IADIS International Journal on Computer Science and Information Systems
Vol. 5, No.2, pp. 21-36
ISSN: 1646-3692

SENSE-MAKING AND ARGUMENTATION-BASED KNOWLEDGE: THE LOST ON THE MOON EXPERIMENT

Teppo Räisänen. *School of Business and Information Management, Oulu University of Applied Sciences, Finland*
Teuvo Pakkalan katu 19, 90130 Oulu.
Teppo.Raisanen@oamk.fi

Harri Oinas-Kukkonen. *Department of Information Processing Science, University of Oulu, Finland*
P.O.Box 3000, 90014 University of Oulu.
Harri.Oinas-Kukkonen@oulu.fi

ABSTRACT

This paper discusses the usability and applicability of argumentation-based knowledge. We are especially interested to find out how the sense-making of argumentation-based knowledge by contemporary Web users could be supported. We use the Lost on the Moon experiment for studying this and compare the achieved results between three experimental groups and a control group. These results demonstrate that support for explicating information and voting over proposed solutions may help in the sense-making process, in particular with issues which are in a true sense open for debate. The sense-making was partially improved in all three experimental versions implemented. The experiment emphasizes that in order to support complex problems it is the simple solutions with easy to use interfaces that are needed in the Web2.0 era.

KEYWORDS

Sense-making, comprehension, Web 2.0, rationale.

1. INTRODUCTION

Today, technologically-supported informal social networks are one of the major models of knowledge creation and exchange (Novak & Wurst, 2004). They are often referred to as virtual communities (Rheingold, 1993), communities of practice (Wegner & Snyder, 2000), or knowledge communities (Novak & Wurst, 2004).

These communities are groups of people who share a concern or a set of problems, and who deepen their knowledge and learn by spontaneously interacting on an ongoing basis (Wegner & Snyder, 2002). Discussion forums where people discuss about cooking or strategies on how to play video game are examples of such communities. They can be physically located, locally networked (e.g., via an Intranet), virtual (i.e., networked across distance), or a combination of these (Preece, 2004).

Within these communities individuals ask and answer questions, and they talk and discuss with other members. They also get support, reassurance, insights, and exposure to different value systems and beliefs (Preece, 2004). Thus both explicit and tacit knowledge can be exchanged. As members of the community interact with each other "gradually shared solutions and insights emerge that contribute to a common store of knowledge that accumulates over time" (Preece, 2004).

In a broad sense, there are two types of members in the Web communities: information providers and information users (Fisher et al., 2003). Information providers are those users who write Wikipedia articles or upload videos to YouTube. Information users most often only read articles or watch the videos.

In an online community information providers are a minority. Indeed more than 90% of participants can be called lurkers (Katz, 1998; Mason, 1999), i.e. members who very seldom post or take part in the conversations. Yet, even if lurkers do not take actively part in activities they not only benefit from the community but they can also benefit the community (Takahashi et al., 2003). For example, the more lurkers who contribute every once and a while (even if rarely) the more diversity and opinions the group will have. This is one of the key criteria of a smart group (Surowiecki, 2004). If only a handful of users contribute to the group's discussions there is a danger of the group becoming too homogeneous. Most of knowledge creation research in the Web2.0 context focuses on content creation (i.e. information providers' perspective), whereas we will focus here on information users' point of view.

One crucial skill that information users need is comprehension (Oinas-Kukkonen, 2004), i.e. the ability to make sense of the content found in the Web. Using sense-making users can internalize new knowledge they discover. One suggested way to promote sense-making is argumentation (Buckingham Shum et al., 1997). Argumentation can support both personal and shared cognition. Personal cognition can be supported by allowing user to see where the argumentation differs from his/her point of view. And shared cognition can be supported when a single user can see what the group as a whole has argued and which opinions have gained the most support.

The paper is organized as follows. The next chapter will present the background focusing on argumentation. Chapter 3 presents the experiment. Chapter 4 will present the results from the experiment. Chapter 5 discusses the research carried out and chapter 6 draws conclusions based on this research.

2. BACKGROUND

In Web-based communities discussions are hampered by the fact that reaching shared understanding seems to be a rare thing. It is much more common to see discussions turn into verbal wars between users having differing opinions. To tackle this, the argumentation approach (Conklin & Begeman, 1987) has been used to depersonalize conflict. For example,

in situations where "there are competing agendas, it helps participants clarify the nature of their disagreement" (Buckinham Shum et al., 2006).

Buckingham Shum (1996) analyzed the usability of the notation of one of the main argumentation approaches, namely the design rationale. He found out that users must learn to manage four interleaving cognitive tasks. These are unbundling, classification, naming, and structuring.

Unbundling is "identifying and separating constituent elements of ideas which have been 'bundled together' when they were initially expressed, but which from an argumentation perspective need to be teased apart". Classification is deciding whether a contribution is e.g. a question, option, or criterion. Naming is labeling the new contribution succinctly but meaningfully, and structuring is linking in a new element to other ideas (Buckingham Shum et al., 2006). To make matters even more difficult the reverse is often true when rationale is to be used. Ideas have to be bundled into explicit forms so that they can be applied to the problem at hand.

Buckingham Shum et al. (1997) conclude that "the basis on which [concept mapping tools] work is that deeper understanding of a domain comes through the discipline of expressing knowledge within a structural framework, working to articulate important distinctions and relationships." In other words, effort must be invested to get the benefits of rationale systems.

Argumentation has also been suggested as a way of achieving shared understanding (Deshpande, de Vries & van Leewen, 2005). Even though it could be seen as a way of reducing the costs related to understanding, using argumentation seems to have high formulation cost. This could explain the relative lack of success of argumentation sites in the Web2.0 era (Buckinham Shum et al., 2006).

Using structured argumentation has also been suggested as a viable means of supporting knowledge creation (Räisänen & Oinas-Kukkonen, 2007). It captures the rationale behind the decision-making process, and this can be later utilized for making decisions in somehow similar situations. The captured rationale could help in other knowledge creation sub-processes as well. For example, users can learn something new through reading the already captured rationale. For knowledge creation purposes merely identifying the pros and cons for different options is not enough (Klein, 2007). Consensus-making should be supported or at least be made discernible in some way by the system itself.

When using argumentation tools the users should structure and summarize their reasoning in such a manner that the user of the knowledge may read and understand it (Oinas-Kukkonen, 1998). If another user argues for a differing opinion the process may be described as a conversation among the stakeholders, in which they bring their expertise and viewpoints to bear on the resolution of issues. The goal of the discussion is for each of the stakeholders to try to understand the specific elements of each others' proposals, and perhaps to persuade others to accept their viewpoint. This kind of argumentation makes it harder to make unconstructive rhetorical moves and supports other more constructive moves, such as seeking the central question, asking questions as much as giving answers, and being specific about the supporting evidence for one's viewpoint. Any problem or concern may require discussion, if not agreement, in order for the work to go on. This kind of argumentation may be used in monologues, e.g. expressing an individual's diversified viewpoints or various roles, and in dialogues between stakeholders, e.g. in a development team.

The content creation in Web communities usually happens through communication or uploads. The communication can contain hyperlinks to other sites, copy and pasted texts, or direct exchange of posts created by the members. It is through the dialogues that take place

between the users that a common ground (Clark & Brennan, 1991) or shared understanding is reached. In order for a user to utilize the knowledge created in the Web communities he or she must internalize the knowledge stored in the community's content. The process through which internalization happens is called sense-making (Dervin, 1993) or comprehension (Oinas-Kukkonen, 2004).

Sense-making can be defined as a process of bridging gaps in knowledge that prevent user from moving forward in a time-space situations (Dervin, 1998). Russell et al. (1993) define it as "a process of searching for a representation and encoding data in that representation to answer task-specific questions." Comprehension is defined as "process of surveying and interacting with the external environment (...) in order to identify problems, needs and opportunities" (Oinas-Kukkonen, 2004). Together with communication, conceptualization and collaboration it is one of the central sub-processes of knowledge creation model called the 7C model (Oinas-Kukkonen, 2004). While communication and collaboration are some what self-evident, conceptualization a collective reflection process that produces new explicit concepts that work as a vehicle for collaboration (Oinas-Kukkonen, 2004).

The interesting thing with sense-making - and comprehension - is that there is no single correct way of doing it (Savolainen, 2006). We all have our own methods and ways of doing this kind of work. This also means that it is very difficult to design solutions that would support the sense-making processes of all users.

One way to support sense-making in argumentation is to visualize the argumentation. It offers two advantages. First, it provides structure, and secondly it provides support for algorithmic decision models (Introne, 2009). Structure helps users to focus on critical areas as well as see the overall view of the issue at hand more clearly. There are some ways how this can be done. One example of this is "argument-as-balance" metaphor (Johnson, 1987). In argument-as-balance rational arguments are understood as weights on either side of a scale. The weight on either side represents the strength of the arguments on each side of a question (McGinn & Picking, 2003). If the arguments on one side weigh more than on the other side the scale dips towards that direction. Visualizing argumentation using the argument-as-balance metaphor helps the users to quickly see which side of the question has stronger arguments. The problem with this kind of visualization is that it is difficult to represent the magnitudes of different factors (McGinn & Picking, 2003). For example, age and weight are not comparable directly so representing their magnitudes against each other would be problematic. A more common way of visualization is to separate issues (or questions that need to be answered), positions (possible answers) and arguments, which support or oppose the positions (Kunz & Rittel, 1970).

In this paper we utilize the Question-Answer-aRgument (QAR) method for argumentation visualization (Oinas-Kukkonen, 1998). It provides a regulated discussion of a proposition between stakeholders, i.e. capturing the argumentation behind concepts. The basic notions are nodes, links and knowledge space. The discussion is expressed using three kinds of nodes: Questions, Answers, and aRguments. The QAR method focuses on the articulation of the key questions. Each question may have many answers. An answer is a statement or assertion that resolves the question. Often answers will be mutually exclusive, but that is not required. Each answer may have one or more arguments that either support that answer or object to it. Thus, each separate question is the root of a discussion tree, with the children of the question being answers and the children of the answers being arguments. There is also a particular way of registering that a question has been resolved by selecting and presenting one of the suggested

answers as a decision. All nodes contain information on the creator of the node, timestamp of creation and other meta-information.

The focus of this paper is studying argumentation from the sense-making point-of-view. The aim is to compare various solutions that could help users in making sense of argumentation-based knowledge. In the Web2.0 spirit we will focus on simple and easy to use solutions. The experiment and the solutions are described in the next section.

3. EXPERIMENT

We implemented a graphical argumentation tool called the Debate Tool. It uses the QAR-notation. The implementation was done using HTML and AJAX allowing the users e.g. insert new questions, answers and arguments, and to move the arguments. Figure 1 shows a screenshot of the system. In the upper part of the screenshot (below the gray bar) there is a question ("What should be the rank for the matches?") and under it there are five different answers and seven arguments related to the answers. All arguments but one are connected to one answer each. One of the arguments is connected to two answers. It opposes one answer (the darker line) and supports the other. Note also that most of the arguments are on the right side of the screen and some answers have not received any arguments.

In addition, we also implemented three experimental versions of the system. The experimental version A calculated the average answer for each question from the available argumentation. It assumed that all the arguments were equal in order to be able to count how many arguments each answer had and to calculate the mean. (Of course this is not true in all cases because an argument can be the correct one and thus it can cancel all the other arguments.) We displayed the mean in a text box under the arguments. The reason why this might improve the shared understanding relies on two things. Firstly, displaying the mean of others' answers allows the participants to apply the rule of social proof (Cialdini, 1993) more easily. In other words, what other participants have done is made more explicit. Displaying the mean also lessens the cognitive effort (i.e. reduces the costs related to understanding and receiving) required by the user. She can see more easily if her answer is somewhat different from the other users' answers. Thus she can correct her answers if needed. So the experimental version A supports sense-making indirectly by showing the users where they might have answered differently than others. According to Cialdini (1993) "we will make fewer mistakes by acting in accord with social evidence than by acting contrary to it".

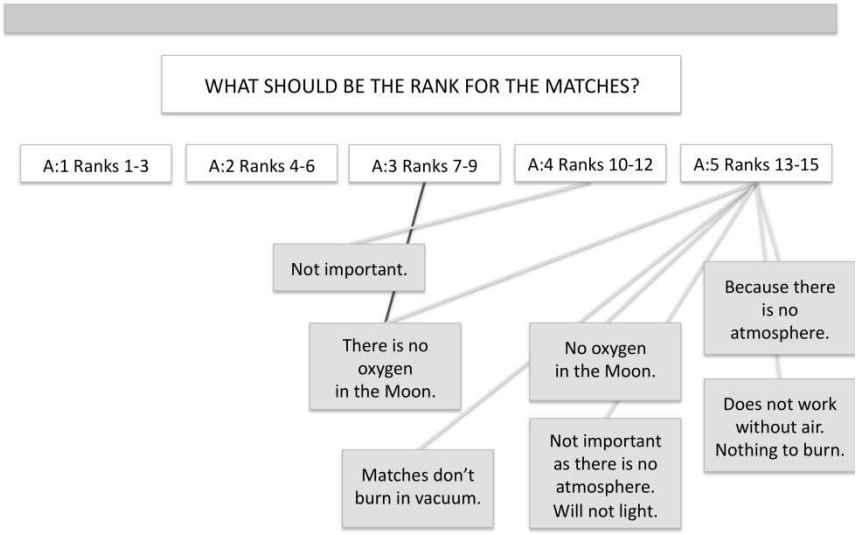

Figure 1. Screenshot from the Debate Tool

The version B used a solution that is commonly found in the Web, namely voting. The pilot test group had the chance of giving each argument a plus or a minus based on how important the user thought the argument was. In this way the users of the second experimental version could see which arguments others had found either important or unimportant. This may help the participants to lessen their cognitive efforts as they don't necessarily have to pay as much attention to arguments that have received many negative votes as those that have received many positive votes. By concentrating on the "important" arguments (i.e. those that have received a lot of plus votes) the user can more easily rank the items.

In the QAR method, each argument can support an answer (argument 'for') or oppose an answer (argument 'against'). As the pilot study users entered their argumentation into the system no one defined a single argument against as part of their reasoning. When asked why, most of the users simply replied that "they did not see any reason for it." A few said that sometimes it is difficult to understand it: "If I use an argument against and I use the negative in the text what happens then?" Since nobody used arguments against an answer, we decided to use them in one experimental version (version C) to see how they would affect the sense-making. In version C, with each question there was one argument that was supporting one answer but also against some other answers. We chose one of the existing arguments and used that. For example one of the arguments in the experiment was: "Food is important but not as important as water". In version C we used this argument to oppose one answer (i.e. the answer "1-3") and to support another one (i.e. the answer "4-6"). Similar arguments were used with other questions. The reason for doing this was that this might possibly help in sense-making because it makes some information more explicit. But it might also increase the cognitive costs related to understanding.

All of the experimental versions displayed the same argumentation as in the control version. The only difference was that the experimental versions each had one new feature. In

total we used four different versions, namely the control version and three experimental versions. See Table 1 for a description of the different versions.

Table 1. Different versions and their descriptions

Version	Description
Control group	Displays only the argumentation
Version A	Displays the argumentation and the mean of the answers.
Version B	Displays the argumentation and users' voting information (i.e. plus and minus votes cast to the arguments)
Version C	Displays the argumentation with one argument defined to oppose one of the answer.

We used the Lost on the Moon exercise (see Hall & Watson, 1970) in the study. In it the participants have to rank 15 items based on how important they feel the items would be if the participant crash-landed on the moon. Typically the exercise is used to study group decision-making (Ramachandran & Canny, 2008).

We made a pilot study with 28 people to see how they ranked the items. Ten of them also provided us with arguments for their answers. We modified the argumentation so that it matched with the official NASA ranking of the items. This way we had an argumentation that argued for the proposed ranking by respected experts, but with some room for individual interpretation.

The argumentation was inputted into the argumentation-based Web-service that we had implemented for the study. The exact same argumentation components were used in each of the version of the tool. In this manner the argumentation used could not affect the results and any results found had to be due to the experimental features in the system and/or interaction of the users.

The control version displayed only argumentation from the pilot study. We entered one question on each of the items in the Lost on the Moon problem. The idea behind the question was always the same: 'What should be the rank for the item?' Answer categories were also the same each time. We did not choose precise answers (i.e. that 'oxygen' could be marked to be 1) since we had argumentation from only ten participants and since we wanted to have some variance in the post-treatment answers. So, the answers were grouped into sets of three: 1-3, 4-6, 7-9, 10-12 and 13-15. A user could argue that 'oxygen' should be in the top 3 (answer 1-3) in the list, or that 'pistols' should be somewhere around 10-12, for example.

The study participants were all Finnish university students who were recruited from the faculty of science and from the faculty of technology. We used convenience sampling to recruit the participants and gave them $5 food coupons as an incentive. A total of 107 students participated in the study. They were divided into a control group (n=24) and three experiment groups (32, 26, and 25 students, respectively). The users were divided into the groups randomly, one group at a time.

First, the participants were given 15 minutes to complete the Lost on the Moon questionnaire. After that they had another 15 minutes to use the system, after which they had 5 minutes to fill in the questionnaire again. Finally, they had 10 minutes to fill in the last questionnaire asking demographic information, usability questions, and study-related questions. All questionnaires were in Finnish. The assumption here was that the better the

participant made sense of the rationale, the better she would answer in the second questionnaire.

Measuring sense-making, mental models or shared cognition can be very tricky (Langan-Fox et al., 2001). A way to tackle this is to find similarities between group members' answers. This is used by Langan-Fox et al. (2001) to measure shared understanding. In our study, we were seeking information about how similarly and accurately each group would answer. The accuracy of answers was measured by comparing individual answers to the official NASA ranking while the similarity was measured by comparing individual answers with the group mean and median. The accuracy measures how well the individuals comprehended the argumentation. The similarity between group members' answers means that individuals within a group possess a similar cognitive representation of the situation (Cannon-Bowers et al., 1995), i.e. a shared mental model.

The following hypotheses were tested:

H1: The members in experimental groups answer more similarly with each other than in the control group.

H2: The members in experimental groups answer more accurately than in the control group.

H1 and H2 were investigated using the Lost on the Moon questionnaire and information collected by a survey at the end of the experiment. In this survey, users were asked how they perceived the system and whether it in their judgment supports sense-making.

4. RESULTS

Table 2 displays the median answers of each group as well as NASA experts ranking (considered here to be "correct"). As can be seen from the table the answers are quite similar between the control group and the experimental groups. Especially evident is that each group has identified the most and least important items quite well (e.g. oxygen, water, matches). Still even if the median answers are quite similar, there can be some differences in how much variation there is within groups.

Table 2. Median answers of the control group and the experimental groups.

Item	Control	Experiment 1	Experiment 2	Experiment 3	NASA
Matches	14	15	15	14	15
Food	4.5	4	4	4	4
Rope	6	6	7	8	6
Heating unit	8	8.5	9	8	8
Parachute silk	10	7	9	7	13
Pistols	13.5	14	13	12	11
Pet milk	11.5	11	11.5	11	12
Oxygen	1	1	1	1	1
Stellar map	4.5	5	4	4	3
Life raft	9.5	10	10	10	9
Compass	13	13	13	14	14
Water	2	2	2	2	2
Signal flares	8.5	10	9	9	10
First-aid kit	7.5	8	8	7	7
FM-reveiver/trasmitter	5	4.5	5	5	8

The median answer in itself does not tell us enough. We are also interested in distribution of the answers and how similarly each group answers. Rather than doing a single measurement, multiple rank measures have been suggested to analyze the similarity of rankings (Diagonis & Graham, 1977). There are also statistical problems with using parametric tests with non-parametric ranking data. We acknowledge this and we will not use t-tests, for instance. To compare the standard deviations of each group we will use both the Levene's test (Levene, 1960) and the modified Levene's test (Montgomery, 2004).

Figure 2 displays the frequencies of the answers related to the item 'parachute silk'. Control group is in the top left, the experiment 1 is in the top right, the experiment 2 is in the bottom left and experiment 3 in the bottom right. From the figure we can see that the control group's answers are scattered more than the answers in the experiments groups. We find similar distribution with other items but the figures are omitted here in order to save space. In addition to figures, we conducted Levene's Test for Equality of Variance to see if there are any statistical differences in the variances.

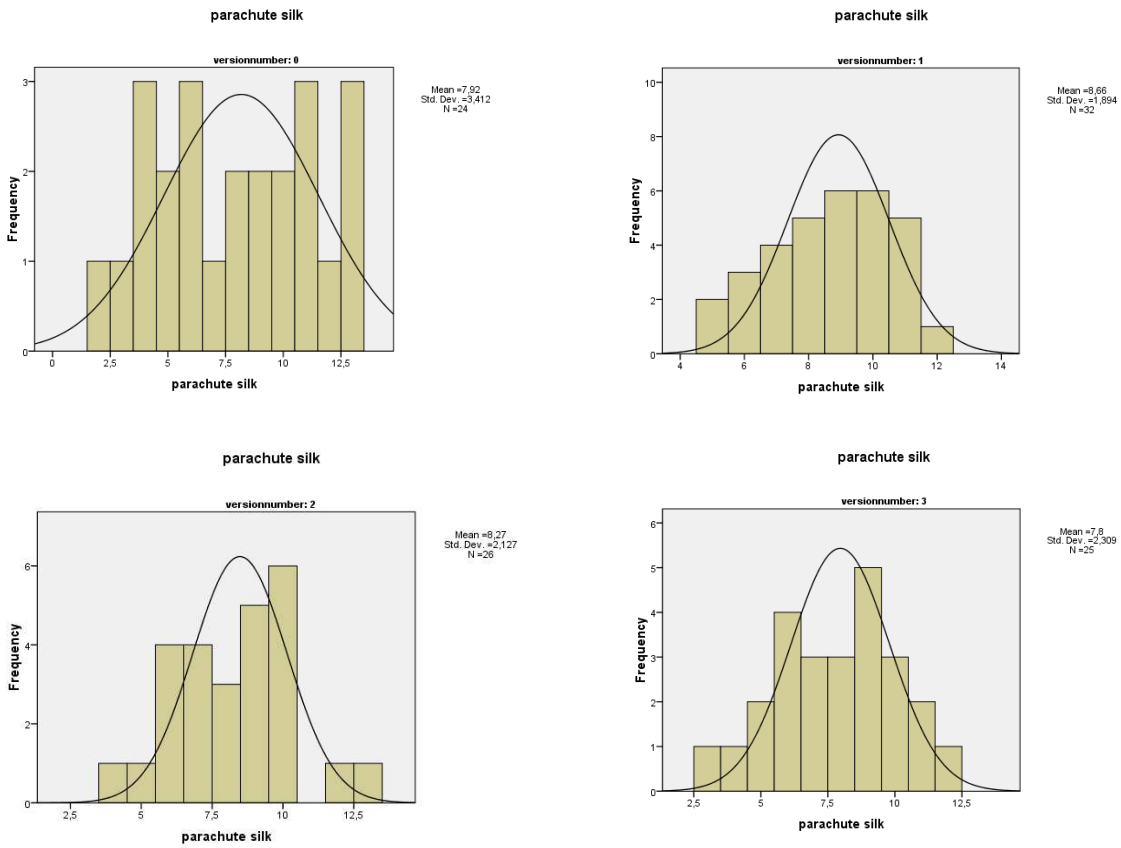

Figure 2. Frequency of answers with parachute silk. Top left is the control version, top right is the experiment 1, bottom left is the experiment 2 and bottom right is the experiment 3.

In Table 3, it can be seen how each version affected the variances with the item rankings. Version A reduced the variance with five items but it also increased the variance with one item. Version B reduced the variance of three items. Version C reduced the variance of three items and increased the variance of one item. The items in parenthesis had greater variance.

As can be seen from Table 3 all of the experimental versions did manage to produce smaller variances than the control version. Experimental version A seemed to have the biggest affect followed by B and C. The most interesting items are the ones where variance was actually bigger than in the control group (like heating unit for version A and flares for version C).

Table 3. The variance affected by experimental versions

Version	N	Item	Levene's Test for Equality of Variance
Version A	32	matches	F=9.904, p=0.003
		food	F=7.076, p=0.010
		rope	F=5.075, p=0.040
		parachute silk	F=8.545, p=0.000
		compass	F=4.973, p=0,030
		(heating unit)	F=4.567, p=0.037
Version B	26	matches	F=9.974, p=0.003
		parachute silk	F=8.704, p=0.005
		Oxygen	F=5.387, p=0.025
Version C	25	parachute silk	F=5.979, p=0.018
		life raft	F=5.596, p=0.022
		compass	F=4.475, p=0.040
		(flares)	F=5.340, p=0,025

However, since the standard Levene's test might not be best suited to study non-parametric data, we also conducted a modified Levene's test. The test is constructed by calculating the absolute deviation from the sample median for each observation, and then using ANOVA to test that the means of this quantity are the same for all of the populations. The test revealed that there was a significant difference with 'parachute silk' and with 'life raft.' See Table 4. It should be noted that all these items seem to be much harder to rate than e.g. water or matches. Thus it is probably understandable that these items also have more variance.

Table 4. Modified Levene's test

Version	N	Item	Modified Levene's Test for Equality of Variance
Version A	32	parachute silk	F=13.694, p=0.001
Version B	26	parachute silk	F=8.387, p=0.006
Version C	25	parachute silk	F=5.920, p=0.019
		life raft	F=6.321, p=0.015
		(flares)	F=4.580, p=0,038

While the standard Levene's test shows a much greater reduction of variances, the modified test is stricter in this sense. By combining both tests, we can conclude that there are small improvements with some of the items. These seem to be the items that are not easy to

rank, in other words they are more open for debate. However, this is not always the case as there can also be an increase in the variance.

We also compared each group's answers to NASA's expert ranking of the items. In the control version, participants had an average of 32 error points, i.e. if an item was ranked 5th by NASA and it was ranked 6th by the participant, (s)he would get one error point. The experimental versions did not improve the overall performance of the participants. The experimental groups did have fewer errors than the control group, but the difference was not statistically significant.

The version that did worst in terms of errors was the one displaying the means, while the other two versions were equal in their error points. This seems somewhat logical as the first version did not provide any new knowledge to the users, whereas the other two versions did. Version B provided the voting information and version C provided extra arguments against some answers. Nevertheless it is interesting to see that small changes to user interface can affect the results.

We also conducted usability tests and compared the answers from the control group to the experimental versions. The data was collected at the end of the experiment. We asked 18 questions with 5-point Likert-scale related to usability of the system, and its support for learning and knowledge creation. We performed one-way ANOVA tests to compare the data from control group with each of the experiment versions. See Table 5.

Table 5. Compared to the control version the experimental versions were perceived to offer less support for these aspects

Version	n	Findings	ANOVA
Version A	32	Comprehension	F=4.836, p=0.032
		Learning	F=21.557, p=0.000
Version B	26	Communication	F=4.678, p=0.036
		Orientation	F=4.328, p=0.043
Version C	25	Learning	F=4.898, p=0.032

The users of version A perceived that they did not comprehend or learn by using the system as much as the control group. If they answered according to the mean information displayed to them this seems logical. Instead of thinking by themselves they may have just copied what the group had answered. Thus they would not have learned very much.

The users of version B perceived that the system did not support communication or orientation (it is closely related to navigation and it refers to functionalities meant to help users find their way in hyperdocuments; See Thüring et al., 1995) as much as the control version. This version allowed users to vote on the arguments. Perhaps this made the lack of communication between the users more explicit than in the control version. No logical explanation why orientation was not perceived to be supported was found.

The users of version C perceived to learn less than the users of the control version. This might relate to the fact that many users felt that the usage of arguments against an answer was difficult and thus they perceived to system as being worse than it actually was.

It is important to notice that even if the experimental versions had only small differences compared with the control group, user perceptions were somewhat different. Even a small extra functionality may change what the users think about the system. However, the differences may also be a result from small sample size. A more rigorous study on this is needed.

5. DISCUSSION

Experimental version A displayed the mean of the answers for the participants. The success of this version is consistent with the ease-of-use emphasis with the Web2.0 users: *The easier it is to do something the more likely users will do it*. In this case the participants could see what other people had answered (on average) and they could easily copy that answer. To confirm this, one could implement an even more persuasive version of the experiment. For example, we could say to the users "Others have answered xyz to this question so maybe you should too." In this way the users' cognitive load could be reduced as the "social proof" would legitimate them to simply copy what others have answered. However, this might defeat the point of creating a space for arguments and consensus. Thus, a balance between how much to influence the users and how much work is required from them should be found.

Version B allowed users to vote for and against the arguments. This seems to work well in Web communities as only a handful of users normally produce content whereas many more are prone to click the plus or minus buttons to vote. It should also be noted that this is information about the content, too. Content receiving a lot of plus votes might be deemed worth reading whereas content receiving many minus votes might be ignored.

The version C displayed extra information in the form of counter-arguments. An argument that was supporting some answer was also placed to be against with another answer. This version did reasonably well in reducing the variance and it performed well when compared with NASA rankings. However, to some extent this version seems to refer to the possible problems with the argumentation-based applications and possibly also with other kinds of Web applications which are not so easy to use. Nobody was willing to use the arguments against – even if they would improve performance. Thus, this should be taken into account when designing applications in the era of Web2.0.

The Levene's test showed some improvement with all experimental versions on the variance of the answers. However, the modified version showed improvements only with a few items. A different experimental setup might show more consistent results. As of now, we can only conclude that the different versions can have an effect on situations where there is a lot of variance to begin with. In our experiment some items were more open to debate than others. For these reasons, hypothesis H1 is supported only partially.

The experimental versions did not improve very much the accuracy of answers. This finding is some what logical as the experimental versions were designed with the similarity of answers in mind. In addition, the experimental versions – with the exception of version C – did not really offer much new information. They simply made the existing information more visible. In fact, this is rather common with the highly successful Web2.0 solutions. Still, it can be concluded that hypothesis H2 was not supported.

The new functionalities seemed to make the system perform little better in terms of the users' results. But quite surprisingly, they also made users perceive the system as being worse than the control version. This might relate to the importance of ease-of-use. Even if the new functionalities do make the system perform better users might perceive it as being worse due to the new functionalities requiring more effort from them. In another words, the new functionalities can increase relevant cognitive costs, e.g. when a user had to think how to define compelling arguments and how to separate arguments from answers. Even if many system features were beneficial they may actually reduce the usage of the system. There seems to be a heavy emphasis on the ease-of-use in the era of Web2.0.

This could indicate that perhaps better sense-making for contemporary Web-users does not come through functionalities supporting deeper thinking (i.e. from solutions requiring high understanding costs) but rather from low-cost and easy to use solutions (such as versions A and B in this study). It could be the constant exposure to the content, i.e. repeated usage of the system that could trigger sense-making. And for users to use a system repeatedly, the system must be very easy to use. This is indeed in the core of the Web2.0 phenomenon. Such solutions could also better take advantage of the wisdom of crowds (Surowiecki, 2004) principles as easy to use solutions are more likely to gain larger user populations than high-cost solutions requiring deep thinking. Confirming this will require future research, though.

The perceived ease-of-use and usefulness of the system are crucial for the success of any Web2.0 applications. But with applications aimed at supporting knowledge creation both ease-of-use and usefulness might be difficult to achieve. Making an application easier to use might require removing some functionalities or options thus reducing its usefulness. And making an application more useful might come at the expense of ease-of-use. This is emphasized even more on contemporary Web-environment where users themselves decide which applications they use and prefer. Within organizations this is little bit different as employees can be ordered to use certain solutions.

So in order to design knowledge creation tool for the Web a careful balance must be found. The tool must support the cognitive processes of knowledge creation (i.e. it must be useful). But at the same time it must not require too much effort from the users (i.e. it must still be easy to use). For example, one reason why the argumentation approach has not been very successful could be that while it is useful it is not always easy to use as "deeper understanding of a domain comes through the discipline of expressing knowledge within a structural framework, working to articulate important distinctions and relationships" (Buckingham Shum et al., 1997).

To tackle this problem persuasive system design (Oinas-Kukkonen & Harjumaa, 2009) could be considered. Persuasion can be defined as (Fogg, 2003) "an attempt to change attitudes or behaviors or both (without using coercion or deception)". The goal of persuasion is to motivate or to influence individual's attitude or behavior in a predetermined way. Thus it could help in getting more users to use more complicated design solutions (e.g. through motivation). For example, the system could utilize various persuasive design principles to get users to contribute to the discussions – or to use argumentation tool. One way of motivating could be to offer users suggestions followed by rewards (rewards are given after user has followed the given suggestion).

The use of persuasive design also raises some interesting issues, especially with knowledge creation. While trying to increase motivation is generally beneficial (i.e. users can be made more motivated to use a system) trying to change behavior can have some problems, too. For example if we try to persuade users with social proof or social learning (i.e. we can show to the user that previously everybody else has done something in a certain way) there is the danger that some new innovative way of doing it will not be discovered as users prefer to do it using a proven method. Thus, as with ease-of-use and usefulness, a balance must be found when using persuasive design. However the results of this study hint that we should indeed use persuasive design in knowledge creation applications, even if it could have some drawbacks. The reason for this is that if indeed knowledge creation can be triggered also by repeated use (i.e. exposure to knowledge) it does not matter if we persuade the users a little bit. As long as users keep using the system new knowledge will emerge. To confirm this will require future research, however.

6. CONCLUSION

In this paper we studied ways to improve sense-making in Web-based argumentation systems. First, we demonstrated that various tools can help users make sense of this kind of knowledge rationale. Smaller deviance with each experimental group was obtained in the moon landing exercise but only with some of the items. The hypotheses were partially supported. By comparing the actual answers with a post-experiment questionnaire we found out that the versions that were most useful were not perceived as useful. This may be related to one of the core ideas behind Web2.0, namely the ease-of-use. For example in Web2.0 learning applications, ease-of-use has been identified as absolutely necessary (Ebner et al., 2007). The same might be true with knowledge creation, too.

Measuring sense-making is difficult. We took the similarity approach by comparing how similarly users ranked the items after the treatment. Admittedly, it can be argued that this does not actually measure sense-making. But the similarity of answers means that individuals within a group possess a shared mental model of the situation. So producing similar answers indicates that the users have made sense of the argumentation the same way.

A major drawback in the argumentation approach is that users sometimes find it time-consuming and effortful in the cognitive sense. Persuasive system design might help overcome some of the challenges regarding this. Persuasion might help e.g. with the cognitive costs of content production and understanding in argumentation-based Web applications. Persuasion could also make a difference on how users perceive argumentation. If, for example, argumentation could be made more game-like, then users might become more interested in using such systems in a true sense. There are still many issues that need to be resolved when integrating persuasive systems design with the argumentation approach. Thus, in-depth studies on the integration of them should be conducted.

Future research should tackle the idea that sense-making in the Web2.0 environment could be triggered by repetitive use, i.e. constant exposure to knowledge. This may have a great impact on how practitioners and researchers perceive sense-making in contemporary Web environments. Another line of future research should perform similar experiment as described here but use face-to-face group as a control group. This way we could gain valuable knowledge on how sense-making in web environment differs from sense-making in more contemporary environments (e.g. classroom).

REFERENCES

Buckingham Shum, S., MacLean, A., Bellotti, V. M. E. & Hammond, N. V. (1997). Graphical argumentation and design cognition. *Human-computer interaction*, vol 12, iss. 3, pp. 267-300.

Buckingham Shum, S., Selvin, A. M., Sierhuis, M., Conklin, J., Haley, C. B. & Nuseibeh, B. (2006). Hypermedia Support for Argumentation-Based Rationale: 15 Years on from gIBIS and QOC. In Dutoit, A. H., McCall, R., Mistrik, I & Paech. B. (Eds.): *Rationale Management in Software Engineering*, Springer-Verlag/Computer Science Editorial.

Cannon-Bowers, J.A., Tannenbaum, S.I., Salas, E., Volpe, C.E., (1995). Defining team competencies and establishing training requirements. In: Guzzo, R., Salas, E. (Eds.), *Team Effectiveness and Decision Making in Organizations*. Jossey-Bass, San Francisco, CA, pp. 333–380.

Cialdini, R. (1993). *Influence: Science and practice* (3rd ed), New York: HarperCollins.

Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. In L. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 127-149). Washington, DC: APA.

Conklin, J. & Begeman, M. L. (1987): gIBIS: A Hypertext Tool for Team Design Deliberation. In: Weiss, S. & Schwartz, M. (eds.): *Proceedings of ACM Hypertext 87 Conference*, November 13-15, 1987, Chapel Hill, North Carolina. pp. 247-251.

Dervin, B. (1998). Sense-Making Theory and Practice: An overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, 2(2), pp. 36-46.

Deshpande, N., de Vries, B. & van Leeuwen J.P. (2005). Building and Supporting Shared Understanding in Collaborative Problem-solving. *Proceeding of the Ninth International Conference on Information Visualisation*, 6-8 July 2005, London, UK.

Ebner, M., Holzinger, A. & Maurer, H. (2007). Web 2.0 Technology: Future Interfaces for Technology Echanced Learning? *In proceedings of the 4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007.* Held as Part of HCI International 2007 Beijing, China, July 22-27, 2007.

Fogg, B. J. (2003). *Persuasive Technology. Using Computers to Change What We Think and Do*. San Francisco, Morgan Kaufmann Publishers.

Hall, J., Watson, W.H. (1970). The Effects of a Normative Intervention on Group Decision-Making Performance. *Human Relations* 23, 299.

Introne, J. E. (2009) Supporting Group Decision by Mediating Deliberation to Improve Information Pooling. *Group 2009 Conference*, May 10-13, Sanibel Island, Florida, USA, pp. 189-198.

Johnson, M. (1987) *The Body in the Mind*, University of Chicago Press.

Kunz, W. & Rittel, H. (1970) Issues as Elements of Information Systems, Working paper No. 131, *Studiengruppe für Systemforschung*, Heidelberg, Germany, July 1970.

Langan-Fox, J., Wirth, A., Code, S., Langfield-Smith, K. & Wirth. A. (2001) Analyzing shared and team mental models. *International Journal of Industrial Ergonomics*, 28(2), pp. 99-112.

Levene, H. (1960). *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press, pp. 278-292.

McGinn, J. & Picking, R. (2003) The Argument-as Metaphor in Decision-making visualisation. *The proceedings of the Seventh International Conference on Information Visualization (IV'03)*, 16-18 July 2003, London, UK.

Montgomery D. C. (2004). *Design and Analysis of Experiments*, Sixth Edition. John Wiley & Sons, Inc.

Novak, J. & Wurst, M. (2004). Supporting Knowledge Creation and Sharing in Communities Based on Mapping Implicit Knowledge. *Journal of Universal Computer Science*, vol. 10, no. 3, pp. 235-251.

Oinas-Kukkonen, H (1998). Evaluating the Usefulness of Design Rationale in CASE. *European Journal of Information Systems*, 9(3), 201-207, September 1998.

Oinas-Kukkonen H. (2004) The 7C Model for Organizational Knowledge Sharing, Learning and Management. *Proceedings of the Fifth European Conference on Organizational Knowledge, Learning and Capabilities (OKLC ' 04)*, Innsbruck, Austria, April 2-3, 2004.

Oinas-Kukkonen, H. & Harjumaa M, (2009) Persuasive Systems Design: Key Issues, Process Model, and System Features. *Communications of the Association for Information Systems*, Vol. 24, Article 28, March 2009, pp. 485-500.

Preece, J., Maloney-Krichmar, D. (2003) Online Communities: Focusing on sociability and usability. In: J. Jacko, A. Sears (eds.) *Handbook of Human-Computer Interaction*. Lawrence Erlbaum Associates Inc. Publishers. Mawhah: NJ. pp. 596 – 620.

Ramachandran, D. & Canny, J (2008). The Persuasive Power of Human-Machine Dialogue. In: Oinas-Kukkonen et al. (Eds): *Persuasive 2008*, LNCS 5033, pp. 189-200.

Russell, D. M., Card, S., Pirolli, P., Stefik, M. (1993). The cost structure of sensemaking, Proceeding of CHI 1993, pp. 269-276.

Räisänen, T & Oinas-Kukkonen, H. (2007). A System Architecture for the 7C Knowledge Environment. *17th European-Japanese Conference on Information Modeling and Knowledge Bases*, June 4-8, Pori, Finland.

Räisänen, T. (2009). Supporting the Sense-Making Processes of Web Users by Using a Proxy Server. *42th Hawaii International Conference on Systems Science*, Waikoloa, Big Island, Hawaii, January 5-8, 2009.

Savolainen, R. (2006). Information use as gap-bridging: the viewpoint of Sense-Making methodology. *Journal of the American Society for Information Science and Technology*, Vol. 57, 8, pp. 1116-1125.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations?* New York, NY: Doubleday.

Thüring, M., Hannemann, J. & Haake, J. M. (1995). Hypermedia and Cognition: Designing for Comprehension. *Communications of the ACM*, August 1995/Vol. 38, No. 8, pp. 57-66.

Wenger, E., C. & Snyder, W. M. (2000). Communities of Practice: The Organizational Frontier. *Harvard Business Review*, January-February 2000, pp. 139-145.