Karlsruhe Reports in Informatics 2013,2
Edited by Karlsruhe Institute of Technology,
Faculty of Informatics
ISSN 2190-4782

A New Approach to Large-Scale Deliberation

Sanja Tanasijevic, Klemens Böhm

2013

KIT – University of the State of Baden-Wuerttemberg and National Research Center of the Helmholtz Association

A New Approach to Large-Scale Deliberation

Sanja Tanasijevic, Klemens Böhm
KIT, Karlsruhe, Germany
sannya.tanasijevic@kit.edu, klemens.boehm@kit.edu

Abstract. In this article, we propose a novel approach for identifying and evaluation of different solutions to discussed issues in online settings, based on the structure of a discussion of the topic in question. Our approach consists of three steps: (1) assigning weights to participants based on formal criteria such as degree of engagement in the discussion; (2) assigning scores to comments, taking the weights of authors and raters into account; (3) assigning scores to proposals, based on the scores of the pro and contra arguments. So an important idea is that individuals whose behavior is in line with our formal criteria have a higher influence on the decisions. Having built a respective online platform, we have evaluated the proposed model by means of an experiment with more than 100 participants who have discussed several topics relevant to them and a subsequent survey. In the survey, the majority of participants has expressed satisfaction with our forum model, including our weighting scheme. In particular, they have been fond of it regarding respect of the opinions of others.

Introduction

The question how communities can come to decisions and solutions that are satisfying for most of their members continues to be fundamentally important. There currently are various experiments and online projects trying to foster *deliberation*, i.e., the thoughtful consideration of all sides of an issue. This includes discussions and voting on budget planning for the German cities of Essen or Stuttgart (essen-kriegt-die-kurve.de, buergerhaushalt-stuttgart.de), to give some examples. The limitations of these two projects, however, are exemplary of the

ones of many other initiatives. For instance, essen-kriegt-die-kurve lets individuals propose concrete budget cuts and discuss these proposals. This project also tries to come to some conclusions from the discussion, by using rather simple quantitative measures such as the number of pro arguments regarding a proposal. However, this does not say much about the importance and relevance of the various arguments. In particular, people have started discussing issues not related to the proposal and have repeated arguments; this has affected those measures nevertheless. With Bürgerhaushalt Stuttgart in turn, individuals can come up with proposals others can then vote on. Our perspective is that, in such contexts, vested interests and priorities of individuals affect the outcome much more than argument quality. Our starting point is the hypothesis that making the community deliberate on the issues in question should increase its satisfaction with subsequent decisions – if the scheme that identifies and selects solutions to the issues discussed takes this deliberation into account. In a nutshell, the question investigated here is how online deliberation can be organized so that satisfying decisions can be derived from it.

Challenges

Our objective is the design of a platform that allows deriving satisfying decisions from the discussion; this is not obvious. In real life and in other studies, e.g., ConsiderIt (Kriplean et al., 2012), one usually takes pro and contra arguments into account when making a decision. We do the same, for each proposal: Proposals are discussed in different threads where people can provide pro and contra arguments for each one, and an automated scheme selects a winner proposal in the end. In other words, decision-making is mainly based on the structure of the arguments, as opposed to voting. Thus, a first challenge is to decide which information to collect from individuals. At first sight, information that is useful includes whether an individual agrees or disagrees with a comment, or feedback on which comments he deems off-topic, repetitions etc. However, we need to flesh out which information is indeed collected. A subsequent challenge then is how we use this information to come to a decision. The decision-making scheme must be understandable and non-ambiguous, to enable participants to provide reliable meta-information and feedback. Finally, evaluating any approach that claims to foster deliberation is challenging as well.

Design decisions

To shed more light on our approach, we now list out our main design decisions, at different levels of abstraction.

The look-and-feel of our deliberation forum is the one of a conventional forum wherever possible. Design and interaction features of our forum mainly are the ones of a classical forum. An alternative would have been an

entirely new design. However, user acceptance is crucial in our context, and our choice is likely to be better in this respect. Further, we can leverage existing technology and, hence, the host of comfort features provided by current implementations.

Comments are typed, and the typing mimics common argumentation structures. We have introduced comment types such as pro and contra arguments. In our forum, each proposal corresponds to a separate thread containing the respective arguments, so that their discussions are separated from each other.

The forum model should be simple and intuitive. We value simplicity of the model more than exactness and comprehensiveness. Literature has proposed various argumentation schemes with a high degree of sophistication, e.g., (Walton, Reed 2002; Parelman, Tyteca 1969; Toulmin 1958; Verheij 2006; Walton, Godden 2005; Restall et al. 2005). However, instead of having a model that is comprehensive but overly complicated for non-experts, and familiarization with it requires a lot of effort, we have limited our model to elementary comment and rating types.

Community members have different weights, according to formal criteria. In order to incentivize community members to follow our guidelines when deliberating, we have decided to assign them weights, and a higher weight gives an individual more influence on the decisions which will be taken. Next, a weight solely depends on formal criteria such as number of arguments provided, or share of arguments not flagged as repetitions by the community, subsequently referred to *indicators*. Note that individuals with low weights can still influence the outcome by coming up with proposals or arguments that a weight-based majority is in favor of. While each argument meets a certain degree of agreement in the community, the weight of an individual does not depend on this degree of agreement of his arguments. The rationale is to not discriminate against minority opinions.

The weight of a participant is the minimum of all his indicator values. An indicator value reveals the extent of a participant obeying the respective formal criterion. We deem it important that participants observe all of our criteria, and we want to incentivize such behavior. To illustrate, we do not want to give a high weight to someone who has issued many arguments if the community labels many of them as off-topic. Hence, the weight is the minimum of all indicator values.

Individuals can give feedback on contributions by others, feedback is typed, and it is used according to its type. Participants may issue feedback on contributions by others, which is then used in different ways. For instance, participants can state that they agree or disagree with an argument issued by someone else or can mark it as off-topic or as a repetition of a previous argument. Given this feedback, an idea might be to combine the various feedback items of different type into one argument score. However, we have found this too undifferentiated.

For instance, agreement/disagreement ratings are used to quantify the acceptance by the community, while off-topic/repetition feedback is used to reject comments.

Weights of individuals are published in the community. The alternative would be to not show this information so that participants are not influenced by it. Our decision has been to display current indicator values to give the participants an idea how their behavior so far has affected their weights. The rationale has been that this might stimulate the behavior desired.

We evaluate our approach experimentally. An alternative to experiments would have been a formal analysis or simulations. A difficulty with these alleged alternatives – at this stage of the project – is that they require various assumptions, e.g., how the number of arguments generated by different individuals is distributed, what is the ratio of off-topic arguments etc.

Contributions

Our contributions are as follows: First, we motivate and propose various criteria that constitute desirable behavior of community members, e.g., originality of arguments, focus on the topic in question etc., and propose formalizations of each of them. To stimulate desirable behavior, each community member has a weight that depends on the degree of adherence to our criteria. The weight determines his influence on the decision to be taken. Next, we propose a decision-making scheme that is argument-based. With our scheme, each argument is assigned a score that depends on the degree of agreement it has obtained from the community and on the weights of the respective individuals. We also formalize when an argument is rejected, i.e., ignored by the decision-making scheme. Our scheme assigns each proposal a score that depends on the share of pro and contra arguments and their scores. In our setup, proposals are alternatives to each other, and the proposal with the highest score will be the winner proposal. Finally, we evaluate our approach in a setting that is very close to a real one, with more than 100 participants. Students of the database course at our university have deliberated on various topics relevant to them. An important result is that the majority has expressed satisfaction with the weighting criteria and decision-making scheme, and they have given preference to our forum model over plain voting in terms of quality of decisions taken, mutual respect of opinion etc.

Related Work

Deliberation is a form of discussion where participants share their considerations in order to make decisions of higher quality and legitimacy (Chambers, 1996; Cohen, 1989; Carpini, Cook, Jacobs, 2004; Fearson, 1998; Fishkin, 1991, 1995; Gastil, 2000; Gutmann, Thompson, 1996). There are several problems with deliberation such as the diversity of the views of participants (Mutz, Marting,

2001), their lack of willingness of respecting deliberative rules (Conover, Searing, and Crewe 2002) or the nature of content allowed in deliberation. Various projects have studied how to nudge discussants towards more balanced considerations. This includes reflexive examination of own reasoning as well as of others. Effective moderation is considered crucial (Edwards, 2008). Otherwise, the perceived anonymity in forums can lead to 'flame wars', polarized debates and dominant minorities. With our approach, the working hypothesis is that our formal criteria are effective in keeping such behavior off.

Related work has identified patterns in deliberative discussions and motives for participation, one of which is true interest in the issues to be deliberated (Habermas; Carpini et al., 2004; Freelon, 2010). In the experiment in (Iyengar, Luskin, 2004), deliberation has yielded a significant increase in the informedness and engagement of participants. The Agora project (Muhlberger, 2004) has addressed the conflict between intensified grouping, opinion polarization, and normative conformity of the group (Kiesler et al., 1984; Kiesler, Sproull, 1992). The E-Deliberate project has tried to impose a certain structure on the decision-making process, but this has been hard to obtain online (Schuler 2009). Thus, the types of the postings we will provide allows to mimic the nature of the deliberative discussion, as described in those articles.

Another group of projects has explored the potential of deliberation using argumentation schemes, e.g., Issue-Based Information Systems (IBIS) (Isenmann et al., 1997). The tool resulting from the Cohere project tries to establish a system of social networking and reputation in the community through idea linking (Shum, 2008). Bart Verheij has proposed various argument-assistance systems such as 'ArguMed', 'Argue!' etc. Still users experience difficulties when formalization is required. In our approach, we have targeted at an argumentation model that is intuitive, rather than exhaustive.

Some recent projects have focused on expanding perspectives on the issues in question. The NewsCube project has tried to broaden views on news by giving several viewpoints (Park et al., 2009). Reflect is a system that engages and motivates discussants to restate, identify and share common grounds (Kriplean, Morgan, 2011). Opinion Space is an online interface incorporating ideas of deliberative polling, collaborative filtering for visualization and navigation through diverse comments (Faribani, Bitton et al., 2010). In our forum model we have tried to mitigate forming groups by enforcing anonymity, and by not explicitly displaying the types of the postings.

Deliberation forum model

We now describe our forum model in detail. We will elaborate on the interface, the discussion structure, the argumentation model, and the implementation.

Discussion structure and comments types

The following is a comprehensive description of the discussion structure and its representation in our model. The discussion structure has the following elements:
Forum (issue). A forum corresponds to the subject of discussion, e.g., 'How should EUR 500 be spent?'.
Thread. Each thread within its forum discusses one specific suggestion on how the issue in question could be solved.
Comments. Comments are the constituents of a thread, i.e., a comment is always part of a specific thread. Comments are typed, e.g., pro argument or contra argument. A comment can refer to another comment.
Ratings. A rating expresses the perspective of an individual on a comment posted by someone else. In our context, a rating is a complex structure consisting of various attributes, e.g., whether the individual agrees or disagrees with the comment, how he evaluates the writing style or the tone of the comment etc.

There are different comment types:

A **proposal** is a suggestion how to solve a forum issue. To illustrate, one issue in our study has been which criterion should be used to give away an iPad. One proposal has been to give it to the student with the highest number of points in the exercise in the current semester.
Extension of a proposal. Individuals can extend a proposal by means of a comment (in contrast to issuing a new proposal). To illustrate, an extension of the proposal just mentioned has been to use the number of points in the exercise to assign a certain number of lots to individuals, and more points increase the number of lots and the probability of winning the iPad.
A **pro argument** is a comment in favor of a proposal.
A **contra argument** is a comment against a proposal.
Other is a comment which the author does not want to classify as one of the types just mentioned.

Rating model

Participants can express their opinion on comments by others by means of a rating. In our context, a rating consists of the following attributes:
Content. Individuals can assess a comment by content using one of the following options: agreement, disagreement, repetition, and off-topic.
Writing style. Writing style can be evaluated using the grading scale from 1 to 5. Rate (5) represents clear, concise, argumentative writing style as opposed to unclear, confusing, incomplete text (1).
Tone. Analogously, tone can be (5) balanced and polite, as opposed to provocative and offensive (1).

Comment type. To ensure that comment types as specified by the authors are correct other users can state the type of a comment as part of a rating as well. The possible values are proposal extension, pro argument, contra argument, and other.

Weighting scheme

Participants have a weight that is based on the formal criteria which describe the desired behavior of individuals. As mentioned, the weight of participants depends on indicators being quantifications of these criteria and determines their influence on the decisions taken eventually. We provide an overview of the criteria before giving the respective formal definitions:

Originality. This indicator has a high value if few comments issued by the participant in question are rated as repetitions by many others.

Focus. The fewer comments by the participant are rated as off-topic, the higher will be the value of the indicator.

Style. The value of this indicator directly depends on the writing-style ratings of her/his comments.

Tone. The value of the tone indicator directly depends on the tone ratings of the comments by the participants.

Engagement. This indicator comprises the number of comments and ratings issued by the participant.

Individuality. The rationale behind this criterion is to make collusion attacks and team-ups of individuals more difficult and to curb the influence of herding behavior. Individuality is the share of participants whom the participant in question agrees with in some context and disagrees with in some other context. To illustrate, a participant being a perfect match with many other participants regarding comments and ratings has a low value regarding this criterion.

Breadth. We postulate that participants engaged in many discussion threads should be rewarded. The rationale is to curb the influence of participants with vested interests who only put attention to their specific issue.

Honesty. The rationale here is to ensure honest behavior of participants. In recent years, economic literature has proposed a number of methods to maximize the reward for individuals answering questions truthfully, even in the absence of an objective truth criterion, so-called *honest feedback mechanisms (HFM).* For instance, the so-called *peer-prediction method* applies scoring rules to the posterior belief on ratings by others, and honest reporting turns out to be a Nash Equilibrium (Miller et al., 2005). The *Bayesian truth serum* in turn does not assume a probabilistic relationship between different responses. It uses ‘the surprisingly common’ criterion as truth criterion. According to Prelec (2004) the premise behind this is as follows: If people have a certain belief they tend to believe that this belief is more common than it actually is. We for our part use the

peer-prediction method; it assigns scores for each rating based on its probability compared to the reference rating (Jurca et al., 2006).

While this is the list of criteria we have come up with after lengthy considerations, we do not claim at this point to have indeed covered all aspects of desirable behavior. However, we are confident not to face major difficulties when coming up with further criteria, redefining ours or even omitting some.

Formulae and notation

P is the set of all participants. $K^{create}(j)$ is the set of all comments Participant j has posted. K is the set of all comments posted in all forums. T is the set of all threads, and $K(t)$ is the set of comments in Thread t. $K^{create}(j,t)$ contains the comments posted by Participant j in t. $F \subset T$ represents a forum. $K(F) = \bigcup_{t \in F} K(t)$ is the set of all comments in Forum F. $R^{create}(j)$ is the set of ratings which Participant j has posted. A rating consists of the following information: content rating, writing style and tone rating, type of the comment and the rater. The type of 'content' is the enumeration that takes values from {agree, disagree, off-topic, repetition}. Writing style and writing tone can take values from 1 to 5. The type of 'comment type' is the enumeration with the following values: extension of a proposal, pro argumentation, contra argumentation and other. In our system, a rating does not have to be complete, i.e., participants can leave open individual values. Each rater can submit only one rating of a comment. R is a set of all ratings, irrespective of who has issued them. $R(k)$ is the set of ratings on Comment k, $R^{subject}(j)$ is the set of ratings on comments issued by Participant j, while $R^{subject}_{off-topic}(j)$ is the set of 'off-topic' ratings of comments of Participant j.

Definition. *$T^{create}(j)$ is the set of all threads Participant j has actively participated in by posting a relatively high number of comments.*

$$T^{create}(j) := \left\{ t \in T \mid \left| K^{create}(j,t) \right| > \underset{i \in P}{avg}\left(\left| K^{create}(i,t) \right| \right) / 2 \right\}$$

Here, we only count threads where the participant has at least posted half of the average number of comments. The rationale has been to have a certain level of engagement as a prerequisite for active participation. The threshold value itself is ad-hoc.

Definition: *Breadth of Participant j.*

$$breadth(j) := \frac{\left| T^{create}(j) \right|}{|T|}$$

Definition: *Focus of Participant j.*

$$focus(j) := 1 - \frac{\left| R^{subject}_{off-topic}(j) \right|}{\left| R^{subject}(j) \right|}$$

Definition: *Originality of Participant j*.

$$orig(j) := 1 - \frac{\left|R^{subject}_{repetition}(j)\right|}{\left|R^{subject}(j)\right|}$$

Definition: *A comment is useful when less than 50% of its ratings are 'off-topic' and 'repetition'*. The set of useful comments posted by Participant *j* is $K^{create}_{useful}(j)$, while the set of all these comments unrelated to a specific author is K_{useful}.

Definition: *Engagement of Participant j*.

$$engage(j) := \frac{\left|K^{create}_{useful}(j)\right|}{avg_{i \in P}\left(\left|K^{create}_{useful}(i)\right|\right)} + \alpha_{engage} \cdot \frac{\left|R^{create}(j)\right|}{avg_{i \in P}\left(\left|R^{create}(i)\right|\right)}$$

The weight α_{engage} is used to give different weights to comments and ratings. Since writing a comment requires more time and effort than submitting a rating, we have set ponder to 0.25 haphazardly. An alleged alternative has been to use $\alpha_{engage} = \frac{\left|K_{useful}\right|}{\left|R\right|}$. However, this value is not known *a priori*.

Definition: *A tone rating is bad when a tone attribute has a value of 1 or 2*. $R^{create}_{tone-}(j)$ is the set of bad tone ratings of the comments that Participant *j* has posted.

Definition: *Tone of Participant j*.

$$tone(j) := 1 - \frac{\left|R^{subject}_{tone-}(j)\right|}{\left|R^{subject}(j)\right|}$$

Definition: *A writing style rating is bad if it has a value of 1 or 2*. $R^{create}_{style-}(j)$ is the set of bad style ratings of the comments Participant *j* has posted.

Definition: Writing style of Participant *j*.

$$style(j) := 1 - \frac{\left|R^{subject}_{style-}(j)\right|}{\left|R^{subject}(j)\right|}$$

Definition: *HFM score of Participant j*. For each rating Participant *j* has posted, he receives a score based on the probability distribution of the given rating and the scoring function used by the peer prediction method (Miller, Resnick, Zeckhauser, 2005). In our implementation, the scoring function is the linear programming function proposed by Jurca (2006). It maximizes the payoff when a participant is honest. The indicator value *hfm(j)* is the average of all HFM scores Participant *j* has received for his ratings according to the method.

Definition: *Similar ratings*. Two ratings are similar when they refer to the same comment, and both either have value ’agreement‘ or ’disagreement’.

Definition: *Set of pairs of similar ratings posted by Participants i and j*. $R_{simil}(i, j)$ is the maximal set of tuples of similar ratings (r_1, r_2) where Participant *i* has posted r_1 and Participant *j* has posted r_2.

Definition: *Similar comments*. Two comments are similar when they are in the same proposal thread, and both either are of type ’pro argument‘ or ’contra argument’.

Definition: *Set of similar pro comments for participants i and j in the thread t.*
$K_{simil}^{pro}(i, j, t)$:= max(number of pro-argument comments posted by Participant *i* in Thread *t*, number of pro-argument comments posted by Participant *j* in a thread *t*).

We define the number of pairs of contra argument for Participant *i* and Participant *j* in Thread t analogously. The number of pairs of similar comments for Participants *i* and *j* is the sum of $K_{simil}^{pro}(i, j, t)$ and $K_{simil}^{contra}(i, j, t)$. $K_{simil}(i, j)$ is the sum of $K_{simil}(i, j)$ over all threads. We define the number of tuples of different ratings and comments ($R_{dissimil}$, $K_{disimil}$) analogously.

Definition: *Consensus of participants i and j*.

$$cons(i, j) := \frac{R_{simil}(i, j) + K_{simil}(i, j)}{R_{simil}(i, j) + R_{dissimil}(i, j) + K_{simil}(i, j) + K_{dissimil}(i, j)}$$

Definition: *Non-consensus of Participants i and j*.

$$noncons(i, j) := \frac{R_{dissimil}(i, j) + K_{dissimil}(i, j)}{R_{simil}(i, j) + R_{dissimil}(i, j) + K_{simil}(i, j) + K_{dissimil}(i, j)}$$

Definition: *Participants partly different from Participant j*.

$$P^{partlyDiff}(j) := \{i \in P \mid cons(i, j) > 0.3 \wedge noncons(i, j) > 0.3\}$$

A participant has a high individuality if there are many participants he has consensus and non-consensus with at the same time. Again, we have set the threshold value to 30% somewhat haphazardly.

Definition: *Individuality of Participant j*.

$$indiv(j) := \frac{\left|P^{partlyDiff}(j)\right|}{|P|}$$

All indicator values are in the range [0, 1]. We have seen two alternatives to normalize these values. Here, normalization does not only take the values, but also their distribution in the community into account. The normalized value of an indicator is the share of participants who have an indicator value lower than the one of the current participant. To illustrate, if only 20% of the community have performed better than Participant *j* regarding criterion breadth, *j*’s normalized

value of indicator breadth is 0.8. The advantage of this kind of normalization is that it distributes the participants over the entire [0, 1] range and makes criteria comparable. The disadvantage is when the majority performs similarly. Then slight deviations can have a significant effect. This is why we have not normalized indicators focus, originality, and style in this way. We have assumed that only a few participants would post off-topic or repetition comments, and if someone has a value slightly worse than average, this kind of normalization would have really set him back. The remaining indicators however are normalized in this way.

Definition: *Normalization of an indicator by frequency distribution*. The normalized value of an indicator of Participant *j* is the share of participants whose indicator value is less than or equal to the value of *j*. We use the notation $indicator^{norm}$, e.g., $indiv^{norm}(j)$, for normalized indicator values.

Definition: *Weight of a participant*.

$$WEIGHT(j) := \min\begin{pmatrix} focus(j), orig(j), style(j), tone(j), breadth^{norm}(j), \\ engage^{norm}(j), indiv^{norm}(j), hfmscore^{norm}(j) \end{pmatrix}$$

A participant must perform well regarding all criteria in order to have a high weight. One reason why we use the minimum function here is that this becomes clear to the user as well. It should now be obvious to him which aspects of his behavior he needs to devote more attention to in order to receive a higher weight.

Decision-making scheme

Our decision-making scheme is argument-based. Each argument receives a score dependent on the degree of agreement it has obtained from the community and the weights of the respective individuals. Next, our scheme assigns each proposal a score that depends on the pro and contra arguments and their scores. In our setup, proposals are alternatives to each other, and the one with the highest score will be the winner proposal.

Formulae and notation

$K_{ref}(p)$ is the set of comments in the thread belonging to Proposal *p*. $K^{+}_{ref}(p)$ is the set of pro arguments related to *p*, $K^{-}_{ref}(p)$ the set of contra arguments. The author of Comment *k* is denoted by *author(k)*. $R_{ref}(k)$ is the set of all ratings of Comment *k*. $R^{+}_{ref}(k)$ is the set of ratings of type 'agreement' while $R^{-}_{ref}(k)$ is the set of 'disagreement' ratings for Comment *k*.

Definition: *Comment score.*

$$score(k) := \left(\frac{weight(author(k)) + \sum_{r \in R^{+}_{ref}(k)} weight(issuer(r))}{weight(author(k)) + \sum_{r \in \left(R^{+}_{ref}(k) \cup R^{-}_{ref}(k) \right)} weight(issuer(r))} - 0{,}5 \right) \cdot w_1(k)$$

$$w_1(k) =: \frac{weight(author(k)) + \sum_{r \in \left(R^{+}_{ref}(k) \cup R^{-}_{ref}(k) \right)} weight(issuer(r))}{\max_{k' \in K(F)} \left(weight(author(k')) + \sum_{r \in \left(R^{+}_{ref}(k) \cup R^{-}_{ref}(k) \right)} weight(issuer(r)) \right)}$$

The score of a comment depends on the weight of its author and raters, and on the share of agreement ratings in the set of all ratings it has received. In addition, Weight w_1 takes into account the number of participants having issued ratings of Comment *k* and normalizes the scores in the forum thread, using the maximum sum of weights of author and raters.

Definition: *Proposal score, pscore.*

$$pscore(p) = \frac{\sum_{k \in \left(K^{+}_{ref}(p) \cup \{p\} \right)} score_k - \sum_{k \in K^{-}_{ref}(p)} score_k}{\max_{p' \in F} \left(\left| \sum_{k \in \left(K^{+}_{ref}(p') \cup \{p'\} \right)} score_k - \sum_{k \in K^{-}_{ref}(p')} score_k \right| \right)}$$

The score of a proposal depends on the scores of its pro and contra arguments. The more pro arguments there are, and the higher their scores are, the higher is the proposal score. Similarly, the fewer contra arguments there are, and the lower their scores are, the higher is the proposal score. Additionally, scores are normalized on the forum level, to make scores in different forums comparable.

The evaluation of proposal extensions is a difficult issue considering that the context of these extensions is not bounded in any way. In particular, extensions can address different perspectives of the proposal; they can mutually exclude each other or not. We have left the question how to score them as future work and have evaluated them by hand in this current study.

Anonymity

Our forum is anonymous. The names of comments' authors or raters are not visible. The rationale has been to indeed put the focus on the comments and the argumentation and not on the persons involved. Further, the type of a comment as specified by its author is not displayed. For instance, if a person is strongly in favor of a certain proposal, he might rate the contra arguments negative a priori

without even bothering to read. Similarly, summaries of ratings of comments issued so far are not shown either to avoid influencing participants.

Hypotheses

We have evaluated our forum model by means of an extensive user study. Before describing it, we list some of our hypotheses, together with their rationale.

H1: Participants have deemed our weighting scheme fair. We are interested in the perception of the fairness of the model by participants, including our choice of criteria and the technical details of the indicator calculation.

H2: The perception of usefulness of decision-making scheme is positively correlated with the perceived fairness of the weighting model. The fairer the weights are perceived, the better is the evaluation of the decision-making scheme.

H3: The perceived fairness of the weighting scheme is positively correlated with the degree of respect for the opinions of others. This hypothesis evaluates the effects of the weighting scheme on the evaluation of proposed solutions to the discussed issues. By assigning weights to the participants, they have different degrees of influence on the decisions.

H4: The higher the perceived usefulness of decision-making scheme, the more satisfied is the community with the winner proposals. If participants perceive the decision-making scheme as useful, it should have a positive effect on their attitude towards winner proposals.

H5: The higher the evaluation of the decision-making scheme, the higher is the perceived quality of the decisions. This claim is similar to Hypothesis H4, but with the distinction that the perceived quality of the decisions is affected.

H6: The perceived quality of the decisions is positively correlated with the participants' feeling that their opinion is respected. If participants think that their opinion is respected in the community, this should affect their evaluation of the quality of decisions in a positive way.

H7: The degree of adherence to our criteria is positively correlated with the fairness perceived. The rationale behind this hypothesis is to gather further evidence whether our design works at all.

H8: Displaying weights of participants affects their behavior. In other words, the indicator values displayed to all participants will influence their behavior in a way that is desirable.

Experimental setup

Our implementation of the forum model proposed so far is based on the open-source forum software *phpBB*. It is written in php and uses the MySQL database for persistent data storage. *phpBB* is listed in relevant blogs and forums as one of the top ten open-source forum projects. We have extended the existing *phpBB* platform with the specifics of our model: comment types, ratings and weighting

and decision-making scheme. Additionally, we have adapted the interface in order to anonymize the data, i.e., we did not want to display information such as the names of comment authors.

To evaluate the proposed model, we have run an experiment with 250 participants. The participants were students in the database course in the fourth semester of the KIT Bachelor program in computer science. The experiment was running for four weeks. In this time period, students have discussed several issues relevant to them. To illustrate, the list of topics is following:

What should be the topic of the last session of this database course? We have come up with the following three proposals ourselves, in line with the knowledge and interests of the instructor, and have made them available for discussion: data management in the cloud, introduction to database security, introduction to the development of database applications.

How should a budget of EUR 500 be spent on behalf of the students? Only proposals which are in line with the German regulations on how public money may be spent are admitted by the moderator. 'Beer' is an example of a proposal that is not acceptable.

We have procured a new iPad (3rd generation, Wi-Fi, 16 GB) to give away; who should receive it. Proposals containing the names of individuals or circumscriptions of concrete individuals are not accepted, only abstract specifications such as 'the best student in the class'.

Assuming that the computer-science department is able to fund a new chair, what should be its research direction? The winner proposal (and only the winner proposal, in order to avoid information overload) will be brought to the attention of the dean of the department.

What should be the topic of a new course in the area of databases/information systems in the next academic year? We have promised that the winner proposal will indeed materialize.

Given the current criteria for the selection of students for the KIT master program in computer science, which one should get a higher weight? The winner proposal will be brought to the attention of the dean.

Regarding the current KIT bachelor program in computer science, what is the most urgent reform? The winner proposal will be brought to the attention of the dean.

We point out that we have announced that the decisions by the group are binding to us. For instance, we have promised that we will indeed offer the course with the highest degree of agreement in the subsequent academic year, and that we will try hard to find a lecturer in case we are not able to teach it ourselves (analogously with the iPad or the EUR 500). Regarding the issues where the winner proposal is brought to the attention of the dean, we also deem this a real incentive, since it should be of interest to the management of the department to get to know the preferences of an entire age group.

To illustrate the effects of moderation, we have discarded the suggestion that EUR 500 should be used to buy cake to throw at each other. However, once a proposal had been approved, we have not filtered any arguments referring to it. For all issues, we have made it clear that there will only be one winner proposal, e.g., the EUR 500 will not be split. The rationale has been that we indeed wanted to study how the community deals with the situation where proposals compete with each other.

In our specific setup, a further incentive for taking part in the forum discussions were bonus points for the final exam, as follows: A participant must have posted 5 comments, none of them off-topic or repetition, and 20 ratings in order to receive a bonus of 5% of the points one could earn in the exam. With fewer comments and ratings, the bonus has been proportionally smaller. Obviously, an urgent question now is whether this bonus is the only rationale for participation. However, statements in the questionnaire and participation statistics indicate that a significant number of participants have been interested in the forum discussions themselves. Out of 163 participants who have posted at least one comment, 74 have posted more than five comments; out of 156 participants who have submitted at least one rating, 103 participants have generated more than 20 ratings. Thus, while that bonus might have influenced participant behavior, it obviously is not the only stimulus for participation.

We have decided to evaluate our forum model by means of a questionnaire. At an early stage of the project, we had considered forming a committee of experts who would assess the various proposals. However, it is difficult to impossible to decide which proposal actually is good, and which one is not. To illustrate, even ‘beer’ might actually be a good proposal, since it fosters socializing within that community – even though the organizers of this experiment might not like it. Further, our research question has been how to arrive at decisions satisfying to most of the community members and not necessarily at good decisions. We point out that privacy is valued highly in Germany, and we have done the evaluation anonymously (and actually had to go through significant effort to facilitate that bonus-point regulation). In consequence, we could not relate questionnaire answers to user behavior in our system. We do plan to analyze the user data collected from our system in detail, but such a study exceeds the scope of this article.

Results

In total, 250 participants have registered. 163 of them have generated at least one comment, and 156 have issued at least one rating. 116 participants have filled out the questionnaire. As described earlier, there have been seven different forum issues, and participants could generate proposals for six of them. The moderator had approved 88 proposals altogether, and 963 comments were generated in total.

We now say which hypotheses we have been able to validate in our setting.

H1: Participants have deemed our weighting scheme fair. Looking at the absolute numbers, 19 participants out of the total number of 116 participants have rated the fairness of the model as moderate. Recall that the grading scale ranges from 1 (not fair at all) up to 5 (very fair). Here, 'moderate' means Rates 1 or 2. Thus, the hypothesis is confirmed. – The criteria with the highest correlation with the perceived fairness are the following ones: focus (7 moderate out of 116), tone (16 moderate out of 116), and honesty (15 out of 116). The highest positive correlation between the perception of the fairness of the weighting scheme and the fairness of the criteria is observed for the following criteria: tone ($r = 0.356591298$, $p < 0.001$), individuality ($r = 0.349014637$, $p < 0.001$), originality ($r = 0.335690504$, $p < 0.001$).

H2: The perception of usefulness of decision-making scheme is positively correlated with the perceived fairness of the weighting model. Our analysis of the questionnaire data shows that there is a significant correlation ($r = 0.543288363$, $p<0.001$). In absolute numbers, only 11 participants out of 116 have rated the decision-making scheme as moderate, 32 were neutral.

H3: The perceived fairness of the weighting scheme is positively correlated with the degree of respect for the opinions of others. We have not observed a significant correlation. One possible explanation is that participants have not seen/understood how their weights affect comment scores and the evaluation of suggested solutions.

H4: The higher the perceived usefulness of decision-making scheme, the more satisfied is the community with the winner proposals. We have not observed a significant correlation that confirms the relationship from the hypothesis. . Leaving aside that we have not been able to confirm that correlation, the usefulness of the decision-making scheme is high: 11 participants out of 116 have rated the decision-making scheme as moderate, 32 were neutral. Furthermore, there is evidence that people think that their opinion is respected. Out of 114 participants who have answered the question on the respect of opinion in the forum, 73 participants have given high rates, 24 were neutral and only 11 participants have found it unsatisfactory.

H5: The higher the evaluation of the decision-making scheme, the higher is the perceived quality of the decisions. Data analysis has shown a certain correlation ($r = 0.201883911$, $p < 0.05$). Still, the probability of misinterpreting the correlation is lower than 5% ($p<0.05$), but not smaller, and this leaves some uncertainty from a statistics point of view.

H6: The perceived quality of the decisions is positively correlated with the participants' feeling that their opinion is respected. The correlation is significant $r = 0.23265865$, $p < 0.02$. The quality of the final decision is closely related to the perceived respect of the opinion of others in the forum.

H7: The degree of adherence to our criteria is positively correlated with the fairness perceived. We have not discovered any significant correlation. Note that this does not mean that the relationship that forms the basis of the hypothesis does not exist; it is just that we have not been able to validate it in our setup and with our questionnaire. In absolute numbers, 31 participants out of 116 have stated that the criteria and the weights have influenced their behavior in the forum. See Section 6.1 for a respective discussion.

H8: Displaying weights of participants affects their behavior. Again, we cannot confirm this hypothesis here. Only 13 participants out of 116 have stated that their weights or the ones of their peers have affected them. We stress that we have communicated our criteria in detail; still, according to most participants, this has not influenced their behavior. Again, see Section 6.1 for a discussion.

A further point is that participants were honest when rating contributions of others. Out of 116 participants 110 claimed that they had behaved honestly. Additionally, in the control question more than 65% of participants have estimated that more than 70% of participants had behaved honestly. In our opinion, such a high percentage of participants deeming a rather large group of other participants honest in many situations is a positive result. The correlation between self-reported honesty and the perceived honesty of others is significant ($r = 0.360109957$, $p < 0.001$).

Discussion

Although the questionnaire results have been helpful to answer some of our questions, there are some results that leave room for different interpretations.

Questionnaire results

Looking at the free-text answers in the questionnaire, we for our part have gained the impression that the judgments on some points were sometimes based on superficial interpretations rather than on a thorough understanding of the issues. According to our web-access statistics, the majority of participants has not fully read the documentation of the weighting and the decision-making scheme. E.g., participants have evaluated criterion ’honesty‘ high, although most of them probably have not understood the peer-prediction method, and a few individuals have complained about their scores. Another example is that the decision-making scheme, though rated highly, has not yielded more acceptable decisions from the community point of view. The participants have acknowledged that respect of opinions of others is higher, and that decisions are of higher quality, but not higher tolerance towards decisions taken. Furthermore, participants rarely have given full-text answers containing constructive comments on how the model could be improved, or explaining why they have not been satisfied.

Democratic principle

Clearly, weighting participants based on their behavior means that participants have different influence on the decisions. The advantage is that this should serve as an incentive to take part in the deliberation in a constructive fashion. However, when trying to convince another community to adopt our approach in order to come to decisions, there has been some resentment that our scheme was ‘not democratic’ because of that reason. However, our approach does not violate the principle of equality according to the German constitution since it treats all participants equally; we have consulted with legal experts on this issue. Further, our perspective is that the criteria are clear and well-documented. In addition, one's opinion does not affect the weight since our criteria are purely formal and do not include the degree of agreement/disagreement of the community with the arguments. Further, participants with a low weight can still influence the decisions, by coming up with arguments that are well received by most community members.

Forum model

The motivation behind our work has been to foster deliberation and to give way to decisions widely accepted by the community. We have conducted our evaluation with the audience of a university course. This has some differences to other communities: First, an age group of university students, being roughly of the same age and sharing similar academic interests, is a relatively homogeneous group of individuals, compared to other settings. For instance, think of public or political discussions which gather different groups of individuals regarding motivation, interests, educational and social background. Second, in our context, while bonus points have been an important incentive, they have certainly not been the only stimulus for participation. Students have shown interest for the topics discussed, i.e., two third of the students who have posted at least one rating have posted more ratings than required to receive the full bonus. Finally, our rules for earning the bonus points have affected the behavior of participants. They should have posted a certain number of comments and ratings in order to earn this reward. These settings have advantages and disadvantages. While it might seem at first sight that this lets our approach appear in a better light, this is not necessarily the case. In particular, individuals who have only been interested in the bonus, but not in the issues to be deliberated had to generate comments and ratings. One would expect this ‘noise’ to curb the satisfaction of the rest of the community with our approach. Nevertheless, the satisfaction rate has been high, as described earlier. This gives way to the expectation that our approach will also work in settings without any external incentives such as bonus points. However, investigating this is future work.

Another issue is that the system is to some extent vulnerable to attacks such as the following ones: Individuals can team up, earn high weights by

deliberating issues of little interest to them, and then use their weights to influence decisions relevant to them. However, our criterion ’breadth’, while not ruling out this attack completely, does make it more difficult. Further, while it does not mean that this behavior pattern does not occur, participants in our study have not observed this kind of attack, at least according to the questionnaire. The question how to make this attack even more difficult is future work. Another problem is that we have observed that some comments did not have any relevance for the discussion; still they have not been marked as off-topic. By finding ways to reduce the number of or eliminate this kind of comment, the overall quality of the arguments would increase. One way to deal with this problem could be to introduce another category next to ‘repetition’ or ‘off-topic’, namely ‘irrelevant’ and to have a respective new criterion, i.e., participants must not post irrelevant comments. Another solution might be to leave aside arguments without any ratings or follow-up comments when computing proposal scores. This item is a specific example of a larger issue, namely that our model can still be improved. As mentioned, our model is ad-hoc, and improvements are likely to be possible. However, note that this is not in contradiction to our contributions. In a nutshell, our concern has been to check whether our specific model is useful.

As stated before, the evaluation of proposal extensions is an open issue which is very difficult to solve, considering the diversity of extensions. For instance, we do not see at this point how to decide whether two proposal extensions mutually exclude each other or could both be implemented. Further, even if we could answer this question, we would have to decide how to select the extensions to be implemented. Addressing these questions exceeds the scope of this current study and is future work. As mentioned, we have evaluated the extensions by hand in our current study. The fact that nobody from the community has brought up any concerns regarding this could indicate that participants might already be happy with a moderator/elected representative choosing the extensions to be implemented, as long as the proposal with the highest score will be carried out.

Conclusions

In this article, we have proposed a novel approach for identifying and evaluation of problem solutions in online settings, based on the discussion itself and its structure. The decision-making process we have proposed relies on the deliberative manner of collecting and exchanging arguments in order to weight solution options. In order to achieve a discussion structure that gives way to an evaluation of solution options we have come up with various extensions of conventional forums and discussion structures. We have evaluated our approach by conducting an experiment with individuals discussing topics of relevance for this particular community. Our overall impression of the discussions is that the

individuals have addressed the issues very well. The results we have presented here are from the survey conducted after the experiment. They suggested that that particular community had been satisfied with our forum model and the respective decisions.

References

Kriplean, Morgan, Freelon et al. (2012): 'Supporting reflective public thought with considerate', *ACM 2012 conference on Computer Supported Cooperative Work*

Walton, Reed (2002): Argumentation Schemes and Defeasible Inferences, ECAI'2002 Workshop on Computational Models of Natural Argument, 2002, pp. 45-55,

Verheij (2006): Evaluating Arguments Based on Toulmin's Scheme, *Argumentation*, vol. 19, 2006

Restall, July, Walton (2005): Justification of Argumentation Schemes, *The Australasian Journal of Logic*, 2005, pp. 1-13

Edwards (2008): Moderation in Government-Run Online Fora, *Encyclopedia of Information Science and Technology*, 2008

Freelon (2010): Analyzing online political discussion using three models of democratic communication, *Media & Society communication*, vol. 12, Nov. 2010, pp. 1172-1190

Iyengar, Luskin (2004): Deliberative Public Opinion in Presidential Primaries: Evidence from the Online Deliberative Poll, *In Voice and Citizenship Conference*, 2004.

Muhlberger (2004): The virtual agora project, *J. of Public Deliberation*, vol. 1, 2005

Schuler (2009): Online civic deliberation with E-liberate, in *Online Deliberation: Design, Research, & Practice*, edited by Davies, Center for the Study of Language and Information (CLSI), Stanford, California, November 2009, pp. 293-303

Isenmann, Reuter, Wolf (1997): IBIS - a Convincing Concept . . . But a Lousy Instrument?, *Conference on Designing interactive systems processes, practices, methods, and techniques*, 1997

Shum (2008): Cohere: Towards web 2.0 argumentation, *International Conference on Computational Models of Argument*, 2008

http://www.ai.rug.nl/~verheij/aaa/

Park, Kang, Chung et al. (2009): NewsCube : Delivering Multiple Aspects of News to Mitigate Media Bias, *Conference on Human Factors in Computing Systems*, 2009, pp. 443-452

Kriplean, Morgan (2011): REFLECT: Supporting Active Listening and Grounding on the Web through Restatement, Conference on Computer Supported Cooperative Work, 2011

Faribani, Bitton et al. (2010): Opinion space: a scalable tool for browsing online comments, *SIGCHI Conference on Human Factors in Computer Systems*, 2010, pp. 1175-1184

Miller, Resnick, Zeckhauser (2005): Eliciting Informative Feedback: The Peer-Prediction Method, *Management Science*, vol. 5, 2005

Prelec (2004): A Bayesian truth serum for subjective data, *Science*, vol. 306, no 5695, 15 October 2004, pp. 462-466

Jurca, Faltings (2006): Minimum payments that reward honest reputation feedback, *ACM conference on Electronic commerce*, 2006

https://www.phpbb.com/

http://shakuras.ipd.uni-karlsruhe.de/dbsforum/

http://shakuras.ipd.uni-karlsruhe.de/dbsforum/pdf/Questionnaire.pdf