Perceived Task Similarities for Task Recommendation in Crowdsourcing Systems

Steffen Schnitzer Svenja Neitzel Sebastian Schmidt

Christoph Rensing

{Steffen.Schnitzer|Svenja.Neitzel|Sebastian.Schmidt|Christoph.Rensing}@kom.tu-darmstadt.de
Multimedia Communications Lab - Technische Universität Darmstadt, Germany

ABSTRACT

Crowdsourcing platforms support the assignment of jobs while relying on the workers' search capabilities. Recommenders can support the workers' decisions to improve quality and outcome for both worker and requester. A precedent study showed, that many workers expect to get tasks recommended, which are similar to previously finished ones. In order to create genuine task recommendation, similarities between tasks have to be identified and analyzed. Therefore, this work provides an empirical study about how workers perceive task similarities. The perceived task similarities may vary between workers with different cultural background and may depend e.g. on the complexity, required action or the requester of the task.

Keywords

Crowdsourcing, Recommender Systems, User Survey

1. INTRODUCTION

Crowdsourcing platforms are used to outsource certain tasks to workers over the internet. Platforms like *Amazon Mechanical Turk*[1] and *Microworkers*[2] allow requesters to publish their work as tasks or campaigns on these platforms, where workers can find the tasks, finish them, and get paid by the requester. Up to now, such micro-task markets rely on the selection capabilities of the workers in order to assign tasks to their workforce.

The high rejection rate this study and other studies report on [9] [7] is an indicator that under-qualified workers get assigned to tasks. Workers who are not motivated to do the work or rather minimize their effort by giving ratings without reading the questions (spamming) are a part of this problem. We also assume, that over-qualified workers are assigned to tasks. This leaves the workers as well as the requesters unsatisfied with the outcomes. Recommender systems, which take the requirements of the tasks and the preferences of the users into account, can support the task selection process and help the workers to find matching tasks more easily. In order to create genuine task recommendation, the workers' preferences and expectation towards such a system must be understood. Previous work has shown, that many workers on a micro-task market expect a recommender to provide similar tasks compared to their recently finished ones [9]. Task similarity is therefore to be seen as a crucial metric to design task recommender systems for crowdsourcing platforms. Therefore, we ran a survey with crowd workers focusing on how they perceive similarities between tasks, yielding qualitative and quantitative results. The results help to understand the requirements which are imposed by the workers towards a task recommendation system and therefore guide the design methodologies of recommender systems for crowdsourcing platforms.

In Section 2 the current state of recommender systems for micro-task markets and the gained insights about workers' task preferences are presented. The design and the execution of the survey are given in Section 3, while the results are presented afterwards in Section 4. Section 5 summarizes the paper at the end and provides ideas about further research.

2. STATE OF THE ART

Crowdsourcing platforms rely on the selection capabilities of the workers and leave the task selection process mostly to the worker. Such platforms provide lists of tasks only supporting the decision with basic filters for category or sorting e.g. for "newest first". Task recommendation can support the workers' decision and help creating appropriate task assignments to improve the quality of the outcome for the requester and for the worker. Recommender systems usually follow two different methodologies. Either they are content-based, relying on the historic user profile to find similar items, or they are collaborative, relying on the preferences of similar users. Hybrid systems are focus of research as well [3].

Several approaches for task recommendation in crowdsourcing systems have been published in recent years, which follow those methodologies. Ambati et al. [1] describe a content-based recommendation system, which relies on a classification of all available tasks based on the history of a worker.

[1] www.mturk.com
[2] www.microworkers.com

WWW '16 April 11–15, 2016, Montreal, Canada
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890087.

This approach uses the bag-of-word scheme to calculate similarities between tasks. Geiger [5] also describes a recommendation system focused on the history of a single worker. He takes the requester and the keywords (tags or categories) of a task into account to calculate a preference estimate. The "TaskRec" approach of Yuen et al. [11] was stepwise developed further into a recommender system, which uses collaborative approaches. Also depending on categories, they include a user-category preference matrix to finally judge the worker's preference towards a certain task. None of these approaches rely on the requirements gathered from the workers but assume that similarity measures based on bag-of-words or keywords suffices to model user behaviour. Therefore, we want to explore the possibilities of more sophisticated similarity measures, based on the task descriptions and taking the needs of the workers into account.
The presented survey indirectly targets the motivation of the worker to choose a certain task. Others have presented insights to the motivation of workers in crowdsourcing already, such as Brabham et al. [2] who discuss motivations to participate in crowdsourcing and detects several different motivators. Kaufmann et al. [8] focus their study on the micro-task market *Amazon Mechanical Turk* and also analyze the motivation of a worker behind preferring one task over another and distinguishing the workers based on their demographic background. Further studies examine the task selection behaviour of workers and their search strategies by analyzing individual worker characteristics or their task processing history [4] [10] [6]. A precedent study showed, that besides money-related criteria, many workers expect to get recommendations based on similarity to the most recently completed task [9]. This survey provides more insights on how this criterion of "similarity" is to be interpreted. Focusing on a task recommendation scheme and a deeper analysis of how workers perceive task similarities is what distinguishes our work from the previously mentioned studies and allows insights on what kind of similarity measures are needed in genuine recommender systems for crowdsourcing platforms.

3. METHODOLOGY

3.1 Survey design

The questionnaire is provided to the workers in a simple structure consisting of four parts shown in Figure 1. At first we motivate the ideas behind task recommendation and explain the necessity of the survey to have the workers understand our goals. We make sure that the workers take their time to read our introduction by not allowing them to move on for exactly one minute. Afterwards, we pose questions about the workers' demographic background and their experience within crowdsourcing platforms. The third page holds the main part with the questions focusing on the opinions of the workers about task similarities. At the end the workers are asked to provide further opinions about task recommendation in general. To detect spammers, some questions in part two and four are used as *consistency questions* [7]. The decision whether to reject a submission or not is mainly based on an introduced test question in the main part of the questionnaire. The introduction of the questionnaire points out, that the questionnaire contains such kinds of attentiveness checks. In the main part of the questionnaire 14 questions (+1 test question in the middle) have to be answered, which follow the same style. The workers are advised to assume they successfully completed a task A and have to determine the similarity of another task B. For each question a certain attribute of the task is pointed out and the workers have to judge the usefulness of the attribute towards determining the similarity. The workers answer by selecting from a Likert scale with five options between "not useful at all" and "very useful". A sample question is depicted in Figure 2. The questions are all given as a statement of similarity between the two tasks which has to be judged. They also include a description with further advice or examples of the mentioned similarity measure. The first five similarity attributes include information which could be received from the task descriptions. They are included to find out, whether more sophisticated similarity measures are needed for task recommendation. The next six similarity attributes handle basic criteria of the tasks related to payment and time. These attributes can easily be received without further analysis of the task description. The last three attributes are actually about the task requester, which can also be of interest in task similarity as shown in [5]. The attributes and the statements are given below, for further details on the descriptions and the test question please refer to the provided detailed description of the survey.[3]

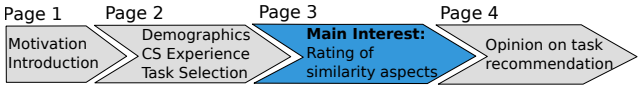

Figure 1: Design of the survey

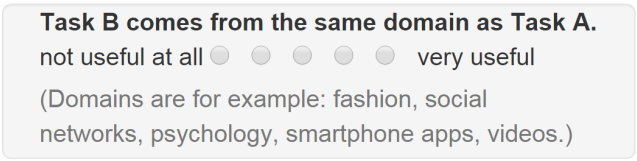

Figure 2: Example question from the main part

domain: The tasks come from the same domain, where domains can be for example social networks or mobile applications.
action: The tasks require the same actions, where actions can be for example writing, searching or voting.
complexity: The tasks have the same complexity, which refers to the requirements needed for successful completion.
comprehensibility: The task's description have the same comprehensibility, which refers to the quality of language or the structure of the instructions.
purpose: The tasks have the same complexity, where the purpose can be for example scientific or commercial.
payment: The tasks have the same payment, which means that a worker is being paid the same amount of money after successfully completion of the tasks.
time: The tasks require the same time for completion.
payment/time: The tasks have the same payment per

[3] `http://www.kom.tu-darmstadt.de/~schnitze/files/msm_www16_survey.pdf`

time, which means that a worker receives the same amount of money per minute of his working time on the tasks.
time to rate: The tasks have the same time to rate, which is the maximal time the employer has to rate the task before it is rated satisfied automatically.
success rate: The tasks have the success rate, which is the ratio of submitted tasks rated as satisfied among all submitted tasks.
nr of open tasks: Task A's and B's campaigns have the same number of open tasks left.
employer experience: Task A's and B's employers are registered for the same time or are equally active on the platform.
employer country: Task A's and B's employers have the same country of residence.
employer type: Task A's and B's employers have an equal type, where type can be for example commercial, scientific or well-known.
In order to gain further insights beyond our predefined questionnaire, the workers are able to provide their remarks about further aspects of similarity in a free text field on the bottom of the same page.
The last part of the questionnaire poses questions about the general acceptance of recommender systems for crowdsourcing platforms and asks for opinions about using task similarity.

3.2 Survey execution

The survey was available on the micro-task market platform *Microworkers* between November 18th and December 4th 2015. The survey was available in English and German. A previous study [9] suggests, that there are different task recommendation preferences depending on the region the workers are coming from. Therefore, the task was made available to workers in five different regions and 100 submissions were gathered for each region. Some of the included countries were chosen due to their importance for the platform in terms of worker count (the top ten countries are included). The in advance defined region of German speaking countries (AT, CH, DE) did not yield enough submissions and was expanded to become the "Europe, West" region. Table 1 shows the five different regions together with their wage for a submission and the amount of submissions per country of residence of the workers (as filled in the questionnaire). Additionally the spam rate is given, which shows how many submissions had to be rejected until enough valid submissions were gathered. For example in the "Europe, East" region, 153 submissions were needed to gather 100 valid ones, which yields a spam rate of 35%. In total, 500 valid submissions were gathered equally divided between the countries to base our results on representative data. For some regions we gathered a few more results but only use the first 100 valid ones for the ease of comparison.
In order to identify spammers, test questions and an attentiveness check were included for quality control, as mentioned before. The attentiveness check was provided together with the other questions in the main part of the questionnaire. The question was obviously nonsense, including non-existent words, while the description stated: "This is an attentiveness check: Please select 'very useful' here.". Still more than 47% of the workers failed to follow this instruction.

Table 1: Submissions per region

Region	Wage	Spam Rate	Residence Country (Code: ISO 3166-1)
Asia, South	\$0.30	63%	BD(77), IN(13), LK(6), NP(3), PK(1)
Asia, South East	\$0.30	52%	ID(34), MY(28), PH(27), VN(6), SG(3), TH(2)
English speaking	\$0.50	35%	US(62), UK(21), CA(13), AU(4)
Europe, East	\$0.40	35%	RS(34), RO(12), MK(12), BA(11), BG(10), HR(7), PL(4), LT(3), AU(2), TR(2), SI(1), HU(1), CZ(1)
Europe, West	\$0.40	42%	IT(28), BE(19), FR(16), PT(14), ES(9), DE(6), FI(3), CH(2), IE(1), DK(1), AT(1)

4. RESULTS

In this section the results from the survey are presented. At first, the answers to the general questions about task recommendation in crowdsourcing platforms are presented in Section 4.1, which sought to show the attitude of the workers towards task recommendation. Section 4.2 presents the collective results of the overall 500 participants, while Section 4.3 describes the differences between the regions in detail. In Section 4.4 other criteria like the experience or the activity of the workers are considered.

4.1 Is task recommendation wanted?

To gather data about the general acceptance and expectations of workers towards task recommendation in crowdsourcing platforms, the workers had to answer two questions. Whether they think it is easy to find enjoyable tasks and whether they want task recommendations on the platform. The first question asks: "Do you think it is easy to find tasks that are interesting and enjoyable to work on?" and was positioned in the second part of the survey. Overall 61.2% said, that it is easy to find enjoyable tasks, while 33.0% answered that it is not easy and 5.8% gave no statement (ns). That means that at least a third of the workers have difficulties and need further support in the task selection process.
Figure 3 shows additionally the voting behaviour for the different regions. Remarkable is the difference of 36 percentage points between "Asia, South" where 78% voted for yes and "Europe, East", where only 42% think it is easy to find enjoyable tasks. The second question, whether workers want task recommendation, was answered positively by 74.6% of all the workers, while the rest voted for 'No'. This question was positioned in the fourth part, right after the different similarities had to be judged and was stated as "Would you like to receive task recommendations on the platform?". Again the voting behavior between the regions shown in Figure 4 shows differences of up to 25 percentage points. In South East Asia, 90% of the workers want to get task recommended, while in English speaking countries still 65% voted for the recommendation of tasks. The workers who voted for "Yes" on this question were asked, whether they want task recommendation based on such similarities

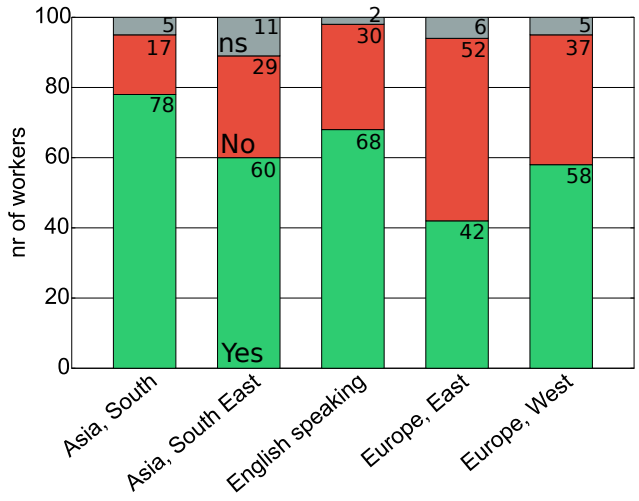

Figure 3: Is it easy to find enjoyable tasks?

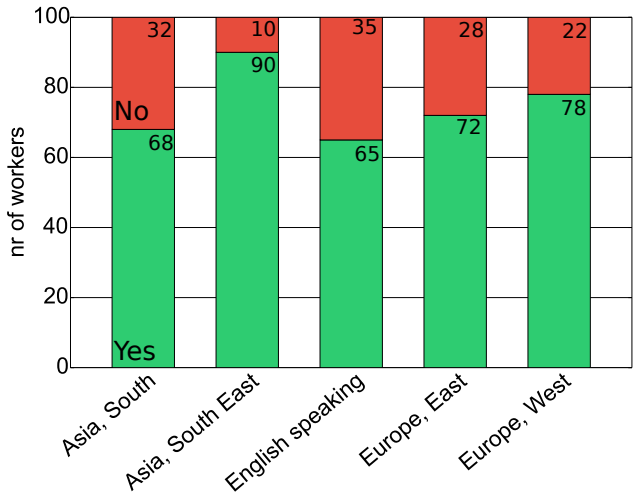

Figure 4: Do you want recommendations?

they had to judge on the page before. Here the positive answers range between 97.2% and 98.9%. These throughout positive attitudes towards task recommendation based on similarities make a strong point, that this needs to be taken into account for genuine task recommendation systems.

4.2 Overall results

The overall impression of the results are shown in Figure 5. The figure shows the ratings of the Likert scale for all the different similarity aspects. The ratings are given between '0' for "not useful at all" and '4' for "very useful". The mean for the weighted rating of each aspect is additionally given in Table 2. The aspect of *action* is obviously ranked the highest and represents the most important similarity feature for workers according to this survey. In general, all the aspects are rated with a positive tendency (with a mean above 2.0) and the *employer country* as an exception. For every aspect (except *employer country*), at least 351 workers and therefore 70% judge the aspect as of having neutral to very

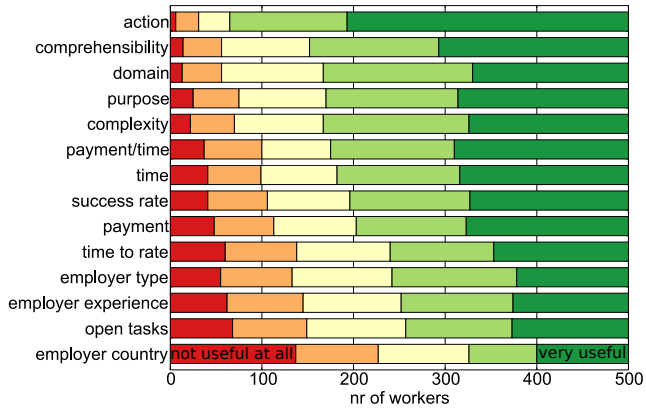

Figure 5: The similarity aspects judged on a Likert scale from "not useful at all" to "very useful" by 500 workers (ordered by average rating).

positive value. For the first 11 of 14 aspects, more than 50% of the workers judge the aspects to be useful or very useful. Figure 5 also reveals that most aspects neighbouring in rank are very close together. However, this picture of indifference is partly dissolved when analyzing the different regions.

The results for judging the different similarity aspects depending on their usefulness are presented in Table 2. For the overall results and the regions the ratings were averaged over all corresponding submissions. In one cell the average rating for the respective similarity aspect is displayed, as well as the rank in the corresponding region in brackets. It is clear to see, that the similarity aspect *action* was voted in the overall analysis with an average rating of 3.41 to the first rank. In "Asia, South" it was voted to the first rank as well with a rating of 3.25. For South Asia one can see, that the aspect *comprehensibility*, which is overall in second position, was voted to the sixth rank with a rating of 2.96. To easily find the first 5 ranked aspects per region, they are given in bold font. As explained in Section 3.1, the different similarity aspects are grouped into the groups of "textual", "basic criteria" and "employer related" aspects. The first five ranks of the overall rating are occupied by the textual similarities. Afterwards, the basic criteria can be found and the employer related ones are found on the last ranks (with exception to the "Nr. of open tasks" aspect). This proves our hypothesis, that sophisticated similarity aspects are relevant for the workers on the platform.

4.3 Results depending on region

The differences between the regions are also quite interesting. Table 2 can be used to find the differences in detail, while only the most important differences found to be significant (rejecting the null hypothesis that the samples come from the same population with at least $p<0.05$) are referred to in the following. The region of South Asia differs the most from the ranking of the others, where *success rate*, *employer type* and *employer experience* are rated to rank on 2,3 and 5 while these aspects come in the overall ranking in places 8, 11 and 12 respectively. Especially the *succes rate* distinguishes Asian regions from the others with a significance value of $p<0.005$. But also the aspects of *open tasks*, *employer experience*, *employer type* and *employer country* follow this pat-

Table 2: Similarity aspects with average rating

Similarity Aspect	Overall Average Rating and (Rank)		Average Rating and Rank by Region									
			Asia, S.		Asia, S. E.		English sp.		Europe, E.		Europe, W.	
action	**3.41**	**(1)**	**3.25**	**(1)**	**3.43**	**(1)**	**3.52**	**(1)**	**3.36**	**(1)**	**3.49**	**(1)**
comprehensibility	**2.97**	**(2)**	2.96	(6)	**3.16**	**(2)**	**3.07**	**(2)**	**2.79**	**(4)**	**2.87**	**(2)**
domain	**2.87**	**(3)**	2.92	(7)	**3.13**	**(3)**	2.73	(7)	**2.87**	**(2)**	**2.69**	**(4)**
purpose	**2.83**	**(4)**	**2.99**	**(4)**	**3.05**	**(5)**	**2.84**	**(4)**	**2.74**	**(5)**	2.54	(6)
complexity	**2.83**	**(5)**	2.74	(13)	2.97	(6)	**2.89**	**(3)**	**2.81**	**(3)**	**2.74**	**(3)**
payment per time	2.76	(6)	2.83	(11)	2.83	(10)	**2.79**	**(5)**	2.73	(6)	**2.60**	**(5)**
time	2.72	(7)	2.87	(8)	2.97	(6)	2.78	(6)	2.62	(8)	2.38	(7)
success rate	2.66	(8)	**3.10**	**(2)**	**3.13**	**(3)**	2.40	(9)	2.47	(9)	2.20	(8)
payment	2.63	(9)	2.84	(10)	2.88	(8)	2.53	(8)	2.71	(7)	2.17	(9)
time to rate	2.42	(10)	2.81	(12)	2.83	(10)	2.26	(10)	2.08	(13)	2.11	(10)
employer type	2.38	(11)	**3.02**	**(3)**	2.70	(13)	2.16	(11)	2.22	(10)	1.82	(11)
employer experience	2.33	(12)	**2.97**	**(5)**	2.84	(9)	1.92	(13)	2.20	(12)	1.74	(12)
nr of open tasks	2.31	(13)	2.87	(8)	2.73	(12)	2.06	(12)	2.22	(10)	1.65	(13)
employer country	1.82	(14)	2.60	(14)	2.37	(14)	1.38	(14)	1.41	(14)	1.34	(14)

tern of difference between Asian regions and the rest. The rating of the similarity aspect *same domain* within the South East Asia region is significantly higher than in the regions of English speaking countries or Western Europe. Also, the *comprehensibility* is valued more by the South East Asian region than by Eastern Europe. For workers from the Asian regions, the *purpose* of the task is a higher ranked similarity aspect than for the region of Western Europe. This holds true for the aspect of *payment* and *time to rate* (even with $p<0.001$). The aspect of *time* is significantly less of interest for workers from Western Europe, than for the Asian and the English speaking regions. It is notable that no significant differences between the regions could be found for the aspects of *action*, *complexity* and *payment per time*.

4.4 Results depending on other criteria

The results were not only analyzed considering different regions but also considering the worker characteristics age, activity, experience and payment. The information about these characteristics came from the platform and were not gathered through the questionnaire. Age was given as date of birth. The date of membership, the overall earnings and the overall number of tasks per worker were used to calculate the other three characteristics. Activity divides the number of tasks by the membership time (more tasks done in less time means a higher activity). The experience is given as the overall number of tasks and the payment was calculated by the overall earnings divided by the overall number of tasks done. For each characteristic the population was splitted by quartiles into four equally sized sub samples described in Table 3. This was done for the whole set of submission, leaving 125 submissions in each part and for each region, leaving 25 submissions in each part.

For the overall analysis there were hardly any differences between the described characteristics. For the characteristic of age we could observe a broadening of the opinions towards the "older" quartiles. That means that the range of means of the similarity aspects changed from (2.09 - 3.38) to (1.48 - 3.5), while the means are more spread throughout the range in the last quartile. When it comes to activity, the more active workers tend to generally rate most of the similarity aspects lower than the less active workers.

Table 3: Quartiles for the different characteristics

quartile	age (years)	activity (task/day)	payment (USD)	experience (tasks)
1	< 23	< 0.450	$< \$0.110$	< 57
2	< 27	< 1.611	$< \$0.138$	< 387
3	< 35	< 4.070	$< \$0.200$	< 1809
4	< 69	< 46.040	$< \$5.274$	< 50322

The worker characteristics were also analyzed between the different regions but there were no results to draw any conclusions from with respect to the rating of similarity aspects.

4.5 Insights from free text comments

A lot of workers emphasize their opinion by mentioning one or more of the already given aspects in the free text answering field. Many workers suggest to recommend tasks based on the skills or qualifications of a worker. Others want to get tasks recommended from requesters they already worked for. Some suggestions also include the possibility of recommending tasks that were done by similar workers or a recommendation based on the popularity among all workers.

5. CONCLUSION AND OUTLOOK

This work motivates and describes the design and execution of a survey to gather the requirements of the workers towards task recommendation in crowdsourcing systems. The results show, that task recommendation would be welcomed by the workers, although to varying extends between different regions of residence. It was also shown, that the perceived task similarity is dependent on many different aspects.

It is remarkable, that all the aspects that were introduced as "textual" occupy the first ranks in the overall analysis. According to the precedent study [9] these similarity based criteria are valued less than money related criteria. Now in this work money and time related similarity aspects are rated lower, showing that such factual task characteristics are required for recommendation, but the characteristics we want to focus on are very relevant as well.

Looking at the big differences between the regions, it is also an interesting result, that the single aspect of *action* was rated to be the most important one throughout all the regions. The significant differences between the regions are also very interesting especially the differences between Asian regions and the others. South East Asia seems to lie in between South Asia and Europe as well as English speaking countries.

All in all the results of this survey show, that sophisticated similarity measures are required for task recommendation in crowdsourcing systems. The workers agree to a certain extend, that semantic similarity aspects like the required action or the domain are more important than factual aspects like time and money. However, this perceived task similarity varies significantly between different world regions, revealing that genuine task recommendation has to be personalized and go beyond the approaches which have been proposed so far. Therefore, we want to examine the possibilities of using semantic similarity features derived from task descriptions in order to build task recommendation schemes. Classifying or clustering the tasks depending on such similarities helps to improve existing task recommendation approaches, which use categories of the platform or employer information.

6. ACKNOWLEDGEMENTS

This work is supported by the Deutsche Forschungsgemeinschaft (DFG) under Grants STE 866/9-1, RE 2593/3-1, in the project "Design und Bewertung neuer Mechanismen für Crowdsourcing".

7. REFERENCES

[1] V. Ambati, S. Vogel, and J. G. Carbonell. Towards Task Recommendation in Micro-Task Markets. In *Human Computation*, pages 1–4, 2011.

[2] D. C. Brabham. Moving the crowd at Threadless: Motivations for participation in a crowdsourcing application. *Information, Communication & Society*, 13(8):1122–1145, 2010.

[3] G. Chartron and G. Kembellec. General Introduction to Recommender Systems. In G. Kembellec, G. Chartron, and I. Saleh, editors, *Recommender Systems*, pages 1–23. John Wiley & Sons, Inc., 2014.

[4] L. B. Chilton, J. J. Horton, R. C. Miller, and S. Azenkot. Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 1–9. ACM, 2010.

[5] D. Geiger. *Personalized Task Recommendation in Crowdsourcing Systems*. Progress in IS. Springer International Publishing, Cham, 2016.

[6] J. K. Goodman, C. E. Cryder, and A. Cheema. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, 26(3):213–224, July 2013.

[7] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *Multimedia, IEEE Transactions on*, 16(2):541–558, 2014.

[8] N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. In *AMCIS*, volume 11, pages 1–11, 2011.

[9] S. Schnitzer, C. Rensing, S. Schmidt, K. Borchert, M. Hirth, and P. Tran-Gia. Demands on task recommendation in crowdsourcing platforms - the worker's perspective. In *ACM RecSys 2015 CrowdRec Workshop*, Vienna, 2015.

[10] M.-C. Yuen, I. King, and K.-S. Leung. Task recommendation in crowdsourcing systems. In *Proceedings of the First International Workshop on Crowdsourcing and Data Mining*, pages 22–26. ACM, 2012.

[11] M.-C. Yuen, I. King, and K.-S. Leung. TaskRec: A Task Recommendation Framework in Crowdsourcing Systems. *Neural Processing Letters*, pages 1–16, 2014.