Karlsruhe Reports in Informatics 2013,11
Edited by Karlsruhe Institute of Technology,
Faculty of Informatics
ISSN 2190-4782

Towards Effective Structure-Based Assessment of Arguments and Proposals in Online Deliberation

Sanja Tanasijevic, Klemens Böhm

2013

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

Towards Effective Structure-Based Assessment of Arguments and Proposals in Online Deliberation

Sanja Tanasijevic
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

+49 721 608 45433
sannya.tanasijevic@kit.edu

Klemens Böhm
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

+49 721 605043968
klemens.boehm@kit.edu

ABSTRACT
Deliberation, i.e., discussing and ranking different proposals and making decisions, is an important issue for many communities, be they political, be they boards of experts for a scientific issue. Online deliberation however has issues, such as unorganized content, off-topic or repetition postings, or aggressive and conflicting behavior of participants. To address these issues, based on a relatively simple argumentation model and on feedback of different type, we propose to weight community members in an elaborate manner; this in turn is used to score arguments and proposals. Given such a scoring scheme, it is important to examine to which extent individuals have understood and accepted the approach, to identify characteristics of 'good' discussants and of strong arguments and proposals, and to study the robustness of the approach with regard to minor changes. To this end, we have carried out an experiment with a real-world community which had to make subjective decisions on issues relevant to them, and we have analyzed the data generated by it systematically, covering the different layers of our approach. Our takeaway is that the approach proposed here is promising to improve deliberation in many settings.

Keywords
Online deliberation, forum, social data analysis

1. INTRODUCTION
Deliberation is the act where communities identify possible solutions for a problem and the one(s) from this space that best meet their needs [1][2]. The spectrum of communities whose discussions rely on reasons and arguments is broad: It not only includes groups of citizens, from (small) municipalities to much larger administrative units. It also ranges from communities in science and technology, including the teams developing software, and communities of online gamers to groups of experts within large companies or organizations. Many communities are small, consisting of about, say, 100 or 200 individuals.

In practice, deliberation faces problems: Major flaws of group discussions are poorly organized content, repetitions, off-topic comments, bad wording and aggressive and conflicting behavior of participants. Some recent projects, e.g., Deliberatorium [3], have tried to apply a very formal argumentation model to bring structure to online discussions and to facilitate content evaluation. However, such rigid formalisms often undermine the natural discussion flow and require a lot of effort from participants. The question we want to investigate here is whether a simple, intuitive argumentation model, but together with ratings by participants, possibly of different type, allows to identify useful points, arguments and convincing proposals.

Designing such a scoring scheme is not obvious. We for our part propose to weight participants based on the adherence to criteria which correspond to efficient discussion behavior, such as the absence of repetition or off-topic comments, clarity of argumentation etc. However, identifying arguments and deriving conclusions and decisions from a discussion still is difficult. Thus, a question we address is to what extent such a derivation can be based on the structure of the discussion. Further, in the discussions foreseen here, there is no objective truth criterion. Instead, criteria we target at include community satisfaction and consensus of opinions. This makes the assessment of approaches such as the one proposed here more difficult. Finally, the broad variety of communities relying on deliberation will make it necessary to accommodate small changes of the scoring scheme, targeting at specific communities. This means that our approach must be robust to such changes.

We have proposed a relatively simple argumentation model to categorize content and different rating types to assess its quality. The rationale has been to give a clear structure to the discussion and to nudge discussants towards deliberation. In more detail, participants discuss different proposals, each one in a separate thread (mainly by posting arguments in favor or against it). Participants also categorize their comments based on its content; examples of respective comment types are 'pro argument' or 'contra argument'. They can also assess comments by other participants, by giving feedback regarding the argumentation presented, post comments that explicitly express agreement or disagreement etc. The assessment can also refer to the clarity of writing, to the tonality of comments, or to the types of the comments. Based on all this information, our approach assesses potential solutions to discussion subjects which participants have proposed in the course of the discussion. With our approach, collecting ideas for solutions is as important as their evaluation. This is in slight contrast to other recent deliberation projects such as ConsiderIt [4], which focuses on the collection of pro and contra arguments.

The contributions of this paper are as follows: Firstly, we motivate and describe our criteria for efficiency of deliberation, e.g., originality of posts, comments focused on the topic etc. Adherence of participants to these criteria results in different participant weights. To incentivize such favorable behavior, an important design decision of ours has been to give participants different degrees of influence on the evaluation of the argumentation, contingent on their weights. Secondly we describe our argumentation model and its expected effects on the discussion structure. Next, we explain the different rating types, as mentioned in the previous paragraph. In contrast to, say,

Facebook's like option or up and down votes on Reddit [5] whose aim is to identify popular content, our rating system serves the deliberative nature of the discussion, tries to keep the discussion streamlined without repetitions, off-topic postings, or conflicting behavior and to support the evaluation of comments and proposals. We then present our scoring scheme for evaluating comments and proposals. Comment scores rely on community consensus, on agreements and disagreements received, and on weights of author and raters. Subsequently, proposals are scored based on the scores of comments referring to them.

We then present results gained in a comprehensive analysis of the data gathered in an experimental study with around 200 participants. In a four-week discussion, the participants have generated 954 comments and 3849 ratings. We have found forum participation to be satisfying. 164 participants have posted at least one comment and 175 have posted at least one rating. To illustrate further, we have been able to conclude the following points from the data: Participants have adopted our argumentation model quite easily and without any serious flaws. Few comments were uncategorized, and comment types reported by the authors and other community members have rarely been mismatched. There have been very few off-topic or repetition comments. We for our part come to the conclusion that the community has acknowledged that higher-weighted participants have made more interesting contributions. Finally, the scoring scheme has shown to be robust against minor modifications, as motivated earlier.

2. Related Work

In this section we introduce related work, as follows. First, we present projects that study specifics of social networks such as characteristics of online conversations, online participation and reputation mechanisms. Next, we outline work on group-decision making. Finally, we review work on community detection and examining the link structure of social networks (SNs).

A first group of related projects analyzes specifics of online conversations, i.e., social interactions mainly in the form of comments and feedback. Wang et al. have examined online conversations over a period of time [6]. They have proposed a model to predict growth and structural properties of conversation threads. They examine properties of online conversations such as user attention for new items, patterns in commenting behavior and social propagation. They have gathered results on three datasets: Digg, Reddit and Epinions. Other projects propose mathematical models to generate conversational structures [7] or for disagreement expression, discussion topics and finally for the intrinsic nature of discussion debates [8]. – A core difference to our work is that they study existing discussion platforms. We for our part aim to design an approach for online deliberation, which we then evaluate. Further, important questions relevant in our context are specific to the nature of our approach/our platform.

Numerous projects have explored reputation mechanisms, their connection with participation and user performance. Reputation is comparable to our weighting of participants based on their adherence to formal criteria such as breadth in interest for topics, number of posts and ratings etc., and it determines the influence of participants on the discussion. Tausczik and Pennebaker have shown that building reputation is a very important incentive for online participation [9]. The significance of a properly designed reputation mechanism (karma) has been shown for Reddit [5], a social voting website where the incoming stream of links is voted up or down. 'karma' is built upon posting links that others like and vote for. Similarly, Slashdot [10], a website for sharing technology-related news, builds user reputation (or karma) through a number of activities, including moderation of comments, or posting comments that get high scores. The aim of Slashdot and Reddit has been to nudge discussants to active participation. Nevertheless, Slashdot has not confirmed the relationship between high karma and the posting of top level comments or early postings. On the other side, these approaches have also revealed some issues in moderating content. User-based moderation with Slashdot has problems with overlooking or late detection of comments that are either very good or bad [10]. Voting up on Reddit has led to overlooking half of the most popular links when they were first submitted [5]. Their 'open' voting is different from 'closed' voting where users do not see any score [11]. We have relied on such experiences when designing our approach. To avoid influencing the 'opinion' of participants, we do not display comment ratings and scores.

Up and down votes or user moderation are ways to categorize online content. In online discussions, argumentation systems are often used to bring structure to conversations. Argumentation systems claim to address this issue, by providing a systematic structure that reduces redundancy and encourages clarity. Deliberatorium [3] uses the well-known IBIS argumentation formalism [12]. Members of a community build deliberation maps, tree-structured networks of posts each representing a unique issue (question to be answered), idea (possible answer to a question), or argument (pro or con for an idea or another argument). The Cohere project has aimed for a tool for distributed and asynchronous argumentation [13]. The IBIS formalism and other ones have limited application in real-life scenarios, due to acceptance of discourse and classification problems related to the completeness, comprehensiveness and pedantry of the classification [12]. Thus, we have targeted at a rather simple, not necessarily exhaustive argumentation model that a community could easily establish and accept.

Next, the problem of group decision-making is addressed in the literature. It is the process of arriving at a conclusion regarding a specific issue based on the opinions of multiple individuals; consensus of the participants is an important indication of group agreement or reliability [14]. The project described in [14] proposes a value-function approach to transforming verbal opinions into values on an interval scale and measuring group consensus based on value variability. Murrel [15] has explored the impacts of computer-based communication and group performance depending on the structure of communication systems. Two synchronous systems with different immediacy of interactions and feedback on group-decision making have been examined. Even with these early discussion platforms, results are that groups produce decisions superior to average initial individual solutions each user has submitted before joining the discussion. The comparison of results relies on a ground truth (expert opinions). In our settings there is no objective truth criterion, so the assessment is more complex and challenging. Hilmer and Dennis confirm that groupware increases the exchange of information for groups, but additional comments do not necessarily lead to better decisions in their context [16]. The study explores groupware processes that require group members to categorize information. Different groupware processes have different effects on attention to and integration of information, and ultimately on decision quality. Our motivation for the typing of comments has been to improve the way opinions are communicated.

Detecting communities, overlapping communities and their tracking over time in large SNs continues to be an important research issue. Although forum communities in our context tend to be much smaller than SNs, it could be interesting to identify

communities within them, based on, say, interests, topics, and social links and track them over time. Sophisticated algorithms tailored to different communities, discriminating criteria for differentiating communities etc. abound. Yhou et al. present an algorithm for community detection based on topics and social links in order to build multi-layer communities [17]. Coscia et al. propose an algorithm for community detection which extends existing clustering algorithms, combined with community networks identified by users [18]. Abrahao et al. propose analysis of community properties by means of a class separability framework [19]. Their approach assesses the structural dissimilarity among the output of multiple community detection algorithms. Others attempt to find communities in dynamic SNs [20]. Since these approaches have exclusively been tested on large data sets, it is unclear whether they would perform equally well for smaller communities. One study has explicitly mentioned this problem, due to the underlying statistics relying on large numbers.

3. Towards a Forum for Deliberation

A design objective of ours has been that the platform envisioned has the interface of a conventional discussion board as much as possible and incorporates its functionality. The rationale is that participants more readily accept our innovations if they occur in a context already familiar to them (we argue). In addition to the usual parts of online discussion models (posts and references to previous posts), our approach has several new features: (1) **comment types**, i.e., an author is supposed to categorize his content according to our argumentation model, e.g., 'pro argument' or 'contra argument' comment; (2) **multi-facet ratings**, e.g., participants can give their feedback on whether they agree or disagree with the content of a comment and rate its writing style, tone, or type. Quite naturally, these features require some modifications of the look-and-feel of a conventional discussion board. Further constituents of our approach are as follows: (3) a **weighting scheme** for participants reflects their adherence to our formal criteria for constructive deliberation and rewards them accordingly with different degrees of influence in the discussion; (4) a **scoring scheme** for the evaluation of comments; (5) a **scoring scheme for proposals** based on the argumentation presented. In the subsequent section, we describe these features in detail and motivate them.

3.1 Comment Types

To facilitate categorization of a post based on its argumentation and to support evaluation of the arguments, authors can classify comments in different types, namely proposal, proposal extension, pro argument, contra argument or other, according to the argumentation model we are proposing. Thus, we have slightly adjusted the discussion structure of conventional discussion boards to encompass these various categories, as follows:

A proposal represents an idea or suggestion how a discussion issue can be solved. In our study with a community of computer-science students (and this is also envisioned to be the case in future studies), there have been several discussion issues such as "How to spend a EUR 500,-- budget on behalf of the students?" or "Which student should an iPad be given to?". Each discussion issue forms a separate forum thread. The number of proposals for each discussion issue, which represent separate proposal threads within a forum thread, is arbitrary. Examples of proposals regarding one of the issues just mentioned have been "The EUR 500,-- can be spent to support a project for the live streaming of lectures." or "Improve WLAN at important spots on the university campus.".

A new proposal extension is a comment referring to proposal suggesting some improvements/extensions. I.e., one of the proposals for the issue "Which student should an iPad be given to?" has to relate the chance of winning the iPad with exam points, and a proposed extension has been to organize a lottery and assign lots proportional to the number of exam points earned.

A pro comment is a comment which contains argumentation in favor of a certain proposal.

A contra comment is a comment containing arguments/reasons against a certain proposal.

A comment of type 'other' is a comment which does not match the categorization just presented, or its author does not want to assign it to one of these categories.

The comment types presented and the underlying argumentation model have affected the discussion structure as follows. Discussion issues form separate **forum threads**. Within each forum thread there are different **proposal threads**. Authors post comments in proposal threads directly referring to the discussed proposal. This is the case even when they are posted as follow-ups of other comments.

3.2 Multi-facet Ratings

Our setting allows for feedback addressing different characteristics of a comment (of any type):

Content. The object of a content rating is the content of a comment. A rater can express his agreement or disagreement with an argument expressed or address structural characteristics of a comment, e.g., marking it as repetition or off-topic. In these cases, the grading scales are binary, e.g., a participant agrees with a comment or not.

Writing style. A rating for writing style reflects how a comment is written on the grading scale from 1 to 5, e.g., Rate 5 stands for clear and concise writing, Rate 1 for unclear, fragmentary input.

Tone. Tone ratings address the tonality of comments on the grading scale between 1 and 5. Rate 5 corresponds to comments which are balanced and polite, whereas provocative and confrontational comments are rated with 1.

Comment type. The object of a comment type rating is the comment type. In order to verify comment types, raters are invited to classify the argumentation of comments themselves.

3.3 Weighting Scheme

A core objective is to facilitate constructive deliberation without repetitions, off-topic comments, offensive and harsh behavior. To this end, we propose formalizations of unwanted behavior. So-called indicators quantify the degree of adherence of a participant to each criterion. As an incentive to refrain from such behavior, we confine the influence of participants with such behavior in the forum. – Our list of criteria is as follows:

Originality. A participant performs well regarding this criterion if he has posted no or very few repetitions of already existing comments. This criterion should decrease repetitions in the discussion.

Focus. The value of this indicator will be high if a participant has received no or a very few off-topic ratings for his comments. Our motivation is to lower the number of off-topic comments and make the discussion more efficient.

Style. If raters rate the writing style of the comments authored by an individual high this will affect his performance regarding this criterion. The rationale is to increase the clarity of comments.

Tone. The better the tone of an author is rated, the higher will be his value for the Tone criterion. The aim with this criterion is to keep the discussion friendly and balanced.

Engagement. The value of this criterion will be larger, the larger the number of posted comments and ratings by the author in question is. The rationale is to reward above-average active discussants.

Individuality. A participant performs better regarding this criterion if he has an individual opinion and is not exclusively expressing the same views as others. The objective also is to make teaming up of individuals and collusion attacks more difficult.

Breadth. The higher the engagement of a participant in different discussions, the higher is the value of the Breadth criterion. Our perspective is that participants with broad interests should be rewarded. Additionally, this criterion should make it more difficult for groups of individuals with specific, narrow interests to collude.

Honesty. The value of the Honesty criterion is larger, the larger the score of participant ratings, the so-called *hfmscore*, as assigned by the so-called peer prediction method [21]. The aim of this and similar approaches [22] is to maximize the reward for honest answers in the absence of an objective truth criterion. Since ratings are an important part of our approach, a mechanism to assess them is needed.

The criteria presented above are the result of intensive discussion between the authors of this article. Our objective with this current study is not to arrive at a list of criteria that is final and covers all aspects of desirable behavior in online deliberation. (Instead, we have aimed at coming up with one concrete proposal and then evaluate it subsequently.) Nevertheless, we hypothesize that we have identified the most important points for constructive and efficient discussions, taking into account problems that previous projects have faced.

3.4 Formulae and Notations

In this subsection we give a more rigorous introduction to our scoring scheme. First, to illustrate, we will elaborate on the formal criteria used to assign weights to participants, along with our formula for calculating this weight. Next, we introduce our formulae for calculating the comment and proposal scores.

Weighting scheme. Our weighting scheme relies on eight different indicators, as described informally in the previous subsection.

As stated already, our approach features ratings of different type. A rating consists of rates for: content {agreement, disagreement, off-topic, repetition, other}, writing style and tone, both presented by a grading scale between (1 – poor) and (5 – good), comment type {pro argument, contra argument, proposal extension, other}. R is the set of all ratings. R(k) is the set of ratings posted for Comment k, and $R^{create}(j)$ is the set of ratings posted by Participant j. The set of all ratings posted for comments of Author j is $R^{subject}(j)$, analogously, $R^{subject}_{off-topic}(j)$, $R^{subject}_{repetition}(j)$ are the sets of off-topic, repetition ratings respectively.

Focus. The indicator focus is defined as the ratio of the number of all off-topic ratings received for comments posted by Participant j over the same number of all ratings received.

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

Originality. The originality indicator is calculated mainly based on the share of off-topic ratings referring to comments issued by Participant j compared to all ratings referring to these comments.

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

Accordingly, $R^{subject}_{style-}(j)$, $R^{subject}_{tone-}(j)$ are the sets of negative ratings (Rate 1 and 2 on the grading scale 1 to 5) for style and tone for comments authored by Participant j respectively.

Style. The style indicator is calculated as follows:

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

Tone. The formula for the tone indicator of Participant j is as follows:

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

$K^{create}(j,t)$ is the set of all useful comments posted by Participant j in discussion thread t. Similarly, $K^{create}(j)$ is the set of useful comments by j in all threads. A useful comment is one that has less than 50% of off-topic or repetition ratings. P is the set of all participants.

Engagement. This indicator is calculated based on the number of ratings and comments issued by Participant j compared to average numbers over all participants.

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

Individuality. To compute the indicator for individuality, we rely on the following auxiliary measures: similarity of posting and of rating behavior of participants. Both of these measures rely on opinion. For instance, participants who agree on the same comments or who post pro comments for the same proposal have a similar opinion. $K^{pro}_{simil}(i,j,t)$ is the maximum of the number of pro comments authored by Participant i in Thread t and of the number of such comments authored by Participant j. $K^{contra}_{simil}(i,j,t)$ is defined analogously for contra comments. $K_{simil}(i,j)$ is the sum of those numbers over all threads. $K_{disimil}(i,j)$ is defined analogously. $R_{simil}(i,j)$ is the set of tuples of ratings (r1, r2) posted by Participant i and j for a comment that are both either agreement or disagreement. $R_{disimil}(i,j)$ in turn is the set of all tuples of ratings expressing different opinions for a comment issued by Participants i and j. Consensus and contention of Participants i and j now is defined as follows:

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

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

Next, based on consensus and contention of Participant j with other participants, we find $P^{partlyDiff}(j)$, the set of participants that sometimes match and sometimes do not match the opinion of j.

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

Finally, the individuality indicator is the ratio of these individuals over the whole set of participants.

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

T is the set of all forum threads, and $T^{create}(j)$ is the set of all forum threads Participant j has actively participated in. Active participation is given if the number of useful posts is at least half of the average number of useful posts by all participants in that forum, as formalized below.

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

Breadth. The value for breadth is the ratio of the number of forum threads where the participant has actively taken part in and the total number of forum threads:

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

Honesty. The honesty indicator is calculated using scores assigned by the peer prediction method [21]. Based on the probability distribution of the given rating and the scoring function, a score for the rating is assigned accordingly. hfmscore(j) is the average of all rating values issued by Participant j.

The first four indicators (originality, focus, style, tone) are not normalized. These indicators refer to a minority behavior such as posting repetitions or off-topic comments, and the rationale behind not normalizing is to demarcate minority behavior from regular behavior. On the other hand, engagement, individuality, breadth and hfmscore are normalized based on frequency. For instance, if 20% of the community has performed better than Participant j regarding the breadth criterion, j's normalized value of the breadth indicator is 0.8. The advantage of this normalization is that the distribution is uniform in the range [0, 1], and values for different criteria now are comparable.

The weight of a participant is the minimum of the different indicator values. We could have used another function here as well, but we have decided to use minimum function. This is to reward participants with higher weights when they obey all criteria.

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

Comment scoring scheme. To assess comments and their argumentation we propose a respective scoring scheme. It takes the following criteria into account: author weight, rater weights and agreement status in the community, e.g., the ratio of 'agreement' ratings received and all ratings. In what follows, *k* is a comment, $weight(author(k))$ is the weight of its author, and $weight(issuer(r))$ is the weight of the individual who has generated rating *r*. $R_{ref}(k)$ 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*. $K(F)$ is the set of all comments in Forum *F*.

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

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

Comment scores are normalized using Weight w_1. It is the ratio of the sum of weights of the author and the raters of Comment *k* and the maximum sum of weights of author and raters for any comment in the Forum F. The rationale is to normalize, i.e., to make comment score comparable on the forum level.

Proposal scoring scheme. To rank proposals regarding a certain discussion issue, we have weighed the argumentation posted based on the scores of the pro and contra arguments. $K_{ref}(p)$ 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.

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

To make proposal scores mutually comparable in a forum, they are normalized as well. We accomplish this by using the maximum difference between pro and contra arguments of a proposal in that particular forum.

4. Experimental Setup

To evaluate our approach we have conducted an experiment. Alternatives would have been to perform formal analyses or simulations. A downside of these altenative methods is that they require certain simplifications and assumptions such as the distribution of posts or ratings etc., which are unknown at this point. On the other hand, experimental studies are not easily repeatable at all, for various reasons: Recruiting a community is difficult and time-consuming, and if the experiment was repeated with another parameter setting, new issues to be discussed would have to be identified. Further, solutions/decisions drawn from the discussion need to be effectuated in order to make the experiment as realistic as possible, and this tends to be costly. (In our case, starting with the implementation of the platform and ending with the realization of the decisions of the community, the study has lasted about 15 months.)

The community, we have conducted our study with is second-year computer-science students of our university. In order to incentivize participation in principle, we have announced a small bonus for the final exam of the database course, which is mandatory in the fourth semester (5% of the exam points that could be earned in total for five comments (no repetitions or off-topic comments) and 20 ratings). Beyond that, we have announced that decisions are binding (for us, the organizers of the experiment), i.e., we will implement the winner proposals. In other, more general settings, an alternative way to avoid coldstart effects could be to 'purchase' proposals and arguments at Amazon

MTurk or some other crowdworking marketplace; exploring this is future work.

Some discussion topics have been as follows:

What should be the topic of the last session of the current database course? We have proposed three different lecture topics, which participants then discussed.

How should a budget of EUR 500,-- be spent on behalf of the students? We have required that the money must be spent as a whole and in a way that does not violate German regulations for spending public money.

We have offered a new iPad to be given away according to a criterion proposed by the community. A proposed criterion should be objectively measurable, so a proposal such as "John Doe" is not acceptable in this respect.

Assuming that the computer-science department had funding for a new chair, what should be its research direction?

What should be the topic of a new course in the area of database/information systems in the next academic year?

Considering the existing selection criteria for the KIT master program in computer science, which one should be given higher priority?

What is the most urgent reform of the KIT bachelor program in computer science?

We have promised to implement the winning proposals or to bring them to the attention of the higher instance within KIT that is responsible. More specifically, the outcomes of the discussion of the three topics described last, i.e., the winning proposals, are presented to the dean of the department.

The discussion itself has lasted four weeks. After that, we have indeed, say, spent EUR 500,-- (to support a project for the video streaming of lectures) and offered a course on NoSQL databases.

On the technical level, we have developed the discussion portal as an extension of the well-known open-source forum software phpbb [23]. It is originally written in php and supports various database systems such as MySQL, which we have used. Extensions we have introduced are new interface features such as specifying the type of a comment; multi-facet rating options, e.g., content/writing style ratings. See our forum [24]. Other important new features are the implementation of the weighting scheme to profile the influence of participants in the discussion and the scoring scheme described in Section 3.4. A preliminary paper of ours that also describes our approach [25] features an evaluation that is based on a questionnaire which we have given out after the forum discussion. According to the questionnaires, participants were satisfied with the approach, and they have given preference to it over plain voting in terms of quality of decisions taken, mutual respect of opinion etc.

5. Data Analysis

Beyond the questionnaire results, we expect an analysis of the data collected during the experiment to give us a broader perspective in terms of performance of our approach, its robustness and assessment even in the absence of an objective truth criterion. In this section we present results from an analysis of the experimental data. The structure of what follows is in line with the main points of our approach: weighting scheme, then scoring schemes for comments and proposals. To begin with, however, to gain further insight in the discussion board, lead discussions and board community, the posting behavior plays a significant role. These fundamentals can acknowledge the performance of our approach in a setting very close to the real world. Additionally, the analysis of posting behavior can provide indications of collusion or misuse attempts.

5.1 Posting Behavior

Number of Contributions. In the four-week experiment, 954 posts and 3849 ratings were generated. 198 participants were registered, and 169 (84%) have posted at least one comment or one rating. When we observe registration of participants, we can conclude that the distribution of registrations has peaks at the beginning of the discussion period and at one point of time when we announced that the discussion period was extended. Half of the total number of participants has registered in first four days of the experiment. Still, if we compare 'early' registered users to 'later' ones, the distribution of posts and ratings is quite uniform.

Figure 1 is a histogram of the share of registered participants with a certain number of comments posted. 70% of these participants posted five or more comments, whereas 38% of participants posted more than five comments (which has been the limit for receiving the small exam bonus in full). Similarly, as presented in Figure 2, 67% of the participants posted 20 ratings or more, and 55% of the participants posted more ratings than required for receiving the bonus. So there has been a significant intrinsic motivation to participate.

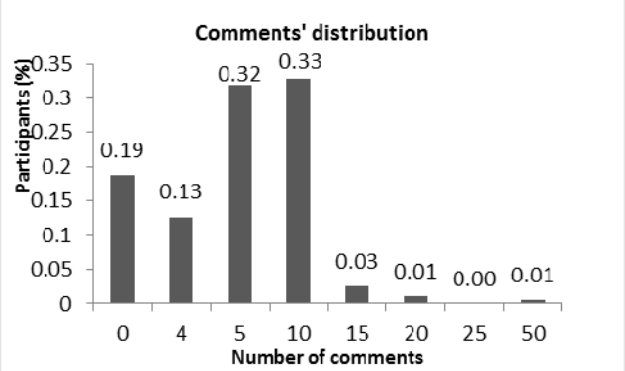

Figure 1. Distribution of comments

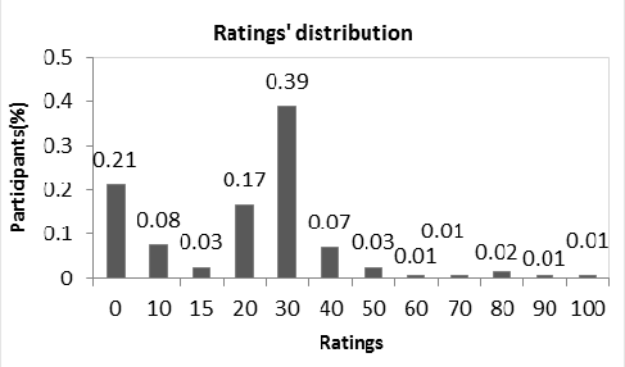

Figure 2. Distribution of ratings

Posting Behavior. According to the experimental data, we can confirm a significant correlation between the number of posts and the number of ratings per participant ($r=0.5748$, $p<0.0005$). Additionally, we discover a significant correlation between the number of follow-up comments, i.e., comments referring to other comments, and the number of ratings ($r=0.3821$, $p<0.001$). In other words, participants who posted more comments have also posted more ratings and more follow-up comments.

In the experiment, 161 participants (81% of registered participants) posted at least one comment. Out of a total number of 954 comments there are 88 proposals (9.22% of the comments), 162 extensions (16.98%), 84 'other' comments (8.8%), 241 pro arguments (25.26%), 379 contra arguments (39.73%). The type Other is rarely used, and this speaks in favor of the acceptance of our argumentation model. Furthermore, the participants who had more posts also had more posts of type Other, and the type is used by 54 different participants. The correlation between the number of posts per participant and the number of 'other' posts is significant, r=0.4668, p<0.001. The number of 'other' comments is fairly small compared to the total number of comments. On the other side, more than a quarter of all registered participants posted them, so it is not confined to a small group of participants. A possible explanation is that participants fail to categorize these comments in rare cases. Thus, according to our interpretation, they are exceptions, rather than indications of misuse attempts or of participants having misunderstood some underlying notions.

We have observed a significant correlation between the number of pro and contra comments by the date of post (r=0.886057, p<0.001). This correlation indicates that participants were involved in the discussion in a differentiated manner, responding to arguments with pro and contra arguments.

Rating Behavior. We now look at the number of ratings per post. In total there are 3849 ratings posted by 156 participants. Out of these, 2364 (61.42%) are agree ratings, 1058 (27.49%) disagree ratings, 141 repetition ratings (3.66%), off-topic (3.35%). So the majority of participants has used ratings to express agreement/disagreement with posted comments.

606 (out of 954) comments have received at least one agreement rating, while 375 comments received at least one disagree rating. Next, 95 comments received at least one repetition rating, and 28 of them received more than 50% of repetition ratings. 43 comments were rated as off-topic at least once, and 18 comments received more than 50% off-topic ratings. Thus, we can conclude that the number of comments marked as off-topic or repetition was small. We had asked two individuals not involved in the experiment to sort out the comments by hand; a result has been that a significant share of repetitions and off-topic comments actually is detected.

To summarize 729 comments received at least one rating; this represents 76% of all comments. This serves as an indication that our proposed approach was well accepted, and that the level of participation was rather satisfying. Agreement and disagreement ratings have shown that participants have followed our suggestion that ratings represent opinion expression, and we can see that ratings options were used extensively.

Qualitative Assessment of Comments. In order to introduce further qualitative measures for the evaluation of comments we define two notions: *rated* and *relevant* comments. A rated comment is one which has received at least one rating. A relevant comment is one which has received off-topic/repetition ratings in less than 50% of all its ratings. Comments with no ratings are relevant by definition. We use these notions to examine the potential effects of 'stricter' weighting and scoring schemes. The necessity of such stricter rules has come up when examining comments manually. There are some borderline comments close to being repetitions or featuring an argument with a somewhat loose connection to the discussion topic. Quite a number of them are unrated – a reader might have found these comments too difficult/too tedious to rate. Out of 954 comments, 913 are relevant (95.7%), 685 are rated and relevant (71.8%). The data presented reveals the significant share of rated comments. Thus, weights and scores as defined so far would be meaningful even with the stricter selection of comments.

Next, we have observed a significant correlation between the number of posts per participant and the number of his posts that have been rated (r=0.8482, p<0.001). There also is a significant correlation between the number of posts per participant and the relevance of his posts (<50% off-topic/repetition ratings) (r=0.9765, p<0.001). When combining these two measures for rated and at the same time relevant posts, we can also confirm a significant correlation between the number of posts and the one of relevant and rated posts (r=0.7936, p<0.001). In Figure 3 we can see the number of participants with a certain share of rated and relevant posts in all their posts. We conclude that participants who have participated in the discussion more actively also received significant attention from the community. In other words, more engaged participants were also better discussants, according to the ratings. Furthermore, 80% of the participants have posted relevant/rated comments with 80% of their posts. This speaks in favor of the relevance of the discussion and of a large share of good discussants with meaningful posts.

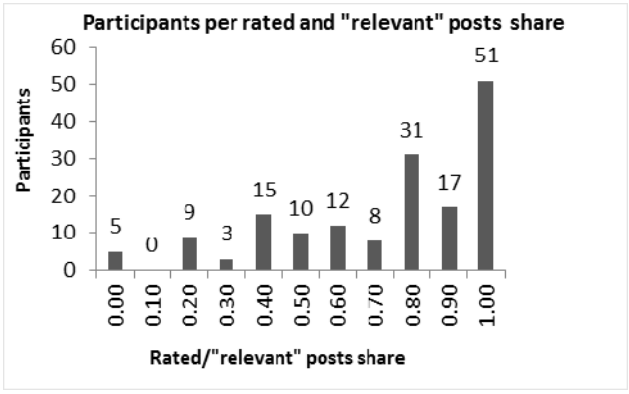

Figure 3. Participants per rated and "relevant" post share

Additionally, we have counted the numbers of rated and "relevant" comments per type. Figure 4 shows the distribution of rated, "relevant" and rated/"relevant" comments per type.

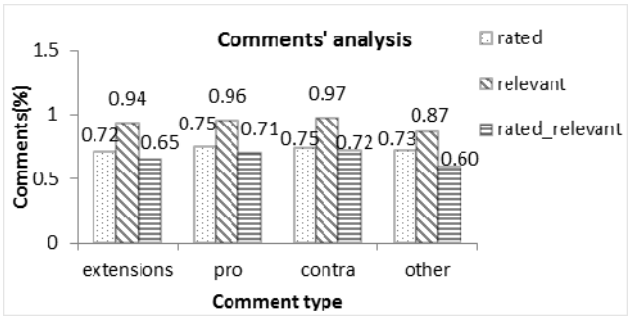

Figure 4. Analysis of comments

In the Figure 5 graphs the coverage of comments with ratings of different types. E.g., 606 posts have received at least one agreement rating, 375 comments at least one disagreement rating.

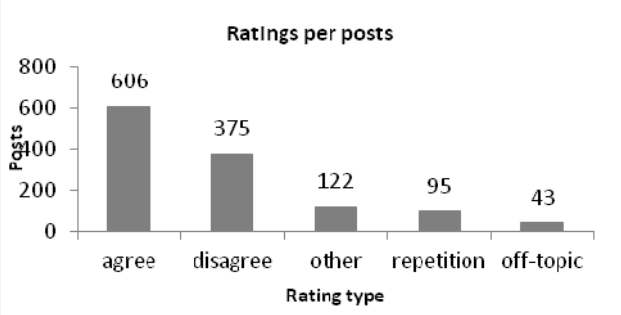

Figure 5. Ratings per posts

Finally, we have compared the comments types specified by authors to feedback information on comment types by other participants by hand. This has not always matched well. In particular, proposal extensions have sometimes been perceived as pro or contra argument comments and vice versa. When studying the robustness of our approach in a later subsection, we will check to which extent these mismatches play a role. Anticipating the respective result, it turns out that the difference is not large. Nevertheless, a potential improvement of our approach could be to allow for such 'reasonable' mismatches, while taking 'real' mismatches (e.g., a fraction of participants stating that a posting is a pro argument and another fraction stating that it is a contra argument) into account. With such a refinement, it would still be plausible why participants should strive to identify the correct type of a comment.

engagement, and individuality are normalized to arrive at a uniform distribution of their values, in contrast to indicators that refer to a minority behavior. As expected, there are only few participants who have performed very badly regarding writing style, cf. Figure 3. On the other hand, normalized indicators, except for individuality, are grouped in ranges by performance and are less differentiated. Individuality is highly discriminative for participants, and, thus, prevents misuse of the system and herd behavior.

The fact that individuality is discriminative can also be seen in Figure 7. The chart shows the number of participants where the respective indicator value was minimal, compared to all other indicator values of the participant. I.e., when looking at the distribution of the engagement indicator, one can see that 20 participants had performed in the range [0.9, 1] and [0.8, 0.9], the value of 24 participants has been between 0.7 and 0.8 and so on. – By redefining indicators or normalizing them differently, the results could be quite different. The rationale behind discussing the concrete indicators values in our setting has been to show the performance of the experiment community and its effect on the evaluation of comments and proposals.

Given our data, we can confirm a significant correlation between the number of comments posted by a participant and his weight ($r=0.6783$, $p<0.0005$). I.e., an engaged participants with broad interests and cooperative behavior should be recognized as a good discussant.

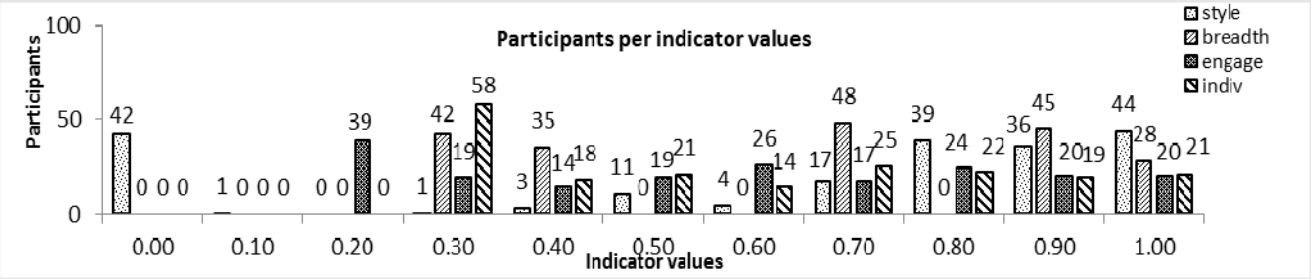

Figure 6. Participants per indicator values

Summary. Participants have been quite active and constant in participating in the discussion. When we look at the participation level and especially the share of participants with a large share of rated and relevant posts, we can conclude that participants have adopted our deliberation approach quite well. The discussion flow has been continuous and lively, responding to the arguments with pro and contra arguments. The majority of comments has received ratings; this is good because they are a prerequisite for the evaluation of comments and proposals.

5.2 Weighting Scheme

Our weighting scheme is based on formal criteria to facilitate efficient discussions without off-topic or repetition comments, offensive and aggressive tone etc.

In Figure 6, a set of indicators and the distribution of their values among participants are shown. In the course of the experiment, we have observed problems with our implementation of hfmscore, the score assigned by the peer prediction method, so we had omitted it as an argument of the minimum function. We had announced to the participants that we intend to fix it in short time, but ultimately have not been able to do so within the four-week discussion time, mainly because administrative issues have required a lot of our attention. As stated in Section 3.4, indicators such as breadth,

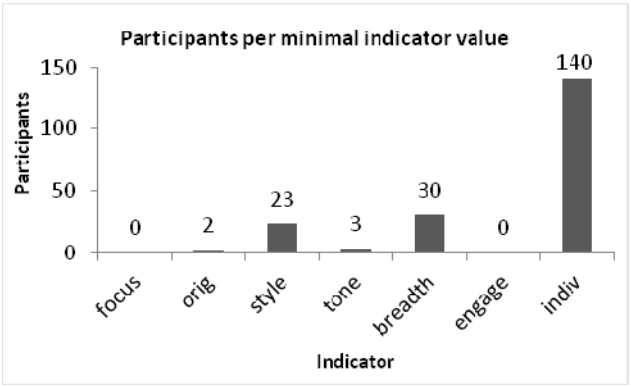

Figure 7. Participants per minimal indicator value

5.3 Comment Scoring Scheme

Comments are evaluated based on the respective scoring scheme. We clearly expect that comments that have received many agreement ratings are likely to be rated higher that the ones with only a few agreements or even a lot of disagreement ratings. The respective correlation is $r=0.46265$, $p<0.0005$. We also expect

comments rated or posted by high-weight participants to score higher than in case of low-weight raters or authors. The correlation between comment scores and author weights is p=0.4530, p<0.0005. The correlation between comment scores and rater weights is lower, with r=0.2625, but it is confirmed with p<0.0005 due to the sample size. Furthermore, posters with higher weights have received more ratings for their comments, and the correlation is significant, with r=0.4826, p<0.001. When we compare the weights of posters and the sum or average of their comment scores, we see significant correlations, r=0.5771, 0.4149 respectively with p<0.001. Based on this, we conclude that discussants with higher weights post comments rated with higher scores. The average score of the comments of these participants was higher, and they received more ratings for their comments. Thus, the community has perceived higher-weighted participants as discussants with more interesting contributions.

The listed results indicate that the comment scores reflect the criteria that we deem important in the evaluation, such as number of ratings received, the reputation of the author and the raters as well as agreement status in the community.

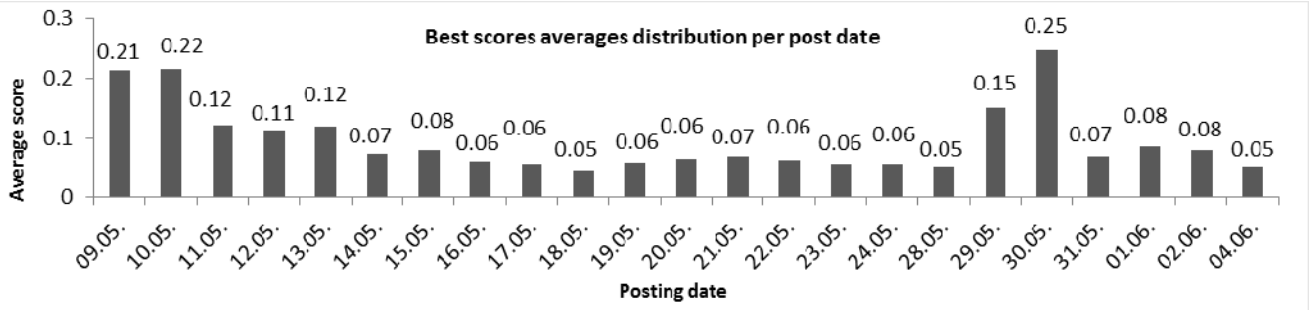

Figure 8 shows the average scores of best posts by post date

Best posts are the top quarter of all posts based on score. When analyzing the connection between the posting date and the final score, we can see that, as expected, early posts in general receive more attention and thus, the scores are higher. Still, if we look at the distribution of the best posts (the top quarter of the scores of all posts), there are peaks related to forum participation, i.e., the opening of the forum discussion and the point of time when the discussion has been extended. Latter comments can also achieve high scores. This speaks in favor of the scoring scheme.

5.4 Proposal Scoring Scheme

To analyze the effects of the proposal scoring scheme, we look at the outcome of proposal ratings and their common characteristics. An objective truth criterion does not exist in this case, and this represents a big challenge for assessing the scoring scheme. There are some rather obvious results, e.g., there is a significant correlation between the proposal score and the number of pro arguments, r=0.5899, p<0.001. Our analysis has not confirmed a significant correlation between the score of the poster and the number of positive proposals, e.g., proposals that have received more pro arguments than contra arguments. The highest number of positive proposals posted by one poster was three. The weight of this participant is in the best fifth of the ones of all participants. Six other participants posted two proposals each, and their weights vary from 0.402 in the top 45% of the participants to 0.8826, the second best participant. Although we cannot directly connect participant weights and proposal scores, we see some indications that participants who performed better regarding our formal criteria posted comparably better proposals. These indications include the correlation between author weights and comment scores or even between average comment scores and author weights.

In order to confirm the robustness of our approach in terms of its sensitivity to changes and different evaluation criteria, we have tried out several alternative scoring schemes and have examined their effect on the outcome (proposal scores in particular). We deem this necessary, since we had observed some imperfections of our approach in the course of our experiment, as mentioned earlier. In particular, there are mismatches of comment types by authors and other participants as well as unrated and irrelevant comments. Since carrying out another experiment clearly exceeds the scope of one publishable unit in terms of time and cost, we now examine how slight modifications of the approach affect the outcome based on the data collected in this current experiment. Ideally, we can verify that these slight changes do not influence the output significantly. We have studied the following alternatives:

(1) Use simple count of agreements and disagreements as up and down votes instead of using weights to calculate comment scores. These calculated comment scores are then used as input of the proposal scoring scheme.

(2) Stricter rules for useful comments, namely, comments that are not rated or not relevant are ignored.

(3) Resolve comment types based on type ratings with 60% and 80% certainty. In our current proposal-scoring scheme, we take the argument type as specified by the author at face value. To include comment-type checking e.g., if a comment is a pro argument or a contra argument, we have taken the type ratings received for comments into account. Here, we propose a threshold of 60% (80%) certainty when resolving the type. Comments whose type is not verified are ignored.

(4) Exclude low-weighted participants. We have omitted all participants whose weight is in the bottom third of all participant weights. Their comments and ratings were also ignored.

The correlation of the original proposal scores and the ones with the modified scoring schemes are significant: 0.9888, 0.9981, 0.9405 (60% certainty), 0.9731 (80% certainty), and 0.9970 respectively.

For the Alternative (1) the similarity of results is rather expected, since the correlation between comment scores and the number of ratings received is significant, r=.46265, p<0.0005. The stricter rule for including comments in the second scenario has not influenced the results much. This is because the comments with the lowest score were excluded. The certainty check for resolving comments types has not made a significant difference either. Finally, the last scenario that expels low-weighted participants, their comments and ratings can be interpreted as an indication that potential collusion attacks from these participants would not seriously affect the outcome either. – To summarize, we can

conclude that the scoring scheme is robust and resistant to changes in the evaluation scheme. Certain misuses that one might expect such as false comment types or misbehavior of low-weighted participants have not influenced the outcome by much.

6. Conclusions

We described and evaluated a new approach to facilitate efficient online deliberation. Our primary aim has been to organize the discussion around the deliberative principle of carefully considering pro and contra points. The essence of our approach is a three-step evaluation: First, based on a set of formal criteria, we rate users and assign them weights. Next, comments are evaluated based on the agreement/disagreement ratings of the argumentation referring to them and the weights of raters and authors. Finally, solutions of discussion issues are evaluated based on the scores of pro and contra arguments posted. To assess our approach, we have systematically analyzed the data collected in an experiment with around 200 participants. We can verify active participation during the four weeks of forum discussion. Significant numbers of comments and ratings have been generated, and a large share of participants has been active. We conclude that our proposed approach for online deliberation was well accepted in the community. In particular, the analysis of participant weights and of comment and proposal scores has demonstrated compliance with our assessment criteria such as community consensus, observance of the deliberation principle and fulfillment of requirements for efficient discussions. In our specific setting, the approach has shown to be very effective when dealing with repetitions, off-topic comments, or aggressive tone. We see this as an indication that it might perform similarly well in other settings.

7. REFERENCES

[1] D. N. Walton, and E. C. W. Krabbe, 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, Albany, NY.

[2] F. H. v. Eemeren, and R. Grootendorst, 2003. *A Systematic Theory of Argumentation: The Pragma- dialectical Approach*. Cambridge University Press.

[3] M. Klein, 2011. *How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium*. CCI working paper, Massachusetts Institute of Technology.

[4] T. Kriplean, J. Morgan, D. Freelon, A. Borning, L. Bennett, 2012. Supporting reflective public thought with ConsiderIt. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (Seattle, WA, USA, February 11 – 15, 2012). CSCW '12. ACM, New York, NY, 265-274. DOI= http://doi.acm.org/10.1145/2145204.2145249.

[5] E. Gilbert, 2013. Widerspread underprovision on Reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work* (San Antonio, Texas, USA, February 23 – 27, 2013). CSCW '13. ACM, New York, NY, 803-808. DOI= http://doi.acm.org/10.1145/2441776.2441866.

[6] C. Wang, M. Ye, and B.A. Huberman, 2012. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (Beijing, China, August 12 – 16, 2012). KDD '12. ACM, New York, NY, 244 - 252. DOI= http://doi.acm.org/10.1145/2339530.2339573.

[7] R. Kuvar, M. Mahdian, M. McGlohan, 2010. Dynamics of Conversations. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (Washington, USA, July 25 – 28, 2010). KDD '10. ACM, New York, NY, 553 - 562. DOI= http://doi.acm.org/10.1145/1835804.1835875.

[8] A. Mukherjee, B. Liu, 2012. Mining Contentions from Discussions and Debates. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (Beijing, China, August 12 – 16, 2012). KDD '12. ACM, New York, NY, 841 - 849. DOI= http://doi.acm.org/10.1145/2339530.2339664.

[9] Y. R. Tausczik, and J. W. Pennebaker, 2012. Participation in an Online Mathematics Community: Differentiating Motivations to Add. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (Seattle, WA, USA, February 11 – 15, 2012). CSCW '12. ACM, New York, NY, 207 - 216. DOI= http://doi.acm.org/10.1145/2145204.2145237.

[10] C. Lampe, E. Johnston, P. Resnick, 2007. Follow the reader: filtering comments on slashdot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, April 28 – May 03, 2007). CHI '07. ACM, New York, NY, 1253 - 1262. DOI= http://doi.acm.org/10.1145/1240624.1240815.

[11] P. Mieghem, 2011. Human Psychology of Common Appraisal: The Reddit Score, *IEEE Transactions on Multimedia*, 13 (2011), 1404-1406.

[12] S. Isenmann, and W. Reuter, 1997. IBIS - a Convincing Concept . . . But a Lousy Instrument?. In *Proceedings of the 2nd conference on Designing interactive systems processes, practices, methods, and techniques* (The Netherlands, Netherlands, August 18-20, 1997). DIS '97. ACM, New York, NY, 1253 - 1262. DOI= http://doi.acm.org/10.1145/263552.263602.

[13] S. Shum, 2008. Cohere: Towards web 2.0 argumentation. In *Proceedings of the 2008 conference on Computational Models of Argument: Proceedings of COMMA 2008* (Toulouse, France, May 28 – 30, 2008), IOS Press Amsterdam, The Netherlands, 97 – 108.

[14] Hsio, Lin, Chang, 2005. Value-based Consensus Measure on Verbal Opinions. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (Hawaii, USA, January 03 – 06, 2005). HICSS '05. IEE CS, Washington, DC, 9.3.

[15] S. Murrel, 1983. Computer communication system design affects group decision making. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '83. ACM, New York, NY, 63 - 67. DOI= http://doi.acm.org/10.1145/800045.801582.

[16] Hilmer, Dennis, 2000. Stimulating Thinking in Group Decision Making. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences* (Hawaii, USA, January 04 – 07, 2005). HICSS '00. IEE CS, Washington, DC, 1028.

[17] Zhou, Jin, Liu, 2012. Community discovery and profiling with Social Messages. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (Beijing, China, August 12 – 16, 2012). KDD '12. ACM, New York, NY, 388 - 396. DOI= http://doi.acm.org/10.1145/2339530.2339593.

[18] M. Coscia, G. Rosseti, F. Giannotti, D. Pedreschi, 2012. DEMON: a Local-First Discovery Method for Overlapping

Communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (Beijing, China, August 12 – 16, 2012). KDD '12. ACM, New York, NY, 615 - 623. DOI= http://doi.acm.org/10.1145/2339530.2339630.

[19] B. Abrahao, S. Soundarajan, J. Hopcroft, R. Kleinberg, 2012. On the Separability of Structural Classes of Communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (Beijing, China, August 12 – 16, 2012). KDD '12. ACM, New York, NY, 624 - 632. DOI= http://doi.acm.org/10.1145/2339530.2339631.

[20] C. Tantipathananandh, T. Y. Berger-Wolf, 2011. Finding Communities in Dynamic Social Networks. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining* (Vancouver, Canada, December 11- 14, 2011). ICDM '11, IEE CS, Washington, DC, 1236 – 1241. DOI= http://doi.acm.org/10.1109/ICDM.2011.67.

[21] R. Jurca, B. Faltings, 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce* (Ann Arbor, MI, USA, June 11 – 15, 2006). EC'11. ACM, New York, NY, 190 - 199. DOI= http://doi.acm.org/10.1145/1134707.1134728.

[22] N. Miller, P. Resnick, R. Zeckhauser, 2005. Eliciting Informative Feedback: The Peer-Prediction Method. Management Science, 51 (September 2011), 1359 – 1373.

[23] https://www.phpbb.com/

[24] http://shakuras.ipd.uni-karlsruhe.de/dbsforum/

[25] S. Tanasijevic, K. Böhm, 2012. A new approach to Large-Scale Deliberation. In *Proceedings of the third international Conference on Cloud and Green Computing* (Karlsruhe, Germany, September 30 – October 02). CGC '13. IEEE CS Press, Los Alamitos, CA