Investigating Crowdsourcing as an Evaluation Method for TEL Recommender Systems

Mojisola Erdt, Florian Jomrich, Katja Schüler, and Christoph Rensing

Multimedia Communications Lab,
Technische Universität Darmstadt, Germany
{erdt,jomrich,schueler,rensing}@kom.tu-darmstadt.de
http://www.kom.tu-darmstadt.de

Abstract. Offline evaluations using historical data offer a fast and repeatable way to evaluate TEL recommender systems. However, this is only possible if historical datasets contain all particular information needed by the recommender algorithm. Another challenge is that users must have indicated interest in the recommended resource in the past for a resource to be evaluated as relevant. This however does not mean the user would not be interested in this newly recommended resource. User experiments help to complement offline evaluations but due to the effort and costs of performing these experiments, very few are conducted. Crowdsourcing is a solution to this challenge as it gives access to sufficient willing users. This paper investigates the evaluation of a graph-based recommender system for TEL using crowdsourcing. Initial results show that crowdsourcing can indeed be used as an evaluation method for TEL recommender systems.

Keywords: recommender systems, evaluation, crowdsourcing

1 Introduction

At the work place, it is increasingly common to learn on-the-job in order to accomplish a certain task or to learn about a new topic needed to solve a particular problem. These days, most of the knowledge is gained from resources found on the Web e.g. from videos on YouTube (www.youtube.com), slides on SlideShare (www.slideshare.net) or forums on LinkedIn (www.linkedin.com). Recommender systems help by suggesting resources fitting the task the person is presently trying to solve or gain knowledge about. Various kinds of recommender systems have been proposed for TEL, each having their particular aims and advantages [7].

A lot of research has gone into the evaluation of TEL recommender systems based on standard methods from information retrieval (IR) which are mostly based on determining the precision of such algorithms using cross-validation on historical or synthetically created datasets. These offline evaluation methods are fast to conduct once the datasets exist and can be repeated and easily compared to other evaluation results [7]. However getting datasets that have exactly the information needed for a specific algorithm remains a challenge. For example, in

order to evaluate our graph-based recommender approach *AScore* [3], a hierarchical activity structure is required. Activities are learning goals or tasks defined by the learner in a hierarchical structure [2]. When the learner finds resources that are needed to achieve a learning goal or to solve a task, he attaches them to the respective activities. Activities thus support the learner during his learning process by helping him plan and organize his tasks and learning resources. AScore exploits these activity structures to recommend learning resources to the learner or to other learners working on related activities. There are however very few datasets that have such hierarchical activity structures [3]. Consequently, the offline evaluation of this approach based on historical data is limited. This motivated us to search for an alternative evaluation method.

Another challenge that arises when evaluating using historical datasets is if new resources are recommended to a user who did not have or know these resources in the past, there is no way of judging if the user would like this resource in the future. There have been attempts to complement offline evaluations by conducting user experiments [7]. However due to the high effort required to perform user experiments not many have been conducted thus far. There therefore exists a gap between the fast, easy-to-conduct offline evaluations and the online experiments. An attempt to bridge this gap is the online evaluation approach using crowdsourcing [5], [4], [1]. Certainly doubts arise regarding the quality of results from an evaluation performed by unknown crowdworkers for a few cents. Experiments however do show that results from crowdsourcing are just as good as from traditional user experiments [6], depending of course on the design of the task to solve [1].

In this paper, we investigate using crowdsourcing to evaluate our TEL recommender system *AScore* comparing it to the state-of-art *FolkRank*. Our goal is to test for relevance, novelty and diversity.

2 Related Work

Crowdsourcing can be described as an open call to online users from a very large community to contribute to solve a problem or to perform a human intelligent task in exchange for payments, social recognition or entertainment [6]. Advantages of crowdsourcing are the fast access to a vast population, the low cost, high quality and flexibility [1]. Limitations are the artificiality of the task, the unknown population and the need for quality control to detect spammers [1]. Crowdsourcing has been used in research to solve various tasks in many different domains e.g. for surveys, usability testing, classification or translation tasks [1]. An example in IR is TERC - Technique for Evaluating Relevance by Crowdsourcing [1], developed to test the effectiveness of IR systems. Recommender strategies have also been evaluated using crowdsourcing [5] to determine the relevance of the recommendations made. Other measures such as novelty, redundancy and diversity have also been measured using crowdsourcing where the crowdworkers state their preference judgements for certain items [4].

3 Crowdsourcing Evaluation Concept and Results

In the crowdsourcing user experiment we investigate these 3 hypotheses:
H1.Relevance: AScore recommends more relevant resources than FolkRank.
H2.Novelty: AScore recommends more unknown or new resources than FolkRank.
H3.Diversity: AScore recommends more diverse resources than FolkRank.

In order to generate recommendations for the experiment, an initial research on the topic of *"Climate Change"* was needed to create a basis graph structure (an extended folksonomy) [3] to run the recommender algorithms on. We selected climate change as this is a topic currently being debated world-wide and it can thus be assumed that the recommended resources to this topic can be understood and evaluated by most participants of the survey. Hence, prior to the experiment, we asked 5 experts using CROKODIL [2] to research for resources on the Web pertaining to specified activities and sub-activities relating to climate change - about 70 resources were found and attached to 8 activities. The graph structure thus created comprising the users, resources, tags, and activities was then used to generate recommendations with the two algorithms AScore and FolkRank. Such a limited dataset would be inadequate for an offline evaluation but it is sufficient to prepare an online user experiment.

In each questionnaire, 5 resources were recommended to the more general activity: *"Understanding Climate Change"* or to the more specific sub-activity *"Analyze the catastrophes which are currently happening or going to happen because of the higher worldwide temperature"*. These resources were either recommended by AScore or FolkRank. To each resource recommended, 10 questions were asked (see Fig. 1): 3 questions to each hypothesis (answered on a 7-point Likert scale) and one control question to help us detect spammers [1]. The participants were asked to first research on the Web for resources relating to the general topic of climate change in order to be able to judge the relevance, novelty and diversity of the recommendations following.

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. 1. Questions asked in the Questionnaire to each Hypothesis and Control Questions

The evaluation jobs were placed on two crowdsourcing platforms: 60 jobs on microWorkers[1] and 40 jobs on CrowdFlower[2]. We had results from all over the world, most of the crowdworkers however came from USA and Bangladesh. After eliminating spammers, we had a total of 68 fully answered questionnaires from paid crowdworkers. We additionally invited 57 voluntary non-crowdworkers (mostly students) to take part in the survey in order to be able to compare the quality of results with those from crowdworkers. In total, 125 fully answered questionnaires were considered for the evaluation. The results of the experiment are shown in Fig. 2. where AScore (left in grey) is compared to FolkRank (right in red). The average answers given on the Likert scale (from 1 - 7) are shown. For each question, AScore receives a better average score than FolkRank. We conducted a two sample Student's t-test for each of the hypotheses. Table 1 gives an overview of the results. Hypothesis 1: Relevance is supported as the t-test gives a p value less than 0.05. This means the answers to questions Q1, Q2 and Q3 support the hypothesis that AScore does recommend more relevant resources than FolkRank. Hypothesis 2: Novelty is supported as well as the p value from the t-test is also less than 0.05, this shows that Q4, Q5, Q6 support the hypothesis that AScore recommends more novel resources than FolkRank . Hypothesis 3: Diversity measured by Q7, Q8 and Q9, is however not supported as the p value is greater than 0.05. Therefore it is not possible to say that AScore recommends more diverse resources than FolkRank. This could be an indication that diversity is harder to evaluate. In conclusion, the results of the experiment support the first two hypotheses: the recommendations made by AScore are more relevant and novel than those recommended by FolkRank.

Table 1. Results of t-Tests

	T-test	Inference
Hypothesis 1: Relevance	$p = 0.0065 < 0.05$	Hypothesis supported
Hypothesis 2: Novelty	$p = 0.0042 < 0.05$	Hypothesis supported
Hypothesis 3: Diversity	$p = 0.0677 > 0.05$	Hypothesis not supported

4 Conclusion and Future Work

In this paper, we argue the need for an alternative evaluation method for TEL recommender systems and propose using crowdsourcing. Initial results show this is possible, concluding that AScore provides more relevant and novel recommendations than FolkRank. We plan to further analyse the data collected to determine the impact of activity hierarchies - comparing the results of recommendations made to a sub-activity with those made to an activity higher in the hierarchy. We hypothesis that recommendations should increase in novelty

[1] `http://www.microworkers.com` (retrieved 19.06.2013)
[2] `http://crowdflower.com` (retrieved 19.06.2013)

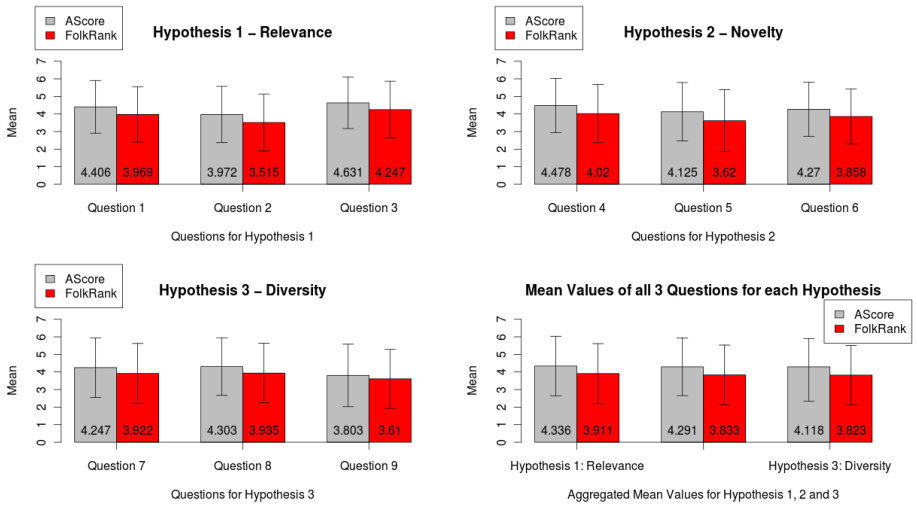

Fig. 2. Hypothesis 1: Relevance (upper left), Hypothesis 2: Novelty (upper right), Hypothesis 3: Diversity (lower left) and All Hypotheses: 1, 2 and 3 (lower right)

the further down the hierarchy. We plan to compare the results between crowdworkers and non-crowdworkers and with these insights improve our proposed crowdsourcing evaluation concept and apply it to further scenarios like evaluating recommendations of learning resources from external sources.

References

1. Alonso, O., Baeza-Yates, R.: Design and implementation of relevance assessments using crowdsourcing. In: Advances in Information Retrieval. LNCS, vol. 6611, pp. 153–164. Springer (2011)
2. Anjorin, M., Rensing, C., Bischoff, K., Bogner, C., Lehmann, L., Reger, A., Faltin, N., Steinacker, A., Lüdemann, A., Domínguez García, R.: CROKODIL - A Platform for Collaborative Resource-Based Learning. In: Towards Ubiquitious Learning. pp. 29–42. LNCS, Springer (2011)
3. Anjorin, M., Rodenhausen, T., García, R.D., Rensing, C.: Exploiting semantic information for graph-based recommendations of learning resources. In: 21st Century Learning for 21st Century Skills, pp. 9–22. Springer (2012)
4. Chandar, P., Carterette, B.: Using preference judgments for novel document retrieval. In: Research and development in IR. pp. 861–870. SIGIR, ACM (2012)
5. Habibi, M., Popescu-Belis, A.: Using crowdsourcing to compare document recommendation strategies for conversations. In: Workshop on Recommendation Utility Evaluation: Beyond RMSE. p. 15 (2012)
6. Kazai, G.: In search of quality in crowdsourcing for search engine evaluation. In: Advances in Information Retrieval. LNCS, vol. 6611, pp. 165–176. Springer (2011)
7. Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., Koper, R.: Recommender Systems in TEL. In: Rec. Sys. Handbook, pp. 387–415. Springer (2011)