Eliciting Structured Knowledge from Situated Crowd Markets

JORGE GONCALVES, Center for Ubiquitous Computing, University of Oulu and Department of Computing and Information Systems, The University of Melbourne
SIMO HOSIO, Center for Ubiquitous Computing, University of Oulu
VASSILIS KOSTAKOS, Center for Ubiquitous Computing, University of Oulu and Department of Computing and Information Systems, The University of Melbourne

We present a crowdsourcing methodology to elicit highly structured knowledge for arbitrary questions. The method elicits potential answers ("options"), criteria against which those options should be evaluated, and a ranking of the top "options." Our study shows that situated crowdsourcing markets can reliably elicit/moderate knowledge to generate a ranking of options based on different criteria that correlate with established online platforms. Our evaluation also shows that local crowds can generate knowledge that is missing from online platforms and on how a local crowd perceives a certain issue. Finally, we discuss the benefits and challenges of eliciting structured knowledge from local crowds.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**

Additional Key Words and Phrases: Crowdsourcing, structured knowledge, situated, questions, options, criteria, performance, accuracy, quality, local crowds

ACM Reference Format:
Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Eliciting structured knowledge from situated crowd markets. ACM Trans. Internet Technol. 17, 2, Article 14 (March 2017), 21 pages.
DOI: http://dx.doi.org/10.1145/3007900

1. INTRODUCTION

In this article, we present a crowdsourcing approach to generate highly structured knowledge on a set of topics. A plethora of online platforms today enable crowds to contribute to shared pools of data or knowledge. For instance, IMDb and Metacritic aggregate reviews and ratings for movies from moviegoers around the world (as well as critics), while TripAdvisor aggregates reviews and ratings from travellers regarding travel-related content. A limitation of such platforms, however, is that they are focused on a particular topic, such as movies or restaurants. Further, while these platforms present knowledge in a structured fashion, they lack the flexibility needed to order options depending on different characteristics (e.g., movies' visual effects or a restaurants' atmosphere).

Most similar to our work are question-and-answer websites, where arbitrary questions can be posed. Quora is such a question-and-answer system, where a wide range of

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 285062-iCYCLE, 286386-CPDSS, and 285459-iSCIENCE) and the European Commission (Grants PCIG11-GA-2012-322138, 645706-GRAGE, and 6AIKA-A71143-AKAI).
Authors' addresses: J. Goncalves, S. Hosio, and V. Kostakos, Erkki Koiso-Kanttilan katu 3, door E, P.O Box 4500, FI-90014 Oulu, Finland; emails: firstname.lastname@oulu.fi.

DOI: http://dx.doi.org/10.1145/3007900

questions is created, answered, edited, and organized by users. However, due to its forumlike functionality that relies on free-text questions and answers, it is far from ideal when dealing with questions that have multiple potential answers and characteristics that might appeal more or less to certain individuals.

Thus, while such systems are excellent at answering fact-based questions ("Where did Led Zeppelin's 1977 tour start?"), they can be challenging for users to find the appropriate answer for them when subjective questions ("Which car should I buy?"). In the latter case, users would need to scour the free-text answers in hopes of finding different answers with pros and cons for each. In our case, when generating a *highly* structured knowledge base, there is a clear separation between options and criteria, instead of the common model of providing free form text answers to a posed question.

In order to build this highly structured knowledge base, we chose to use a situated crowdsourcing market [Hosio et al. 2014a]. A situated crowdsourcing market allows workers to complete a variety of crowdsourcing tasks to earn rewards. However, unlike online labour markets, the interaction happens on embedded input mechanisms (e.g., public displays, tablets) in a physical space in order to leverage users' serendipitous availability [Müller et al. 2010] or idle time ("cognitive surplus" [Shirky 2010]). As highlighted in the literature, situated crowdsourcing deployments are sustained through a "self-renewable workforce" by steadily attracting new workers [Goncalves et al. 2013; Heimerl et al. 2012]. In other words, the serendipitous nature of situated technologies can help build a useful long-term knowledge base due to a constant stream of new contributors. Further, such local crowds have been shown to provide reliable crowd data [Goncalves et al. 2013; Heimerl et al. 2012] and more in-depth information on locally relevant tasks [Goncalves et al. 2014a]. Previous work has also shown how situated crowdsourcing can be leveraged to directly study a particular community [Goncalves et al. 2014b] due to its geofenced crowdsourcing environment in which only those within a location can contribute, while with mobile crowdsourcing location information can be faked [Guo et al. 2015; Goncalves et al. 2016].

To assess the suitability and benefits of leveraging local crowds to elicit highly structured knowledge, we selected five different questions:

—Two general questions aimed at assessing if a local crowd can reliably solve problems when compared to reference online platforms (IMDb and Numbeo).
—One local and one semi-local question aimed at demonstrating that local crowds can provide more in-depth knowledge not available on online platforms (TripAdvisor and Numbeo).
—One general question with potential local ramifications aimed at demonstrating that structured knowledge from local crowds can provide useful information on how they perceive a certain issue.

Through our analysis, we show that local crowds can be effective recommenders and moderators, and highlight the benefits of eliciting structured knowledge using a situated crowdsourcing platform.

2. RELATED WORK

2.1. Problem Structuring

In developing a systematic way to generate structured knowledge for arbitrary questions, we choose to only consider subjective questions ("Which car should I buy?") that have multiple answers and tradeoff criteria. Research in psychology has suggested that artifacts under consideration (or competing answers) can be represented as sets of features and that similarity/differences between these artifacts can be expressed as operations on their respective feature sets. For instance, the contrast model proposed by

Tversky [1977] models the similarity between artifacts as a function of their features. Depending on context, the theory suggests that different features will be assigned a different weight when contributing towards an overall similarity assessment. In our work, we adopt this approach to enable the crowd to systematically consider the various tradeoffs between various possible answers to a question.

Lee et al. [2005] applied a similar approach and theory to model document similarity with poor results, while Navarro and Lee [2004] proposed using linear combinations of features with varying weights to model similarity. This is a key insight, because it suggests that to model the tradeoffs to a solution and then different weights should be considered. Previous work in the crowdsourcing domain has split up tasks into fault-tolerant subtasks, which has been shown to improve worker performance. For example, both Soylent [Bernstein et al. 2010] and PlateMate [Noronha et al. 2011] had an input decomposition phase at the start. In a more high-level approach, Kittur et al. [2011] proposed a general-purpose framework named CrowdForge for distributed processing and providing scaffolding for complex crowdsourcing tasks.

Another research domain with similarities to our work is "games with purpose" [von Ahn and Dabbish 2004], which provide a practical method to collect descriptive features for arbitrary images, albeit the weight of a particular feature is not explicitly indicated by users. Further similarities can be observed in critiquing-based recommender systems that explore problem domains by letting end users to assign different weights to descriptive attributes and get recommendations based on their preferences [Chen and Pu 2011]. However, critiquing-based systems rely on the system admins designing the desired attributes (not users) and are typically seen as suitable for only high-involvement items.

In terms of actually collecting the attributes and their values from end users, for instance, WikiData [2015] has emerged as a highly popular collective effort to present "everything" with attributes and their numerical values. WikiData allows end users to create attributes, enabling the database of things to grow practically endlessly and thus become more accurate over time. Finally, Cheng and Bernstein [2015] recently introduced Flock, a system leveraging crowdsourcing in machine learning. In practice, the crowd's role is to come up with human understandable features from any items online, and the system then uses machine learning and automation (using CrowdFlower microtask market) to aggregate the features.

In our case, we ask workers to annotate a particular problem statement in terms of potential answers and potential tradeoff criteria. Then, drawing on Tversky's [1977] theory of similarity and Navarro and Lee's [2004] findings, we ask workers to consider the relationship between solutions and criteria by estimating the weight of each solution-criterion relationship. Our approach does not limit the amount of items or weights on how an item is modelled, and, at the same time, the system is suitable for all kinds of items and not only on, for example, low-priced commodity items or high-involvement items.

2.2. Situated Crowdsourcing

Situated crowdsourcing is relatively under-explored when compared to online and mobile crowdsourcing. It allows for a geofenced and more contextually controlled crowdsourcing environment, thus enabling targeting of certain individuals, leveraging people's local knowledge, or simply reaching an untapped source of potential workers [Hosio et al. 2014a; Hosio et al. 2015]. Although with potentially fewer "workers" than its online crowdsourcing counterpart, this approach has been shown to reduce noise and bias in "crowd-data" [Goncalves et al. 2013]. Furthermore, situated crowdsourcing does not require any promotion, installing of dedicated software by the worker, and, in many cases, no sign-ups or logins are needed as people serendipitously encounter these

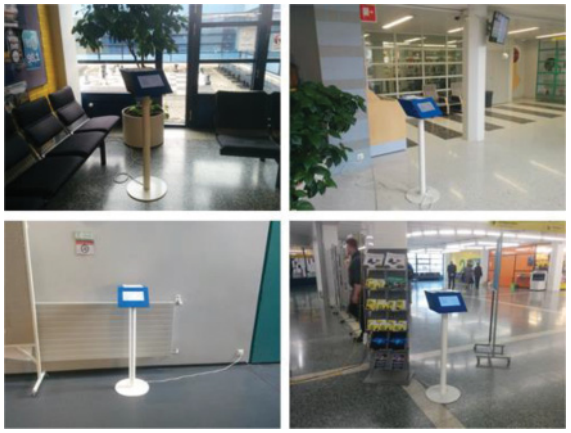

Fig. 1. The four kiosks used during our deployment.

devices and complete tasks to "kill" time [Goncalves et al. 2013; Huang 2015]. Situated crowdsourcing differs from other types of crowdsourcing substantially, offering a complementary—not replacement—means of enabling crowd work [Hosio et al. 2014].

An important limitation of crowdsourcing using such embedded input mechanisms, however, is that they need to rely on simple user interfaces and need to be effortless to use. They also need to support "walk up and use," so users can learn from others or start using them immediately [Brignull and Rogers 2003; Kukka et al. 2013; Hosio et al. 2014b]. On the other hand, a benefit of these technologies is that people approach them when they have free time and use them without clear motives [Müller et al. 2010]. Thus, situated technologies, like the kiosks used in our study, provide a crowdsourcing opportunity for people to spend their time and earn rewards. One such example of leveraging situated technologies for crowdsourcing purposes was Umati [Heimerl et al. 2012]. Umati used a vending machine with a touch display for locally relevant tasks and gave out snacks as rewards on task completion. Similarly, Hosio et al. [2014a] investigated the feasibility of a situated crowdsourcing market with a variety of tasks and worker payment. Their results showed that a situated crowdsourcing market can attract a populous workforce with comparable quality of contributions to its online and mobile counterparts while maintaining higher task uptake. We adopted their platform, Bazaar, due to its flexibility and the existence of a login mechanism that enables assigning virtual currency to workers that they can then exchange for tangible goods. This way we could focus on the design of the experiment instead of spending a great deal of time developing a platform.

3. PLATFORM DESCRIPTION

Our study was conducted on Bazaar, a situated crowdsourcing market [Hosio et al. 2014a]. A full description of the market is beyond the scope of our article, yet we include all the necessary details relevant to our study and findings. Bazaar has a virtual currency ("HexaCoins") that can be redeemed for goods or cash. It consists of a grid of physical crowdsourcing "kiosks" coordinated by a single network server that records in detail all user actions and completed tasks (Figure 1).

Each kiosk contains an Android tablet with a 10.1″ touch-screen, and a charger to keep the tablet always on, and uses WiFi to connect to the server. The tablets are set to "kiosk mode" [SureLock 2015] to ensure that the crowdsourcing software is always visible on screen, it recovers from crashes, and unwanted Operating System functionality

(notification bars, etc.) is disabled. The physical buttons of the tablet are physically obscured by the kiosk's enclosure. The welcome screen of the kiosks contains a brief introduction to the system and prompts users to login or create an account. Registration requires just a username and password. On login, users can work on new tasks and see whether their previous work has been approved. They can also review their HexaCoin balance, transfer them to another user, or exchange them for goods/cash.

3.1. Virtual Currency

Bazaar workers are rewarded with HexaCoins, which they can in turn exchange for goods or cash. When completing tasks, users receive HexaCoins subject to moderation by administrators. Moderation and rewarding take place in chunks. The value of HexaCoins is approximately 3,600 HexaCoins per hour of work, that is, workers expect to receive one HexaCoin per second of work. This value is influenced by contextual and cultural factors of the location where the platform is deployed, and therefore those do not follow online prices (e.g., Mechanical Turk) [Hosio et al. 2014a]. Users can ultimately exchange HexaCoins for goods, using a rough exchange rate of 360 HexaCoins per 1€. They can obtain cash in 10€ or 25€ packs and various other goods, including coffee vouchers and movie tickets. Previous work has shown that cash and movie tickets are typically the most popular items on this platform [Hosio et al. 2014a]. Workers email the administrators to schedule a pickup of the items, and these opportunities are used to conduct interviews with the workers.

4. STUDY

The aim of the study is to investigate whether we can successfully leverage local crowd contributions to obtain structured knowledge for arbitrary subjective questions. Our study focuses on questions that have multiple potential answers rather than fact-finding questions. The structured knowledge we wish to obtain for subjective questions includes the following:

—potential answers to the question (called "options"),
—criteria against which those options should be evaluated,
—a ranking of the top "options" (e.g., "top restaurants in a city").

Using a situated crowdsourcing market platform, we have built an automated analysis pipeline that generates tasks and collects crowd contributions in a series of steps. To verify the quality of the contributions, we compare the ultimate ranking of top "options" generated by our system to online repositories of similar nature and, where necessary, evaluate it by experts.

An important consideration regarding this study is that while the earlier work on Bazaar showed that it is possible to elicit a high amount of input rapidly using situated crowdsourcing, the system proposed here does not rely on this characteristic. It is designed to build a rich knowledge base of information over longer periods of time, and therefore it is not meant to address time-critical issues.

4.1. Methodology

The kiosks used in our experiment were distributed in four different locations (i.e., different Faculties) within our University campus. The four chosen locations had a steady flow of people passing by and were effectively busy walkways (i.e., main corridors). The campus has about 12,000 registered students and staff, but we expect that a subset of these visit the university on a daily basis. We did not actively promote Bazaar except by attaching an A4-sized poster on each of the kiosks. We specifically avoided the use of email lists, Facebook and Twitter, to minimise participation bias and relied mostly on word-to-mouth promotion and people serendipitously encountering the kiosks.

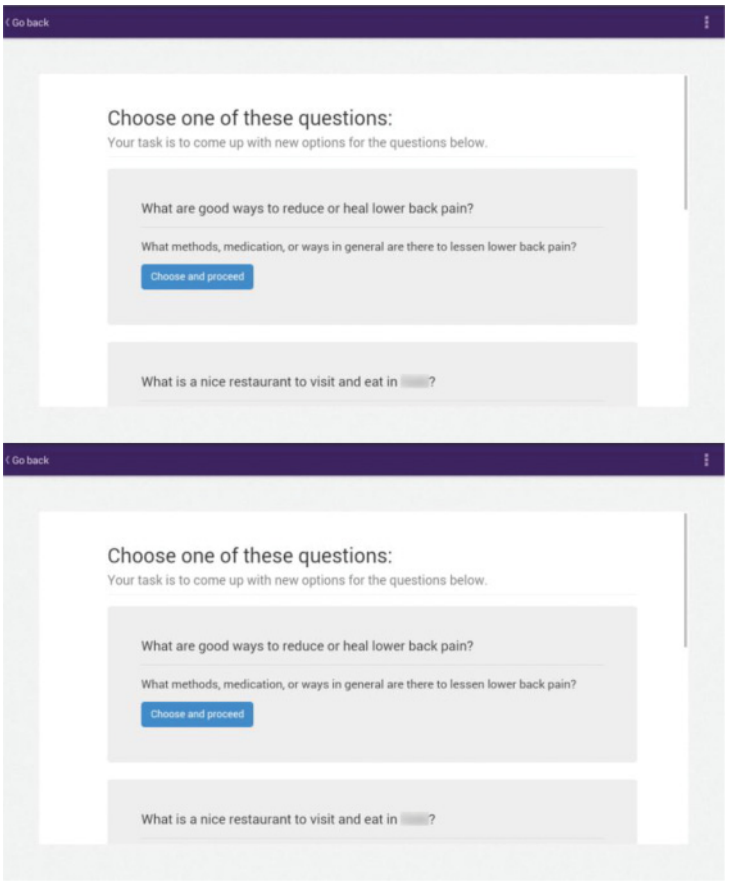

Fig. 2. Top: The choose question screen after selection the "Add new option" task. Bottom: After choosing the question, workers were presented with this screen to add their option and description. These two screens were very similar for the "Add new criterion" task with minor tweaks to the instructions.

Our experiment consisted of three different stages as follows: (1) collection of structured knowledge, (2) crowd moderation of generated options and criteria, and (3) collection of ratings for the different option and criteria pairs. Workers were presented with tasks from each stage in a sequential fashion. During this first stage, workers of Bazaar were presented with two tasks after logging into the market: "Add new option" and "Add new criterion." Both of these tasks rewarded the same amount of HexaCoins. After selecting either of these tasks, workers were asked to choose one of the five questions to which they would like to contribute (Figure 2, top). The questions were presented in random order to mitigate any selection bias beyond worker preference. The actual task required free-text entry of an "option"/"criterion" and a short explanation for it (Figure 2, bottom). The text in the figure is purposefully generic, because it was automatically generated by the system at runtime. Our goal during this stage was to gather at least 100 responses (options+criteria) and their respective descriptions for each of the five questions in the system. When that quota was completed, the system progressed to the next stage.

Previous work has shown that crowdsourcing deployments can reliably leverage the crowd to both provide and evaluate input [Callison-Burch 2009; Ipeirotis et al. 2010; Goncalves et al. 2014a]. In Stage 2, workers of Bazaar were presented only with a "Moderate" task on logging in. After selecting the task, workers had to choose which

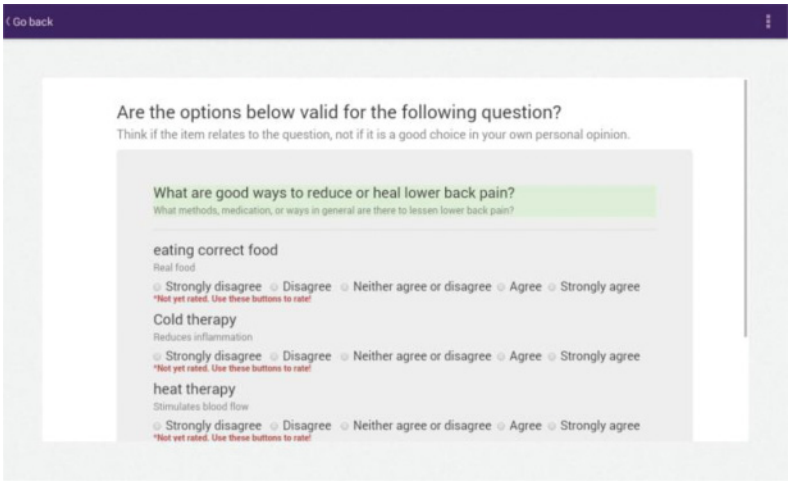

Fig. 3. Screen for moderating options for Q5 (back pain).

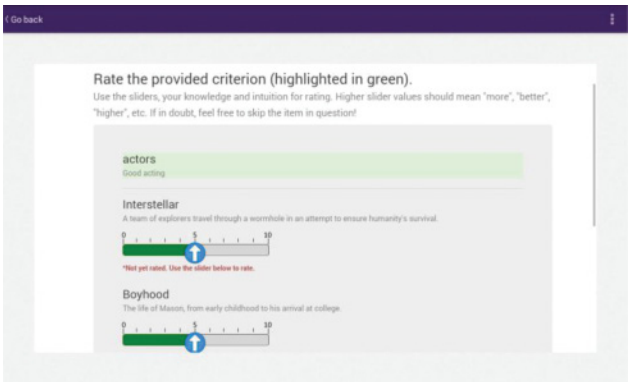

Fig. 4. Screen for rating different options based on certain criteria.

of the five questions they wanted to moderate and whether they wanted to moderate either their options or criteria. As with Stage 1, the questions were presented in random order to mitigate any selection bias beyond worker preference.

In Figure 3, we show the moderation screen for a particular question. Workers had to rate each criterion or option in terms of their validity using a 5-point Likert scale (Strongly disagree to Strongly Agree) for the shown question. Workers could potentially moderate all items from Stage 1, but each criterion or option could only be moderated once. The task description was generated dynamically by the system, and the system prioritised items with the least amount of moderation.

The final stage of the deployment closely mimics the rating functionality of online systems like TripAdvisor and IMDb. At this stage, the system had a list of potential options (i.e., solutions to the question), and it asked workers to rate those solutions in terms of the criteria generated by the crowd.

When workers logged into Bazaar, they could choose which question they wanted to rate. The questions were presented in random order to mitigate any selection bias beyond worker preference. Figure 4 shows a screenshot where workers could rate the acting quality for different movies. Workers were presented with three options (i.e., movies) at a time for each criterion, until all options had been rated for that criterion.

We choose not to present all options at the same time in order to avoid worker fatigue and the potential for workers to quit mid-task without submitting any ratings. This highlights an important tradeoff: If workers can either see all tasks simultaneously, then they can calibrate their ratings but they may find the task daunting. Alternatively, when workers see small subsets of the task, then they may not be able to calibrate their answers, but the work will appear to be more manageable. We opted for making the work manageable to attract sufficient contributions.

The rating interface showed sliders set by default to the middle (5), and red text stated, "*Not yet rated. Use the slider below to rate." If a worker did not move the slider, then this was considered a "skip." While Figure 4 shows ticks for integers between 0 and 10, the workers could place the slider in any position, giving the scale more granularity. The back end maps the values from 1 to 100. This flow is implemented to give the system and its users more flexibility in rating items while making it still visually easy to differentiate between two options that are very similar in the user's opinion (e.g., ratings 9.7 and 10). After all pairs were rated for a criterion, the next criterion with the least amount of ratings was presented, and the process would continue until no more pairs were available or until the worker decided to stop.

Finally, in addition to the tasks pertaining to these stages, each worker could complete a survey once they had completed at least 30 tasks in the system. It contained a standardised System Usability Scale (SUS) [Bangor et al. 2008] and an open-ended text field for users to provide comments on the tasks. Furthermore, we also interviewed face-to-face users who came to collect their prizes. We held semi-structured interviews [Rogers et al. 2011] driven by a pre-drafted protocol that elicited open-ended responses on themes such as preferred task, preferred question(s), and understandability of the tasks.

4.2. The Questions

We chose five questions for which the workers generated highly structured knowledge. Two of these questions were of local and semi-local nature, designed to demonstrate that local crowds can provide knowledge that does not exist in online platforms (TripAdvisor and Numbeo). Two other questions were of more general nature and were designed to demonstrate that local crowds can provide reliable knowledge when compared to reference online platforms (IMDb and Numbeo). The final question was aimed at demonstrating that structured knowledge from local crowds can provide useful information on how they perceive a certain issue. Furthermore, for quality check purposes we pre-added options and criteria to four of these questions. This ultimately allowed us to check how the local crowd would treat these during moderation, and to also check whether they would appear "on top" of the crowd's recommendations. These pre-added entries were gathered from specialised or specific websites (Numbeo, IMDb, American Movie Awards, restaurant websites), local experts, or from Wikipedia.

—*Question 1: What is a good city to live in <country>?*

Topic: Semi-local
Pre-added options: The three biggest cities in <country> plus the city where the deployment was conducted: <city1>, <city2>, <city3>, <city4>. We also added a short description for each, using a couple of sentences from each city's Wikipedia page.
Pre-added criteria: Four criteria randomly selected from Numbeo, the world's largest database of user-contributed data about cities and countries worldwide: Cost of Living, Crime, Quality of Life, and Traffic. A short explanation for each was generated from a couple of sentences of each criterion's Wikipedia page.

Table I. Number of Crowd-Generated Options and Criteria Added in the System for Each Question

	Q1	Q2	Q3	Q4	Q5	Total
Crowd-generated options	66	60	174	60	77	437
Crowd-generated criteria	35	40	38	40	32	185
Total	101	100	212	100	102	661

—*Question 2: What is a nice restaurant to visit and eat in <city>?*

Topic: Local
Pre-added options: The four top-rated and the four bottom-rated restaurants with reviews from TripAdvisor for the city of the deployment. A description was added from each restaurant's website or from a local expert.
Pre-added criteria: Four criteria used by TripAdvisor to rate restaurants: Food, Service, Value, and Atmosphere, along with their explanation from TripAdvisor.

—*Question 3: To which country should I move?*

Topic: General
Pre-added options: Four well-known countries in the world: United Kingdom, United States, India, and Russia, along with a couple of sentences from each country's Wikipedia page.
Pre-added criteria: Quality of Life, Economic Equality, and Gender Equality, along with a couple of sentences from each criterion's Wikipedia page.

—*Question 4: What is a good movie to watch from the last 2 years?*

Topic: General
Pre-added options: Four top-rated movies of the past two years on IMDb, and four hyped movies that were generally considered disappointments. We avoided the bottom-rated movies, as it would be unlikely that the targeted crowd would have watched them.
Top: Interstellar, The Wolf of Wall Street, Boyhood, and Gone Girl.
Hyped, but disappointing: The Lone Ranger, The Amazing Spiderman 2, Godzilla, and Teenage Mutant Ninja Turtles.
Pre-added criteria: Taken from the American Movie Awards judging criteria [American Movie Awards 2017]: Plot, Characters, Originality, and Entertainment Value, along with a couple of sentences from each criterion's Wikipedia page.

—*Question 5: What are good ways to reduce lower back pain?*

Topic: General
Pre-added options and criteria: We decided not to add any expert options or criteria to this question, as we wanted to build a knowledge base solely from crowd contributions.

5. ANALYSIS AND RESULTS

In total, 72 unique workers contributed 22,077 unique tasks during our deployment. The majority of tasks were completed between 8am and 6pm, which coincides with lecture periods at our University. Session length varied between a few seconds to a couple of hours of completing tasks. Next, we describe in detail the results of each of the deployment.

5.1. Stage 1: Collection of Structured Knowledge

As expected, some questions received more attention than others. This led us to remove a question from the system when it had achieved its target of 100 entries, thus ensuring that all questions reached their quota eventually. Table I shows the number of collected

Table II. Number of Crowd-Generated Options and Criteria Added in the System for Each Question

	Q1	Q2	Q3	Q4	Q5	Total
Crowd moderations (options)	439	546	952	353	439	2729
Crowd moderations (criteria)	173	310	266	306	152	1207
Total	612	856	1218	659	591	3936

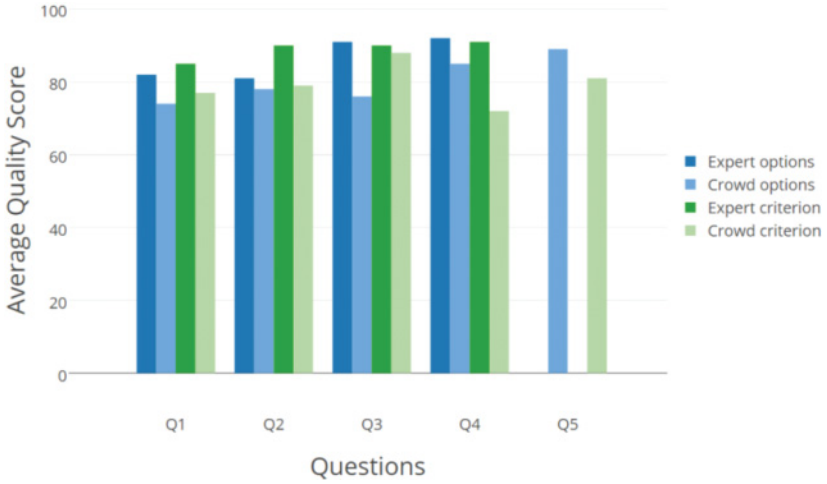

Fig. 5. Average quality score of expert and crowd-generated options and criteria for each question.

options and criteria for each question. A total of 31 workers contributed towards this stage during 3 days of deployment. We gathered more contributions for Question 3 since workers leaned heavily towards this question on one afternoon, and questions were removed only at midnight.

Some examples of options added to questions are as follows: Question 3 (Germany, Brazil, Sweden), Question 4 (*Frozen*, *Annabelle*, *The Lego Movie*), Question 5 (Massage, Saddle Chair, Exercise). Examples of criteria include the following: Question 1 (Size, Accessibility, Nature), Question 2 (Decoration, Selection, Location), Question 3 (Weather, Language, Employment), Question 4 (Cast, Dialogue, Realism), and Question 5 (Effort, Affordability, Effectiveness).

5.2. Stage 2: Crowd Moderation of Generated Options and Criteria

A total of 35 workers contributed towards this stage, of which 14 were returning workers from Stage 1. This stage lasted 5 days of deployment (2 of which were during the weekend, with limited activity). Table II shows the number of moderation tasks completed per question.

Using these contributions, we calculated a "quality score" for each criterion and option. The Likert scale ratings were mapped to 0, 25, 50, 75, and 100, respectively, for back-end consistency purposes. The quality score for a criterion or option was the average of the scores given by the workers. Figure 5 shows the average quality score for all items. In this figure, we group separately those options and criteria that we pre-added to the system: We label all those as "expert" contributions, while the rest are labelled as crowd contributions. As expected, both the expert options ($U = 6616.5$, $p = 0.02$) and criteria ($U = 1916$, $p < 0.01$) had significantly higher-quality scores.

In Table III, we show some examples of options and criteria that were of high (>85) or low (<60) quality. Workers downvoted the validity of gibberish inputs (bbbb, vvvv). In addition, wrongly allocated items were also downvoted by the crowd. This could be

Table III. Examples of High- and Low-Validity Options and Criteria for Each Question (Cities and Restaurant Have Been Anonymized)

	Q1	Q2	Q3	Q4	Q5
High validity options	CityA CityB CityC	RestaurantA RestaurantB RestaurantC	Japan Canada S. Korea	Fast & Furious 7 Interstellar Frozen	Massage Foam roller Good posture
Low validity options	bbbb Pokka	vvvv Sweden	Nature Culture	Music Blood Heart	Lada Talk therapy
High validity criteria	Healthcare Population Prices	Cleanliness Food Selection	Social Benefits Security	Characters Director Sound	Price Ease Effectiveness
Low validity criteria	Beer Price "Free time are important for students"	RestaurantX RestaurantY	Estonia	Interstellar Wu Ze Tian	Exercise Sleep Acupuncture

Table IV. Number of Options and Criteria Remaining in the System after Filtering. In Parentheses Is the Number of Removed Options and Criteria Due to Poor Quality or Duplication

	Q1	Q2	Q3	Q4	Q5	Total
Number of options	20 (46)	38 (22)	53 (121)	48 (12)	33 (44)	192 (245)
Number of criteria	23 (12)	18 (22)	28 (10)	19 (21)	11 (21)	86)

options added as criteria (or vice versa) or options/criteria added to the wrong question. For example, for Q3, Estonia constitutes a valid option but was wrongly inserted as a criterion, and Sweden was wrongly added as an option to Q2 instead of Q3. These results demonstrate that the moderation process worked well, appropriately filtering erroneous inputs from the crowd.

Based on the crowd moderation, we excluded all criteria and options that obtained a quality score below 75 as this marks the value for the "Agree" selection on the used Likert scale. Next, the system removes duplicate entries by applying Approximate String Matching to the remaining criteria/options using a generalized Levenshtein edit distance: the minimal possibly weighted number of insertions, deletions, and substitutions needed to transform one string into another [Navarro 2001]. If an option or criterion had duplicate entries at this point, then the system retained the one with the highest quality score. The assumption behind this decision is that we expect the description of that particular entry to be better, which we manually confirmed was the case for the majority of cases. Pre-added entries tended to get selected over their crowd generated counterparts (similar options and criteria, likely due to their more thorough descriptions. The amount of options and criteria remaining in the system after this process is reported in Table IV below. The number in parentheses denotes the total number of options and criteria removed by moderation and filtered out due to repetition.

5.3. Collection of Ratings

A total of 37 workers contributed towards this stage, of which 17 were returning workers from the previous stages. Table V below sums up the key results from this stage, including the number of unique pairs (options x criteria) available per question and the number of ratings submitted.

The final output of the system is a ranked list of *options* vis-à-vis the criteria identified by the crowd. To evaluate the quality of this ranked list, we compare it to lists obtained

Table V. Number of Available Unique Pairs and Rated Pairs for Each Question

	Q1	Q2	Q3	Q4	Q5	Total
Available pairs	460	684	1484	912	363	3903
Rated pairs	3653	3405	5850	2573	1999	17480

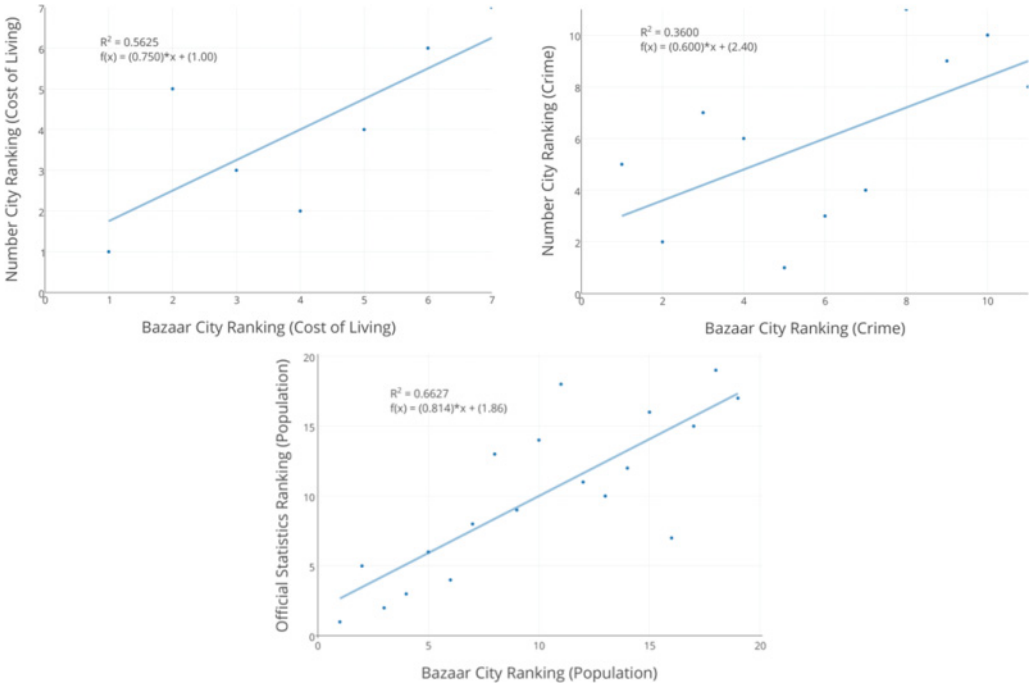

Fig. 6. Comparison of our system's city rankings vs. Numbeo and official statistics based on three different criteria (Cost of Living, Crime, Population).

from online systems. There are multiple ways to carry out such an assessment, for example, by considering whether the list is exhaustive (i.e., all potential answers have been taken into account) or whether the true best answer has been ranked at the top. However, both of these approaches test the capability of the system to exhaust the solution space rather than the quality of the recommendations. To evaluate the quality of the ranking generated by our system, we consider the set of options included in our system's ranking and check how *they* were ranked by our system vs. an online system. For instance, for the list of movies recommended by our system, we check the rank of each movie in our system vs the rank of the movie in IMDb (considering only those movies included in our system). Effectively, this is a Spearman's rank correlation procedure, which considers the same set of movies across two systems. This enables us to validate the quality of the ranking but does not explicitly account for how exhaustive the list may be.

Below we show scatterplots of the ranking of options generated by our system vs. those found in various online systems. For example, we show a scatterplot of city rankings (Q1), demonstrating how cities were ranked in our system vs. Numbeo based on certain criteria (Figure 6). The scatterplots only consider options and criteria that are found in both systems. This is because some cities and criteria in our system were actually not present in Numbeo and vice versa. In addition, in some cases, while the options and criteria existed in both systems, there was not enough data on Numbeo.

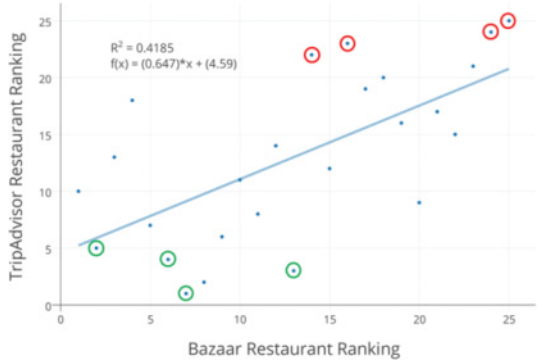

Fig. 7. Comparison of our system's restaurant rankings vs. TripAdvisor. Points marked on the bottom left (green circles) are the pre-added restaurants from the top of the list in TripAdvisor, while points market on the top right are pre-added restaurants from the bottom of the list in TripAdvisor.

Hence, we chose only the criteria that had sufficient contributions on Numbeo (Cost of Living, Crime), as well as "Population" and compared it to official statistics. Similarly, the scatterplots (Figure 8) with country rankings (Q3) also rely on Numbeo indexes where again we only compare criteria which are presented on both systems (Quality of Life, Cost of Living, Crime). For Q2 (restaurants) and Q4 (movies), we generated a single overall ranking by aggregating the ratings of individual criteria. This enabled us to directly compare the restaurant and movie rankings produced by our system to the rankings from TripAdvisor (Figure 7) and IMDb rankings (Figure 9), respectively. In addition, we also show where our pre-added options ended up. We note that some of the rankings in TripAdvisor and IMDb changed slightly from when we initially choose the movies and restaurants to when we conducted the ranking analysis.

For Q1 there was a positive correlation between our system's ranking and Numbeo for the Quality of Life and Crime criteria ($R^2 = 0.56$ and 0.36). Similarly, when comparing the cities rankings regarding the population criterion to official statistics there was a strong positive correlation ($R^2 = 0.66$). Regarding Q2, the overall ranking of restaurants in our system positively correlated with the rankings obtained from TripAdvisor ($R^2 = 0.42$). Similarly to Q1, we compared the ranking of options depending on multiple criteria of Q3 to Numbeo. We found a strong positive correlation between the two systems for Quality of Life ($R^2 = 0.69$) and Cost of Living ($R^2 = 0.60$). However, there was no correlation between both systems in regards to crime rankings ($R^2 = 0.05$). Finally, for Q4, we compared the overall ranking of movies on our system and IMDb and found a positive correlation ($R^2 = 0.57$).

For Q5 (back pain), we relied on the expert assessment of a physician that specialises on back-related ailments. We held a half-day workshop with him in an attempt to better understand how structure contributions from a local community, in this case university students, could provide useful information on how they perceive a certain issue. During the workshop we provided an overview of the system and its purpose, a summary of the obtained results, conducted a card sorting activity, followed by a lengthy discussion.

During card sorting we showed him index cards, with one per option added by the crowd ($N = 33$). We then ask the physician to cluster them in terms of their effectiveness. We chose this particular criterion because the remaining criteria chosen by the crowd were easily verifiable without in-depth knowledge from an expert (e.g., affordability, equipment, independence). The expert immediately excluded certain options such as Vitamin D, Vitamin B12, Magnesium, and Creams, since they do not help with back-pain-related issues. These options were ranked #30, #31, #32, and #21 in

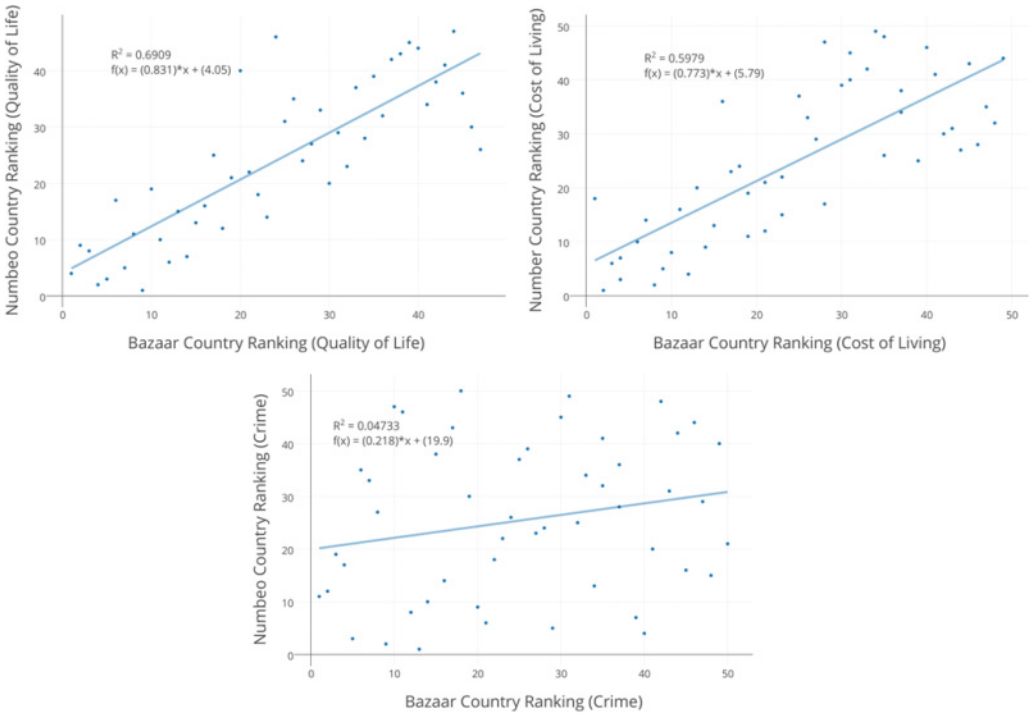

Fig. 8. Comparison of our system's country rankings vs. Numbeo based on three different criteria (Quality of Life, Cost of Living, Crime).

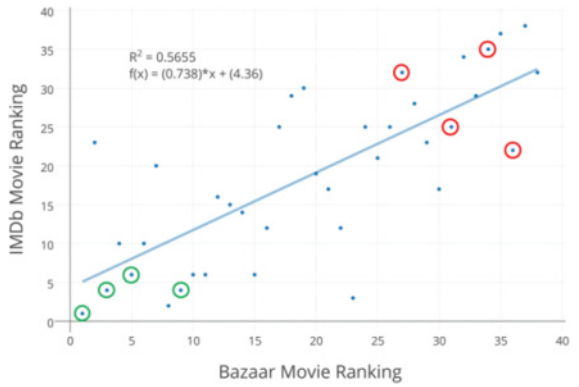

Fig. 9. Comparison of our system's movie rankings vs. IMDb. Points marked on the bottom left (green circle) are the pre-added movies from the top of the list in IMDb, while points market on the top right are pre-added movies that were hyped but ultimately considered a disappointment.

our system, respectively. He mentioned that these options do not help in the slightest with back-pain-related issues. Certain options were clustered together as only having a marginal benefit: "Heat Therapy," "Cold Therapy," "Foam Roller," and "Hot Packs." These options were ranked #10, #18, #19, and #20 in our system, respectively. The remainder of the options were classified in two groups: as effective back pain preventers (e.g., exercise, saddle chair, maintaining a good posture) or as effective back pain symptom relievers (e.g., painkillers, massage, acupuncture), all of which ranked highly

in our system. Interestingly, the crowd ranked "Surgery" as the most effective option. While the expert agreed that it can be very effective in certain cases, it should only be considered in extreme cases due to the associated risks. He expressed concern that the community members might think that surgery is a magical pill that can instantly solve back pain issues, which is definitely not the case.

After completing this card sorting activity, the back pain expert stated that it would be useful to know what local people think are good ways to alleviate back pain and in what way. Back pain issues are an increasingly common problem with students and young people in general due to long periods of sitting and computer use.

In the end, the expert expressed that by learning about common misconceptions related to back pain within this local community regarding several treatments and their characteristics (effectiveness, independence, price, etc.), health officials can then tackle this issue with more confidence. As an example, the physician mentioned that with such information they could make better-informed interventions within the community (pamphlets, posters, guest lectures, etc.).

5.4. Survey and Interviews

In total, 46 workers completed the survey task. Analysis of the SUS revealed an overall score of 77 (SD = 17.6) on a scale from 0 to 100. There were some complaints regarding loading times due to occasional poor WiFI connectivity, which negatively impacted the scores given in the SUS survey. The positive statement with the lowest value was whether users found the various functions in the system to be well integrated (M = 3.8, SD = 1.0). Other values showed that users did not consider that they needed to learn a lot of things before they could get going with the system (M = 1.7, SD = 1.0), found that it was easy to use (M = 4.2, SD = 0.8), can quickly be learned (M = 4.3, SD = 0.8), and requires no technical support (M = 1.6, SD = 0.9). Overall, users stated that the system was good and that the idea was interesting. Some users mentioned that the system lacked task diversity, which led to them being bored with it after a while. Others enjoyed the fact that the tasks kept changing and were building on each other. Finally, 15 workers (9 male, 6 female) were interviewed when they picked up their prizes. Their average age was 24.7 (SD = 2.6). The key findings from the interviews are used to support our discussion.

6. DISCUSSION

6.1. Obtaining Structured Crowd Contributions

There are two crucial aspects to obtaining crowd contributions: quantity and quality. In line with previous work, our deployment was able to rapidly gather a substantial amount of input across all stages. During Stage 1, we collected over 600 items in just 2 days, with a clear preference towards adding new options as opposed to criteria (Table Stage1). In our interviews, the majority of workers reported that they felt that options were much easier to think of than criteria (14 of 15 interviewees).

> *"Options is much easier, has many potential answers making it easier to do fast. Criteria is more difficult to think about." - P2*

> *"Options. They 'exist' so it's easy to come up with them. The criterion one demands much more creativity and is harder to come up with new ones to." - P6*

This result was expected, as for subjective questions there is a larger number of potential options than criteria. We purposely had both these tasks reward the same amount of coins and expected that workers would lean more towards adding options. This is consistent with the claim of cognitive load theory [Sweller 1988] that, *ceteris paribus*, tasks with higher cognitive load are less likely to be completed than tasks with

lower cognitive load. Literature on crowdsourcing has also shown that fewer difficult tasks are more likely to be completed when considering the same reward [Horton et al. 2011].

Regarding quality, given that we were offering rewards, our tasks were more susceptible to gaming behaviour, particularly the open-ended inputs in Stage 1. Gaming behaviour is usually exhibited by workers in an attempt to reap rewards with minimum effort [Downs et al. 2010]. We adopted a crowd-moderation strategy, which has been proven effective in the past [Lampe et al. 2014; Goncalves et al. 2014a]. In just 3 days, we gathered almost 4000 crowd moderations that successfully filtered out "noisy" input from the crowd. Gibberish words, items added to the wrong question, and options wrongfully added as criteria (and vice versa) were all identified by the crowd as invalid contributions.

During our interviews, workers reported learning new things during Stage 2 and 3 when they could see others' contributions. Earlier, similar remarks have been made for example, in the context of providing technical support for open-source software (Apache) online, where some of the more active participants reported gaining noticeable learning benefits by simply browsing the content that others had already contributed [Lakhani and von Hoppel 2003]. Some of the most relevant examples given by workers in our case include:

"Some of the options and criteria people came up with surprised me as I had never thought about, like Zumba for back pain. Did not even know what Zumba was." - P11

"It was nice learning that there is a Nepalese restaurant in town that I did not know about. Tried to look for Nepalese food on TripAdvisor but did not find any." - P14 (from Nepal)

"Many movies I never heard about and some of them had interesting description, so I might look them up." - P15

Similarly, Numbeo lacked indexes for certain criteria for many of the cities and countries added in our system, while IMDb only offers general rankings. Hence, we demonstrate that trough a structured knowledge system like ours, people can arrive at specific rankings that cater more to their own personal preferences.

6.2. Generating Structured Knowledge In Situ

An important feature of our deployment is that it was conducted using a situated technology with all the advantages and disadvantages that come with it [Goncalves et al. 2013; Goncalves et al. 2015]. Online crowdsourced systems like TripAdvisor and Numbeo can sometimes lack information from certain locations. For instance, on TripAdvisor, smaller cities with less touristic activity may lack contributors or some proprietors might not have listed themselves on the platform. Our decision to use a situated crowdsourcing platform meant that we were able to gather knowledge that does not exist in these online platforms. In Q2 (restaurants), there were 13 restaurants in our system that do not exist in TripAdvisor. They were student-focused pizzerias, kebab joints, and old family businesses. Further, the majority of cities that were added to the system in Q1 had very little or no information on Numbeo. Also, since we demonstrate that a situated crowdsourcing platform can reliably gather structured knowledge, this opens up the possibility to explore highly localized questions with more confidence ("In which garage should I service my car?" and "Which park is the best to walk my dog").

Another aspect we were interested in was how workers would approach and choose to complete the various questions. A few of the local workers reported that they had an easier time completing the tasks that were locally relevant (Q1: cities, and Q2: restaurants) in Stages 1 and 3.

"Local questions were the easiest (city, restaurants) as I have a lot of knowledge about it. I got everything only from my head." - P7

"City and restaurant one was very easy for me because I am local. Did a few of the rest." - P11

On the other hand, some of the foreigners (i.e., non-native workers) gravitated more towards other questions (Q3: countries and Q4: movies). This finding strongly resonates with the theory of psychological distance, which suggests that people are more likely to engage in tasks that are "close" to them psychologically [Liberman et al. 2007]. However, the same effect was not observed in Stage 2 (moderation), as workers reported having no preference in terms of questions, as it only required them to rate the validity of options and criteria added by others.

"Most of them were very good so it was easy. I ended up doing a bit of everything." - P2

"It was about if they are valid options or criterion. No preference in what question. Sometimes challenging to choose which rating (descriptions changed a lot)." -P7

Finally, several workers reported that they would enjoy getting suggestions from such a structured knowledge system. However, they would only trust a generic crowd with more banal decisions (e.g., what movie to watch) but not with more important decisions like deciding to what city or country to move. In those cases, they would need to be assured that the opinions came from local people or people that have more concrete knowledge.

"I would use this for non-important decision (movies for example), not for important decisions. For more specific I would like specific people to give the criteria/rate (e.g., people that visited or live in a country)" - P2

6.3. Quality of Ratings

In the final stage, we investigated the quality of ratings given the crowd-generated and crowd-moderated options and criteria. Workers reported that the task description and/or the sliders made it clear that what was being asked was their own subjective opinion towards the option/criteria pairs. As hypothesized, pre-added options for Q2 and Q4 gravitated towards opposite sides of the quality spectrum (i.e., top and bottom rankings, Figures 7 and 9).

Overall, the rankings produced by our system correlated well with rankings in other crowd-powered systems (TripAdvisor, IMDb, Numbeo). Interestingly, the more local question (Q2) had a relatively low correlation with the reference online platform. A potential explanation for this is that locals and tourists perceive places differently [Wu et al. 2011]. For example, the #1 ranked restaurant on TripAdvisor is a fine-dining establishment serving local dishes. Users of our system ranked this restaurant much lower, as they most likely perceive it as a touristy restaurant.

It is very interesting to consider the shortcomings of our results. In addition to the lower correlation in Q2, the crime ranking for countries also did not correlate well with Numbeo, meaning that the perceptions of the crowd differed from the numbers shown in Numbeo. Similarly, most scatterplots in our results contain substantial outliers, that

is, options for which our crowd had a much different opinion. One explanation for these results could be a "calibration bias" that workers most likely exhibited. Since each worker was asked to rate a subset of items at any given time, we can expect that some of their early ratings may not have been calibrated. However, we also did observe that for some criteria where crowd contributions could be compared to objective ground-truth measurements (e.g., the population of cities), the results were not poor overall ($R^2 = 0.66$).

What is more interesting to investigate is particular outliers in the scatterplots. To the extent that the scatterplots represent our crowd's assessment, outliers in the scatterplots represent crowd preferences towards particular options or criteria. These preferences may, in some case, actually be useful information to have about the crowd. For instance, the physician who evaluated our results from Q5 noted that while clinical care guidelines inform healthcare professionals, they are one sided and do not account for patients' perceptions. In that regard, it was extremely insightful for him to identify treatment options that patients (incorrectly) prefer or value. Similarly, he noted that patients may want to hear the opinions of other patients, and that is why online patient forums thrive. However, while forums allow patients to share perceptions in an unstructured manner, our crowdsourcing approach can impose structure, which in turn can reveal interesting insights.

Another example of preferences of this crowd was observed in Q4 (movies). One outlier in our results was *Hunger Games*, which was rated much higher in our system than IMDb. This can be attributed to the age demographic of our crowd, which consisted mostly of young adults. Conversely, *Song of the Sea* faired highly in IMDb but not our system, most likely because it did not appeal to our crowd. This suggests that by positioning situated crowdsourcing kiosks in different locations, it is possible to gather demographic-specific structured knowledge.

Finally, while several workers reported that Q5 (back pain) was the hardest one to come up with options and criteria (Stage 1), it was the easiest to rate (Stage 3). An explanation for this was mentioned in our interviews in that this question does not require them to have physically visited certain places or watched specific movies to be able to give ratings. Hence, it was easier for workers to simply rate objects (foam roller, creams) and activities (yoga, massage, exercise) based on the presented criteria:

> *"I found the back pain one the simplest on this stage (was most difficult for me on first stage). No need to have knowledge of specific movies or countries." - P2*

> *"I knew how to answer all of these (back pain), but the others I missed some knowledge to do them properly." - P10*

6.4. Limitations

While we collected a large amount of options and criteria, the produced lists were not exhaustive (e.g., not all movies from the past 2 years or restaurants in the city were added). Hence, our analysis was based on the quality of the inputted contributions and not on how exhaustive the lists were. Furthermore, it is entirely possible that certain pairs performed badly due to lack of knowledge from the workers ultimately influencing our ranked lists (e.g., Crime criterion for Q3). We also acknowledge that the mechanism presented in this article is not ideal for time-critical questions that require a quick turnaround. This is mainly caused by the fact that the pool of immediately available workers is smaller than the workforce present in online crowdsourcing markets. Nevertheless, situated crowdsourcing brings a set of key affordances that effectively complement other means of crowd work.

Also, we note that a larger crowd would most likely lead to improved results. In this regard, a key limitation of our approach is that certain option-criterion pairs may have had limited input, for example, rated only by four workers. It is not clear to what extent this could lead to error in the rankings. Therefore, we should determine some contribution threshold below which option-criterion pairs should not be considered. However, we note that overall the system performed well when compared to other crowd-powered platforms.

Finally, even though we took steps to make the instructions as clear as possible in the different stages, there might have been instances where workers misunderstood what was being asked. However, the majority of interviewed workers ($N = 14$) stated that what was being asked was clear, and the crowd-moderations in Stage 2 appropriately filtered out any mistakes from Stage 1. Even so, we acknowledge that some workers may have answered Stage 3 in a non-serious way or may have been influenced by the discrepancy in descriptions of the options and criteria between selected pre-added entries and crowd-generated entries, which may have had an impact on our results. Finally, the WiFi connectivity of the tablets within their enclosures occasionally failed, which led to some frustration.

7. CONCLUSION AND ONGOING WORK

This article has outlined a method to systematically elicit structured knowledge from situated crowds, which can be used in a variety of contexts such as decision support [Hosio et al. 2016]. Our crowdsourcing approach involves a series of steps whereby situated crowds elicit, moderate, and rate options and criteria for arbitrary questions. We have shown that this is a flexible approach that can generate recommendations for a variety of issues. Our evaluation has contrasted our results against those of online recommendation systems. While the performance of the system is acceptable, we have highlighted that it may also reflect crowd preferences, most likely due to the situated nature of the crowd. In our ongoing work, we are investigating a number of important issues that this article has not addressed. First, we are exploring whether the user interface for data entry is likely to bias the quality of data and ratings collected from workers. Furthermore, we are investigating the use of classifiers in eliciting recommendations: In the present article, we have only considered the generation of ranked lists, but it is also possible to implement and compare multiple classification methods to elicit recommendations. We are also taking steps to further automate the workflow, for instance, developing a more intelligent algorithm to transition a question from one stage to the other (e.g., when the added items reach a certain point of saturation). Finally, we are exploring whether users perceive such recommendations as trustworthy, given that they are produced from an elaborate crowdsourcing process.

REFERENCES

Aaron Bangor, Philip Kortum, and James Miller. 2008. An empirical evaluation of the system usability scale. *Intl. J. Hum.–Comput. Interact.* 24, 6, 574–594.

American Movie Awards. 2017. Judging criteria. Retrieved from https://www.americanmovieawards.com/enter-now/judging-criteria-sub.

Michael S. Bernstein, Greg Little, Robert Miller, and Björn Hartmann, et al. 2010. Soylent: A word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology* (UIST'10), 313–322.

Harry Brignull and Yvonne Rogers. 2003. Enticing people to interact with large public displays in public spaces. In *Proceedings of INTERACT*, 17–24.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, 286–295.

Li Chen and Pearl Pu. 2011. Critiquing-based recommenders: survey and emerging trends. *User Model. User-Adapt. Interac.* 22, 1–2, 125–150.

Justin Cheng and Michael Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 600–611.

Julie Downs, Mandy Holbrook, Steve Sheng, and Lorrie L. F. Cranor. 2010. Are your participants gaming the system?: Screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, 2399–2402.

Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. 2013. Crowdsourcing on the spot: Altruistic use of public displays, feasibility, performance, and behaviours. In *Proceedings of UbiComp'13*, 753–762.

Jorge Goncalves, Simo Hosio, Denzil Ferreira, and Vassilis Kostakos. 2014a. Game of words: Tagging places through crowdsourcing on public displays. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS'14)*, 705–714.

Jorge Goncalves, Pratyush Pandab, Denzil Ferreira, Mohammad Ghahramani, Guoying Zhao, and Vassilis Kostakos. 2014b. Projective testing of diurnal collective emotion. In *Proceedings of UbiComp'14*, 487–497.

Jorge Goncalves, Simo Hosio, Jakob Rogstadius, Evangelos Karapanos, and Vassilis Kostakos. 2015. Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. *Comput. Netw.* 90, 34–48.

Jorge Goncalves, Hannu Kukka, Iván Sánchez, and Vassilis Kostakos. 2016. Crowdsourcing queue estimations in situ. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW'16)*, 1040–1051.

Bin Guo, Zhu Wang, Zhiwen Yu, Yu Wang, Neil Y. Yen, Runhe Huang, and Xingshe Zhou. 2016. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Comput. Surv.* 48, 1, 7.

Kurtis Heimerl, Brian Gawalt, Kuang Chen, Tapan Parikh, and Björn Hartmann. 2012. CommunitySourcing: Engaging local crowds to perform expert work via physical kiosks. In *Proceedings of the SIGCHI* Conference *on Human Factors in Computing Systems (CHI'12)*, 1539–1548.

John Horton, David Rand, and Richard Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Exp. Econ.* 14, 3, 399–425.

Simo Hosio, Jorge Goncalves, Vili Lehdonvirta, Denzil Ferreira, and Vassilis Kostakos. 2014a. Situated crowdsourcing using a market model. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST'14)*, 55–64.

Simo Hosio, Jorge Goncalves, Vassilis Kostakos, and Jukka Riekki. 2014b. Exploring civic engagement on public displays. In *User-Centric Technology Design for Nonprofit and Civic Engagements*, Saqib Saeed (Ed.). Springer, 91–111.

Simo Hosio, Jorge Goncalves, Vassilis Kostakos, and Jukka Riekki. 2015. Crowdsourcing public opinion using urban pervasive technologies: Lessons from real-life experiments in oulu. *Pol. Internet* 7, 2, 203–222.

Simo Hosio, Jorge Goncalves, Theodoros Anagnostopoulos, and Vassilis Kostakos. 2016. Leveraging the wisdom of the crowd for decision support. In *Proceedings of the 2016 British HCI Conference* (British HCI'16).

Yi-Ching Huang. 2015. Designing a micro-volunteering platform for situated crowdsourcing. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (CSCW'15)*, 73–76.

Panagiotis Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP'10)*, 64–67.

Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert R. E. Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST'11)*, 43–52.

Hannu Kukka, Heidi Oja, Vassilis Kostakos, Jorge Goncalves, and Timo Ojala. 2013. What makes you click: Exploring visual signals to entice interaction on public displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1699–1708.

Karim R. Lakhani and Eric von Hippel. 2003. How open source software works: "free" user-to-user assistance. *Res. Policy* 32, 6, 923–943.

Cliff Lampe, Paul Zube, Jusil Lee, Chul C. H. Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Gov. Inf. Quart.* 31, 2, 317–326.

M. D. Lee, B. M. Pincombe, and M. B. Welsh. 2005. An empirical evaluation of models of text document similarity. *Cogn. Sci.* 1254–1259.

Nira Liberman, Yaacov Trope, and Elena Stephan. 2007. Psychological distance. *Soc. Psychol.: Handb. Basic Princip.* 2, 353–383.

Jörg Müller, Florian Alt, Daniel Michelis, and Albrecht Schmidt. 2010. Requirements and design space for interactive public displays. In *Proceedings of the International Conference on Multimedia*, 1285–1294.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.* 33, 1, 31–88.

D. J. Navarro and M. D. Lee. 2004. Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonom. Bull. Rev.* 11, 6, 961–974.

Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Gajos. 2011. Platemate: Crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST'11)*, 1–12.

Yvonne Rogers, Helen Sharp, and Jenny Preece. 2011. *Interaction Design: Beyond Human-Computer Interaction.* John Wiley & Sons.

Clay Shirky. 2010. Cognitive surplus: how technology makes consumers into collaborators. *Penguin*. SureLock. Retrieved from www.42gears.com/surelock/

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* 12, 2, 257–285.

A. Tversky. 1977. Features of similarity. *Psychol. Rev.* 84, 4, 327–352.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*, 319–326.

Wikidata. Retrieved July 22, 2015 from https://www.wikidata.org/wiki/Wikidata:Main_Page

Shaomei Wu, Shenwei Liu, Dan Cosley, and Michael Macy. 2011. Mining collective local knowledge from google MyMaps. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW'11)*, 151–152.

Received February 2016; revised June 2016; accepted October 2016