Demands on task recommendation in crowdsourcing platforms - the worker's perspective

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

Kathrin Borchert Matthias Hirth Phuoc Tran-Gia
{Kathrin.Borchert|Matthias.Hirth|trangia}@informatik.uni-wuerzburg.de
Chair of Communication Networks - University of Würzburg, Germany

ABSTRACT

Crowdsourcing platforms support the assignment of jobs to help requesters in their project completion and allow workers to earn money. Most crowdsourcing platforms apply simple schemes in order to filter the tasks a worker can choose from or rely on the workers' search capabilities. Using genuine task recommendation within such crowdsourcing platforms opens promising opportunities. Such recommendation schemes will only be effective if the workers are confident that they are used towards their own good. In order to gain insights on what kind of recommendations the workers would expect and accept, this work provides an empirical study about the demands of the workers.

Keywords

Crowdsourcing, Recommender Systems, User Survey

1. INTRODUCTION

In crowdsourcing platforms currently the selection of tasks by workers is done based on lists of tasks which can be filtered or sorted using different criteria. Usually a preselection is done so that only tasks which match the worker's competences are shown. To ensure a fast processing of campaigns and to support the selection done by the user, recommendations are considered suitable. In general, two kinds of recommendations can be distinguished: the recommendation of tasks to workers or vice versa workers to a new campaign of an employer. Before developing new recommender systems as part of a crowdsourcing platform it is necessary to know the criteria based on which workers select tasks in existing systems and which kind of recommendations they would prefer for future systems. Designing recommender systems without a clear understanding of workers' behaviour and preferences can result in low acceptance of the implemented recommender system. Therefore, we ran a qualitative and quantitative survey with crowd workers focusing on task selection and task recommendation taking demographic characteristics of the workers into account. This distinguishes our work from the widespread use of crowdsourcing to collect information about items aiming to use this information to recommend the respective items [4]. Our hypothesis that the preferences of workers are inhomogeneous and that criteria which are not available for selection of tasks in current platforms are also relevant have been supported. In Section 2 we summarize the current state of recommender systems in crowdsourcing platforms and existing insights in workers preferences and motivate our survey in detail. Subsequent in Section 3 we describe the design and execution of the study. Section 4 introduces the results. The paper ends with a summary and outlook in Section 5.

CrowdRec 2015 19. Sept., Vienna, Austria

2. RECOMMENDER SYSTEMS FOR CROWDSOURCING PLATFORMS

Existing crowdsourcing platforms rely on the selection of tasks by the worker. Following this pattern the selection of tasks turns into a challenge for the workers [3]. Recommender systems in crowdsourcing platforms should pursue different goals besides the support of the worker's task selection. One central aim is to reduce the time needed for the complete processing of a campaign or to increase the processing quality [9]. Therefore researchers work on the design of recommender systems. Basak et al. [2] present a framework to experiment with recommendation techniques. The framework provides rich information about worker and task properties. Geiger et al. have identified six different approaches for recommendation in crowd processing platforms [5]. All these approaches use knowledge of the tasks completed so far by the workers to generate recommendations. Our overall goal is to use additional information for the calculation of recommendations, e.g. the description of tasks which is a basis for the recommendation of similar as well as different tasks. But for designing a useful recommender system for crowdsourcing platforms, there is also a need to understand how workers select tasks and furthermore what they expect from a recommender system. There are different investigations about the workers' task selection. Schulze et al. investigate which task properties influence worker's task selection and differentiate the results regarding the worker's

demographic background and additional characteristics [8]. Goodman et al. analyzes individual worker characteristics such as their personality and motivation which might also influence task selection strategies [6]. Furthermore, there are different studies which examine how workers search for tasks on crowdsourcing platforms based on an analysis of their task processing and task selection behaviour [3],[9]. Often the task design itself is object of the investigation. Geiger and Schader [5] summarize the findings "that contributors choose tasks according to a complex multidimensional construct of motivational factors, which are weighted different among individuals". Common to the existing investigations is the main orientation along the functional possibilities of the existing platforms and the limitation on platforms mainly used for micro-tasks. In particular, task characteristics and user preferences which might only be implicitly inferred from the task descriptions respective the user profile need to be considered for the goal of better task recommendations [1].

3. METHODOLOGY

The focus of this study is to find out what kind of recommendation workers prefer while performing tasks on a micro-task-market platform. Therefore, a survey was chosen including qualitative as well as quantitative elements to explore different aspects and opinions of the workers.
The overall design of the survey is so that we first introduce the workers to the idea of task recommendation, then ask the workers to choose their most important recommendation criteria and afterwards rank the criteria by importance. The presented recommendation criteria consist of six standard measurements, which consider e.g. payment and time, as well as three criteria which consider similarity and are of interest for our research. Section 3.1 elaborates more on the survey's design.
As a representative set of survey submissions was aimed for, the task was posted openly on the commercial crowdsourcing platform *Microworkers*[1]. To gain insight about different preferences among world regions, the task was posted three times with restrictions to the country, grouping the submissions into Asia, Western (English speaking countries) and Europe. More details about the execution are given in Section 3.2.

3.1 Survey Design

The questionnaire itself was divided into five sections, as shown in Figure 1. On the introduction page, the idea behind recommendations in micro-task-markets was introduced first and the workers were forced to stay on this page for at least 1 minute before being able to move on to the actual questionnaire. This provided enough time for the workers to get familiar with the topic. The first and the last section requested demographic and personal information such as the age, gender and questions about the activity on the platform, including *consistency questions* as a quality assurance measurement as described in [7] to identify spammers. The three main sections in the middle of the survey specifically asked about the type of recommendation preferred by the worker, which was the main interest of the survey. Here the workers were presented with the nine recommendation criteria including six standard criteria like "most money" and our three additional criteria "similar", "different" and "similar worker".
In the first of these sections the worker had to choose the four most important recommendation criteria from a given set of nine. In the second section, the worker had to rank four out of the nine given recommendation criteria according to their importance. The order of the recommendation criteria was randomly changed (also changing between choice and ranking section) to prevent a bias e.g. by workers who are likely to select the first items in a list.
In the third main section the workers had to answer via a free text field the question about their current selection criteria. This open question was placed after revealing the chosen criteria to allow the workers to get more used to the idea of task recommendation first. It also allowed the workers to either come up with additional criteria or chose one of the presented to emphasize their opinion.

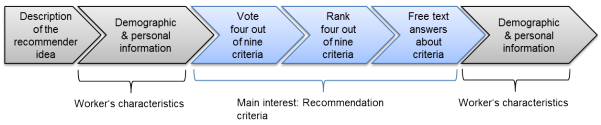

Figure 1: Design of the survey

The following recommendation criteria were given to the worker by description and short explanation as shown below. The first three criteria represent the similarity measurements while the last six criteria represent standard measurements.
similar: Task similar to your previous tasks: This task bears similar characteristics like those you recently completed.
different: Task different to your previous task: A task which is different to the one you recently completed (e.g. different category, to achieve some diversion).
similar worker: Task done by workers similar to you: A task which was completed by workers who completed similar jobs like you.
best requester: Task offered by best rated requester: A task from a requester who was rated by other workers to be the best one.
most money: Task with highest payment: This task offers the highest payment.
least time: Task taking the least time: This task will be completed the fastest.
payment per time: Task with highest payment per time ratio: A task where you get the most money for your invested time.
time to rate: Task with fastest time to rate: With this task you will be paid very soon.
best rated: Task best rated by others: Other workers rated this task to be the best one.
The two sections of choosing and ranking the same nine recommendations were deliberately designed redundant in order to identify spammers and inconsistent answers. From each submission a set of four chosen and as well as a list of four ranked recommendation criteria was retrieved. In this process it was assumed, that the worker should have chosen and ranked the same four recommendation criteria. For each worker the amount of matches between choice and ranking

[1] `http://www.microworkers.com`

Table 1: Different regions, the countries from which submissions were received, the number of submissions and the number of votes after filtering out the spammed votes

region	countries	submissions	votes
Asia	BD, NP, PH	37	104
EU - West	FR, DE, ES, IE, IT, NL, PT, SE	45	133
Western	US, UK, CA, AU	48	140

was counted. Though this yielded less data, each submission with less than two matches was considered spam and not taken into account for data analysis, in order to improve the quality of the data.
Furthermore, for a single submission with two and more matches between choice and ranking, only the actual matches were taken into account for data analysis. Therefore, the submission of a worker can contribute between two and four ranks to the gathered data. That means in a submission with votes for criteria A, B, C, D and ranks for 1:A 2:X 3:Y 4:B only the ranks 1:A and 4:B are taken into account.
For further insights about the survey, we documented the different steps through the sections on our website[2].

3.2 Survey Execution

The crowdsourcing platform *Microworkers* was chosen, because it was possible to have unfiltered access to international workers without stringent pre-selection as seen for *Amazon Mechanical Turk*[3] or *Crowdflower*[4]. The analysis of the survey results is supposed to also run on reliable worker characteristics like activity on the platform and average payment per task and Microworkers provided the most of them. The survey was published via a self implemented system on our own website and the questions were available in English and German. The submissions for the survey were gathered in the mid of April and the mid of May 2015, running about two days overall. There were 151 submissions with 21 identified as spammers and filtering the submissions and ranks as described before, 130 submissions and 377 votes are used for the analysis. The workers were paid $0.25 in the Asia region and $0.50 in Europe and Western regions.
In order to find differences in recommendation preference within the gathered data, several characteristics that were reliably available through the platform, were chosen to be analyzed:
Region The survey was made available separately to workers in the regions defined by *Microworkers*. Table 1 shows the three chosen regions and provides the countries where submissions actually came from.
Gender The characteristic of the gender was retrieved through the questionnaire.
Experience The overall number of tasks done was available for each of the worker and we use this metric in order to analyze the preferences among different experience levels.
Payment The average payment per task is calculated by dividing the overall gained money by the number of the worker's tasks done to the requester's satisfaction.

[2]`http://www.kom.tu-darmstadt.de/~schnitze/files/recsys15_survey.pdf`
[3]`http://www.mturk.com`
[4]`http://www.crowdflower.com`

Table 2: Ouartiles for the different characteristics

	activity		payment		experience	
	task/day	votes	USD	v.	tasks	v.
1	< 0.968	96	$<$ \$0.146	103	< 206	95
2	< 2.712	91	$<$ \$0.182	92	< 914	96
3	< 7.001	97	$<$ \$0.226	95	< 2732	97
4	< 25.122	93	$<$ \$0.642	87	< 35154	89

Activity The activity of the workers is calculated by dividing the number of overall tasks done by the worker by the days of membership in the platform. A higher value stands for a higher activity.

4. RESULTS

As described in Section 3.1, each filtered survey submission provides only those ranks for the recommendation criteria, which have also been chosen by the same worker in the previous step. Therefore, each submission accounts for 2-4 ranked votes for certain recommendation criteria. In order to calculate an overall ranking among the recommendation criteria, the votes are weighted with respect to the chosen rank with weights from 4 to 1, so that a higher rank results in a higher weight. Where unmatched ranks left a gap, the ranks were not moved up. Those weighted votes are then summed up for each recommendation criteria and divided by the overall sum of weighted votes. This *average weighted ranking (**awr**)* provides a relative value between 0 and 1 for each recommendation criteria, where the sum of all values of the criteria sums up to 1.0. As a full submission with the weighted ranks 1-4 consists of ten points for the weighted ranking, and the highest rank for a single submission contributes 4 points, an *awr* of 0.4 for one criteria means that every worker voted this criteria to the highest rank. An *awr* of 0.3, 0.2 or 0.1 on the other hand, means that an average worker ranked the criteria to the second, third or fourth rank respectively.
For the nominal analysis criteria of region and gender the set of votes is naturally divided. For activity, age, experience and payment there is no natural division given and therefore the data for those criteria is split into quartiles. As those criteria are based on numeric attributes of the workers, it was only meaningful to split the set of workers into four equal sized quartiles. As the submission of a worker provides 2-4 votes, the actual number of votes is therefore only roughly equally distributed among the quartiles. Table 2 gives the borders of the quartiles and the number of votes.
One task of the survey was to answer the free text question about the current selection criteria for the worker. From the very different answers of each worker we manually derived categories of similar answers to gather them and count how often they are mentioned throughout the set of workers. Often a worker mentioned more than one of those categories in his answer which increases the count for all of them.

4.1 Overall results

Before going deeper into the analysis of different groups and clusters within the data, the overall picture of the votes distribution is of high interest. Figure 2 shows the overall results for the preferred recommendation criteria while Ta-

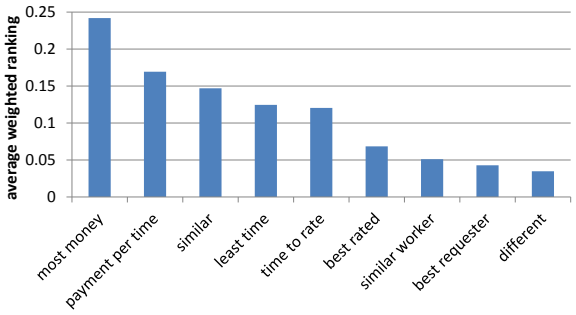

Figure 2: Overall preferred recommendation criteria

Table 3: Overall votes and *awr*

criteria	*awr*	rank				sum
		1	2	3	4	
most money	0.242	37	17	13	12	79
payment per time	0.169	19	20	9	12	60
similar	0.147	18	12	11	14	55
least time	0.124	6	19	17	7	49
time to rate	0.120	5	16	21	8	50
best rated	0.068	5	7	8	10	30
similar worker	0.051	4	7	4	5	20
best requester	0.043	6	1	4	7	18
different	0.035	2	3	6	5	16
sum	1	102	102	93	80	377

ble 3 gives the detailed values of the *awr* and the actual rank vote distribution.
Unsurprisingly the two most wanted criteria are the money-related recommendation criteria of *most money* and *payment per time* where *most money* obviously dominates the whole statistic with an *awr* of 0.242 which means that on average every worker almost ranked this criteria to the rank of 2.5. Besides the money-related criteria coming first and second, there are the time-related criteria *least time* and *time to rate* coming fourth and fifth with *awr* values of 0.124 and 0.120 respectively. From a worker's perspective this focus on money and time as criteria is comprehensible. Those criteria are values which are stored for each of the tasks on a micro-task-market and filtering by one of those criteria is easily possible but should also be considered when designing recommendation systems for micro-task-market platforms.
However, in the third place in our overall results, just between the money-related and the time-related criteria, there is the criterion *similar* with an *awr* of 0.147. A worker who is focused on a certain kind of task probably performs better when being able to repeat this kind of task several times in a row. Also, the worker has put effort into finding a task which fits his requirements and skills and would therefore prefer a task with similar attributes. For a recommendation system this criterion is of high interest as there are many different possibilities of calculating similarities between different tasks. As this criterion is ranked relatively high, it encourages us to further detailed research on design of recommendation systems for crowdsourcing platforms, besides concentrating on the obvious measures of optimizing money- and time-related criteria.

As working on very similar tasks for a certain time can be boresome, we expected the workers to also vote for the criterion *different*, but it got the least votes of all the criteria with an *awr* of 0.035. However, the scenario provided within the survey was not designed to find out whether workers would prefer a change now and then, which is probably necessary for further insights on this criterion.
The criteria *best rated*, *similar worker* and *best requester* all have an *awr* around 0.05, what means that about half of the workers did not consider them worth ranking. On the other hand, about half of the workers must have ranked them at least on rank four, leaving five other criteria behind, what shows that a sophisticated recommendation system should also take them into account.

4.2 Results depending on region

The recommendation criteria preference is very individual and is probably formed by the cultural background of the worker. Therefore, Figure 3 presents the results in dependency of the regions the workers came from and Tabel 1 shows the distribution of submissions and votes between the regions.
The order of the recommendation criterion correlates naturally with the overall results and following the results for the *EU* region throughout the criteria, its *awr* distribution is closest to the overall results. The most significant differences between the regions and the overall result, is found for the criterion of *similar*. For the *Asia* region it returned the smallest *awr* (0.053) while for the *Western* region it returned the highest *awr* (0.227). Also very interesting is, that *Asia* and *EU* agree on the importance of the *most money* criterion, while for the *Western* region it is ranked second behind *similar*. Besides the peak and valley for *similar* in the *Western* and *Asia* region, they also follow mostly the distribution of the overall results. One interesting aspect of the results for the *Asia* region in contrast to the overall results is, that while *similar* is voted so low, the criterion of *different* is voted relatively high (rank six out of nine instead of coming last).

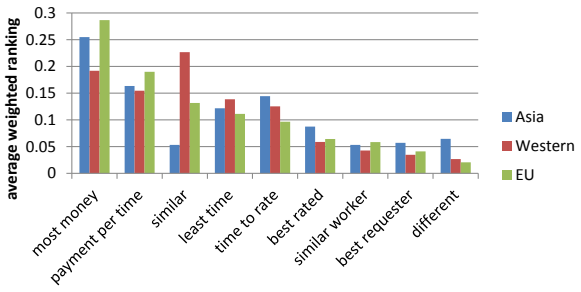

Figure 3: Results with respect to worker's region

4.3 Result in respect to other characteristics

Besides the main results, presented in the previous sections, the survey data was also analyzed with respect to the characteristics of *gender*, *activity*, *age*, *experience* and *payment*. Most of the results here show, that the preferred

Table 4: Amount of votes between genders

gender	amount of votes
female	129
male	248

recommendations are of a very individual nature and that such characteristics bear almost no support to certain conclusions whether a recommendation criteria is preferred or not. Therefore, only the obvious cases, where conclusions can be drawn within this data are mentioned.

Table 4 shows the very imbalanced distribution of male and female workers within the survey. Figure 4 depicts the different preferences between the genders and shows once more, that the *similar* criterion is the most controversial one.

Figure 5 gives the development of recommendation criteria preference along the four quartiles of worker activity. A pattern, which is also seen for the other characteristics, is very clearly depicted by the third quartile in the activity chart. There is the very dominant *most money* criterion, followed by a cluster of the four criteria *payment per time*, *similar*, *least time* and *time to rate* and another cluster of the very low ranked other four criteria. This shows, that the overall results actually give a good impression about the importance of the different criteria to the worker.

In Figure 6 the results for the recommendation preference

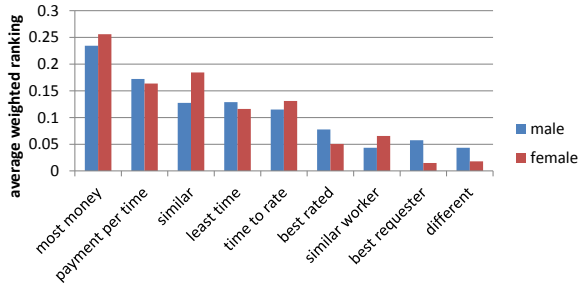

Figure 4: Results with respect to workers' gender

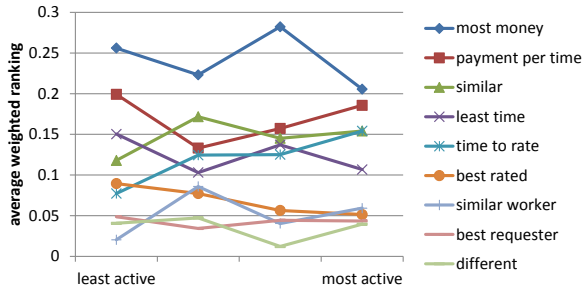

Figure 5: Results with respect to workers' activity

in dependency from the experience is given. It appears, that the clustering found for the activity becomes clearer the more experienced the workers are. Also interesting is the falling preference for *most money* and *best rated* as well as the increasing preference for *time to rate*.

Figure 7 depicts the changes in preferences from a low towards a high average payment. *Similar* and *time to rate* appear to increase together with the payment and surprisingly the *most money* criterion is the lowest for the highest average payment class.

The description of the results in consideration of the age is left out since they adduce no further insight.

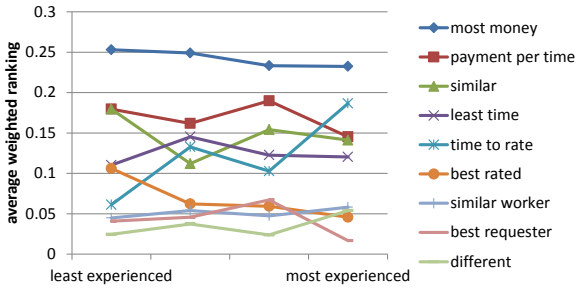

Figure 6: Results w.r.t. workers' experience

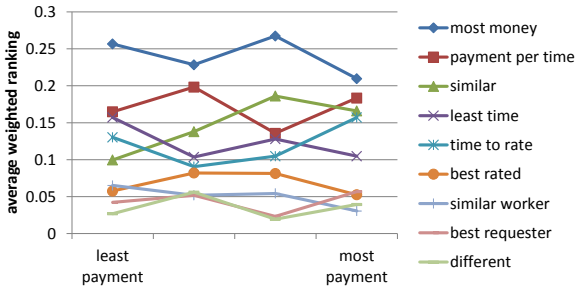

Figure 7: Results w.r.t. workers' average payment

4.4 Qualitative results from free text

The answers from the free text field were manually clustered into criteria with the same meaning. Table 5 shows the ten most mentioned criteria clusters from the answers. Some of the answers reflect the results from the rest of the survey, such as that three out of the four most mentioned categories are related to money and time. Also other criteria that were used within the survey are mentioned freely, like *similar* or *best rated*. Other answers introduce new concepts such as *simplicity*, *category* and *skills*. Simplicity is again a very subjective individual measurement which is not represented in the attribute set of tasks and depends on the skills of the worker, which is also underrepresented in most microtask-marketmodels. Category may be interpreted as similar. Since the categories in the existing platforms are defined by the platform provider and often broad and since similarity goes beyond these categories it needs further investigation. This shows again that more complex recommender systems are required for proper task recommendation in crowdsourcing systems.

Table 5: Free text answer categories

criteria	amount of mentions
payment	36
simplicity	27
payment-time-ratio	24
time	23
category	16
skills	11
similar	8
requester	6
fast rated	5
best rated	4

4.5 Remarks on the results

As seen in the free text part of the results, many workers chose their tasks depending on the category. Therefore, the results of this survey, which was available on the platform within the category "survey" is possibly biased towards such workers, which prefer the category of surveys.
The result showed a large difference for certain criteria between the three regions. The results presented for the other characteristics of the workers, like average payment and experience is presented independent from the region. A survey, which analyses the different characteristics for different regions separately might reveal more conclusive dependencies. Splitting the data of this survey by region and additionally by the characteristics would yield not enough data to expect it to be representative any longer and this was therefore not feasible in our analysis.

5. CONCLUSION AND OUTLOOK

This paper presents the result of a survey, which was performed to gain insights into the workers' preference of recommendation criteria within micro-task-markets. The survey was designed to find qualitative and quantitative answers to these questions. On the one hand, the results show as expected, that workers are focused towards the criteria of time and money. On the other hand, the results show, that less strong criteria like similarity and simplicity are also of high interest for the workers and should be analyzed in a more sophisticated manner. The survey also showed that the criteria preference can vary significantly between regions and other characteristics of the worker. Summarizing, the survey revealed that recommender systems for micro-task-market platforms do not only have to take the usual metrics into account, but also need to dig deeper into characteristics like task similarity and simplicity in order to provide acceptable recommendations for the workers. This encourages us to further reasearch towards our goal of using additional information such as the task description for the recommendation of similar tasks or the origin of the worker.
As mentioned at the end of the results section, more insights can be gained by repeating the survey under different circumstances. Additionally to gathering more data from the different regions in order to create a region-dependent analysis for the worker characteristics, gathering submissions from more than one micro-task-market platform would also reveal, whether and which of those insights can be concluded generally and platform-independent.
The focus of this survey was very specific towards micro-task-markets. However, the preference of recommendation criteria within crowdsourcing platforms in general is of interest. Therefore, further similar surveys focusing on other kind of crowdsourcing platforms could be executed, where we would expect varying results.
As the importance of the recommendation criteria of *similar* and *simplicity* was shown, further research is necessary to drill down which of the task characteristics a worker would prefer to be e.g. similar such as similar description, similar category, similar payment, time, requester, etc..
The given criteria of *different* was voted down in the overall results, probably a survey which is more focused towards the need of variety for the worker will allow better conclusions about the requirement of diverse tasks in micro-task-markets.

6. ACKNOWLEDGMENTS

This work is supported by the Deutsche Forschungsgemeinschaft (DFG) under Grants STE 866/9-1, RE 2593/3-1, HO4770/2-1 and TR257/38-1 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 *Proceedings of The 25th AAAI Workshop in Human Computation*. AAAI Publications, 2011.

[2] D. Basak, B. Loni, and A. Bozzon. A Platform for Task Recommendation in Human Computation. In *RecSys 2014 CrowdRec Workshop*. ACM, 2014.

[3] 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*. ACM, 2010.

[4] A. Felfernig, S. Haas, G. Ninaus, M. Schwarz, T. Ulz, M. Stettinger, K. Isak, M. Jeran, and S. Reiterer. Recturk: Constraint-based Recommendation based on Human Computation. In *RecSys 2014 CrowdRec Workshop*. ACM, 2014.

[5] D. Geiger and M. Schader. Personalized task recommendation in crowdsourcing information systems - Current state of the art. *Decision Support Systems*, 65, 2014.

[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. *IEEE Transactions on Multimedia*, 16, Feb. 2014.

[8] T. Schulze, S. Seedorf, D. Geiger, N. Kaufmann, and M. Schader. Exploring Task Properties in Crowdsourcing - An Empirical Study on Mechanical Turk. *Proceedings of the European Conference on Information Systems (ECIS '11)*, Oct. 2011.

[9] 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*. ACM, 2012.