Leveraging Wisdom of the Crowd for Decision Support

Simo Hosio
University of Oulu
Pentti Kaiteran katu 1 90570
Oulu, Finland
simo.hosio@ee.oulu.fi

Jorge Goncalves
University of Oulu
Pentti Kaiteran katu 1 90570
Oulu, Finland
jorge.goncalves@ee.oulu.fi

Theodoros Anagnostopoulos
University of Oulu
Pentti Kaiteran katu 1 90570
Oulu, Finland
tanagnos@ee.oulu.fi

Vassilis Kostakos
University of Oulu
Pentti Kaiteran katu 1 90570
Oulu, Finland
vassilis.kostakos@ee.oulu.fi

While Decision Support Systems (DSS) have a long history, their usefulness for non-experts outside specific organisations has not lived up to the promise. A key reason for this is the high cost associated with populating the underlying knowledge bases. In this article, we describe how DSSs can leverage crowds and their wisdom in constructing knowledge bases to overcome this challenge. We also demonstrate how to construct DSSs on-the-fly using the collected data. Our user-driven laboratory studies focus on user perceptions of the concept itself, and motives for contributing and using such a DSS. To the best of our knowledge, our approach and implementation are the first to demonstrate such use of crowdsourcing in building DSSs.

Decision Support Systems, evaluation, crowdsourcing, wisdom of the crowd

1. INTRODUCTION

Decision Support Systems (DSS) (Druzdzel, Flynn 1999) typically combine multiple data sources, expert input, and computational methods, to explore a given problem domain and ultimately help choose from a set of defined options. Early DSSs were aimed especially for organisational contexts where strong financial motives to make informed decisions exist.

So far, DSSs have not been particularly successful with non-expert users outside specific organisational contexts. Just as one example, Personal Decision Support Systems (PDSS) have been proposed for assisting individuals (Shambaugh 2009), but they have not really lived up to their early promise. A key challenge in developing DSSs lies in obtaining adequate amounts of recent and accurate input data, and this process can be difficult, time-consuming, and painstakingly costly (Er 1988; Geurts 1994). We argue that crowdsourcing, or, in this case, leveraging Wisdom of the Crowd, can overcome this challenge in many cases.

In this paper, we demonstrate how DSSs in arbitrary problem domains can be constructed using crowds. We present our Decision Support Platform in the context of two aspects of DSSs: populating knowledge bases (Study 1) and providing decision support by exploiting the knowledge bases (Study 2). These areas are the most important elements for a DSS to function (Druzdzel, Flynn 1999). We also present preliminary evidence of the system's scalability beyond controlled laboratory settings.

Ultimately, our vision is to transform Wisdom of the Crowd into useful DSSs on-the-fly. In doing so, we hope to overcome several acknowledged limitations in traditional DSSs.

2. RELATED WORK

Our approach intersects Wisdom of the Crowd and Decision Support Systems. These domains both have a rich history and rather complementary research challenges.

2.1 Wisdom of the Crowd

Wisdom of the Crowd refers to the aggregated opinions of a crowd. It is a statistical phenomenon with no social or psychological explanation behind it, and it relies on mathematical aggregation methods (Lorenz et al 2011). While its earliest mentions can be traced back to Aristotle, the work by Sir Francis Galton in 1907 on a weight-judging contest of a fat ox at a farmer's fair is widely acknowledged as the first academic investigation of the concept (Galton 1907). Galton observed that the collective knowledge of a crowd (the fair audience) remarkably outperformed the accuracy of expert opinions (butchers). Similar findings have been repeatedly verified by several researchers in other contexts (Page 2008; Surowiecki 2005), and in more recent these findings have motivated leveraging crowds for computationally challenging problems (Kittur et al 2013; Poetz, Schreier 2012).

Surowiecki defines four qualities that make a crowd smart and to likely outperform the individual group members (Surowiecki 2005). First, the crowd needs

Proceedings of British HCI 2016 - Fusion, Bournemouth, UK

to be diverse, so that individuals can offer different pieces of information to the table. Second, the crowd needs to be decentralized, so that no one at the top dictates the collective output. Third, there needs to exist a way to summarize different opinions. Finally, the people in the crowd must be independent, so that they do not consider what others in the group think.

Conversely, factors hampering crowd performance have been identified. For instance, *social influence* refers to how the opinions of one's peers affect individual judgement. This, in turn, undermines Wisdom of the Crowd by reducing the crowd's diversity and individuals' independence (Lorenz et al 2011). Ideally, crowd members should not be aware of each other's opinions – or even the aggregate opinion – as there is evidence that humans seek for consensus (Yaniv, Milyavsky 2007). As such, a small number of extremely vocal users may be able to "over-contribute" their opinions and bias the crowd's opinion (Kostakos 2009).

2.2 Seeking Advice from User-driven Websites

Online services such as Yahoo Answers, Quora and forums on a plethora of topics are popular ways to seek advice for problems online. On these sites, users post questions and topics, or provide open-ended answers. Such websites are well suited for fact-finding questions (e.g., "*Who composed the soundtrack to Braveheart?*") and discussions, but are problematic when users seek structured decision support. The reason is that forum-like architectures do not explicitly account for the distinct sets of options and criteria to support reasoning (Wang et al 2006).

Instead, user-driven advice websites typically rely on text-based input, and optionally some type of voting scheme to identify the best contributions. Consequently, users are required to mentally consolidate multiple opinions, weight their trustworthiness, and finally determine an appropriate solution to their problem. In this context it has been shown that by using persuasive language alone it is possible to greatly affect others' perceptions on issues, and that in such environments the community leaders have more influence than others, despite not necessarily being any wiser than the others (Huffaker 2010).

Lurking, i.e. reading but not contributing, is a common phenomenon on forums (Preece et al 2004). Lurkers typically have difficulties in correctly formulating their thoughts or do not feel welcome in an already established group among forum veterans. Further, the hassle of creating and verifying accounts for a one-off contribution is a great barrier to participation. It is notable that lurkers can make up to 90% of all visitors of user-driven websites (Nonnecke, Preece 2000). It is quite justified then to claim that a great deal of an audience's potential to contribute to a topic on these sites is lost.

So, in light of evidence, the online forum-like environments are not optimal for decision-making -- nor are they designed for it. They are places to discuss topics. Still, they are often used as starting points to look for help when making decisions. Our platform sets to complement (not replace) these existing means.

2.3 Decision Support Systems

Decision Support Systems is a diverse discipline of information systems that assist in making decisions (Arnott, Pervan 2005). While DSSs lack a single accepted definition (Druzdzel, Flynn 1999; Shim et al 2002), Finlay defines a DSS as broadly as "*a computer-based system that aids the process of decision making*" (Finlay 1994). Similarly, Power defines DSSs as "*interactive computer-based systems that help people use computer communications, data, documents, knowledge, and models to solve problems and make decisions*" (Power 2002).

More recently, Recommender Systems have emerged as a related field, drawing influences from DSSs, information retrieval and machine learning, among others (Jannach et al 2012). Used often interchangeably with DSSs in academic literature, they offer users recommendations, typically by observing users' past actions and preferences (Resnick, Varian 1997; Ricci et al 2011). For instance, the suggestions provided by NetFlix or Amazon are offered by a recommender system.

Conceptually, DSSs consist of three main components: the knowledge base, the model and the user interface (Druzdzel, Flynn 1999). The knowledge base stores data relevant to the problem. The model formulates decision based on knowledge base contents, and the user interface enables users to build the models (input data), and obtain decision support by adjusting input parameters.

In our DSS work, we offer users a simple platform capable of providing decision-support based on aggregated data collected from a crowd. In other words, we leverage Wisdom of the Crowd for decision support. Particularly relevant to our work are Model Driven DSSs: systems that rely on quantitative models of the problem-space, offer an end-user interface for manipulating the parameters, and supporting "*what if?*" analysis. We follow the Suggestion Model (Alter 1982) and provide "*suggestions to a person for a defined domain or task*". An overview of Model-driven DSSs and the related research challenges is provided by Power & Sharda (Power, Sharda 2007).

An important goal in our work is to replace the costly process of harvesting accurate input from multiple sources to populate knowledge bases (Er 1988). Another key goal is to devise a means of sustaining the quality (completeness, accuracy, recency) of knowledge bases (Barnett et al 1987; Geurts 1994). To this end, recent work suggests that ubicomp technologies and appropriately designed incentives can help in reaching large numbers of individuals affordably and rapidly (crowdsourcing), to establish and sustain accurate information inventories (Goncalves et al 2014a; Hosio et al 2015; Hosio et al 2014; Hosio et al 2012).

3. SYSTEM DESCRIPTION

Our system – AnswerBot – is an online, Web-based DSS that i) enables any visitor to contribute to the knowledge bases of the hosted problems, and ii) provides decision support for the problem using the knowledge base.

3.1 Populating the Knowledge Bases

AnswerBot is by design geared toward decision-making problems that are defined in terms of i) potential answers to the problem (we call them options), and ii) tradeoff dimensions (we call them criteria). For instance, for the problem "*Where should I go on my Honeymoon?*", some potential options are [Hawaii, Paris] and example criteria are [romantic, nightlife].

AnswerBot facilitates collecting options and criteria from the crowd using an interface that simply has two text input fields, one for the entry title and another for a more detailed description. Using the aforementioned example, an additional entry (criterion) could be: "*friendly for tourists -- how friendly in general is the atmosphere among locals towards tourists?*" In this article, however, the main focus is not on crowdsourcing options and criteria, as in many cases these dimensions are strictly pre-defined (already exists a limited set of options to consider in the light of certain criteria).

After having a set of options and criteria for a question, and to construct a useful DSS, the underlying knowledge base must contain sufficient input to model the relationship between every option-criterion combination. AnswerBot asks the crowd to rate option-criterion pairs using numeric sliders (ranging from 1 to 10), as shown in Figure 1. For instance, if a particular question has 3 options and 4 criteria, then there are 3x4=12 option-criterion pairs that need to be scored by the crowd. To break the task to easier to process chunks, as suggested in (Bernstein et al 2010; Noronha et al 2011), the system displays up to four sliders at a time on screen.

In principle, making decisions can be described as considering available options in light of related criteria (Wang et al 2006). For example: "*How romantic is Hawaii compared to Madeira or Paris?*" However, when eliciting input for option-criterion pairs, or visualising a given problem, it has not been previously explored whether several options should be judged against one criterion (We call this Option-Driven (OD) assessment, as there are more options to consider simultaneously), or several criteria against an option (Criteria-Driven (CD), more criteria to consider simultaneously). Further, content presentation affects how it is perceived by users and how people interact with it (De Angeli et al 2006). Thus, we hypothesized that different ways at looking at the same problem space could yield different results and user experience.

Figure 1 depicts AnswerBot's user interface to elicit input in Option-Driven condition (OD), where several options and one criterion are displayed to the user. The only difference in CD condition is that options and criteria swap places, making the users estimate several criteria against one option at a time. Thus, the UI is identical, and just the descriptions swap places.

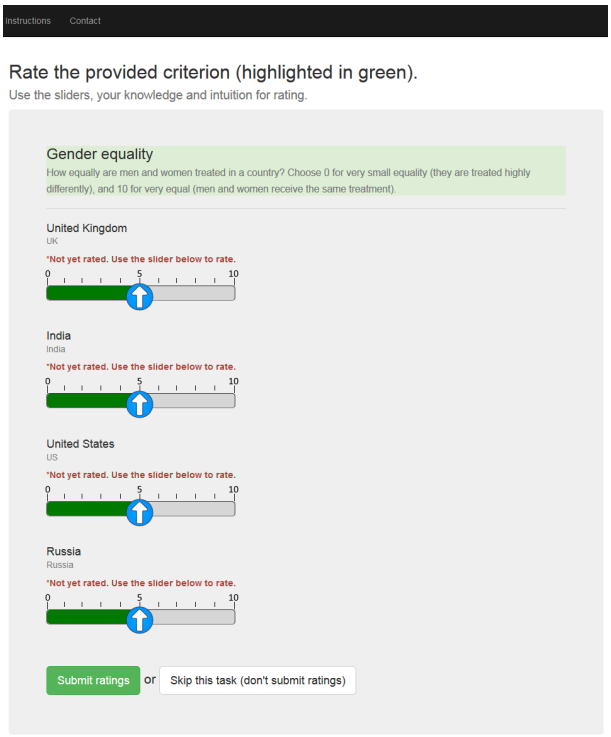

Figure 1. Populating the knowledge base (Option-Driven condition) by rating options in terms of a criterion.

3.2 Obtaining Decision Support

Once the knowledge base of a problem is populated, AnswerBot on-the-go instantiates a DSS capable of providing decision support for the problem. A key feature of model-based DSSs is to support ad hoc "what if" analyses (Power, Sharda

2007), where users manipulate the available model parameters in an attempt to identify ideal solutions. To enable this, in our system users first see a list of all criteria associated to the problem, as depicted in Figure 2. Then, AnswerBot computes and displays recommended solutions based on goodness of fit (explained later), as depicted in Figure 3.

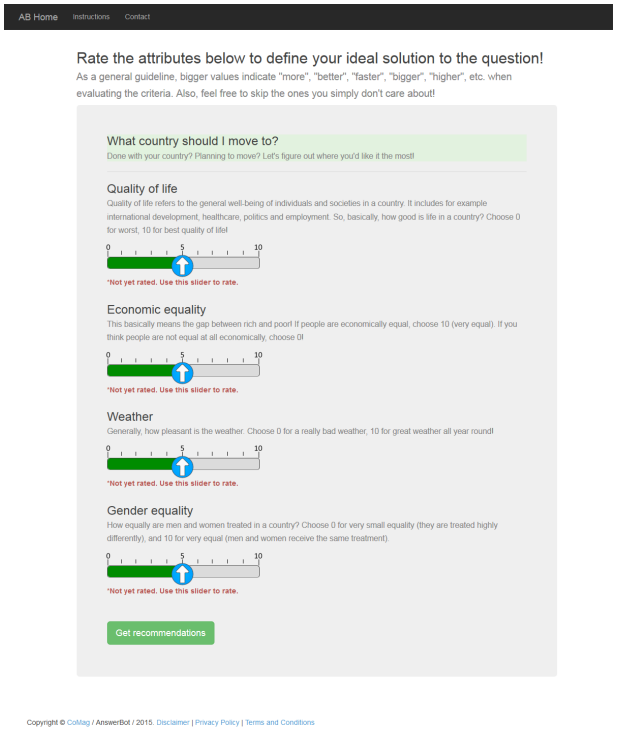

Figure 2. The interface to conduct "what if" analysis by adjusting desired criteria.

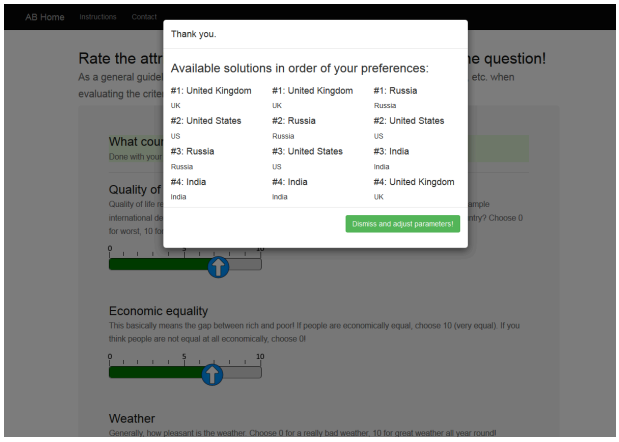

Figure 3. Decision support is offered based on three different recommendation models.

As can be seen in Figure 2, users manipulate the importance of each criterion using simple sliders as input elements. We considered the slider inputs as intuitive elements, as they are already highly popular in many online DSSs. Just one example of such is *VotingAid* (VotingAid 2015), used by Reuters, EuroNews and Al Jazeera, among others. Further, an acknowledged pioneer of the field, Daniel Power, also notes that "*The [what if] analysis is likely to be more complete if an input object like a spinner or a slider is used to change values. Such an approach is much faster and easier than typing in individually new input values*" (Ask Dan! 2015).

In response to the criteria adjusted by the end-user, the system performs its part of the "what if" analysis on server-side. Then, AnswerBot displays the options with best goodness of fit according to the knowledge base and the used decision model. The calculations are performed runtime, enabling users to conduct scenario analysis dynamically. Figure 3 depicts the result interface, where recommendations are displayed in a simple modal popup window. The figure depicts 3 distinct sets of recommendations, because in Study 2 we evaluated multiple approaches in terms of i) building a model from the underlying knowledge base, and ii) end-users' judgement of the accuracy of the solutions offered by each model.

4. STUDY 1: GETTING INPUT FROM A CROWD

Study 1 was a controlled laboratory study designed to assess quantitatively and qualitatively the user interface for populating knowledge bases. In addition to generic usability and user perceptions, we evaluated the two alternative design options for obtaining ratings from participants: Option-Driven and Criterion-Driven. The assessment focused on examining the result data, and if there are any differences in how users perceive the conditions.

4.1 Experimental Task

We recruited 24 participants (20 male, 4 female, with average age of 29.9 years) from our campus using email lists. Each participant arrived to our laboratory for a 45-minute session with a researcher. In such a session, every participant used AnswerBot and contributed to populating the knowledge base for two questions (Q1 and Q2, listed in Table 1).

Table 1. The experimental task requires participants to rate all option-criterion pairs for Q1 and Q2.

Q1: To which country should I move?		**Q2: In which restaurant on our campus should I eat?**	
Criteria	Options	Criteria	Options
Quality of life Economic equality Weather Gender equality	UK US India Russia	Noise level Quality of special foods Queues Staff friendliness	Nick's Joe's Frank's Gregg's

These 2 questions were trivial for methodological purposes. By defining the options and criteria ourselves, we minimised the chance of participants

not understanding the tasks, not knowing how to answer, or perceiving the tasks as too difficult. The restaurants of Q2 are on-campus restaurants, and thus familiar to most participants. Each question had four criteria and four options. Thus, there were 16 option-criterion pairs per question (Table 1), and 32 pairs in total for each participant to rate.

4.2 Procedure

Each participant was first given a short briefing about AnswerBot. Then, users provided input (Figure 1) using a desktop computer provided by the research laboratory. They rated all option-criterion pairs of both questions. The order of conditions was counter-balanced, so that half of the participants first gave ratings for the questions in OD condition and then in CD condition. The other half first used CD and then OD to answer both questions. All participants completed a System Usability Scale (SUS) (Bangor et al 2008) after completing the questions in the first condition.

4.3 Results

Participants gave 1536 unique ratings (24 participants x 2 conditions x 2 questions x 4 options x 4 criteria). Of these ratings, 188 were skipped entries, i.e., the participant did not want to provide a rating for the option-criterion pair. To facilitate analysis, we conducted statistical imputation: skipped entries were replaced with the median of non-skipped entries for the same option-criterion-pair.

4.3.1 On quality and necessary amount of data
Considering our data model, the standard deviation of option-criterion pairs practically indicates how well participants agree on it. We investigate whether the standard deviation differs between the conditions on any pair, and plot the observed values for both OD and CD conditions (Figure 4). For Q2 (restaurant-themed), we identify one criterion ("*quality of special foods*") where participants gave noisier responses under both conditions (SD higher than that of other criteria, denoted with dashed lines in Figure 4).

Given that the standard deviation differs between the pairs, the logical next question becomes how much data do we need before a knowledge base is ready to be used to provide decision support? A common method in analysing DSS data is simulation (Power, Sharda 2007). We simulated the development of our two knowledge bases progressively as more users provide input to the available option-criteria pairs. We simulate the order of ratings arriving to the system, not the ratings themselves. In Figure 5, we depict the expected median rating and standard deviations for an option-criterion pair if only 1, 2, 3, etc. random users of the 24 users in the study have given their ratings. We ran 1000 simulations for each combination of [pair, number_of_ratings], and calculated the mean of all simulations. We used data from users in the CD condition.

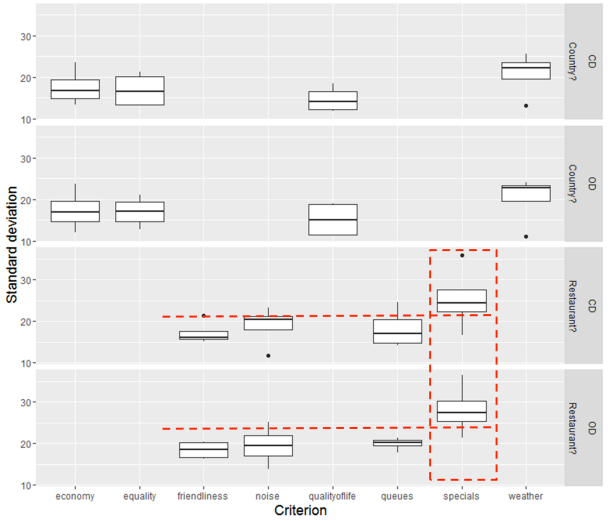

Figure 4. Box plot of the standard deviation values produced by participants. Data is grouped by each option-criterion pair, and separated per question and experimental condition (OD vs CD).

Our results show that some pairs reached consensus (median stays the same, SD is low) quickly. In such cases the crowd has a strong opinion on the pair. On the contrary, certain pairs are noisy, and it takes many users' input to reach consensus. For instance, "*equality*" in "*India*" (marked "A" in Figure 5) stabilises (crowd finds a consensus) with just a few ratings, but more are needed to determine the "quality of special foods" in "*Frank's*" (marked "B" in Figure 5). We see potential in such simulations in optimising data input, by prioritising pairs that have a poor consensus over ones that the crowd quickly agrees on.

4.3.2. Usability and participant feedback
we employed the standardised 10-point System Usability Scale (SUS) (Bangor et al 2008) to assess AnswerBot's data input interface. The CD condition obtained an average score of 83.3 (SD=12.1) while the OD scored 82.7 (SD=8.3). The maximum score in SUS is 100. As expected, the two conditions did not rate substantially different in terms of usability (they look near-identical). However, when participants were explicitly asked whether they preferred to use the CD or OD condition, 19 chose the OD (several options shown simultaneously) while only 5 chose the CD (several criteria simultaneously), indicating clear preference to OD condition.

Drawing on participants' comments during the concluding interview, we verified the previously discussed difficulty in assessing the criteria "*quality*

of special foods". The participants indeed found it hard to rate this criterion because "*it varies day by day a lot*", and it was not clear what exactly does "special" mean. This suggests that poorly worded criteria are inevitably challenging to rate, and that we made a mistake in not defining this one carefully enough.

5. STUDY 2 - MODELING AND DECISION SUPPORT

The purpose of Study 2 was to explore how data bootstrapped in Study 1 transforms into decision support. From the user's perspective, this means if they find the resulting decision support of high quality. With AnswerBot users get decision support simply by adjusting input parameters (criteria), based on which AnswerBot returns an ordered list of the available options, based on goodness of fit.

To begin with, we created three *models* to derive decision support. First, we introduced a baseline model that returns the options in random order. This allowed us to explore if our decision support has any desirable qualities, or is it merely noise that users perceive as trustworthy. After all, it is true that participants modify their responses to please researchers when observed (McCarney et al 2007).

Second, we modelled AnswerBot response to reflect the theory behind the original Galton's experiment on Wisdom of the Crowd (Galton 1907). Applied to AnswerBot's data model, this means that the median of each option-criterion pair ratings in the knowledge base is closest to the "Gold Standard", or *Vox Populi* (voice of the people). Therefore, the sum of Euclidian distances of each input value to the median of the same option-criteria pair in the knowledgebase indicates a given option's goodness of fit. Thus, we could provide results simply ordered by distance to the crowd's collective opinion about which options most closely resemble the input values.

Finally, as our case can be modelled as a classification task, we used machine learning with Logistic Model Trees (LMT) (Landwehr et al 2005) as the classifier. LMT combines a tree structure and logistic regression in a single tree. Every class (option) is made binary and a set of trees is produced for each class. LMT applies logistic regression on the attribute leaves (criteria) to perform classification. We argue LMT as well suited to our study, as the number of classes is small, and thus the number of the produced trees remains small. In addition, the incorporation of logistic regression produces explicit class probability estimates that are useable to rank the options.

A performance analysis of the LMT classifier reveals it performing better in the CD condition (Q1: 72.9%, Q2: 52%) than in OD condition (Q1: 63.5%, Q2: 49%).

5.1. Procedure

Evaluating DSSs can be difficult, given that their benefits are often qualitative in nature (Keen 1981). We evaluated the models and the provided decision support with the help of knowledgeable end-users, a typical approach for evaluating DSSs, as there exists no undisputable ground truth.

We recruited 16 participants (12 male, 4 female, average age 28.4 years) from our campus. None of the participants took part in Study 1. Again, each participant arrived to a 1-on-1 session, where they were briefed about AnswerBot's decision-support

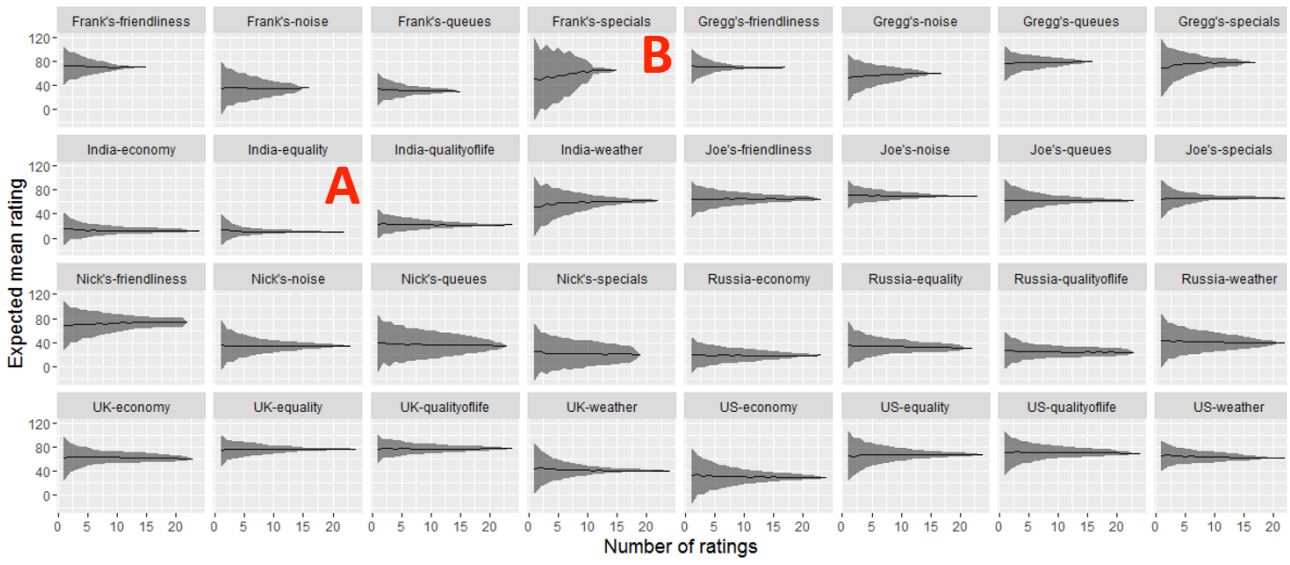

Figure 5. For each option-criterion pair we estimate the mean rating if only a subset of the ratings was used.

concept. Participants then used the interface in Figure 2 on a desktop provided by the researchers. The decision support consisted of 3 different lists, as shown in Figure 3: 1 based on each of the three previously discussed models. We counterbalanced the order of the result lists, to make sure the presentation order did not skew the results.

The participants used AnswerBot to obtain as many sets of suggestions to both of the two questions (countries, restaurants) as they wished to. They were encouraged to explore the system using multiple arbitrary criteria values, and thus effectively conduct "what if" analysis. Once participants felt satisfied with the exploration, they were asked to reason about the 3 presented models: which one they found most accurate, or inaccurate, and why.

In addition to discussing the goodness and perceived usefulness of the models, the concluding interviews focused on topics such as trust or distrust towards AnswerBot and its suggestions, the crowd-based DSS concept in general, incentives of participation, privacy, and limitations of the system.

5.2. Results

Together the 16 participants conducted a total of 313 "what if" analyses using AnswerBot. Typically, a participant explored the questions for approximately 7-10 minutes before being ready to voice her verdict on the worst and best models.

All 16 participants recognised the random model clearly the worst. This indicates that the two other models are capable of providing at least better than random support. The median-based approach was preferred by 12, and the LMT-based by 2 participants. There were 2 split opinions where the participants could not tell if they preferred median-based or LMT-based suggestions.

5.2.1 Overarching interview findings

In the interviews a clear majority of our participants suggested that a typical search for decision support starts with a Google search. Then, they would proceed to forums or any sites ranking high on Goole results: "*I just Google always. Always yahoo answers or similar pops up, and then I read those*" (P1_4) or "*Usually just google and then of course seek for audience opinion on forums also…"* (P1_6). This further verifies that others' opinions play a great role in decision-making.

Further, the interviews revealed users to trust our crowd-based decision support: "*I do trust the people more than e.g. just one expert. I mean they are the people who go there all the time" (restaurants)*" (P2_3), or *"[the system] is really useful, real people, real users of services are what I trust to give the honest truth about things*" (P2_1), or "*I am actually impressed about the knowledge [inside AnswerBot] – like, after I decided which model I like the best and focused on it, I really think it displays me what I expect it to*" (P2_8).

We also enquired about the specific details of the underlying crowd and its composition that influence users' attitude towards AnswerBot's suggestions and overall trustworthiness. Commonly mentioned characteristics here were the combination of age and gender, occupation, expertise, and geographical area: "*Well, first I'd like to know about age, gender. And then, if the topic is something like Climate Change, I would also like to know the occupations... I would want to have researchers in the crowd*" (P2_3), or "*I would like to make sure the people who are giving this information are not some pensioners but about the same age as me*" (P2_7). So, knowing that a relatable crowd is behind the given suggestions increases trust in the system, but also further demographic data should be collected and presented. In addition, our interviewees indicated that a larger crowd feels more trustworthy: "*I think more important is to know the number of people, not so much who exactly are they. I cannot say exact numbers, but more is better obviously*" (P2_5) and "*If the N is small, I need more details about the crowd definitely. If the N is bigger, like hundreds of people, I really don't care at all.*" (P2_4).

Finally, we asked initial thoughts on what would make users contribute to AnswerBot in the future. We received comments, such as "*I would not use it for giving information, would use it for getting information…if the system provides value to my life*." (P2_5), "*A strong motivator for me is to know if others have done it before me. Why would they have done it if not without a reason?*" (P2_1), or "*More societally meaningful questions of course are better for motivating me*" (P2_3). More interview findings are weaved into the following discussion.

6. DISCUSSION

A recent survey of 1093 DSS articles criticises the field's orientation towards theoretical studies and poor identification of the hypothesised DSS users (Arnott, Pervan 2005). Moreover, current systems are built in the context of a specific domain. These are limitations we seek to overcome by turning the crowd's aggregated wisdom into PDSSs, on-the-fly, and for arbitrary decision-making problems.

An intuitive argument can be made that AnswerBot is "too simple" to be called DSS. While we understand such criticism, it is the simplicity, and the fact that it is entirely crowd-driven (even small crowds work remarkably well in decision-making (Surowiecki 2005)), that make it useful for its designed purpose. We acknowledge, that while any question can technically be posed, certain

questions simply do not fit for our model. In the end our PDSS is only suitable for questions where Wisdom of the Crowd works. However, and contrary to what many self-proclaimed experts admit (as it would devalue their opinions), it has been shown to work on a plethora of decision-making domains (Surowiecki 2005).

As mentioned earlier, forums, Q&A sites and similar ones provide valuable yet unstructured decision support (Yaniv, Milyavsky 2007). However, clear structure is important in human decision-making (Shambaugh 2009). Thus, we suggest complementing the somewhat messier but undeniably valuable information sources with a more structured solution like ours. AnswerBot provides each individual in the crowd an equal voice, thus avoiding the undesired effects of social influence and community leadership (Huffaker 2010). It also offers a low barrier to contribute, as users do not need to identify themselves, create accounts, or even formulate their thoughts verbally. These are all identified barriers to participation on current popular sites where people turn for decision support (Nonnecke, Preece 2000; Resnick, Varian 1997).

6.1. Leveraging the Crowd to Bootstrap Knowledge Bases and Develop Models

Advances in crowdsourcing have made distributing tasks to truly global audiences both feasible and practical (Ipeirotis, Gabrilovich 2014; Kittur et al 2013). However, volume alone is not sufficient in creating a usable AnswerBot knowledge base, but also the input quality matters a great deal. The input must represent a clear snapshot of a given crowd's wisdom on a problem. We next examine this through the lens of previously identified quality control approaches: *appropriate task design* and *post-hoc analysis of results* (Kittur et al 2013).

6.1.1. Appropriate Task Design

Successful task design can improve input quality, and especially the perceived difficulty of a task affects a crowd's performance (Rogstadius et al 2011). Workers may give up or provide inaccurate input if the task is too difficult. We designed the user interface of AnswerBot to minimise user burden and explored two variations of how to present the option-criteria pairs to users. Because the SUS score for both conditions was very high, above 80.3 ("Grade A") (Bangor et al 2008), it is likely that users would recommend the system to other users (Sauro). The score also justifies the choice of sliders as the input elements.

Despite the similar SUS scores, users preferred the Option-Driven condition (19 users of 24). The reasons for this varied: "*...the criterion basically represents a question to me, so of course I want to think of one question at a time*" (P1_6), "*...so much easier to think about many options from the same point of view. I get a clearer overview of what's actually being asked*" (P1_8) or "*criteria are more 'abstract' and demand explanation, but solutions are easy to understand. So it's less reading this way*" (P1_11).

The 5 participants who preferred the CD condition indicated being so familiar with the options, that it was straightforward to contextualise to one at a time and rate the associated criteria. More research is called for to uncover how exactly are data affected by the design choice in question. However, when harvesting knowledge from crowds for a problem space such as ours, we argue in favour of design that pitches several options against one criterion at a time, as it was clearly cognitively easier for users.

6.1.2. Post-hoc analysis of results

Another common approach for quality control is post-hoc filtering of contributions. While one of the most commonly used techniques is to adopt a Gold Standard (Downs et al 2010), in our system we must deal with arbitrary tasks and crowds, which, by their very nature, produce subjective knowledge. Thus, this approach may be problematic, although clearly inappropriate options or criteria could be identified by the crowd. Nevertheless, since AnswerBot is designed to host the crowd's collective opinion, or Wisdom of the Crowd, instead of hard truth, about the option-criterion pairs, we should not impose Gold Standards, but rather strive to satisfy the four qualities described by Surowiecki (Surowiecki 2005).

Next, we consider another common method of assessing the performance of DSSs: the end-users themselves. In Study 2 the model using Euclidian distances to medians of the option-criterion pairs in the knowledge base performed well, with 14 users (of 16 users) finding it as the most accurate. They voiced comments such as "*Well, to me the list feels surprisingly accurate, quite correct in my opinion*." (P2_2) or "*9 times out of ten, I feel that [the Euclidian model] is clearly the best for me and gives me what I think it should*" (P2_4).

Another approach in our case is to analyse the extent to which workers agree with each other in their answers. Such analysis has been shown to yield valuable information about data reliability in crowd work contexts (Goncalves et al 2014a). Analysing Study 1 results, we identified certain anomalies in the input. For instance, the criterion "*quality of Special Foods*", in the restaurant question, caused high variance to ratings (Figure 4). Subsequent interviews found the criterion as ambiguous: participants could not be sure what it actually meant.

However, high variance does not always imply ambiguity. It could be just that the given option-

criterion pair splits opinions between e.g. different demographics, gender, or even individuals. A viable way forward here could be surveying all contributors, and offer suggestions based on different underlying crowd compositions. This was indeed also implied in the interviews, where participants indicated that they would find it useful to be able to specify underlying factors about the crowd used for populating the knowledge base.

6.2. The Crowd as Information Source

For AnswerBot to function, the collected knowledge has to be perceived as trustworthy by end-users. Overall, our participants indicated high trust in the system and the suggestions it provided. The fact that the underlying knowledge is harvested from a crowd, instead of a single expert or a static knowledge source clearly helped in creating trust. However, users would have desired to know more about the underlying crowd composition that might also affect the suggestions provided by AnswerBot.

Although the fact that users trust other users is hardly a new phenomenon (think of Amazon or TripAdvisor), leveraging this same trust in a PDSS provides a great opportunity. Just like in online rating platforms, the user base populating a knowledge base in our case should be at least somewhat knowledgeable about the topic. In Study 1 we addressed this issue by using trivial questions, but reaching out to the right crowd per each topic certainly is a challenge with our PDSS in the future.

Also the size of the crowd behind the input is interesting. Studies have shown that revealing or concealing its size affects user perceptions even if the underlying data does not change (Salganik et al 2006). The analysis in Study 1 suggests that we certainly do not need hundreds of contributors to rate an option-criterion pair for it to “stabilise” (as can be observed in Figure 5) and become reliable in the decision-making models. So, ultimately it is a matter of balancing between quantity and cost. If the crowd is a representative sample from a population (in our case, for example, students from our campus) then a small size would suffice. However, as indicated in our interviews, end-users might not trust the system much if they are aware that only a handful of people have contributed their knowledge to the issue.

6.3. Incentivising Contribution

A key challenge in obtaining input from crowds is incentivising honest participation (Chiu et al 2014). Further, an identified challenge for DSSs is rapid and cost-efficient data collection (Er 1988). Paid crowdsourcing offers one potential solution to these problems (Kittur et al 2013). Another cost-effective way to reach and engage users that are highly interested, and thus likely also knowledgeable, on a particular problem or topic is advertising. Platforms such as Taboola, Yahoo Gemini, or Facebook can reach targeted demographics and interests with ease. This approach has been demonstrated in the context of crowdsourcing by Ipeirotis (Ipeirotis, Gabrilovich 2014). Finally, situated crowdsourcing -- both paid and unpaid -- has emerged as a promising means for reaching users with wanted expertise, and has already been explored in a variety of contexts (Goncalves et al 2013; Goncalves et al 2014b; Goncalves et al 2014a; Hosio et al 2014; Hosio et al 2015).

However, an ideal solution for reaching the correct crowd, one can argue, is to provide enough value for individuals to contribute on their own initiative, drawing e.g. on intrinsic motivators (Kaufmann et al 2011). This is the case in online forums that stay alive thanks to their members' dedication to the cause. So, an important challenge in our work as well is motivating users to contribute. To this end, in Study 2 we discussed with participants what would motivate them to return to this system on their own. The consensus was that money, perhaps surprisingly, is not considered as an optimal driver for participation. Instead, intrinsic motivation and sense of community were hypothesized as drivers for participation.

6.4. Scaling AnswerBot Outside the Lab

In the studies analysed in this article, we verified that AnswerBot concept works, is easy to use, and in general makes sense to users. Even so, the question whether it scales and is feasible to be deployed standalone, outside the laboratory settings, remains. To this end, we have also trialled AnswerBot using an existing labour market, Bazaar (Hosio et al 2014), as the user and knowledge source. While the main focus of the Bazaar study is outside the scope of this article, we wish to include key evidence of AnswerBot's feasibility regards to operating *in-the-wild* here.

In the trial we collected 437 options and 185 criteria for 5 different problems. Further, the crowd moderated the entries (3936 moderation votes given), successfully purging irrelevant options and criteria. Then, in the final stage, users arriving via Bazaar contributed 17480 evaluations for the remaining option-criteria pairs. The entire deployment took 11 days and used standard payment rates of Bazaar (Hosio et al 2014). In this trial, no briefing or technical support was given to users who bootstrapped the knowledge bases by completing AnswerBot tasks in the labour market.

In the resulting data, clear differences emerged between the rated pairs, suggesting that the data reflected the crowd is valid for offering decision support. This is exactly what our DSS is designed

for: it builds a model of the proverbial "crowd brain" from the available knowledge on a given matter.

Second, we have started to use our DSS as an educational tool. AnswerBot combines both crowdsourcing and the founding theory behind Wisdom of the Crowd in an easy-to-use interface online. It is suitable for exemplifying both of these topics in a concrete fashion. We regularly organise workshops for senior high-school students to promote our department as a potential institution to study in. In these, we have noticed that it is a particularly captivating experience for the students to bootstrap and play with a DSS that focuses on a topic related to their own hometowns or schools (e.g., "*Who is the best teacher in our school?*", or "*What is the coolest place in [their home town]?*"). These workshops also provide us insights into how to develop the DSS in the future.

6.5. Limitations

The first shortcoming of our work is the limited user base in our studies 1 and 2: we assessed AnswerBot with 40 users (24 input, 16 decision support). While we could have opted for a wider audience (e.g. by using an online crowdsourcing platform), we decided to retain the control provided by a laboratory setting. Specifically, we needed to thoroughly interview people and make sure everyone provides ratings without disruptions. Still, we were able to show that 24 users were enough for providing decision support, although it is clearly not enough to rigorously examine how crowd composition (gender, age, education, etc) affect user perceptions about the quality of the offered support.

We note that neither the purpose nor the claim here is that we simply tap into MTurk, or any similar platform, and solve problems. AnswerBot is designed to harness a given crowd's wisdom. We emphasize that finding the correct crowd is a challenge for crowdsourcing markets or other platforms to solve. Our DSS simply then taps into that source and transforms the collective wisdom into decision support.

7. CONCLUSION AND ONGOING WORK

In this paper we present AnswerBot, a Personal Decision Support System powered by crowds. Whereas most DSSs are designed and fine-tuned for a particular task (Arnott, Pervan 2005) – and most likely do not generalize to other types of problems – we offer a system that can address several problem domains.

By using crowds our approach can help overcome a perennial problem of DSSs: populating the knowledge bases in an easy and cost-effective way (Er 1988; Geurts 1994).

AnswerBot performed well its two most crucial operational stages: in using crowds to create knowledge bases that support decision-making and providing decision support based on the knowledge bases. While our aim is to keep AnswerBot simple and intuitive to use, we constantly test new features that could make AnswerBot more capable. This is how all DSSs evolve (Keen 1981), and AnswerBot is no exception. In particular, we are investigating how to retrieve only relevant results from much larger knowledge bases than explored in this study. We will test methods and metrics, such as Discounted cumulative gain (DCG) (Järvelin, Kekäläinen 2000), from Information Retrieval (IR) literature.

Other issues to explore revolve around human factors: how to cost-effectively reach a crowd, and best transform the collectively intelligent input into trustworthy decision support? To this end, our ongoing research looks way beyond laboratory studies and paid participants.

First, together with physiatrists and physiotherapists we are mapping the ways to alleviate and cure lower back pain. Here, an added benefit is that by separating the expert crowd from the patients, the resulting differences in the knowledge base reveal interesting insights about the conceptions and misconceptions of the patients. This is useful for the practitioners especially when informing patients about optional treatment methods.

Second, we are collecting a knowledge base about how to best prevent everyday racism in Finland. The target crowd here is literally everyone, but we are offering the participants an extensive demographic survey, to enable more granular use and analysis of the result data.

At the time or writing this paper, these two prototypes have in less than a month gathered hundreds of text entries as options and criteria related to the issues and over 20000 ratings from thousands of unpaid visitors. These users act based on their own interest in the topic, and find the projects in word-of-mouth fashion online.

ACKNOWLEDGEMENTS

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 285062-iCYCLE, 286386-CPDSS, 285459-iSCIENCE), and the European Commission (Grants PCIG11-GA-2012-322138, 645706-GRAGE, and 6AIKA-A71143-AKAI)

REFERENCES

Alter, S., 1982, Decision support systems: Current practice and continuing challenges. Reading, Massachusetts: Addison-Wesley Publishing Co., 1980, 316 pp, *Behavioral Science*, 27(1), pp. 91-2.

De Angeli, A., Sutcliffe, A. & Hartmann, J. (2006) Interaction, Usability and Aesthetics: What Influences Users' Preferences? *Proceedings of the 6th Conference on Designing Interactive Systems,* ACM, pp. 271-80.

Arnott, D., Pervan, G., 2005, A critical analysis of decision support systems research, *Journal of information technology*, 20(2), pp. 67-87.

Ask Dan! about DSS - How does sensitivity analysis differ from "What if?" analysis?. Retrieved May 21, 2015, from http://dssresources.com/faq/index.php?action=artikel&id=121

Bangor, A., Kortum, P.T., Miller, J.T., 2008, An Empirical Evaluation of the System Usability Scale, *Intl. Journal of Human–Computer Interaction*, 24(6), pp. 574-94.

Barnett, G.O., Cimino, J.J., Hupp, J.A., Hoffer, E.P., 1987, DXplain. An evolving diagnostic decision-support system, *JAMA*, 258(1), pp. 67-74.

Bernstein, M.S. et al (2010) Soylent: A Word Processor with a Crowd Inside. *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology,* ACM, pp. 313-22.

Chiu, C., Liang, T., Turban, E., 2014, What can crowdsourcing do for decision support? *Decision Support Systems*, 65, pp. 40-9.

Downs, J.S., Holbrook, M.B., Sheng, S. & Cranor, L.F. (2010) Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* ACM, pp. 2399-402.

Druzdzel, M.J. & Flynn, R.R., 1999, *Decision support systems. Encyclopedia of library and information science,* Marcel Dekker, Inc. Last Login,.

Er, M.C., 1988, Decision Support Systems: A Summary, Problems, and Future Trends, *Decis. Support Syst.*, 4(3), pp. 355-63.

Finlay, P.N., 1994, *Introducing decision support systems,* NCC Blackwell ; Cambridge, Mass., USA : Blackwell Publishers, Oxford, UK.

Galton, F., 1907, Vox populi (the wisdom of crowds), *Nature*, 75, pp. 450-1.

Geurts, M.D., (1994) Data problems in decision support systems. *Hawaii International Conference on System Sciences,* IEEE, pp. 155-8.

Goncalves, J., Hosio, S., Ferreira, D. & Kostakos, V. (2014a) Game of Words: Tagging Places through Crowdsourcing on Public Displays. *Designing Interactive Systems,* Vancouver, BC, Canada, pp. 705-14.

Goncalves, J., Hosio, S., Liu, Y., Kostakos, V., 2014b, Eliciting Situated Feedback: A Comparison of Paper, Web Forms and Public Displays, *Displays*, 35(1), pp. 27-37.

Goncalves, J. et al (2013) Crowdsourcing on the spot: altruistic use of public displays, feasibility, performance, and behaviours. *International Joint Conference on Pervasive and Ubiquitous Computing,* ACM, pp. 753-62.

Hosio, S., Goncalves, J., Kostakos, V. & Riekki, J. 2014, Exploring Civic Engagement on Public Displays, in S Saeed (ed), *User-Centric Technology Design for Nonprofit and Civic Engagements,* Springer International Publishing, pp. 91-111.

Hosio, S., Goncalves, J., Kostakos, V., Riekki, J., 2015, Crowdsourcing Public Opinion using Urban Pervasive Technologies: Lessons from Real-Life Experiments in Oulu, *Policy & Internet*, 7(2), pp. 203-22.

Hosio, S. et al (2012) From school food to skate parks in a few clicks: using public displays to bootstrap civic engagement of the young. *International Conference on Pervasive Computing,* Springer, pp. 425-42.

Hosio, S. et al (2014) Situated Crowdsourcing Using a Market Model. *User Interface Software and Technology,* ACM, pp. 55-64.

Huffaker, D., 2010, Dimensions of Leadership and Social Influence in Online Communities, *Human Communication Research*, 36(4), pp. 593-617.

Ipeirotis, P.G. & Gabrilovich, E. (2014) Quizz: Targeted crowdsourcing with a billion (potential) users. *Proceedings of the 23rd international conference on World wide web,* pp. 143-54.

Jannach, D., Zanker, M., Ge, M. & Gröning, M. 2012, Recommender Systems in Computer Science and Information Systems – A Landscape of Research, in *E-Commerce and Web Technologies,* Springer Berlin Heidelberg, pp. 76-87.

Järvelin, K. & Kekäläinen, J. (2000) IR Evaluation Methods for Retrieving Highly Relevant Documents. *Proceedings of the 23rd Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, ACM, pp. 41-8.

Kaufmann, N., Schulze, T. & Veit, D. (2011) More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. *AMCIS, .*

Keen, P.G.W., 1981, Value Analysis: Justifying Decision Support Systems, *MIS Q.*, 5(1), pp. 1-15.

Kittur, A. et al (2013) The future of crowd work. *Proceedings of the 2013 conference on Computer supported cooperative work,* ACM, pp. 1301-18.

Kostakos, V., (2009) Is the Crowd's Wisdom Biased? A Quantitative Analysis of Three Online Communities. *Computational Science and Engineering,* pp. 251-5.

Landwehr, N., Hall, M., Frank, E., 2005, Logistic Model Trees, *Machine Learning*, 59(1-2), pp. 161-205.

Lorenz, J., Rauhut, H., Schweitzer, F., Helbing, D., 2011, How social influence can undermine the wisdom of crowd effect, *Proceedings of the National Academy of Sciences*, 108(22), pp. 9020-5.

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., Fisher, P., 2007, The Hawthorne Effect: a randomised, controlled trial, *BMC medical research methodology*, 7, p. 30.

Nonnecke, B. & Preece, J. (2000) Lurker demographics: counting the silent. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems,* ACM, pp. 73-80.

Noronha, J., Hysen, E., Zhang, H. & Gajos, K.Z. (2011) Platemate: crowdsourcing nutritional analysis from food photographs. *Proceedings of the 24th annual ACM symposium on User interface software and technology,* pp. 1-12.

Page, S.E., 2008, *The difference : how the power of diversity creates better groups, firms, schools, and societies,* Princeton University Press, Princeton, N.J.; Woodstock.

Poetz, M.K., Schreier, M., 2012, The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas? *Journal of Product Innovation Management*, 29(2), pp. 245-56.

Power, D.J., 2002, *Decision support systems : concepts and resources for managers,* Quorum Books, Westport, Conn..

Power, D.J., Sharda, R., 2007, Model-driven decision support systems: Concepts and research directions, *Decision Support Systems*, 43(3), pp. 1044-61.

Preece, J., Nonnecke, B., Andrews, D., 2004, The top five reasons for lurking: improving community experiences for everyone, *Computers in Human Behavior*, 20(2), pp. 201 - 223.

Resnick, P., Varian, H.R., 1997, Recommender Systems, *Commun. ACM*, 40(3), pp. 56-8.

Ricci, F., Rokach, L. & Shapira, B. 2011, Introduction to Recommender Systems Handbook, in *Recommender Systems Handbook,* Springer US, pp. 1-35.

Rogstadius, J. et al (2011) An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *International AAAI Conference on Web and Social Media,* Barcelona, Spain, pp. 321-8.

Salganik, MJ, Dodds PS & Watts DJ 2006, Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market, *Science,*February 2006, pp. 854-6.

Sauro, JMeasuring U, Retrieved September 28, 2014, from http://www.measuringu.com/sus.php

Shambaugh, N. 2009, Personalized Decision Support Systems, in JR Rabunal, J Dorado & A Pazos Sierra (eds), *Encyclopedia of artificial intelligence,* Information Science Reference, Hershey, PA,.

Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., Carlsson, C., 2002, Past, present, and future of decision support technology, *Decision Support Systems*, 33(2), pp. 111-26.

Surowiecki, J., 2005, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations,* 1st ed. Anchor, New York.

VotingAid, *VotingAid.* Retrieved September 16, 2015, from http://zef.fi/votingaid/en/home/

Wang, Y., Wang, Y., Patel, S., Patel, D., 2006, A layered reference model of the brain (LRMB), *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 36(2), pp. 124-33.

Yaniv, I., Milyavsky, M., 2007, Using advice from multiple sources to revise and improve judgments, *Organizational Behavior and Human Decision Processes*, 103(1), pp. 104-20.