Knowledge Crowdsourcing Acceleration

Jie Yang(✉), Alessandro Bozzon, and Geert-Jan Houben

Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
{j.yang-3,a.bozzon,g.j.p.m.houben}@tudelft.nl

Abstract. Crowdsourcing has recently become a powerful computational tool for data collection and augmentation. Although crowdsourcing has been extensively applied in diverse domains, most tasks are of low complexity such that workers are assumed to be endless, anonymous and disposable. By unlocking the value of human knowledge-related features, e.g., experience, expertise and opinion, we envision that crowdsourcing can reach its full potential to solve complex tasks. We aim at creating a comprehensive theory of crowdsourcing for knowledge creation, i.e., *knowledge crowdsourcing*, with a focus on developing methods and tools to control and accelerate knowledge creation process. Inspired by previous work, we describe a reference model of knowledge crowdsourcing acceleration, together with three case studies for model validation and extension. The results of our first case study on on-line knowledge creation demonstrate the potential contribution to web engineering.

Keywords: Crowdsourcing · Knowledge creation · Collaborative question answering · Enterprise expert finding · Urban computing

1 Introduction

Crowdsourcing is the process of sourcing tasks to large online crowds [1]. As a discipline, crowdsoucing recently emerged as a promising form of computation and knowledge generation [2], which provides effective methods for data collection and augmentation.

Motivation. Crowdsourcing has been usually studied as a computational tool, where tasks are assumed to be of low cognitive complexity, and workers to be endless, anonymous and disposable. Such crowdsourcing only takes little advantage of human capabilities, mostly only relying on the availability of workers. However, the rich, knowledge-related features of humans (i.e. expertise and skills), and their subjective perceptions are less considered. By fully unlocking the value of such inherent human abilities, crowdsourcing can reach its full potential, and enable the solution of more complex, cognitive intensive tasks.

Related Work. Although crowdsourcing has drawn much attention from researchers with diverse background, much less studies have been focusing on knowledge-related task crowdsourcing. The few studies addressing crowdsourcing

P. Cimiano et al. (Eds.): ICWE 2015, LNCS 9114, pp. 639–643, 2015.
DOI: 10.1007/978-3-319-19890-3_47

for knowledge creation present results that are difficult to compare, and of difficult generalization, as experiments are performed mostly on an ad-hoc basis. Typical tasks are solved as a bottom-up process. For instance, the processes of building wiki's or collaborative QA (CQA) systems, are not based on systematical crowdsourcing methods, but more on the spontaneus and autonomous contribution of volunteers. A comprehensive theory of *crowdsourcing* for *knowledge creation* is in demand [3].

Goal. Our work focuses on *knowledge crowdsourcing*, i.e. the process of designing, executing and coordinating crowdsourcing tasks that are knowledge intensive. Based on this definition, our goal is to develop the methods and tools required to control and accelerate the process of crowdsourced knowledge creation, by taking into account the rich set of knowledge-related features of humans, like experience, expertise, or opinion. We envision application in several domains, from the acceleration of the process of on-line knowledge creation, to the augmentation of subjective human perception data for better urban computing.

2 Methodology

Our path towards the creation of a more comprehensive theory of knowledge crowdsourcing builds upon a reference model, presented later in this section. Based on the initial model, we then follow an iterative improvement process based on its application to different real-world case studies, which, in turn, validate and extend the theory in each iteration. In this sense the case studies are important, as they serve as testbeds for validation, and basic resources for developing the knowledge crowdsourcing theory.

Initial Model. To control the process of knowledge crowdsourcing, we need to understand how it operates. Therefore a high-level reference model is to be proposed to capture its key steps. Our model follows the generic conceptual framework of human computation systems (HCS) [4]. It is applied in all case studies, such that their results, by conforming to a unified framework, could be better compared and, possibly, generalized.

Our model builds on the following key components: ***1)*** *Worker modeling* techniques to assess the worker skills and expertise; ***2)*** *Task modeling* techniques to represent the knowledge to be created; ***3)*** methods for *task assignment and recommendation*, and for *workflow control and optimization*, to enable quick and high-quality knowledge creation; and ***4)*** *tools* to support all the above. These components correspond to the key facets of conceptual framework of HCS: component 1 considers worker properties and engagement; component 2 defines the goal of a task; component 3 manages the problem-solving process.

Our work aim at showing how, by optimally combining the components described above, it will be possible to accelerate knowledge creation in a systematic and effective way. We envision that the accelerating methods and tools can be demonstrated in a new framework, built upon reference works in the field [5].

Case Studies. We identified 3 representative case studies where the knowledge creation task is complex, and crowdsourcing can be beneficial. Each case covers key components of our theory, while stressing different complexity dimensions.

On-line Knowledge Creation. This case study considers knowledge crowdsourcing in the open Web environment. Typical applications include CQA systems such as Stack Overflow and Yahoo! Answers, content curation systems like Reddit, and diverse on-line forums. Despite the high activeness of such systems, requests are not always satisfied due to the wide range of request difficulty and user expertise levels. In this use case, our research questions are: 1) how to model users with different levels of expertise; 2) how to model requests of different levels of difficulty; and 3) how to reduce time needed for obtaining good solutions/content. Section 3 presents initial results achieved with this case study.

Enterprise Knowledge Creation. This case study addresses the problem of knowledge creation in large companies. Tasks in this case are designed for enterprise interest, thus having narrow and focused knowledge needs. Requesters and workers are employees from the same company, thus offering crowds of a size smaller than that in the on-line case. Consequently, the challenges include: 1) modeling knowledge needs in task design; 2) modeling diverse facets of employees' expertise to match the task requirement; and 3) non-monetary incentive mechanisms (e.g. gamification) to engage workers, to cater for the presence of stronger monetary compensations (i.e. salary).

Urban Knowledge Creation. Crowdsourcing urban perception increases the availability of data to model urban conditions, for a better understanding of cities. In this case, crowds are engaged to source factual perception of urban environment (e.g., traffic congestion), but also subjective perception, such as travelers' emotional reaction to city spots. The later type of perception can be influenced by personalities or cultural backgrounds. Another important feature of this case study is that targeted workers are often very dynamic, yet geographically constrained. To summarize, this case study poses the following research questions: 1) properly embedding subjective element in task model; 2) modeling and eliciting general or subjective user features; and 3) instantly capturing and engaging the dynamic workers in the crowdsourcing workflow.

The three case studies have also important economical and societal value. Online knowledge creation platforms are important as it has been shown that they are reforming the ways people create and share knowledge on the Web. Enterprise knowledge creation, which is less studied, has the potential effect on boosting the performance of enterprise-level task execution. Urban knowledge creation addresses the challenge in optimizing the largest public living environment, i.e. cities, which already accommodate 72% of Europe population.

3 Results

This section presents our research results of the first case study, i.e. online knowledge creation. We focused on Stack Overflow, one of the most active CQA

systems on the Web. Despite its success, a large portion of Stack Overflow questions do not receive good quality answers, and the average time for a question to obtain an answer is at a magnitude of days. Therefore, Stack Overflow is a perfect candidate to study techniques for accelerating knowledge creation.

Edit Suggestion. We first propose to better understand the impact of question quality on knowledge creation, by study methods to suggest the right edit to apply on poorly formulated questions. Based on a qualitative study that reveals the main functions of question edits, we presented a methodology to automatically detect whether a newly posted question needs an edit, and if so, what type of edit it needs. Experiments show that we can reach an F-measure of 0.7 for edit need prediction, and an F-measure of 0.76 for code type prediction. The detail of this research has been published in [6].

Expertise Identification. Identifying user expertise in CQA systems has been widely studied, while most of the research approximates user expertise with user activeness. We show in our recently published paper [7] that activeness does not necessarily strictly correlate with expertise. We proposed a novel expertise metric that is hardly influenced by activeness, based on which we define a group of users to approximate experts. By characterizing their behaviors, we show that they fit the image of experts better than experts defined by other metrics.

Question Routing. We are currently studying how better understanding the engagement of potential answerers can lead to better question routing, i.e. recommend questions to users to reduce answering time. We assume that questions of different topics require different combinations of user roles: questions relating to general skills may require active answerers engaged in discussion to ultimately generate a best answer, while questions relating to specific language or framework may only need one expert user to directly provide the right answer. Preliminary experiments show that optimizing the weights of user roles could improve the accuracy of question routing.

4 Future Works

We have instantiated knowledge crowdsourcing theory in on-line knowledge creation case, and demonstrated its potential benefits in several publications. With the promising results achieved so far, we plan to continue exploring knowledge crowdsourcing acceleration in enterprise (e.g. by collaborating with IBM Netherlands) and urban knowledge creation (e.g. in the context of the Amsterdam AMS initiative) cases, to further develop the theory with methods and tools.

References

1. Howe, J.: The Rise of Crowdsourcing. Wired Magazine **14**(6), 1–4 (2006)
2. Law, E., Ahn, L.V.: Human Computation. Synthesis Lectures on Artificial Intelligence and Machine Learning **5**(3), 1–121 (2011)

3. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: CSCW 2013, pp. 1301–1318. ACM, New York (2013)
4. Malone, T.W., Laubacher, R., Dellarocas, C.: Harnessing Crowds: Mapping the Genome of Collective Intelligence. MIT Sloan Research Paper, No. 4732–09 (2009)
5. Bozzon, A., Brambilla, M., Ceri, S., Mauri, A.: Reactive crowdsourcing. In: WWW 2013, pp. 153–164. ACM, New York (2013)
6. Yang, J., Hauff, C., Bozzon, A., Houben, G.J.: Asking the right question in collaborative Q&A systems. In: Hypertext 2014, pp. 179–189. ACM, New York (2014)
7. Yang, J., Tao, K., Bozzon, A., Houben, G.-J.: Sparrows and owls: characterisation of expert behaviour in stackoverflow. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) UMAP 2014. LNCS, vol. 8538, pp. 266–277. Springer, Heidelberg (2014)