_Super	3.83	1.74			
Experiment Autumn						
Hypothesis	**Algo.**	**M**	**SD**	**t**	**df**	**p-value**
1: Relevance	F_Sub	4.04	1.40	2.44	2412	0.01481
	F_Super	3.90	1.44			
2: Novelty	F_Sub	4.11	1.41	0.38	2412	0.7064
	F_Super	4.09	1.42			
3: Diversity	F_Sub	4.07	1.44	1.06	2413	0.2881
	F_Super	4.01	1.45			

diverse recommendations than the state-of-the-art algorithm FolkRank. Additionally, AScore provides more relevant, novel and diverse recommendations to sub-activities than to activities higher up in the hierarchy thereby providing learners with more support the more precise their research becomes. These results show that the proposed crowdsourcing concept can be successfully applied to evaluate TEL recommender algorithms. One limitation of this approach however, is that crowdworkers, being publicly and randomly invited users, may have different motivations as the typical learners using the system. Their judgements may thus differ and the impact of this on the evaluation of recommender algorithms needs to be considered in future work. Furthermore, as the recommender algorithm only makes up a part of a complete recommender system, it remains a challenge to investigate how crowdsourcing could be used to evaluate other aspects of a TEL recommender system, for example, the effects of the presentation of the recommendations or the usefulness of explanations of the recommendations made to the learner.

REFERENCES

[1] C. Rensing, C. Bogner, T. Prescher, R. D. García, and M. Anjorin, "Aufgabenprototypen zur Unterstützung der Selbststeuerung im Ressourcenbasierten Lernen," in *DeLFI 2011 - Die 9. e-Learning Fachtagung Informatik*. Köllen Verlag, 2011.

[2] N. Manouselis, H. Drachsler, K. Verbert, and E. Duval, *Recommender Systems for Learning*. Springer, 2013.

[3] J. Buder and C. Schwind, "Learning with personalized recommender systems: A psychological view," *Computers in Human Behavior*, vol. 28, no. 1, 2012.

[4] M. Chatti, S. Dakova, H. Thus, and U. Schroeder, "Tag-Based Collaborative Filtering Recommendation in Personal Learning Environments," *IEEE Transactions on Learning Technologies*, vol. 6, no. 4, 2013.

[5] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*. Cambridge University Press, 2010.

[6] B. P. Knijnenburg, M. C. Willemsen, and A. Kobsa, "A Pragmatic Procedure to Support the User-centric Evaluation of Recommender Systems," in *Proceedings of the 5th ACM Conference on Recommender Systems*. ACM, 2011.

[7] G. Shani and A. Gunawardana, "Evaluating Recommendation Systems," in *Recommender Systems Handbook*. Springer, 2011.

[8] O. Alonso and R. Baeza-Yates, "Design and Implementation of Relevance Assessments Using Crowdsourcing," in *Advances in Information Retrieval*, vol. 6611. Springer, 2011.

[9] M. Anjorin, T. Rodenhausen, R. D. García, and C. Rensing, "Exploiting Semantic Information for Graph-based Recommendations of Learning Resources," in *21st Century Learning for 21st Century Skills*. Springer, 2012.

[10] M. Migenda, M. Erdt, M. Gutjahr, and C. Rensing, "Semantische Graph-basierte Empfehlungen zur Unterstutzung des Ressourcenbasierten Lernens," in *Proceedings der Pre-Conference Workshops der 11. e-Learning Fachtagung Informatik - DeLFI 2013*. Logos Verlag, 2013.

[11] M. Erdt, F. Jomrich, K. Schüler, and C. Rensing, "Investigating Crowdsourcing as an Evaluation Method for TEL Recommenders," in *Proceedings of ECTEL meets ECSCW 2013, the Workshop on Collaborative Technologies for Working and Learning*, vol. 1047. CEUR-WS, 2013.

[12] G. Kazai, "In Search of Quality in Crowdsourcing for Search Engine Evaluation," in *Advances in Information Retrieval*, vol. 6611. Springer, 2011.

[13] M. Habibi and A. Popescu-Belis, "Using Crowdsourcing to Compare Document Recommendation Strategies for Conversations," in *Workshop on Recommendation Utility Evaluation: Beyond RMSE*, 2012.

[14] R. Kawase, B. Pereira Nunes, and P. Siehndel, "Content-based Movie Recommendation within Learning Contexts," in *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on*. IEEE Computer Society, 2013.

[15] P. Chandar and B. Carterette, "Using preference judgments for novel document retrieval," in *Research and development in IR*. ACM, 2012.

[16] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information Retrieval in Folksonomies: Search and Ranking," in *Proceedings of the 3rd European Semantic Web Conference on the Semantic Web: Research and Applications*. Springer, 2006.

[17] M. Anjorin, C. Rensing, K. Bischoff, C. Bogner, L. Lehmann, A. Reger, N. Faltin, A. Steinacker, A. Lüdemann, and R. Domínguez García, "CROKODIL - A Platform for Collaborative Resource-Based Learning," in *Towards Ubiquitious Learning*. Springer, 2011.

[18] V. Posea and S. Trausan-Matu, "Bringing the Social Semantic Web to the Personal Learning Environment," in *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, 2010.

[19] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R*. SAGE Publications, 2012.