Self-Generating a Labor Force for Crowdsourcing

Is Worker Confidence a Predictor of Quality?

Julian Jarrett, Larissa Ferreira da Silva, Laerte Mello, Sadallo Andere, Gustavo Cruz
Department of Computer Science
University of Miami
Miami, Florida, USA
{j.jarrett, lcf44, lxm789, sxa845, gho7}@miami.edu

M. Brian Blake
College of Computing and Informatics
Drexel University
Philadelphia, Pennsylvania, USA
MBrian.Blake@drexel.edu

***Abstract* - When leveraging the crowd to perform complex tasks, it is imperative to identify the most effective worker for a particular job. Demographic profiles provided by workers, skill self-assessments by workers, and past performance as captured by employers all represent viable data points available within labor markets. Employers often question the validity of a worker's self-assessment of skills and expertise level when selecting workers in context of other information. More specifically, employers would like to answer the question, "Is worker confidence a predictor of quality?" In this paper, we discuss the state-of-the-art in recommending crowd workers based on assessment information. A major contribution of our work is an architecture, platform, and push/pull process for categorizing and recommending workers based on available self-assessment information. We present a study exploring the validity of skills input by workers in light of their actual performance and other metrics captured by employers. A further contribution of this approach is the extrapolation of a body of workers to describe the nature of the community more broadly. Through experimentation, within the language-processing domain, we demonstrate a new capability of deriving trends that might help future employers to select appropriate workers.**

Keywords—crowdsourcing; recommender systems; human computation; labor markets; recruitment; labor force

I. INTRODUCTION

Labor Markets like Amazon Mechanical Turk [1] and Microworkers [2], often allow workers to autonomously choose human intelligence tasks (HITs) spanning multiple categories from various employers [3][4]. Crowdsourcing, as a paradigm and as an infrastructure, uniquely allows worker biographical information (such as information typically found on resumes), worker past performance (typically captured in employer appraisals and recommendation letters), and job requirements (typically embedded in job descriptions) to be contained in a connected database and machine-interpretable. Based on user and employer inputs, task offerings presented to the worker for selection can be customized to satisfy multiple conditions. In some cases, these are *hard conditions* that cannot change, are difficult to change or cannot be changed in a timely manner. Demographics such as gender, age, place of birth or residence and other constraints qualify as hard conditions. *Softer conditions* are more indicative of worker candidacy by outlining the ideal capability and mastery of required skills and worker interests suitable to perform the tasks at hand. For workers, softer conditions can be met through the acquisition or enhancement of required skills. Previous work shows that worker competence or quality of work is not been consistently guaranteed when workers self-select their tasks indicated [3][4][5]. Consequently, employers question the validity of workers' perception of their own skills when building their skills profiles. *Are there potential measures that labor markets can employ to validate worker profiles and their competences?*

In this paper, we evaluate the state-of-the-art in recommender systems that identify effective crowd workers. Furthermore, through experimentation, we identify the nature of how self-assessment relates to performance. We focus on soft conditions required by employers, primarily worker skills as outlined in their profiles and its validity in light of the worker's actual performance. Augmented by a collaborative filtering approach, we present a fine-grained assessment of workers' skills profiles based on performance feedback from employers. With this assessment, we are able to collectively validate a labor force's competency given their skill profiles and their actual performance. The paper proceeds with an outline of the related literature followed by a comprehensive platform for the crowdsourcing lifecycle and issues with respect to actively building an effective labor force. The preliminary study, findings, discussion, and conclusion finalize the paper.

II. RELATED LITERATURE

A. Task Selection Process

Schulze, Krug and Schader [3], defined the task selection process for crowdsourcing as consisting of 4 main steps. The first of these is the worker's subscription to a labor marketplace or crowdsourcing platform. Influenced by high-level details of tasks, such as the title and compensation, the worker then makes a selection from a list of presented tasks. This list is a subset of all tasks whereas the system filters the tasks (for inclusion or exclusion) based on hard and/or soft conditions [3]. Upon selection, the candidate worker is provided with full details of the task. The candidate worker then opts to work on instances of the task. From this embryonic work experience, the user opts to continue working on more instances of the task, find another task to work on or unsubscribes from the platform.

978-1-4673-9688-2/15 $31.00 © 2015 IEEE
DOI 10.1109/HotWeb.2015.9

B. Worker Profiles and Expertise

The authors in [5] assert that the quality of submissions for crowd-sourced tasks is influenced by a worker's profile. The worker profile consists of a combination of a worker's reputation and expertise. Reputation is seen as a global measure in the community influenced by quality, timeliness of submissions and other metrics as reported through employer evaluation and feedback. It is expected that employers will receive submissions of higher quality from workers with higher reputations. Expertise is inferred from metrics such as credentials and experience. It is dependent on the task at hand and indicates the worker's capability.

C. Job Recommendation Strategies

Recommending jobs to workers extends the more general work for recommending services on a web-based platform [6] [7]. The authors in [3] propose the person-job fit model. This model examines the suitability of workers' abilities and skills to those required by the job. It also evaluates the needs of the worker and the features of a job that satisfies those needs. These needs span human factors including goals, interests, values, payment, supply of tools for the job amongst others.

Authors in [8][9] also used the person-job fit model however as the engine of a CV-recommender. Their motivation comes from application of recommenders in driving the purchase of services, products and marketing on the Internet. Their approach is favored by the increase in electronic CV's and applications. Given these digitized CV's, they propose a bilateral recommender to match people with jobs based on their skills, abilities and individuals which with to collaborate.

There is a line of work that leverages openly available social network information to assist analytics [10]. Lim, Quercia and Finkelstein [11] created StakeSource, a recommender system that works on a friend of a friend type association and aggregates data through social network analysis. It uses stakeholders in a software development project, to make recommendations about other stakeholders to involve in the process. This is done in an attempt to identify all relevant stakeholders to minimize this risk of omitting valuable requirements specification phase of the software development lifecycle.

Paparrizos, Cambazoglu, and Gionis [12] use a machine learning-based job recommender to predict job transitions from the profiles of employees found on the web. The recommender predicts the destination institution and employee's potential job transition given historical data such as past job transitions and other employee demographic information.

Difallah, Demartini and Cudré-Mauroux [4] proposed Pick-a-Crowd, a job recommendation system that utilizes a push methodology as opposed to a first-come-first-serve model. The system is designed with the assumption that workers will perform better on HITs similar to their personal interests. Workers are assigned to HITs suited to their worker profiles. These profiles are built using information from their social network profiles cross-referenced with a taxonomy built from terms from Linked Open Data (LOD) cloud. Once worker profiles are matched to related HITs, they are assigned to workers using category-based, text-based, and graph-based approaches.

The closest comparisons to our work come from the Pick-a-Crowd system [5], the CV-recommender [8][9] and the machine learned job recommender [12]. For [4], our work differs where we use a collaborative filtering based recommender. We pull profiles from social and professional circles on the Internet for recruitment and crowd source viable jobs from labor markets. Based on initial recruitment and the conditions of later job completions, we push jobs to potential workers based on previous performance. In relation to [8][9], our approach evaluates the users' skillsets for crowdsourced tasks based on their job performance history as opposed to skillsets the users profess to possess. We work with the assumption that they had the required skills to perform a given task if they received a score indicative of high quality. With respect to [12], we also use a machine learned job recommender to recommend jobs however [12] uses previous job transitions as a training set while our system satisfactory performance history to similar jobs to the jobs being recommended.

III. Comprehensive Platform for Crowdsourcing Lifecycle

We outline a comprehensive platform for the crowdsourcing lifecycle consisting of 4 major components, the crowd interfaces, employer interfaces, service synchronization and coordination middleware (SSCM) and the integrative internal and external communication channels (Fig. 1). Crowd interfaces consist of specialized interfaces tailored to meet the needs of their respective platforms to include but not limited to professional and social networks, resume repositories and labor markets. Employer interfaces span various types of organizations in need of a proliferation of crowdsourcing services. Furthermore, these services provide answers to tasks where the employer believes the answers exist in the wisdom of the crowd otherwise known as consensus tasks [13].

The SSCM manages external messages from both crowd and employer interfaces as well as the internal messages exchange within its local services. The recruitment manager polls crowd interfaces for profiles and references them locally via a profile bank. Employers push crowd source compatible jobs to the SSCM that internally references them by a job bank. Tasks are assigned through the job allocation manager with contract manager tracking assignments. The solution resolution manager manages all completed tasks and allows employers to accept or reject submissions based on quality, performance or other contractual terms.

A. Recruitment and Retention Management

The community subscription model restricts current labor markets [3], Howe [14] describes it as an open call via the Internet. Comprehensive crowd management requires the efficient recruitment and retention of crowds. This is the primary challenge for the paradigm, and a platform's success hinges on the construction and maintenance of an effective

labor force [15][16]. Many prospective communities for crowdsourcing exist in the form of professional networks (e.g. LinkedIn), social networks (e.g. Facebook), other labor markets (Amazon Mechanical Turk) and crowdsourcing frameworks. Each community's data is tailored for different needs and hence their mechanisms for messaging fundamentally differ. Professional networks explicitly professional credentials, interests and experiences, while social networks may implicitly capture this in social circles and relationships [17]. As opposed to the traditional passive methods for recruitment, we propose a more active approach using an open pull mechanism for recruitment and an open push for retention supported through the SSCM (Fig. 1).

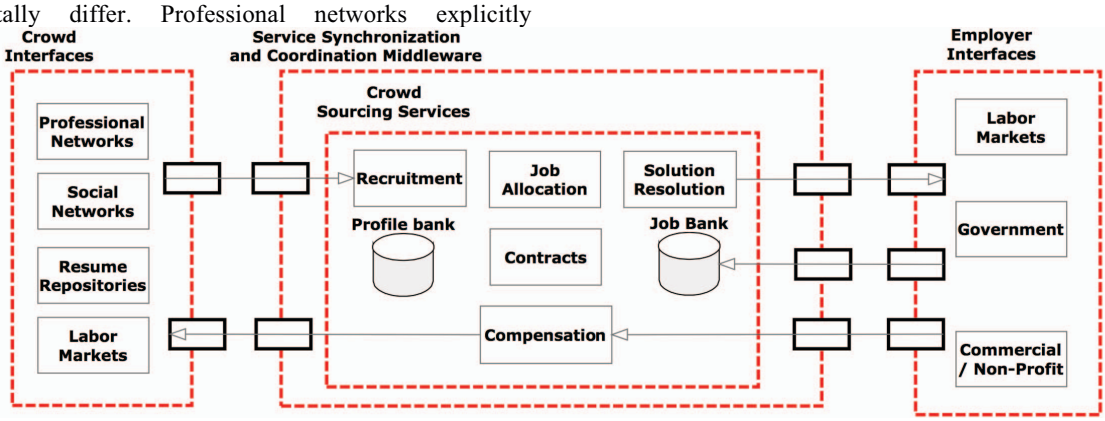

Fig. 1. The Full-Lifecycle Crowdsourcing Platform.

B. Open Pull Mechanism for Recruitment

Given these existing communities and profiles, there exists a need for a service coordination and synchronization middleware, capable of interacting with disparate services and amalgamating profiles into a consistent data structure representing potential workers for assigning tasks. As opposed to the traditional open call, it can be now viewed as an open pull. For this open pull, the middleware queries the existing communities via specialized interfaces, typically some web service or open web API. Upon receiving and converting the data, recruitment management services can employ varying mechanisms or combination thereof to include profile matching, collaborative filtering, case-based reasoning and machine learning [17][18].

C. Open Push Mechanism for Retention

With worker profiles now available from the open pull, and worker history from the upload of jobs, an open push mechanism can be employed to retain the active members of the labor force. Through techniques such as collaborative filtering and recommenders, tasks are now pushed to the worker based on previous task history including performance, skill requirements, difficulty, and similarity amongst other features.

IV. Our Preliminary Study

A. Phenomena Under Investigation

Our preliminary study considers *how to validate worker skills as an indication of their performance*. We seek to investigate the following:

- *Is the workers' self-evaluation of expertise a valid measure for employers to use to determine the actual performance of workers?*
- *What perceived level of workers are most consistent with their actual performance level?*
- *In a labor market, what is the worker trend in opting to do tasks higher than, consistent with or lower than their self-evaluated level and their actual level?*

In this preliminary study, we created a crowdsourcing task consisting of 34 idioms in Portuguese that require translation to English (Fig. 3). The difficulties for the idioms were calculated on a scale of Level 1 to 5, with Level 5 being the highest. They were calculated from the subjective independent evaluation of 4 native Brazilian employers. All 52 workers were native Portuguese speakers of Brazilian decent, some residing in Brazil and others in the United States. Their profiles indicated their respective levels of mastery of the English language in the categories beginner (1), intermediate (2), advanced (3), fluent (4) and native (5) (Fig. 3). All workers were asked to translate a maximum of 5 idioms from the available 34 (Fig. 3). This study extends our previous work [19] by adding a machine-oriented user to perform all translations in the form of *Google Translate* (Portuguese-to-English) for a total of 53 users for evaluation. The average rating of the translations for all users were calculated, again from the subjective, independent evaluation of the 4 employers. Using this information, we applied a collaborative filtering algorithm to recommend $N \leq 10$ translations (R) that were most suited to the users' performance in previously completed translations that had at least 70% similarity to those recommended.

B. Calculations and Analysis for Self Assessment

For each worker we have 2 primary calculations, their performance (P) and the worker's self-perception index (SPI). P (1) for a given worker is calculated as the average rating of all translations (T) previously performed that were at least 70% to similar to the $N \leq 10$ top translations recommended (R) by the system. For discrete categorization in some tables, graphs and charts, we round the P to the nearest whole number. The

SPI (2) is calculated by simply dividing their self-professed mastery of English (S) obtained from their profiles by P followed by subtracting the whole.

$$P(worker) = Avg(P \text{ for } T \mid T \in R) \quad (1)$$

$$SPI(worker) = (S / P) - 1 \quad (2)$$

Given P for each worker, we are able to calculate the community capability index (CCI) (3), the average of the P's for each worker. This is indicative of the labor community's general performance for the tasks in which they opt to undertake. From the SPI's for each worker, we are able to calculate the community perception index (CPI) (4), the average of all workers' SPI's. This metric indicates the labor community's own awareness of it's true capability based on performance feedback from all employers.

$$CCI(community) = Avg(P) \quad (3)$$

$$CPI(community) = Avg(SPI(\forall \text{ worker})) \quad (4)$$

C. Correlating Self-Assessment and Performance

From the set worker {P, SPI}, we are able to observe the worker's actual performance versus their perceived level of expertise. Using this 2-tuple for each worker, we can observe the trend in the community, comparing the workers' self-evaluation to their actual performance. We can also see which groups of workers' self-evaluation were most consistent with their actual performance. Calculating the average P and SPI for all workers allow us to collectively correlate the labor market's self-evaluation to actual performance in the community 2-tuple {CCI, CPI}.

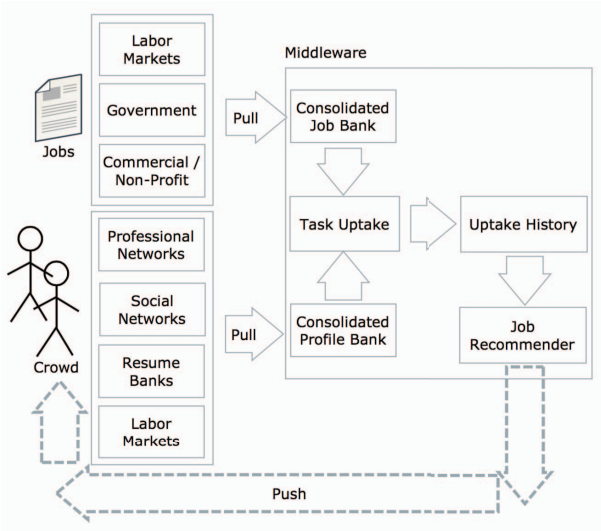

Fig. 2. A Recruitment/Retention Open Push/Pull Mechanism.

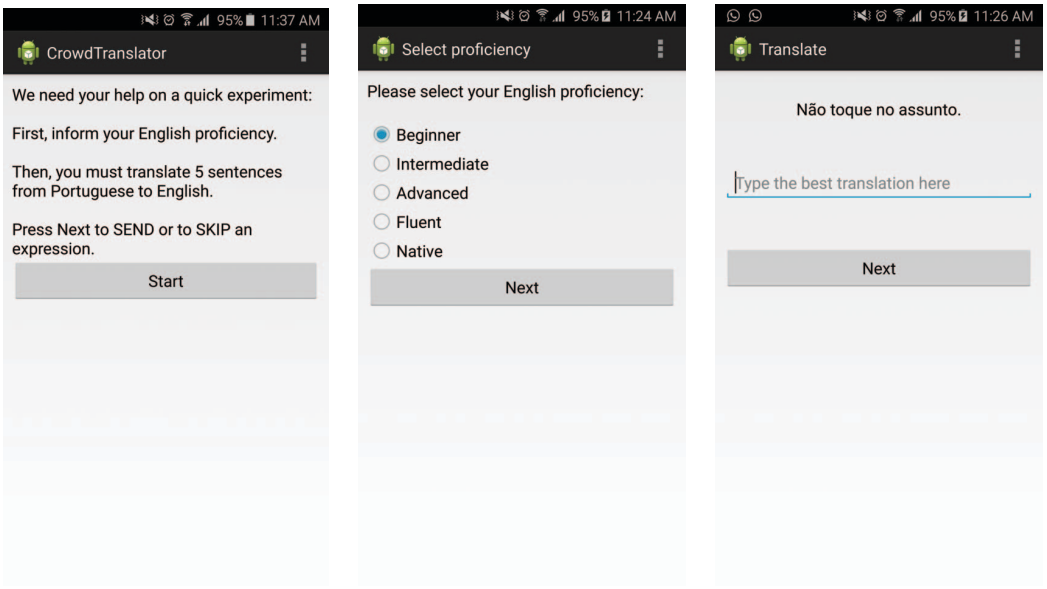

Fig. 3. Screenshots of worker crowdsourcing mobile interface.

D. Preliminary Results

For this community, we primarily observed conservative self-evaluation by the users. Table 1 and Fig. 4 illustrate the users' self-evaluation (S) of their mastery of English pulled from their profiles and their calculated performance (P) based on the translations previously performed. The data shows that workers tend to under estimate their actual capability to perform translation tasks. We found 32 of 53 workers under-estimating their own capability, another 20 correctly estimating their capability and single worker over-estimating. This reflected in an average performance rating of 4.35 to a lower self-evaluated rating of 3.36 (Fig. 5), showing that the labor force is more competent than it perceives itself to be. More than 90% of all workers attempted translations with difficulties consistent with or less than their own self-evaluation (Fig. 6). The quality of the translations for 94.34% of all workers was higher than the difficulty level of the translation task (Fig. 7). This conveys the community's proficiency in the tasks they opted to do.

TABLE I. SELF VS CALCULATED RATINGS FOR ALL USERS

	Self	Calculated
Level 1	2	0
Level 2	7	0
Level 3	16	3
Level 4	26	25
Level 5	2	25

V. DISCUSSION

A. The Most Self-Aware

Our granular analysis of self-evaluation showing in Fig. 7 revealed that 35 of 53 users under-estimated themselves with a self-evaluation lower than their actual performance. Another 17 users correctly characterized their skills, and a lone worker overestimating his/her performance. The 2 workers with a self-evaluation of 1 had higher levels of performance. All 7 workers with a self-evaluation of Level 2 underrepresented their actual performance. 15 of 16 workers who evaluated themselves at Level 3 had higher levels of performance with 1 with a correct self-evaluation. The first and only over-estimation in our study appears in Level 4 with 14 estimating themselves correctly and another 11 underestimating. The 2 workers who evaluated themselves as Level 5 correctly evaluated themselves; this group was the only group matching 100% self-evaluations to performance.

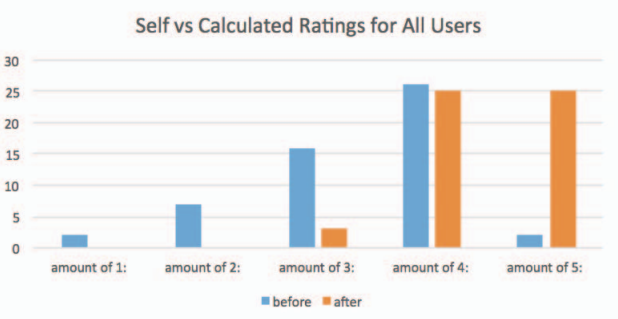

Fig. 4. Self-calculated Ratings vs System-calculated Ratings for all Users.

B. Labor Force Assessment

Using metrics such as the CCI and CPI, we are able to collectively understand the nature of a given labor force. Recall the CCI is a collective indication of their performance capability based on actual performance and the CPI the general community's consensus of themselves. In our case, we observed a conservative community evaluation in a performance to self-evaluation ratio (Fig. 6). This community is perceived as being *overly critical* of its own capability and has negatively represented itself by a CCI of – 19.58%. Positive CCI ratings result in over confidence and over representation of worker capabilities.

Given this type of information, employers are able to understand the true nature of the capabilities of workers in a given labor force relative to another. With a more informed cross-sectional view into the capabilities and worker perception of their own skills in a given labor force, employers can adjust their levels of worker confidence and opt whether to crowd source their tasks through the platform or seek another with a higher CCI index.

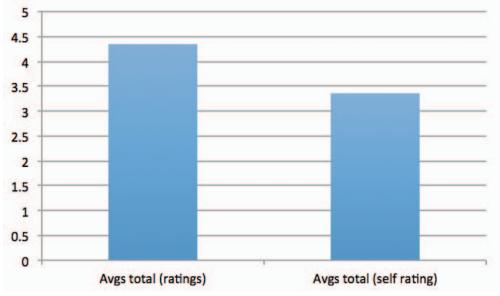

Fig. 5. Average actual rating vs. average self-evaluated rating.

C. Enhanced Recommender: Push with Feedback

Using collaborative filtering driven recommenders, we are able to recommend tasks suited for the workers' unknown profile. Worker skills are assumed based on the feature set of tasks previously completed over time as opposed to matching required skills to worker profiles in filtering tasks. New tasks with the highest predicted performance and matching a similarity threshold in features to previously completed tasks are now pushed for completion given the worker's collective performance history. With this type of push methodology, we recommend tasks to workers in contrast to the current paradigm practices where workers solicit and agree to engage tasks posted by employers. In concept, this method allows a central system to infer skill profiles as it provides an objective evaluation of the workers' skillset as per feedback by the employers and owners of previously completed tasks.

D. Future Infrastructures

Considering these new techniques, we foresee future infrastructure enhancements in full lifecycle crowdsourcing systems as described in Fig. 1. The five core services in need of enhancement are recruitment, job allocation and contracts, compensation, and finally solution resolution.

Recruitment job allocation, contractual enhancements must migrate from the open call subscription model to an active open pull model that constantly searches communities for potential workers. It should also utilize a multitude of adaptive approaches such as profile matching, collaborative filtering, case-based reasoning and machine learning [17] to recommend jobs to workers through an open push. Compensation must be multidimensional [17] to attract varying types of audiences for varying types of tasks. Cash based compensation can conform to pay-for-performance, quota-based or team-based compensation models [20]. Though often implemented as financial [20], compensation can appeal to other aspects human nature such as volunteerism, altruism, humanitarianism and the common good [21].

Mechanisms must be in place for the resolution of submissions from workers. Robust database management strategies are required for the streaming of solutions into repositories for the perusal of employers. These submissions are either rejected based on quality metrics or accepted and the worker compensated [17].

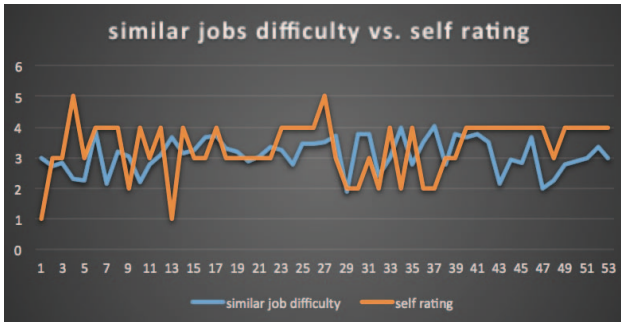

Fig. 6. Job Difficulty vs Worker's Self-evaluation.

VI. Conclusion and Future Work

In this paper, we discuss related work in recommender systems. We have provided metrics that may be used, in conjunction with a collaborative filtering type recommender, to calculate and compare worker self-evaluation and actual performance feedback from employers in a labor market. We proposed the worker 2-tuple {P, SPI} and the community 2 tuple {CCI, CPI}. From our preliminary findings, we find that workers' self-evaluation is not indicative of their actual performance. Fortunately in this case, our findings show that the community was conservative and hence yielded much better performance and had grossly underrepresented themselves from their profiles. Please note that our findings may be constrained to language translation oriented crowdsourced tasks. For future work, we extend our current experiment for other types of languages, with new communities, and other types of crowdsourcing task categories.

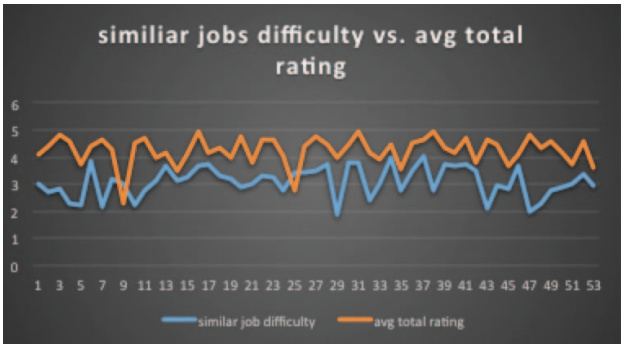

Fig. 7. Difficulty of Jobs Performed vs Worker's Actual Performance.

References

[1] Amazon Mechanical Turk (2015): https://www.mturk.com/mturk/welcome

[2] M. Hirth, T. Hoßfeld, and P. Tran-Gia. "Anatomy of a crowdsourcing platform-using the example of microworkers. com." In *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*, pp. 322-329. IEEE, 2011.

[3] T. Schulze, S. Krug, & M. Schader (2012). Workers' task choice in crowdsourcing and human computation markets.

[4] D. E. Difallah, G. Demartini & P. Cudré-Mauroux. (2013, May). Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 367-374). International World Wide Web Conferences Steering Committee.

[5] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar. "Quality control in crowdsourcing systems: Issues and directions." *IEEE Internet Computing* 2 (2013): 76-81.

[6] M.B. Blake and M.F. Nowlan."A Web Service Recommender System using Enhanced Syntactical Matching", *IEEE International Conference on Web Services (ICWS 2007),* July 2007.

[7] D. Schall, M.B. Blake, S. Dustdar "Programming Human and Software-Based Web Services", *IEEE Computer*, 43,No. 7, pp 82-86, July 2010.

[8] J. Malinowski, T. Keim, O. Wendt, and T. Weitzel. "Matching people and jobs: A bilateral recommendation approach." In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, vol. 6, pp. 137c-137c. IEEE, 2006.

[9] T. Keim. "Extending the applicability of recommender systems: A multilayer framework for matching human resources." In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pp. 169-169. IEEE, 2007.

[10] W. Tan, M.B. Blake, I. Saleh, S. Dustdar. "Social-Network-Sourced Big Data Analytics", *IEEE Internet Computing*, Vol. 17, No. 5, pp. 62-69, Sept/Oct 2013.

[11] S. L. Lim, D. Quercia, and A. Finkelstein. "StakeSource: harnessing the power of crowdsourcing and social networks in stakeholder analysis." *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 2*. ACM, 2010.

[12] I. Paparrizos, B. Barla Cambazoglu, and Aristides Gionis. "Machine learned job recommendation." In *Proceedings of the fifth ACM Conference on Recommender Systems*, pp. 325-328. ACM, 2011.

[13] K. Ece, S. Hacker, and E. Horvitz. "Combining human and machine intelligence in large-scale crowdsourcing." In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems.* pp. 467-474, 2012.

[14] J. Howe. "The rise of crowdsourcing." *Wired Magazine,* 14.6 (2006):1-4.

[15] A. Doan, R. Ramakrishnan, & A. Y. Halevy. "Crowdsourcing systems on the world-wide web." *Comm. of the ACM* 54.4 (2011): 86-96.

[16] A. Quinn, and B. Bederson. "Human computation: a survey and taxonomy of a growing field." 2009. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1403-1412. ACM, 2011.

[17] J. Jarrett, M. B. Blake. Collaborative Infrastructure for On-Demand Crowd sourced Tasks. IEEE WETICE 2015.

[18] S. T. Al-Otaibi, & M. Ykhlef. (2012). A survey of job recommender systems. *Int. Journal of the Physical Sciences*, *7*(29), 5127-5142.

[19] J. Jarrett, I. Saleh, M. B. Blake, R. Malcolm, S. Thorpe and T. Grandison. "Combining Human and Machine Computing Elements for Analysis via Crowdsourcing". *10th IEEE Int. Conf. o n Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2014)* October 22-25, 2014 Miami, FL USA.

[20] O. Scekic, H. Truong, and S. Dustdar. "Incentives and rewarding in social computing." *Communications of the ACM* 56.6 (2013): 72-82.

[21] B. Huberman, D. Romero, and F. Wu. "Crowdsourcing, attention and productivity." *Journal of Information Science* (2009).