Report from Dagstuhl Seminar 15481

Evaluation in the Crowd: Crowdsourcing and Human-Centred Experiments

Edited by

Daniel Archambault[1], Tobias Hoßfeld[2], and Helen C. Purchase[3]

1 **Swansea University, GB,** D.W.Archambault@swansea.ac.uk
2 **University of Duisburg-Essen, DE,** tobias.hossfeld@uni-due.de
3 **University of Glasgow, GB,** helen.purchase@glasgow.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15481 "Evaluation in the Crowd: Crowdsourcing and Human-Centred Experiments". Human-centred empirical evaluations play important roles in the fields of human-computer interaction, visualization, graphics, multimedia, and psychology. The advent of crowdsourcing platforms, such as Amazon Mechanical Turk or Microworkers, has provided a revolutionary methodology to conduct human-centred experiments. Through such platforms, experiments can now collect data from hundreds, even thousands, of participants from a diverse user community over a matter of weeks, greatly increasing the ease with which we can collect data as well as the power and generalizability of experimental results. However, such an experimental platform does not come without its problems: ensuring participant investment in the task, defining experimental controls, and understanding the ethics behind deploying such experiments en-masse.

The major interests of the seminar participants were focused in different working groups on (W1) Crowdsourcing Technology, (W2) Crowdsourcing Community, (W3) Crowdsourcing vs. Lab, (W4) Crowdsourcing & Visualization, (W5) Crowdsourcing & Psychology, (W6) Crowdsourcing & QoE Assessment.

Seminar November 22–27, 2015 – http://www.dagstuhl.de/15481

1998 ACM Subject Classification H.1.2 User/Machine Systems, H.4 Information Systems Applications, H.5 Information Interfaces and Presentation, H.5.1 Multimedia Information Systems, H.5.3 Group and Organization Interfaces, J.0 General Computer Applications, K.4 Computers and Society, K.4.3 Organizational Impacts

Keywords and phrases Crowdsourcing, Human Computation, Crowdsourcing Design, Mechanisms, Engineering, Practical Experience, Computer Graphics, Applied Perception, HCI, Visualization

Digital Object Identifier 10.4230/DagRep.5.11.103

Edited in cooperation with Fintan McGee (Luxembourg Institute of Science and Technology, LU, fintan.mcgee@list.lu)

Evaluation in the Crowd: Crowdsourcing and Human-Centred Experiments, *Dagstuhl Reports*, Vol. 5, Issue 11, pp. 103–126

Editors: Daniel Archambault, Tobias Hoßfeld, and Helen C. Purchase

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Daniel Archambault
Tobias Hoßfeld
Helen C. Purchase

In various areas of computer science like visualization, graphics, or multimedia, it is often required to involve the users, e.g. to measure the performance of the system with respect to users, e.g. to measure the user perceived quality or usability of a system. A popular and scientifically rigorous method for assessing this performance or subjective quality is through formal experimentation, where participants are asked to perform tasks on visual representations and their performance is measured quantitatively (often through response time and errors). For the evaluation of the user perceived quality, users are conducting some experiments with the system under investigation or are completing user surveys. Also in other scientific areas like psychology, such subjective tests and user surveys are required. One approach is to conduct such empirical evaluations in the laboratory, often with the experimenter present, allowing for the controlled collection of quantitative and qualitative data. Crowdsourcing platforms can address these limitations by providing an infrastructure for the deployment of experiments and the collection of data over diverse user populations and often allows for hundreds, sometimes even thousands, of participants to be run in parallel over one or two weeks. However, when running experiments on this platform, it is hard to ensure that participants are actively engaging with the experiment and experimental controls are difficult to implement. Often, qualitative data is difficult, if not impossible, to collect as the experimenter is not present in the room to conduct an exit survey. Finally, and most importantly, the ethics behind running such experiments require further consideration. When we post a job on a crowdsourcing platform, it is often easy to forget that people are completing the job for us on the other side of the machine.

The focus of this Dagstuhl seminar was to discuss experiences and methodological considerations when using crowdsourcing platforms to run human-centred experiments to test the effectiveness of visual representations in these fields. We primarily target members of the human-computer interaction, visualization, and applied perception research as these communities often engage in human-centred experimental methodologies to evaluate their developed technologies and have deployed such technologies on crowdsourcing platforms in the past. Also, we engaged researchers that study the technology that makes crowdsourcing possible. Finally, researchers from psychology, social science and computer science that study the crowdsourcing community participated and brought another perspective on this topic. In total, 40 researchers from 13 different countries participated in the seminar. The seminar was held over one week, and included topic talks, stimulus talks and flash ('late breaking') talks. In a 'madness' session, all participants introduced themselves in a fast-paced session within 1 minutes. The participants stated their areas of interest, their expectations from the seminar, and their view on crowdsourcing science. The major interests of the participants were focused in different working groups:

- Technology to support Crowdsourcing
- Crowdworkers and the Crowdsourcing Community
- Crowdsourcing experiments vs laboratory experiments
- The use of Crowdsourcing in Psychology research
- The use of Crowdsourcing in Visualisation research
- Using Crowdsoursing to assess Quality of Experience

Figure 1 Tag cloud of the keywords from the abstracts.

The abstracts from the different talks, as well as the summary of the working groups can be found on the seminar homepage[1] and this Dagstuhl report. Apart from the report, we will produce an edited volume of articles that will become a primer text on (1) the crowdsourcing technology and methodology, (2) a comparison between crowdsourcing and lab experiments, (3) the use of crowdsourcing for visualization, psychology, and applied perception empirical studies, and (4) the nature of crowdworkers and their work, their motivation and demographic background, as well as the relationships among people forming the crowdsourcing community.

[1] http://www.dagstuhl.de/15481/

2 Table of Contents

3 Topic Talks (60min)

3.1 Crowdsourcing Technology

Matthias Hirth (Universität Würzburg, DE) and Michelle X. Zhou (Juji Inc. – Saratoga, US)

Abstract by Matthias Hirth

Over the past few years, Crowdsourcing has become a valuable tool for researchers to easily access a large number of people in a time and cost-effective manner. This enables new possibilities for all kind of studies involving human judgements or subjective ratings. However, even if Crowdsourcing is already widely used, still open questions remain: "How to transfer lab studies to the Crowdsourcing environment?", "How to optimize the quality of data obtained via Crowdsourcing?", and "How to keep the same ethical standards as in lab experiment?' are just a few of them.

To foster the discussion on possible solutions to these questions, this talk gives an overview of the current realization of the Crowdsourcing approach in commercial platforms and the resulting limitations in terms of scientific user studies. Further, it summarizes best practices and technical solutions to overcome some of those limitations or minimalise their effects on the results of crowdsourced studies.

Discussion

After the talk of Matthias Hirth the question arose, what should actually be considers as crowdsourcing. Especially the example of reCaptcha was controversial, because the users are not aware that they are working. Similarly, the use case of Crowdsensing was discussed, because users do not actively work in this case, instead they monitor passively environmental conditions with their devices. Another discussion arose about how to monitor the surrounding conditions of the crowd workers and how they influence the workers. This includes both technical conditions, e.g., the speed of the internet connection, and the physical work place of the worker, e.g., an Internet cafe. The consensus was that some of parameters, e.g., the hardware setting of the workers, do not influence the test if it is well designed and can also be checked automatically. For assessing non-technical factors, the crowd workers need to be asked explicitly. It was pointed out that also here technical solutions might be applicable, e.g., taking pictures with the worker's web cam. However, in this case privacy issues will arise. Another question related to the workers' background was the language of the tasks. Matthias Hirth mentioned that in the crowdsourcing platform used for his research, the task description is almost always in English. Based on that, the questions arose about whether the quality of tasks might be affected by the language of their description. On one the one hand, different languages of task descriptions might attract different workers, on the other hand, task descriptions in the native language might avoid misunderstandings. Continuing on quality control mechanisms, different methods where discussed to test if a worker is paying attention and being attentive. Here, an important outcome of the discussion was that questions testing the consistency of rating throughout a survey have to be chosen carefully. Otherwise, users might be influenced by the questions and will answer according to what they think they are expected to, instead of being truthful. Further, workers aware of these consistency questions might also communicate them (and the expected answers) to other

workers via external channels, like forums. To overcome this, it was mentioned that optional free text questions might be a good indicator for the truthfulness of a worker.

Abstract by Michelle X. Zhou

As crowdsourcing becomes a more and more popular approach in human-centred studies, it is important to get a deeper understanding of the crowd who participates in such studies. Who are they? What motivates them besides financial gain? How trustworthy are they? What kind of crowd may be best suitable for what type of studies based on their characteristics and qualities? To answer these questions, both traditional and computational psychometrics technologies may be used. In this talk, I describe a computational platform that can automatically infer a crowd worker's intrinsic traits, which can be further used to determine his/her suitability for a task (e.g., image tagging vs. critical tasks) as well as predict his/her task performance.

Discussion

Michelle Zhou's talk fostered a discussion about the reliability of psychometric tests, the reasons for different test set-ups, and their comparability. It was mentioned in the discussion that each test design is based on different underlying theories and that the tests also differ, depending on their purpose. This could include for example reliability tests or utility tests. However, even if psychometric tests normally archive a Cronbach's Alpha within 0.75 and 0.83, reliability might still not be given. Additionally, it was pointed out that in these tests, rank orders may stay the same, but actual values might change. Further it was mentioned that the test-retest reliability is also depending on the test itself and not a general property. Everyone agreed that it is very important to know if someone is willing and able to cheat/deceive while participating in crowd-sourced studies. Still, the question how to detect someone's willingness to cheat (reliably) remained open.

After this discussion, the question arose, how personality tests could be included in a user study and how much time should elapse between the test and the actual task. Here one suggestion was to first run a personality test, then the actual task and one more personality test again at the end. The final question was about the semantic technologies used to analyse the free text questions Michelle Zhou mentioned in her talk.

At the end Michelle Zhou raised two questions for the discussion during the reminder of the seminar:

1. If we have a platform that can measure crowd worker's various psychological traits, such as motivations and trustworthiness, would you use it as crowd selection criteria or additional factors for result analysis?
2. What characteristics of crowd workers you believe are important to measure for your type of tasks (e.g., visualization studies)?

3.2 Crowdsourcing & Visualization

Bongshin Lee (Microsoft Research – Redmond, US) and Rita Borgo (Swansea University, GB)

Abstract

Crowdsourced labour markets represent a powerful new paradigm for accomplishing work. Visualization relies on systematic evaluation to assess effectiveness and quality of designs and tools. Evaluation methodologies however often rely on restricted, specialised user cohorts to produce hypotheses or explanations of patterns and trends in data, which in turn might not yield widely applicable results in practice. Crowdsourcing offers an interesting alternative. Yet, asking users with varying skills, backgrounds, and motivations can as well result in speculative explanations of variable quality. Understanding the crowdsourcing phenomena, social context, and environment could therefore have significant benefits for the visualization community and beyond.

Discussion

Questions:

- How reliable is the task assessment completion measure?
- What do you measure in CS experiments?
- Is there CS not related to visualization?
- What do we understand of CS as a community?
- Are CS studies comparable to lab studies?

Important aspects highlighted through the discussion:

- Characterize visualization tasks that do not require deeper understanding.
- Characterize measures and metrics for both data collection and reporting.
- Characterize and categorize workers to be able to recall them for future experiments, some CS platforms provide this functionality.
- Characterize workers abilities e.g., visual and spatial abilities, pair-match CS samples based on demographics and abilities.
- Consider features such as language differences, e.g, how do translations differ? Compare people translating from the same to the same language.
- Important to focus on: task classification, data collection vs. conditions.
- Important outcomes of the proposed Dagstuhl Book: set of definitions, categories, and an overview of current state of the art.
- There is no reproducibility in CS studies, however lab studies have also their own limitations, for example, most experiments are not repeated.

3.3 Crowdsourcing & Psychology

Edwards, Darren J. (Swansea University, GB)

Abstract

Psychology is a broad field which encompasses cognitive, social, biological, neuroscience, behavioural, clinical, and health psychology. In computer science, areas such as visualization, or Human Computer Interaction, can often rely on psychological theory (particularly cognitive psychology) to inform experimental design and develop a priori hypotheses. Today, more studies are being carried out in computer science and psychology on on-line platforms such as Mechanical Turk and Survey Monkey. In order for this novel on-line platform being accepted by the larger psychological community, replicating studies from the laboratory to the on-line platform is important. Several studies such as the Stroop task, one-shot decision making tasks, inhibition of return studies, supervised categorisation, and even brief clinical mindfulness intervention studies have been replicated. All of these tasks are quite short, and easy to complete. So, what is not known is the reliability of the on-line platform in replicating more complex tasks which may use multi-modal cognition (e.g., a variety of spatial, sequential, decision-making components), or require a large amount of attention, memory and time to complete the task. The replication of these more complex studies are needed in order to assess the potential limitations of using on-line crowdsourced platforms. However, replications can only inform the research community so far, and they do not give any information about very novel studies which have not been replicated. For this reason it is difficult to know whether these new novel study results are an artefact of on-line user environments only, and maybe cannot be replicated in the laboratory. It is assumed that the laboratory should be the gold standard, as historically, laboratory study results have been accepted as transferable to real life scenarios. However, this assumption maybe limited, as the laboratory setting also has its own biases such as participant expectancy bias. These concerns are likely to be further highlighted in the future, as more laboratories choose to use crowdsourcing as an alternative to the laboratory.

Discussion

The discussion centred around several general topics. For example, questions were asked about the nature and structure of psychology student's education, and how it came about that participation in experiments is compulsory. Also, the limitations of this setting were discussed. The problem when using psychology students, for example, is that they often know what the experiment is about, through their knowledge about psychology. They may give the experimenter what they feel he/she wants to find. Also, given the compulsory nature of laboratory experiments in that the psychology students are forced to be participants in order to obtain course credit, raises potential problems in terms of reliability of the findings. So too, do the on-line platforms have their own problems, such as "career turkers", who are individuals seeking to make profit, or support a living from these on-line tasks. These career turkers often have multiple accounts and work on several projects at once. So, there is an obvious loss of control of the environment in which the participant works under, which can produce greater degrees of noise in the dataset. In some situations, such as sequence learning tasks, the task results rely on a strict presentation order of items, and working on multiple screens could cause false results, which are an artefact of the environment and

not the task. Other points were discussed such as whether we can categorise a platform like Survey Monkey with Mechanical Turk. Survey Monkey has a dynamic graphical user interface, and you cannot create dynamic experiments with moving images, to click on, for example, or to measure reaction time. You can however use it more than for just basic surveys, such as simple categorisation studies and one-shot decision making tasks. It also has its own participant pool of users. Finally, some discussion was made about the need for graphical user interfaces being introduced so that a greater number of psychologists could use, for example, Mechanical Turk. In conclusion, both the laboratory and the on-line platform are limited in different ways. As a way forward, greater understanding of the limitations of both environments are required in order to make appropriate predictions for studies.

3.4 Getting To Know The Crowd: The Crowdsourcing Community

David Martin (Xerox Research Centre Europe – Grenoble, FR) and Neha Gupta (University of Nottingham, GB)

Abstract

This talk focused on delivering an understanding of the crowdworkers who carry out the tasks posted on micro-tasking crowdsourcing platforms like Amazon Mechanical Turk (MTurk). We contend that the correct way to view such platforms is as labour marketplaces. We provide details on the crowdworkers through sharing quantitative and qualitative research across disciplines as varied as computer science, law and sociology. We find that workers are primarily motivated by pay but that other aspects of tasks such as learning interest, and enjoyment also count. Workers face a number of problems such as unfair treatment (e.g. rejected work, blocking), low pay, information deficit and lack/asymmetry of power. We discuss some of the challenges and responsibilities for researchers who use MTurk for various types of experiments and data services. We make some remarks on how some of the conditions may be altered such that relationships may be improved between both workers and requesters. In this way it is more likely that both parties can achieve what they want from crowdsourcing while promoting respectful relationships.

Discussion

During the talk we described how we had conducted our different studies of the US Turkers and in India. The focus was very much on Turkers – covering such matters as why they go on the forum, to share all sorts of information and to work as a community, and the fact that it has a vital role in on-boarding and learning. Without the resources like the forum, it would be much harder to earn and learn. We also discussed some of the features of good jobs: good pay, prompt pay, good communication, regular posting. We had some discussion of other platforms as a comparison. There was a focus on the particular problems of outages and infrastructure for Indian Turkers. After there was a conversation that centred around how to label and categorize different types of problematic work and workers with a general feeling that we should be careful about using words like 'malicious'. A next question was around finding the right price for work, with an understanding that very low pay was not good in any way and would attract bad behaviour but that there was no simple relationship

between price and quality. Finally we discussed the nature of MTurk as a market and how this poses new design questions for computer science.

4 Stimulus Talks (20min): Monday

Scribe: *Benjamin Bach (Microsoft Research – Inria Joint Centre, FR)*

This session featured four different talks covering users engagement, and motivation to participate, their background and motivation to malicious behaviour, studies in the wild, and how to improve the effectiveness of paid microtask crowdsourcing.

4.1 Evaluating Engagement and Enjoyment

Stephen G. Kobourov (University of Arizona – Tucson, US)

Abstract

While evaluation studies in InfoVis usually involve traditional performance measurements, there has been a concerted effort to move beyond time and accuracy. Of these alternative aspects, memorability and recall of visualizations have been recently considered, but other aspects such as enjoyment and engagement are not as well explored. We discuss recent studies of these topics and possible directions for future work.

Discussion

The discussion evolved around the concept of engagement and how to measure it. A first question asked about a definition of engagement; whether that meant to attract attention in the first place, or to retain attention, once it has been attracted? The answer was that both are considered, especially for information visualization.

Is it possible to measure engagement in museum-type setting? Museums have been mentioned as a potential place to measure engagement. Measuring engagement would have to take into account the place where engagement is measured, whether it's museums, public spaces, or waiting rooms, whether it's exhibition set-ups or personal devices, whether there are distractions, and so forth.

How to quantity engagement? Possible measures to quantify engagement could involve 1) How long do people stay watching or interacting with an artefact (the infographic)?, 2) Will they come back?, and 3) Will they bring someone with them? Higher values on a higher level would suggest higher overall engagement. Generally, engagement could be quantified as a matter of selection between two choices: two (or more) possibilities given, which one is chosen, and by how many?

In his informal study, the talk author has compared two network representations: node-link diagrams and map diagrams, consisting of node-link diagrams with clusters visualized as regions of a map. Results suggest that people where more "engaged" in the map representation. A question that emerged was about hypothesis or explanations *why* people stopped more often and stare at the map representations and spend more time watching? Is it because

people are more familiar with maps? Is it because maps are more unusual to represent non-spatial information?

Open questions involves what motivated people to come back? Can engagement be measured, and if so, can that happen in a crowd environment given the technical constraints as well as the fact that experimenters and participants are not co-located?

4.2 Measuring Workers' Pre-Task Interactions

Jason Jacques (University of Cambridge, GB)

Abstract

The ability to entice and engage crowd workers to participate in Human Intelligence Tasks (HITs) is critical for many human computation systems and large-scale experiments. We discuss how the conversion rate of workers – the number of potential workers aware of a task that choose to accept the task – can affect the quantity, quality, and validity of any data collected via crowdsourcing. We also contribute a tool – Turkmill – that enables requesters on Amazon Mechanical Turk to easily measure the conversion rate of HITs. We investigate how four HIT design features (value proposition, branding, quality of presentation, and intrinsic motivation) affect conversion rates. Among other things, we find that including a clear value proposition has a strong significant, positive effect on the nominal conversion rate.

Discussion

The discussion evolved around the study reported in the talk. This study tried to asses what are the factors that make a task more likely to be accepted by workers.

A first question asked *Where did the workforce numbers came from?* The respective numbers (650,000 and 825,000 for the two reported studies) came from Amazon itself, but where some years old. However it was estimated that the number of active workers on Amazon Mechanical Turk (AMT) has not changed too much since then.

Did the study followed an between or within subjects design? i.e. if some workers participated in several conditions or not. The experiment was designed as a between subjects study. The experiment was listed as a single task on AMT and used a redirected workers to assign them to their tasks (group). This method has been successfully used in previous studies.

How did you measure intrinsic motivation of participants to accept a task? One of the investigated factors that make workers accepting a task was *intrinsic motivation.* The only measure of whether a task was motivating or not was acceptance, once a worker pressed the button to accept. Workers have not been asked explicitly about their intrinsic motivation. The study reported in the talk was a replication of a previous study.

Are previous study results different from this one? Previous studies have examined whether the tested factors lead to higher task accuracy. This study measured conversion rate, i.e. how the factors influence a worker's decision to accept a task.

Did you use the same AMT account for your studies? The talk reported on several studies testing for different conditions, and which all have used the same account. This

may have lead to workers participating in multiple conditions, biasing the experiment. The answer was that very little overlap was reported, especially since steadily new workers are joining the pool. Further, Turkopticon ratings remained stable throughout the study period.

Did you identify reasons for changes in conversion rate and could you assess popularity? Workers have been posting tasks and on Reddit and turker forums. However, for non US-workers, there have been no reported factors.

4.3 Emotive Visual Display: Design Through Indoor and Outdoor Citizen Science?

Sara Fabrikant (Universität Zürich, CH)

Abstract

We use increasingly dynamic and mobile map displays for every-day decision making tasks (i.e., daily commutes in congested cities), and to find solutions to and communicate about complex global environmental challenges and societal needs (i.e., global climate change). However, we still have a poor understanding on how autonomic nervous activity might influence the already limited perceptual and cognitive resources of display users, for example, in time critical situations or in dilemmatic decision-making contexts (e.g., navigation, disaster mitigation and response, search and rescue, etc.).

In my talk I aim to highlight ongoing empirical research on animated and mobile map display use in the lab and in the wild, capitalizing on ambulatory human behaviour sensing methods (i.e., eye tracking, galvanic skin response, and EEG measurements). With this collected empirical data and supported by cognitive/vision theories we are guiding the process of designing maps for salience and positive engagement, thus aiming to create usable and useful visual analytics tools.

My open questions for this workshop are whether and how we might transfer this kind of direct human psycho-physiological sensing approach to evaluate visual displays into a crowdsourcing evaluation context.

Discussion

What do you consider as expertise? The reported studies investigated people with expertise in various areas. One has to differentiate between different kinds of expertise, e.g., domain expertise in a theme, or tool expertise, etc. Specifically, in the reported lab case, it included air traffic controllers trained in making decisions about aircraft movement with so-called semi-static displays, in which aircraft positions are updated every 4 seconds.

How did you obtain good measures outdoors? In the presented outdoor study, we tested expert navigators from the Swiss Armed Forces using a mobile map shown on smart device to perform a wayfinding task in unknown territory. It is generally more difficult to obtain good eye tracking data outdoors due to many often uncontrollable factors such as, weather conditions (i.e., sunlight reflected on digital displays, traffic and noise levels, and other unforeseen potentially occurring test interruptions, etc. The reported studies therefore always involve multiple measures that need to be triangulated for the analysis. For example, participants are individually shadowed by an experimenter, within a reasonable distance.

The experimenter records participants' behaviours on video and protocols unusual events. Different base line data are recorded prior to the study (i.e., spatial ability or to calibrate psycho-physiological measures) and these are then included in the data analysis.

Is that worth the effort? Could you just ask people about their feelings? Aside from self-reports, additional data collected from participants could include several psychophysiological measures including electrodermal response, eye tracking, and EEG. One can then systematically assess self-reports from questionnaires (e.g., collected on a Likert scale or open questions, etc.) with actual performance measures, to systematically assess what people say or believe, and what they actually do and how they actually perform.

Why has eye tracking been recorded in the studies? Eye tracking has been specifically employed to systematically document overt decision making process behaviour that leads to a decision or a response. This is particularly useful in combination with predetermined research hypotheses, that is, whether participants follow a theoretically predicted response pattern, i.e., formulated in a hypothesis. Eye-tracking thus can be used as process measure to explain behavioural responses.

4.4 Give Me More! Improving the Effectiveness of Paid Microtask Crowdsourcing

Ujwal Gadiraju (Leibniz Universität Hannover, DE)

Abstract

This talk discusses two pivotal aspects that influence the effectiveness of the paid crowdsourcing paradigm: (i) task design, and (ii) crowd workers' behaviour. Leveraging the dynamics of tasks that are crowdsourced on the one hand, and accounting for the behaviour of workers on the other hand, can help in designing tasks efficiently. To improve the overall quality of results, behavioural metrics can be used to measure and counter malicious activity in crowdsourced tasks. I will review some recommended guidelines for the effective design of crowdsourced surveys based on our recent research works. We also briefly discuss methods to train crowd workers to improve their performance in different types of crowdsourced microtasks.

Discussion

In assessing malicious behaviour and rule-breaking in crowdsourcing, what is your notion of rule-breaking? The notion of rule breaking in the study was the authors' one. It meant that if task rules have been violated and results were wrong or invalid. Questions have been inserted to check for workers attention, e.g. "This is an attention check question. Please select the second option". Some answers are syntactically wrong and these are easy to detect. However, answers that are semantically wrong are harder to detect.

Did workers have to pass both tests in the study? Yes, workers had to pass both attention tests in order to successfully finish the task. Tests were installed to filter workers who did not follow the task attentively and were delivering wrong answers. Workers have nevertheless been paid, even for training only.

What where the conclusions about training and tests? Training and tests reduced malicious behaviour and lead to an increased data quality.

Where did training take place in your study designs? Explicit training takes place before the tasks. The experiment design includes two separate blocks: training and experiment. *Implicit* training is inserted within the study, alternating between training and actual task sessions.

You said, the most common malicious people where of type Fast Deceivers (FD). Do people change their behaviour during tasks? The talk reported on FDs as those workers who quickly enter *some* response without checking for correctness. This behaviour is consistent throughout the tasks, i.e. workers do not become FDs throughout the task, but start as such.

What type of malicious workers appeared during training? This was not investigated in the study.

Your study reported that 40% of workers are in some way malicious. What's the motivation for malicious behaviour? We have no feedback about this. Our study wanted to track people's behaviour.

How do reputation systems works? Workers have a score, which defines their reputation. A low reputation score prevents workers from committing to certain tasks. Yet, many CS platforms do not have reputation systems set up. On the other side, the Crowdflower platform tries to create groups of similar skilled workers to specify qualifications, independent of task completion.

Do workers differ across platforms? Yes, they do. Further research is needed to compare platforms and their worker populations. There should also be research about comparing time of the day and continent of workers.

5 Stimulus Talks (20min): Wednesday

Scribe: *Sebastian Egger (AIT Austrian Institute of Technology – Wien, AT)*

This session featured four different talks, covering a framework for delivering crowdsourced evaluation of visualization, real examples of visualization evaluations using crowdsourcing, the motivation and background of crowdsourcing workers, and personal experience of crowdsourcing results in comparison with lab based results.

5.1 A crowd-sourcing framework for automated visualization evaluation

Radu Jianu (Florida International University – Miami, US)

Abstract

Due to recent research advances, the necessary ingredients for automating the process of designing and fielding user studies of data visualizations now exist. First, prescriptive rules formalize how user studies should be designed, what performance data should be collected, and how this data should be analysed. Second, task taxonomies and benchmark data standardize tasks used in evaluations. Third, a combination of advancements in web-technologies and crowd-sourcing allow even complex interactive visualizations to be evaluated

online. The talk builds on previous and current efforts in designing GraphUnit and VisUnit, two online services for crowd-sourced evaluation of visualizations, to describe how we could ultimately design and field quantitative user studies within minutes, and move our field towards benchmark driven visualization development.

Discussion

Radu Jianu reported on the advances he and his team made to develop a platform that is able to crowdsource evaluations of different visualizations without prior knowledge of web development or Java script. The resulting discussion is summarized in bullets below.

- *Is it possible to also use 3D graphs?* Yes, it's up to you which visualizations / graphs you use.
- *You have to provided functions for the evaluations. Is it code?* Indeed you have to provide code snippets
- *Are there standardized methods (ISO, IEEE...)* Rather accepted standards not standardized by an institution.
- *Which graph models do you currently support in terms of tasks: dynamic? multivariate?*
- *Was there confusion in the instructions about the technical term 'highest degree'? Did the workers have problems to identify what 'degree of . . . ' means?* Subjects had training sessions where this was explained. Training session to introduce the terms. There is also a module for language translation available.
- *Do you have a graph model that you use for verification?* Yes, we have.
- *Is it possible to run it on my own server?* Yes, it is Open Source Software.

5.2 Crowdsourcing Experiments in Visualization Research: Data, Perception, and Cognition

Remco Chang (Tufts University – Medford, US)

Abstract

In this talk I present three types of visualization experiments that we have successfully conducted on Amazon mechanical Turk (AMT). The first is on collecting user interaction patterns in a visual search task (Finding Waldo). Second is a perceptional study on modelling the perception of correlations in a scatterplot by using Weber's Law. The last is a set of cognitive experiments on priming the user's emotion (affect) and locus of control. In these experiments we demonstrated that AMT can be used to replicate laboratory studies- We further discuss some caveats and techniques for running these experiments via crowdsourcing.

Discussion

Remco Chang reported on three studies he and his team conducted with CS. One of their main aims was to gather a large quantity of data such that the were able to apply certain modelling approaches to gather either clusters of users (which share certain personality traits) based on their behaviour or to model their task performance as a function of task complexity. The resulting discussion is summarized in bullets below.

- *Did you also use psychometric tests for identifying the personality trait?* Yes, we used standardized questionnaires.
- *How did you capture and create the "locus of control"?* It's based only on mouse movement and map moves. Eye Movement is not tracked.
- *Did you track the pixels where the clicked on?* We track everything in the whole browser window. This is also one of the lessons learned, that this is necessary.
- *Were these charts identical in colour etc. ?* Yes, when both datasets were shown in one visualization we did use the same colours for all visualizations
- *How did you control brightness in the first study?* We didn't do the first study, it's a physical principle.
- *Resolution can be the same for different screen sizes.* Sure, but over 200 subjects you equal such effects out in the end.
- *How did you assess the LOC in these studies (Priming studies)?*
- *Why did you use LOC?* This was standing out well from other principles. It is more powerful than other measures, and you can also manipulate it well

5.3 Motivation of Crowd Workers, does it matter?

Babak Naderi (TU Berlin, DE)

Abstract

In my presentation, I introduced a model explaining relation between workers' motivation, outcomes and influencing factors in crowdsourcing micro-task platform based on self determination theory (Deci, Ryan 1985). On top of that we developed the Crowdsourcing Work Motivation Scale to measure the crowd workers' motivation in domain level. Furthermore, our empirical findings on how to design Trapping questions (gold standard questions) to increase the reliability of outcomes was described. Last but not least, tools provided by Quality and Usability Lab of TU-Berlin described: Crowdee the mobile crowdsourcing platform and Turkmotion , a task rating system for crowd workers.

Discussion

Babak Naderi presented a validated questionnaire for measuring the crowdworker's motivation. Furthermore, he reported on results they gathered for designing reliability questions (noticeable and hidden (=less noticeable)) in CS systems. the results showed that the noticeable reliability questions yielded better results in terms of reliability. Therefore, they concluded that the knowledge of being observed does positively impact reliability of the crowdworkers. Finally, an overview of TU Berlin's crowdee application concluded the talk. The resulting discussion is summarized in bullets below.

- *I had another concept of intrinsic motivation. Is that question (given as an example) what you base your judgement on whether or not somebody is intrinsically motivated?* It is one of the questions we use for intrinsic motivation determination, there are more questions we use to judge for intrinsic motivation.

- *You say 'unnoticeable'. Are you sure that the people do not notice the duplication of questions?* We had 97 items and it questions were separated and very difficult to notice. But perhaps it is better to label it 'less noticeable'.
- *Do you think the length of the questionnaire played also a role? Fatigue etc.* Could be, however, the longer questionnaire would typically lead to less reliable results than. But we have shown that the 'feeling of being controlled' increases the reliability score despite the longer questionnaire.
- *Motivation is to a large extent extrinsic (payment) but engagement is effected by intrinsic factors etc.*

5.4 Crowdsourcing Multimedia Experiences: Horror Stories and Lessons Learnt

Judith Redi (TU Delft, NL)

Abstract

Crowdsourcing gives researchers the opportunity to collect subjective data quickly, in the real-world, and from a very diverse pool of users. In a long-term study on image aesthetic appeal and recognizability, we challenged the crowdsourced assessments with typical lab methodologies in order to identify and analyse the impact of crowdsourcing environment on the reliability of subjective data. We identified and conducted three types of crowdsourcing experiments that helped us perform an in-depth analysis of factors influencing reliability and reproducibility of results in uncontrolled crowdsourcing environments. We provided a set of lessons learnt for future research studies which will try to port lab-based evaluation methodologies into crowdsourcing, towards avoiding the typical pitfalls in design and analysis of crowdsourcing based experiments with users.

Discussion

Judith Redi's talk gave an overview of practical experiences from a number of CS studies conducted. An interesting insight shared, was the fact that rather minor differences in interface design between lab and CS studies lead to considerably different results. Judith concluded the talk with a number of best practices that should be considered in the design of further human centred experiments. The resulting discussion is summarized in bullets below.

- *What is recognizability?* If people can identify what the image is about.
- *Who participated in the lab experiments?* Students and friend of the students that ran the experiments.
- *Recognizability could be used well across populations. Image quality is beauty and technical image quality.*
- *Is the study interface important?* Absolutely! We asked in the 2nd study "beauty" rather than "aesthetic appeal" which seems to be better comprehensible to the subjects.

6 Stimulus Talks (20min): Thursday

Scribe: *Ina Wechsung (TU Berlin, DE)*

The session covered a wide range of topics and disciplines; the first talk by Benjamin Bach was discussing whether or not complex tasks, which are usually evaluated by domain experts, can be evaluated in the crowd. Neha Gupta presented results from an ethnographic study examining crowd workers in India. A technology-driven perspective was given by Christian Keimel, who talked about hybrid devices and the opportunities offered by such devices in the context of crowd sourcing. The fourth talk in the session was held by Brian Fisher who showed how cognitive scientists make use of experimentation in the crowd.

6.1 Bringing Network visualizations to Domain Scientists

Benjamin Bach (Microsoft Research – Inria Joint Centre, FR)

Abstract

Working with domain scientists in neuroscience and history for some years, I wonder what we can learn from current crowdsourcing practices and technology to evaluate visualizations with domain scientists in the wild. Current experiments leveraging crowdsourcing focus on small and independent tasks, performable by untrained users. If we define crowdsourcing as finding and collaborating with many people you have known before, what are the limitations of such a collaboration? Can we contact and work with domain experts? Can we evaluate complex tasks? Can we measure exploration? What is a "useful" visualization? How can experts be encouraged to participate? How does the data must be?

Discussion

After the talk it was discussed how highly specialized and complex visualizations such as EEG scans can be adapted to the crowd-sourcing domain. It was debated whether or not it is possible to decompose such complex visualization as EEG scans in order to make them "crowdsourceable". In addition, questions arose on the definition of (domain) expertise, on how to become an expert, and on the experts which we, as researchers, want to participate in our studies. It was stated that experts may be defined as people who are dealing with the specific problem in question every day.

6.2 Turk-life in India: Supporting the work of crowdwork microtask crowdsourcing

Neha Gupta (University of Nottingham, GB)

Abstract

Previous studies on Amazon Mechanical Turk (AMT), the largest marketplace for microtasks, show that the largest population of workers on AMT is U.S. based, while the second largest is based in India. In this paper, we present insights from an ethnographic study conducted in India to introduce some of these workers or "Turkers", who they are, how they work and what turking means to them. We examine the work they do to maintain their reputations and their work-life balance. In doing this, we illustrate how MT's design practically impacts on turk-work. Understanding the "lived work" of crowdwork is a valuable first step for technology design.

Discussion

The talk raised a number of questions on how researchers as employers can improve the work conditions of Turkers and allow them to have sustainable careers. For example, it was discussed if "employer recommender systems", which are currently not part of the platform, should be integrated. It was assumed that such an internal system may support the workers in getting done their communication and interaction with each other; however, it was also mentioned that a conflict of interest may arise if the system is centralized as such an "employer recommender system" primarily reflects the worker point of view. It was concluded that in order to build better systems more research is needed; we need to get a better understanding of the crowd, and the possible differences between the Indian Turkers and the US Turkers.

6.3 Crowdsourcing and hybrid devices

Christian Keimel (IRT – München, DE)

Abstract

Hybrid devices bring together the broadcast and broadband world. They provide broadcasters for the first time a direct in-program return channel for capturing consumers' feedback, resulting in new enriched "big data". Could crowdsourcing on such devices enable an in-system optimization of content distribution and maybe even content itself by taking human perception/preferences into account? Can these devices be used to bring crowdsourcing into the lean back environment in the living room, enabling new opportunities for linear and non-linear content providers?

Discussion

It was asked if such commercial platform could possibly be used for couch crowdsourcing. While this is technologically possible, as such hybrid devices offer the same possibilities as browsers, the diffusion of such devices and the services may not be sufficient yet. Nevertheless marketing research is already carried out using hybrid devices.

6.4 Psychological perspectives on crowdsourcing

Brian D. Fisher (Simon Fraser University – Surrey, CA)

Abstract

I will describe a distributed cognition perspective on visual analytics with lab studies, field studies, and translational field experiments as methods. I will then speculate a bit about ways in which crowdsourcing methods might integrate with those more developed methods in visual analytics research.

Discussion

The discussion centred around the question which experiments or tasks are (un-)suitable for crowd sourcing. It was argued that the large sample size and the diversity of crowdsourcing studies are traded off for experimental control and that this trade off is working well for some studies (e.g. video coding) but not all (e.g. psychophysiological functions). In addition, crowdsourcing is especially suited for studies which are "too big" for the lab (such as annotations of large amount of videos). It was further discussed how collaboration and related issues such as coordination problems may be studied in the crowd.

7 Flash Talks: A summary of the spontaneous talks from Thursday

Scribe: *Daniel Archambault (Swansea University, GB)*

The flash talks at this seminar provided an opportunity for participants who were inspired by the activities during the week to present small 15 minute presentations to the group for discussion. We had four such presentations by Sheelagh Carpendale, Sebastian Egger, Tatiana von Landesberger, and Fintan McGee.

Sheelagh Carpendale began first with a presentation of an overview of her laboratory's research. In particular, she presented a summary of some of her work on tabletop displays (territory, gestures in context). She also presented work on how participants construct visualizations and a finding that participants tend to construct the visualization image first and subsequently set up the axes.

During the second half of the talk, she described a crowdsourcing study where she was trying to figure out how general people interpreted words such as user, person, researcher, participant, and others by asking them to draw images of these words. In particular, the experiment tried to figure out if the images had trends according to gender and other

attributes. In the first version of the study, she asked participants to draw a person. This first study produced a number of random drawings that were not people at all. By changing the question into a two phase question asking participants to first think of a person and then draw a person sitting down, nearly all participants drew people. By preparing participants for the task, good data was collected. Discussion around this talk centred around this finding and how methods for posing crowdsourcing questions is important.

Sebastian Egger presented a study on determining effective scales for evaluating video quality for use in laboratory settings and crowdsourcing applications. In this experiment, he tried to overcome a bias seen in crowdsourcing studies whereby participants would not use the full scale to judge video quality. Rather, participants would never use the bottom end of a scale $[1, 5]$. The study found that placing unclickable anchoring elements at the maximum and minimum end of the scale encouraged the participants to use the full range.

One of the interesting parts of this talk was the novel consistency/reliability checks used in this experiment. In this work, tasks where the participant was asked to list numbers displayed on the screen in order or click on stars present in a black background helped ensure that the participant was paying attention during the task. The results of these tasks were recorded to gauge participant reliability and engagement. If participant reliability was high, the participant would be put into a second pool whereby they would be called for further experiments. Using this technique allowed the experimenters to produce a more reliable pool of participants faster than previous methods.

Tatiana von Landesberger was interested in what people did with visualization techniques in the wild. She created a system, that relies on natural language processing approaches, to visualize news queries as a graph. The approach first processed news to find relationships between actors and locations and then constructed a very large network. This network could be queried to find an appropriate subgraph that could be visualized, encompassing a topic that surrounded the keyword. The system search functionality had to be tuned as certain terms consistently appeared in all graphs (e.g. Angela Merkel) due to their high connectivity.

Once the query system was created, the authors deployed both text excerpts and graphical representations of the topics to participants in a crowdsourcing study. The study found that annotations were greater in the text condition. The experimenters explained this finding by stating that the visualization helped clarify relationships in the data and therefore elaborate annotations were not needed. In a second study, they deployed the system as part of a Halloween contest. The study did not receive a large number of entries because the topic really wasn't present in the data. However, participants instead constructed graphs of things that scared them (e.g. Vladimir Putin) or constructed faces/images from the data around this theme.

Fintan McGee presented work done during the seminar on trying to determine what crowdsourcing was in terms of human-centred experimentation. Originally it was presented as a definition, but through discussion after the short presentation, it was determined that these were more characteristics of crowdsourcing experiments. The original characteristics were:

1. The participants are selected through an open anonymous call. We do not know the identities of specific participants.
2. The experiment is distributed using an online technical platform. This platform can support participant recruitment and task deployment.

3. The experiment is administered in an uncontrolled environment with limited interaction between the participant and the experimenter.

Discussion was lively and it was determined that these were properties of crowdsourcing experiments and not a definition. The second and third properties seemed integral but many seminar participants were debating if the first property was necessary.

Participants of the Seminar

In total, 40 researchers (28 male, 12 female) from 13 different countries participated in the seminar. 25 % were young researchers who actively brought in their opinions and enriched the discussions especially in the working groups.

Participants

- Daniel Archambault
Swansea University, GB
- Benjamin Bach
Microsoft Research – Inria Joint Centre, FR
- Kathrin Ballweg
TU Darmstadt, DE
- Rita Borgo
Swansea University, GB
- Alessandro Bozzon
TU Delft, NL
- Sheelagh Carpendale
University of Calgary, CA
- Remco Chang
Tufts University – Medford, US
- Min Chen
University of Oxford, GB
- Stephan Diehl
Universität Trier, DE
- Darren J. Edwards
Swansea University, GB
- Sebastian Egger
AIT Austrian Institute of Technology – Wien, AT
- Sara Fabrikant
Universität Zürich, CH
- Brian D. Fisher
Simon Fraser Univ. – Surrey, CA
- Ujwal Gadiraju
Leibniz Univ. Hannover, DE
- Neha Gupta
University of Nottingham, GB
- Matthias Hirth
Universität Würzburg, DE
- Tobias Hoßfeld
Universität Duisburg-Essen, DE
- Jason Jacques
University of Cambridge, GB
- Radu Jianu
Florida International Univ. – Miami, US
- Christian Keimel
IRT – München, DE
- Andreas Kerren
Linnaeus University – Växjö, SE
- Stephen G. Kobourov
Univ. of Arizona – Tucson, US
- Bongshin Lee
Microsoft Res. – Redmond, US
- David Martin
Xerox Research Centre Europe – Grenoble, FR
- Andrea Mauri
Polytechnic Univ. of Milan, IT
- Fintan McGee
Luxembourg Inst. of Science & Technology, LU
- Luana Micallef
HIIT – Helsinki, FI
- Sebastian Möller
TU Berlin, DE
- Babak Naderi
TU Berlin, DE
- Martin Nöllenburg
TU Wien, AT
- Helen C. Purchase
University of Glasgow, GB
- Judith Redi
TU Delft, NL
- Peter Rodgers
University of Kent, GB
- Dietmar Saupe
Universität Konstanz, DE
- Ognjen Scekic
TU Wien, AT
- Paolo Simonetto
Romano d'Ezzelino, IT
- Tatiana von Landesberger
TU Darmstadt, DE
- Ina Wechsung
TU Berlin, DE
- Michael Wybrow
Monash Univ. – Caulfield, AU
- Michelle X. Zhou
Juji Inc. – Saratoga, US