Computer Networks 90 (2015) 110–120

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

A system for scalable and reliable technical-skill testing in online labor markets

Maria Christoforaki*, Panagiotis G. Ipeirotis

New York University, New York, NY, United States

ARTICLE INFO

Article history:
Received 16 October 2014
Revised 1 April 2015
Accepted 15 May 2015
Available online 11 July 2015

Keywords:
Test question generation
Test question evaluation
Technical-skill assessment
Online labor markets

ABSTRACT

The emergence of online labor platforms, online crowdsourcing sites, and even Massive Open Online Courses (MOOCs), has created an increasing need for reliably evaluating the skills of the participating users (e.g., "does a candidate know Java") in a scalable way. Many platforms already allow job candidates to take online tests to assess their competence in a variety of technical topics. However the existing approaches face many problems. First, cheating is very common in online testing without supervision, as the test questions often "leak" and become easily available online along with the answers. Second, technical-skills, such as programming, require the tests to be frequently updated in order to reflect the current state-of-the-art. Third, there is very limited evaluation of the tests themselves, and how effectively they measure the skill that the users are tested for.

In this article we present a platform, which continuously generates test questions and evaluates their quality as predictors of the user skill level. Our platform leverages content that is already available on question answering sites such as Stack Overflow and re-purposes these questions to generate tests. This approach has some major benefits: we continuously generate new questions, decreasing the impact of cheating, and we also create questions that are closer to the real problems that the skill holder is expected to solve in real life. Our platform leverages the use of Item Response Theory to evaluate the quality of the questions. We also use external signals about the quality of the workers to examine the external validity of the generated test questions: questions that have external validity also have a strong predictive ability for identifying early the workers that have the potential to succeed in the online job marketplaces. Our experimental evaluation shows that our system generates questions of comparable or higher quality compared to existing tests, with a cost of approximately \$3–\$5 dollars per question, which is lower than the cost of licensing questions from existing test banks, and an order of magnitude lower than the cost of producing such questions from scratch using experts.

1. Introduction

Today, increasingly more skilled labor activities are carried out online. Online labor markets, such as eLance-oDesk and Freelancer, are platforms that connect workers with relevant employers.[1] These computer-mediated marketplaces can eliminate geographical restrictions, help participants find desirable jobs, guide workers through complex goals, and better understand workers' abilities. Broadly, online labor markets offer participants the opportunity to chart their own careers, pursue work that they find valuable, and

* Corresponding author. Tel.: +1 9174451582.
E-mail addresses: mc3563@nyu.edu (M. Christoforaki), panos@stern.nyu.edu (P.G. Ipeirotis).

[1] Online labor markets require more high level skills than microtask crowdsourcing markets [1,2].

http://dx.doi.org/10.1016/j.comnet.2015.05.020
1389-1286/© 2015 Published by Elsevier B.V.

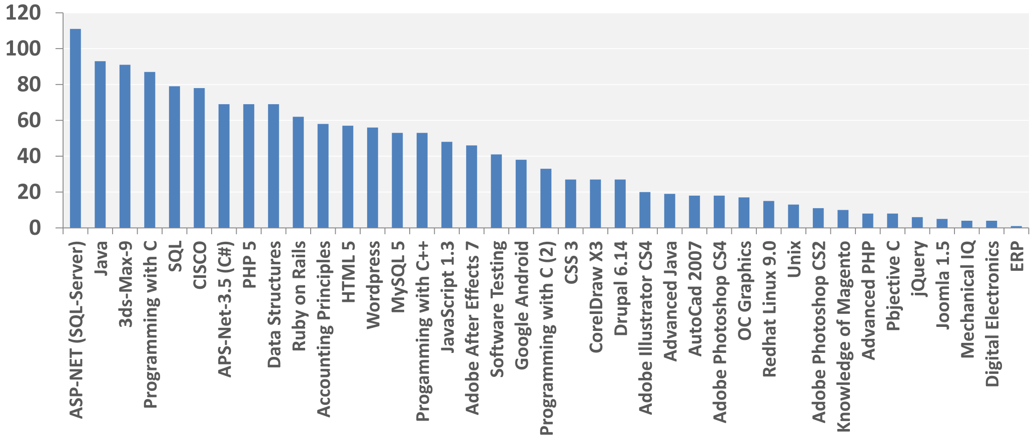

Fig. 1. Number of URLs containing solutions to tests offered by eLance-oDesk and Freelancer (the two biggest online labor marketplaces). Each bar of the *X*-axis represents a test and the *Y*-axis denotes the number of identified URLs that contain the test-questions along with their answers.

do all this at a scale that few companies can today. Spurred by this revolution, some predict that remote work will be the norm rather than the exception within the next decade [3]. One major challenge in this setting is to build skill assessment systems that can evaluate and certify the skills of workers reliably, in order to facilitate the job matching process. Online labor markets currently rely on two forms of assessment mechanisms: *reputation systems* and *skill certification*.

Reputation systems are widely used for instilling trust among the participants [4,5]. A reputation system for an online labor market computes a reputation score for each worker based on a collection of ratings by employers that have hired them in the past. However, existing reputation systems are better-suited for markets where participants engage in a large number of transactions (e.g., selling electronics, where a merchant may sell tens or hundreds of items in a short period of time). Online labor inherently suffers from data sparseness: most work engagements require at least a few hours of work, and many last for weeks or months. As a result, there are many participants that have only minimal number of feedback ratings, which is a very weak reputation signal.[2] Unfortunately, the lack of reputation signals creates a cold-start problem [9]: workers cannot get jobs because they do not have feedback, and therefore cannot get feedback that would help them to get a job. In a worst case scenario, such markets may become “markets for lemons,” [10] forcing the departure of high-quality participants, leaving only low-quality workers as potential entrants.

An alternative approach to instill trust is to use skill certifications. In offline labor markets, educational credentials are often used to signal the quality of the participants and avoid the cold-start problem [11]. In global online markets, credentialing is much trickier: verifying educational background is difficult, and knowledge of the quality of the educational institutions on a global scale is limited. Given the shortcomings of using educational credentials in a global setting, many online labor markets resort to using *skill testing* as means of assessment. So today most online labor markets offer their own certification mechanisms. The goal of these tests is to certify that a given worker indeed possesses a particular skill. For example, eLance-oDesk and vWorker allow workers to take online tests that assess the competency of these contractors across various skills (e.g., Java, Photoshop, Accounting, etc.) and then allow the contractors to display the achieved scores and ranking in their profile. Similarly, crowdsourcing companies, such as CrowdFlower and Mechanical Turk, are certifying the ability of contractors to perform certain tasks (e.g., photo moderation, content writing, translation) and allow employers to restrict recruiting to the population of certified workers. Unfortunately, online certification of skills is still problematic for a number of reasons with cheating being one of the biggest challenges.

The tests currently used by online labor platforms are usually licenced from companies such as ExpertRating[3] that pay domain experts to write test-questions; hence tests are frequently used and are accessible online. That often allows test takers to “leak” the tests and their answers become widely available on the web. Fig. 1 illustrates a number of websites that contain solutions for some of the popular tests[4] available on eLance-oDesk and Freelancer. For example, leaked questions for the ASP-NET test were identified by our crawlers in more than a hundred websites, and, correspondingly, we found more than 90 websites with leaked questions for the Java test. Needless to say, the reliability of the tests for which answers are easily available through a web search is questionable.

Furthermore, it is common, even for expert organizations, to create questions with errors or ambiguities, especially if the test questions have not been properly assessed and calibrated with relatively large samples of test takers [12]. Such problematic questions introduce noise into the

[2] Crowdsourcing research has recently examined the use of peer assessment as an additional form of reputation, focusing on techniques for getting crowd members to evaluate each other [6,7]. The hope is that peer assessment can lead to better learning outcomes as well [8]. Unfortunately, these systems still have large variance in final assessment scores, which makes them a poor match for certification and qualification.

[3] http://www.expertrating.com/.

[4] Sites such as http://1faq.com/ and http://www.livejar.info/, are a couple of examples of the offenders.