Domain Specific Knowledge Base Construction via Crowdsourcing

Ari Kobren, Thomas Logan, Siddarth Sampangi, Andrew McCallum
School of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
{akobren, tlogan, ssampangi, mccallum}@cs.umass.edu

Abstract

Guiding principles for selecting the best crowdsourcing methodology for a given information gathering task remain insufficient. This paper contributes additional experimental evidence and analysis to this problem. Our work focuses on a subset of crowdsourcing problems we term *expert tasks*—tasks that require specific domain knowledge. We experiment with crowdsourcing a knowledge base (KB) of scientists and their institutions using two methods: the first recruits experts who are likely to already know the necessary domain knowledge (using Google Adwords); the second employs non-experts who are incentivized to look up the information (using Amazon Mechanical Turk). We find that responses received through Mechanical Turk are more accurate than those received through Adwords. We analyze this result in terms of the difficulty of recruiting experts for our task and the willingness of Mechanical Turk workers to search the web for information. Our work highlights important considerations for crowdsourcing tasks requiring various types of expertise.

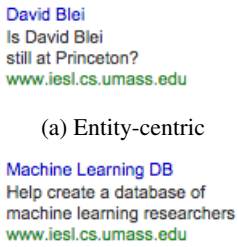
1 Introduction

Crowdsourcing is a method for completing arbitrary tasks by soliciting contributions from a group of human workers. Although workers can be underqualified, practitioners have used redundancy and intelligent recruitment strategies to successfully crowdsource data labeling, collection and maintenance [15]. Indeed, resources such as Wikipedia, Freebase, Duolingo (a language learning tool) and Galaxy Zoo (a citizen science project) have all been built by crowdsourcing.

Paid workers are often recruited for crowdsourcing tasks through *micro-task markets* such as Amazon Mechanical Turk (AMT). A *requester* using AMT offers a small monetary reward (e.g. \$0.05) to any worker who completes his or her tasks. To be eligible for a task, an AMT worker must satisfy a set of minimum requirements defined by the task's requester (e.g. number of tasks previously completed, percentage of tasks approved, etc.). Tasks crowdsourced via AMT are diverse and include image labeling, data collection and word sense disambiguation [12, 13].

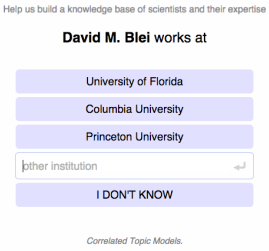
Recently, Ipeirotis and Gabrilovich [4] have presented a new method for unpaid worker recruitment that uses Google Adwords (GA)—an online advertising platform. A requester crowdsourcing with GA creates a set of keywords (and phrases), a daily budget, and a set of advertisements that link to his or her task. When a user issues a web search using one of the specified keywords, the requester's ads bid (against other ads) for the chance to be shown. GA can also place ads on *display network* websites whose contents are similar to that of the ads. GA learns the keywords and websites that generate the most clicks and optimizes the bidding strategy appropriately.

Crowdsourcing with GA has been shown to have multiple advantages [4]. Unlike AMT, GA recruits workers from (potentially) the entire internet. Requesters pay GA instead of the workers and thus



(a) Entity-centric

(b) KB-centric



(c) Landing Page

Figure 1: Two ads and the landing page of our task. The ad in figure 1a targets users who know or are interested in David Blei; the ad in figure 1b targets users who are interested in contributing to a new KB. Clicking either ad redirects a user to the page in 1c.

side-step the problem of workers who seek to maximize profit without regard for the accuracy of their work. Coupled with its keyword-based targeting capabilities and bidding optimization, GA can target workers who are interested in a requester's task and already possess the knowledge necessary to complete it accurately. When compared to AMT on a gamified fact verification task, Ipeirotis and Gabrilovich's results show that GA attracts higher quality users at lower cost. Using GA, they are even able to recruit workers to answer difficult medical questions requiring significant expertise [4].

The most effective way to crowdsource an *expert task*—a task that requires specific domain knowledge—remains an important open question. In this work, we compare worker recruitment via AMT and GA for the expert task of scientific knowledge base (KB) construction. We attempt to build a KB by extracting a set of computer scientists and using each platform to crowdsource the scientists' respective organizational affiliations.

Our results on this task are substantially different from those in the earlier work. Our experiments show that users recruited through GA submit correct answers less than 6% of the time while 59.3% of the answers submitted through AMT are correct. We hypothesize that GA is an ineffective worker recruitment tool for this task because the set of ideal workers is small and difficult to target. Our analysis shows that while AMT workers are not experts who already know the answers to our questions, many are willing to perform a web search for the correct affiliations. With this insight we are able to modify the task instructions to achieve up to a 10% improvement using AMT.

2 Methodology

Our goal is to build a knowledge base (KB) of scientists by crowdsourcing the affiliations of an automatically extracted list of researchers. We extract the authors from a subset of the 2011 DBLP dataset (a database of computer scientists) [7]. For each author we construct a multiple choice question of the form "X works at Y" where X is the author and Y includes three institutions extracted from the Brown Crowdsourcing Assignment Dataset[1] [9]. Users may also submit a free text answer. We use three groups of questions: coauthors of David Blei, Yann LeCun and Noah Smith respectively.

Building a KB of scientists requires specific domain knowledge, therefore we design an advertising campaign that strives to draw users who possess that knowledge. Our campaign includes two ad types: one is entity-centric (Figure 1a) and the other is kb-entric (Figure 1b). We also experiment with three keyword lists; the first is comprised of the coauthor names of a given author, the second is comprised of phrases that GA extracts from relevant paper titles and the third is a hand-tuned list of topical phrases. We set our maximum advertising budget to $75 per day split across all ad and keyword groups and allow GA to automatically tune our bidding strategy. When a user clicks one of our ads, that user is directed to a website on which he or she can complete our tasks (Figure 1c). We allow the campaign to run for one week and measure the accuracy of the submitted responses.

[1]This crowdsourced dataset was collected by students who were in CSCI 2951L at Brown University and contains attributes of computer scientists at the top 50 universities.

For AMT, we require our workers to have completed more than 5000 HITs (i.e. tasks) with at least a 98% approval rating. Each question in our set must be answered by 5 distinct users. As in our GA tasks, each question has 4 potential answers including a free text option. However, for AMT we do not include an option to submit "I don't know." We offer a $0.05 reward for each answer submitted. For most of our tasks, we are able to collect the majority of the data in the task's first 24 hours.

3 Experiments

We find that GA is not an effective recruitment tool for our task. Although we receive a large number of answers, users submit "I don't know" 46.39% of the time, and only 5.94% of the other submissions are correct. Our hand-tuned keyword lists generate the most clicks and our entity ads (Figure 1a) tend to generate more correct answers than the KB-centric ads (Table 1).

Keywords	Ad	Clicks	Responses	IDK	Correct
Hand-tuned	Entity	672	365	187 (51.23%)	31 (17.42%)
Hand-tuned	KB	455	222	114 (51.35%)	7 (6.48%)
Phrases	Entity	107	53	21 (39.62%)	3 (9.375%)
Phrases	KB	84	66	6 (9.09%)	1 (1.67%)
Coauthors	Entity	12	0	0 (0.0%)	0 (0.0%)
Coauthors	KB	2	1	0 (0.0%)	0 (0.0%)
Total	—	1332	707	328 (46.39%)	42 (5.94%)

Table 1: Answers collected over a one week period using Google Adwords. Percent correct calculation does not include "I don't know" (IDK) answers.

We compare the efficacy of AMT and GA as crowd recruitment platforms for the task described above. The outcome of this experiment is summarized in Table 2. Unlike the previous work [4], GA performs worse than AMT in all cases except the Yann LeCun quiz where dramatically fewer users answer more than the first question.

	GA			AMT		
Quiz	Correct	Free Text (%)	Avg time (s)	Correct	Free Text (%)	Avg time (s)
LeCun	45.1%	1.36%	14.00s	37.98%	94.23%	378.84s
Blei	7.02%	0.00%	11.84s	59.29%	87.86%	64.76s
Smith	5.53%	4.37%	7.12s	71.81%	79.19%	82.18s

Table 2: Properties of submitted responses. On the Blei and Smith quizzes, AMT workers outperform GA workers by 8.5x and 13x respectively. GA workers submit relatively few answers for the LeCun quiz partially accounting for the high correctness.

Even though AMT workers are presented with multiple choice answers, 88% of responses are submitted in free text. AMT workers also take more time to complete each question than workers recruited through GA.

4 Discussion

Our experiments with GA highlight the importance of effective user targeting. The answers collected through AMT reveal properties of AMT workers that will be useful for practitioners deciding whether or not the platform is appropriate for their task.

4.1 Error Analysis for Google Adwords

Why does GA perform so poorly in our experiment when it has been shown to recruit expert users in previous work? We observe that three of the top websites (in terms of clicks) on which GA chooses to display our ads are `answers.com` (a question answering forum), `tutorialspoint.com` (a website that teaches basic programming skills) and `obitko.com` (a website that offers AI tutorials). None of these websites are heavily trafficked by users with tremendous knowledge of computer

scientists and their affiliations. This is different from the previous work in which GA advertises medical quizzes on websites that are frequented by medical experts like HealthLine and Mayo Clinic [4]. Are there insufficient websites attracting relevant CS experts or does GA simply fail to find them? Do those websites not participate in GA? It is difficult to answer such questions. This opacity is a disadvantage of the GA approach.

Another reason for GA's inefficacy is that the number of knowledgeable users for this task is extremely small. Some of our questions ask for the affiliations of graduate students and other obscure scientists. Knowledge of the affiliations for these people is limited to a small set of colleagues. Even if they were targetable, this small group of users may be uninterested in our advertisements.

From these results we conclude that the utility of GA for crowdsourcing hinges on the size of the set of ideal users, the difficulty of targeting them with advertisements and their willingness to click on ads. A tool for evaluating these properties could help make crowdsourcing with GA more effective.

4.2 Lessons from Amazon Mechanical Turk

For our task, AMT is a more effective platform than GA, yet AMT workers still make a significant number of errors. Contrary to common models of AMT workers, we find that worker errors are correlated [10]. For example, in a question about the scientist "John William Paisley" (who works at Columbia), 3 out of 5 AMT workers submitted the same incorrect affiliation: "University of Trier." This affiliation is included in the first result returned by a web search for "John William Paisley" (the first result links to John Paisley's DBLP page which is maintained by researchers at the University of Trier). From this observation we present two important conclusions: first, an assumption that AMT workers are independent is often incorrect, and second, AMT workers are willing to search the web to solve certain tasks. Thus, tasks which can be solved by web search are good candidates for crowdsourcing via AMT. Our hypothesis that AMT workers perform web search to solve our tasks is consistent with their lengthy average time per question (Table 2).

With these observations in mind, we repeat the previous experiment with minor variations: we use the same questions but provide a free text option only (i.e. no multiple choice options). Additionally, in the task instructions we note that each researcher is likely a computer scientist and provide an example web query that a user could issue to find the researcher's affiliation. Specifically, we always provide a query built from the researcher's name followed by the phrase "computer science." These minor variations increase overall submission accuracy by 3% (and 10% for the LeCun quiz).

5 Related Work

Previous work has demonstrated the efficacy of GA as a crowd recruitment tool for gamified fact verification [4]. In this work GA outperforms AMT in terms of cost and accuracy. Other work with GA focuses on matching queries to ads [8, 3] and keyword selection [11, 1].

The crowdsourcing of expert tasks is also an active area of research. In one experiment researchers train AMT workers to complete *citizen engineering* tasks and show that aggregated AMT workers and domain experts are similarly accurate [14]. Another proposed approach divides a difficult task into *micro-tasks* and then combines the responses [5, 6]. In addition to expert recruitment, some research focuses on using badge systems to incentivize experts to make many contributions after they have been recruited [2].

6 Conclusion

Our results highlight the importance of identifying the properties of crowd recruitment platforms (e.g. workers on AMT will search the web for answers). A better understanding of these properties will help practitioners address a fundamental crowdsourcing question: given a task, how can it be crowdsourced most efficiently? Our work also raises new directions for research including building classifiers to predict the success of a task on a crowdsourcing platform, designing tools to measure important properties of crowdsourcing tasks and reasoning about the ideal workers for a given task.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by an award from Google, in part by NSF grant #CNS-0958392, and in part by the National Science Foundation under NSF DGE-0907995. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[1] Yifan Chen, Gui-Rong Xue, and Yong Yu. Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the 2008 international conference on web search and data mining*, pages 251–260. ACM, 2008.

[2] David Easley and Arpita Ghosh. Incentives, gamification, and game theory: an economic approach to badge design. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 359–376. ACM, 2013.

[3] Gagan Goel and Aranyak Mehta. Online budgeted matching in random input models with applications to adwords. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 982–991. Society for Industrial and Applied Mathematics, 2008.

[4] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. Quizz: targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on World wide web*, pages 143–154. International World Wide Web Conferences Steering Committee, 2014.

[5] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 43–52. ACM, 2011.

[6] Kazuhiro Kuwabara and Naoki Ohta. Toward a crowdsourcing platform for knowledge base construction. In *eKNOW 2014, The Sixth International Conference on Information, Process, and Knowledge Management*, pages 89–92, 2014.

[7] Michael Ley. Dblp computer science bibliography. 2005.

[8] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22, 2007.

[9] Alexandra Papoutsaki, Hua Guo, Danae Metaxa-Kakavouli, Connor Gramazio, Jeff Rasley, Wenting Xie, Guan Wang, and Jeff Huang. Dataset of 2200 faculty in 50 top us computer science programs, Spring 2014.

[10] Aditya G Parameswaran. *Human-powered Data Management*. PhD thesis, Stanford University, 2013.

[11] Paat Rusmevichientong and David P Williamson. An adaptive algorithm for selecting profitable keywords for search-based advertising services. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 260–269. ACM, 2006.

[12] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[13] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. *Urbana*, 51(61):820, 2008.

[14] Matthew Staffelbach, Peter Sempolinski, David Hachen, Ahsan Kareem, Tracy Kijewski-Correa, Douglas Thain, Daniel Wei, and Greg Madey. Lessons learned from an experiment in crowdsourcing complex citizen engineering tasks with amazon mechanical turk. *CoRR*, abs/1406.7588, 2014.

[15] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.