User Model User-Adap Inter (2006) 16:321–348
DOI 10.1007/s11257-006-9013-6

ORIGINAL PAPER

Design and evaluation of an adaptive incentive mechanism for sustained educational online communities

Ran Cheng · Julita Vassileva

Received: 10 October 2005 / Accepted: 16 April 2006 / Published online: 7 September 2006

Abstract Most online communities, such as discussion forums, file-sharing communities, e-learning communities, and others, suffer from insufficient user participation in their initial phase of development. Therefore, it is important to provide incentives to encourage participation, until the community reaches a critical mass and "takes off". However, too much participation, especially of low-quality can also be detrimental for the community, since it leads to information overload, which makes users leave the community. Therefore, to regulate the quality and the quantity of user contributions and ensure a sustainable level of user participation in the online community, it is important to adapt the rewards for particular forms of participation for individual users depending on their reputation and the current needs of the community. An incentive mechanism with these properties is proposed. The main idea is to measure and reward the desirable user activities and compute a user participation measure, then cluster the users based on their participation measure into different classes, which have different status in the community and enjoy special privileges. For each user, the reward for each type of activity is computed dynamically based on a model of community needs and an individual user model. The model of the community needs predicts what types of contributions (e.g. more new papers or more ratings) are most valuable at the current moment for the community. The individual model predicts the style of contributions of the user based on her past performance (whether the user tends to make high-quality contributions or not, whether she fairly rates the contributions of others). The adaptive rewards are displayed to the user at the beginning of each session and the user can decide what form of contribution to make considering the rewards that she will earn. The mechanism was evaluated in an online class resource-sharing system, Comtella. The results indicate that the mechanism successfully encourages stable and active user participation; it lowers the level of information overload and therefore enhances the sustainability of the community.

R. Cheng (✉) · J. Vassileva
Department of Computer Science, University of Saskatchewan, 178.8 Thorvaldson Bldg., 110 Science Place, Saskatoon, Saskatchewan, S7N 5C9, Canada
e-mail: chengran@gmail.com

Keywords Online communities · Virtual communities · Participation · Ratings · Incentive mechanisms · Personalized rewards

1 Introduction

An online community is a group of people who interact in a virtual environment. They have a common interest or purpose, are supported by technology, and are guided by norms and policies (Preece 2000).

Depending on the type of the community, people can make different contributions by sharing opinions, information, blogs, music and video files, photos, etc. Examples of online communities include discussion boards, like Electric Minds, social networks, like Orkut, blog systems, like Blogger, photo-sharing communities, like Flickr, and file-sharing communities, like KaZaA or BitTorrent.

The proliferation of online communities may suggest that the design of a community for a particular purpose is straightforward. Unfortunately, this is not the case. Although software providing basic community infrastructure is readily available, it is not enough to ensure that the community will "take off" and work well. A critical mass of user participation has to be reached for an online community to survive and be sustainable (Hiltz and Turoff 1978).

Market-based incentive approaches have been proposed to stimulate participation in online communities. For example, (Burgahain et al. 2003; Golle et al. 2001) have proposed game-theoretic incentive mechanisms for peer to peer online communities using micro-payments to reward individual contributions. However, practical applications of market mechanisms using micro currency, like Mojo Nation have not been successful in stimulating participation. Shirky (2003) argued that micro payments involve a high-cognitive cost for users (the decision of whether to carry out the transaction or of "consuming" a shared resource when one needs to pay for it, even a miniscule amount) and that social rewards, such a fame or status in the community can be a stronger motivator for participation.

Comtella (Vassileva 2002) exploits exactly this idea. It is a small-scale online community developed at the MADMUC lab at University of Saskatchewan for sharing URLs of class-related web-resources (bookmarks) among students. Similar to other online communities, it also faced the problem of scarcity of user participation and contributions. To address the problem, we proposed a motivational mechanism (Vassileva et al. 2004), which rewarded contributing users with higher status in the community and better quality of service. The contributions and participation of each user were measured and users were able to earn different classes "memberships" in the Comtella community: bronze, silver, and gold, each associated with a different interface and additional search options. The evaluation showed that while the mechanism was effective in increasing the quantity of contributions, it stimulated some users to try to game the system (Cheng and Vassileva 2005). To maximize their rewards and minimize their effort, these users contributed many resources of medium or low quality. This made it hard for everyone to find good resources in the system, resulting in disappointment and a decrease in the level of user participation.

This phenomenon happens in online communities and is called "information overload" (Shenk 1997). Jones and Rafaeli (1999) found that the users' most common response to it is to reduce or end their participation in the community. Therefore, the abundance of user contributions does not necessarily guarantee sustainability of

an online community. On the contrary, excessive contributions may result in user withdrawal and may impair the sustainability of the community. In order to ensure active and stable user participation, we propose a new *adaptive* mechanism to measure the quality of user contributions, control the overall number of contributions in the community, and motivate users to contribute high-quality resources. On one side, the mechanism should encourage users to rate contributions thus ensuring decentralized community moderation. On the other side, the mechanism should influence the individual users' actions of contributing by adapting the rewards using a model of the current needs of the community and a model the users' individual reputation in contributing quality resources.

The paper is organized as follows: the next section presents related work in the area of collaborative quality evaluation mechanisms. Section 3, presents the proposed mechanism and Sect. 4 describes its implementation in Comtella. Section 5 presents the design and the results of the evaluation. Section 6, discusses the results and raises some more general issues related to the design of incentive mechanisms in online communities. Section 7, concludes the paper and outlines directions for future research.

2 Related Work

To ensure active and sustained user participation and avoid information overload in the online community, the incentive mechanism has to take into account the quality of user contributions, i.e. to reward the contributions with high quality, and inhibit the inferior ones.

It is not easy to measure the quality of each contribution impartially and accurately because quality measures are mostly subjective. Centralized moderation is feasible only for small and narrowly focused communities, where members share similar evaluation criteria. For medium or large online communities (e.g. the ones with more than 100 users), a decentralized moderation for quality measurement is necessary. A real world example of decentralized moderation is the impact factor which measures the quality of journals or papers by counting the times they were cited. Although it is somewhat controversial whether the impact factor is able to represent fairly the quality of research papers (e.g. Merton and Zuckerman 1968), it indicates the extent to which the paper is used by other scholars. In a similar way, one can measure the quality of a posting in an online community by counting the times it was viewed (clicked). However, this method is based on the assumption that people who view a resource hold a positive attitude to its quality, which is not always the case.

Another way of evaluating the quality of resources or comments is through explicit user ratings, like for example, the peer-reviewing process in academia and the rating process in online communities like Slashdot. Since the final evaluations of resources are computed based on ratings from many users, they are more unbiased. However, a study of the Slashdot rating mechanism showed that some deserving comments may receive insufficient attention and end up with an unfair score, especially the ones with lower initial rating and those that were contributed late in the discussion (Lampe and Resnick 2004). The reason for this is that low-rated or late contributions have a lower chance to be read by others (since they select contributions to read based on their rating), and therefore, a lower chance to get rated. The same effect appears in the

majority of decentralized quality measurement systems and it is sometimes referred to the "rich get richer" or "Matthew effect" (Merton and Zuckerman 1968). It is impossible to prevent this effect and ensure a totally fair mechanism in a decentralized rating system since the resources with high ratings inevitably become more visible and tend to be viewed and rated more often. However, knowledge of this effect can help an individual to develop strategies to achieve a better standing. Obviously, the timeliness of contributing resources is important and users should be encouraged to contribute early because late contributions are unable to receive enough attention and are therefore less useful for the community. This is especially relevant in a class-support system like Comtella or I-Help (Greer et al. 2001) since the topics typically change on a weekly basis and late contributions tend to be neglected. Apart from playing a role in the likelihood that a contribution will be read and rated, the timeliness of the contributions reflects the needs of the community. Early in the discussion period, when there are not many contributions, it is important for the community to bring new contributions, so that there are enough materials to be read. Later, when the number of contributions is already high, it is more important for the community that the users rate the contributions, since in this way they provide guidance for other users in finding good contributions. Therefore, the rewards for different types of cooperative activities (e.g. sharing new resources; giving ratings) should be different, depending on the current needs of the community.

A challenge in all systems that rely on decentralized moderation is to ensure that there are enough user ratings. Beenen et al. (2004) proposed to send users e-mail invitations to encourage them to rate movies in MovieLens, which increased the ratings in the system. However, this approach is questionable as a long-term solution since the effect of receiving email-invitations will likely wear off. To stimulate users to rate resources constantly, persistent incentives are necessary.

The evaluation of our hierarchical memberships motivational mechanism (Cheng and Vassileva 2005) showed that different users had different contribution patterns. Some contributed many but poor-quality resources, and others contributed few but high-quality resources. Therefore, the incentive mechanism should be adaptive to the patterns of contributions of different users. The mechanism should stimulate the users with high-reputation, i.e. those contributing few but high-quality resources, to contribute more and it should inhibit contributions from users with low-reputation, who contribute many low-quality resources, unless they improve the quality of their contributions.

In addition, the mechanism should adapt to the total number of resources desirable in the online community. For example, the rewards for contributing new resources should be decreased, if there are already too many resources in the community that could cause information overload. However, how to decide what amount of resources is desirable is an open question. In a context of a class-support community, the instructor may be able to specify a desirable number of contributions, based on previous experience, and knowledge about the availably of resources for each topic.

Next, an incentive mechanism with the qualities discussed above is proposed. It aims to encourage users to rate resources, and depending on the quality evaluation from user ratings, the mechanism is able to adapt the rewards of different forms of participation for individual users, depending on their reputations and the current needs of the community. The goal is to influence the users' contributions in terms of both sharing resources and rating resources so that they benefit the community the most.

3 Proposed Incentive Mechanism

The proposed mechanism consists of two parts: a mechanism encouraging users to rate resources and an adaptive reward mechanism.

3.1 Encouraging users to rate contributions

The proposed collaborative rating mechanism is inspired by the Slashdot moderation system. In order to have a broader source of ratings, all users can rate others' contributions. Each user receives a limited number of rating points to give out. She can award any contribution +1 or −1 point depending on whether she likes or dislikes it, consuming one of her rating points (Fig. 1). The users with higher membership levels receive more points to give out, which makes them more influential in the community. To ensure that all contributions have an equal chance to be read and rated initially, the initial rating for every new contribution is zero regardless of its providers' membership level or the quality of her previous contributions (unlike Slashdot, where the postings of the users with higher karma start at a higher rating). In the end, the final rating for each contribution is calculated as the sum of all the ratings that it has obtained, i.e. it is the summative rating, rather than the average rating. The summative rating ("Earned Ratings" in Fig. 1) for each contribution is displayed in the list of search results, which can be sorted by the user and viewed as a "top 10" list of articles for any topic.

The "Fake" option allows users to option to "report" duplicated links, broken links, or links that require subscription. "Fake count" shows the number of times a link was reported as fake (by default 0) and is used to signal that the user does not need to spend time and click on it.

As a persistent incentive for users to rate contributions, a virtual currency is introduced, called "*c-point*". The *c-points* are different from the rating points mentioned above. Initially, each user has only a limited number of *c-points*. Whenever, the user rates a contribution, she is awarded a certain number of *c-points*, depending on her reputation of giving high-quality ratings. The user can invest the earned *c-points* to increase the initial visibility in the search result list of her contributions. Most users prefer their contributions to appear in salient positions in the search result list, e.g. in the first place or among the top 10, because in those positions they will have a better chance to be read and rated. The Comtella search facility displays by default all the contributions matching a query in a sorted list according to the number of *c-points*

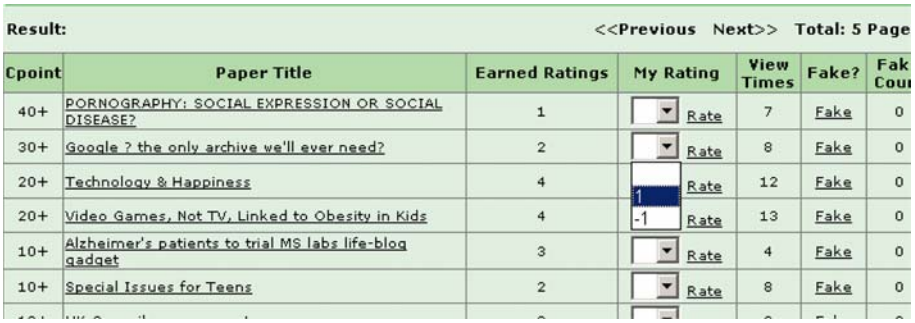


Fig. 1 A segment of a search results list

allocated by the contributors (Fig. 1) and then sorted by the time of submission. To compensate the weakness of Slashdot, the proposed mechanism allows the user the flexibility to invest any number of their *c-points* in a particular posting to increase its visibility, so even if a resource was contributed late in the week, it will have a chance to be seen and rated. However, in order to earn *c-points*, the users have to rate resources contributed by others. The *c-points* are earned immediately after users give a rating, bringing instant gratification to the user. Since the users with higher membership levels get more ratings to give out, they can rate more articles, and are potentially able to earn more *c-points*.

3.2 Overview of the adaptive rewards mechanism

The adaptive reward mechanism is introduced as an improvement of the mechanism of hierarchical memberships, proposed in (Cheng and Vassileva 2005; Vassileva et al. 2004). The basic idea is to adapt the rewards for different forms of participation for individual users to the user's current reputation (based on the quality of their contributions so far) and the current needs of the community. The individual rewards for each type of action are displayed in personalized motivational messages, which the user sees at login, outlining what the community expects from the user in terms of quantity and quality of contributions. As shown in (Beenen et al. 2004), personalized messages stating specific performance goals were effective in influencing and directing the users' behavior of contributing ratings in MovieLens. Therefore, we expect that our approach would also influence the user's behavior in a desirable way. Figure 2, presents an overview of the mechanism. The adaptive rewards are calculated using two models: a community model and an individual model.

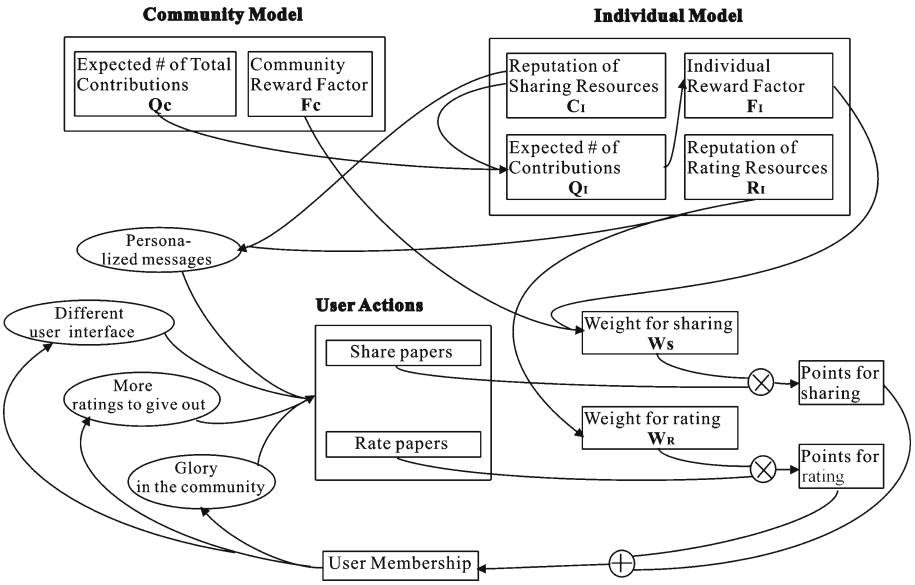


Fig. 2 An overview of adaptive reward mechanism

3.3 Community model

The community model is used to describe the current phase in the community needs. It includes the expected number of resources that should be contributed by all the users for the current topic (Q_C) and the community reward factor (F_C). For each week, when a new class topic is introduced, Q_C is set by the community administrator (e.g. the instructor of the course) for the current topic, depending on her knowledge of certain features of the topic (e.g. how interesting it is expected to be for the students, the relative amount of materials available online that could be shared) and the users' situation (e.g. how much time and energy the students can devote, depending on their coursework, exams, etc.). F_C reflects the extent to which newly contributed resources are useful for the whole community. Generally, new resources are needed as soon as possible after a topic has been announced or opened for discussion and those resources that are contributed late in the period are less useful than the ones contributed early. Therefore, F_C has its maximum value when a new topic is introduced and decreases gradually with the time. After the middle of the discussion period, it decreases faster (Fig. 3).

3.4 Individual model

Each user has an individual model that keeps her reputation of contributing high-quality resources (which we will call "user's resource-reputation" for the sake of brevity) and her reputation of giving high-quality ratings (we will call it "user's rating-reputation"). The individual user model contains also the data describing the user's current membership level. Since the summative rating of a resource denotes its quality, a user's resource reputation (denoted as C_I) is defined in a straightforward way as the average summative rating of all the resources she has shared so far.

However, the quality of the user's ratings cannot be defined so easily, since the ratings are by nature subjective. Although in educational online community instructors or teaching assistants may have the ability to evaluate the quality of ratings, the workload of evaluating all user ratings is overwhelming. Moreover, in other types of online communities there are no such arbiters available. We chose to measure the quality of each rating by the difference between it and the average of all the ratings the resource gets eventually: the smaller the difference, the higher the quality of the rating. Accordingly, a user's rating reputation (R_I) is defined as the reciprocal of the

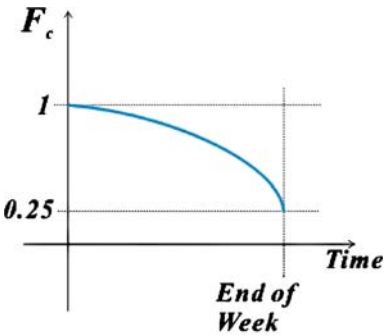


Fig. 3 The changing of the community reward factor (F_C)

average difference between all ratings she has made and the respective average ratings of the rated resources. Formula (1) shows how R_I is calculated.

$$R_I = \frac{N}{\sum_{i=1}^{N} |r_i - \bar{r}_i|} \tag{1}$$

Here $r_i (i = 1, 2, 3, \ldots, N)$ are the ratings the user has made; and $\bar{r}_i (i = 1, 2, 3, \ldots, N)$ are the average ratings of the respective resources. This approach for computing the user's rating-reputation is based on the assumption that the average rating of a resource reflects the opinion of the majority of users and is therefore less biased. Since this metric can be easily skewed, if users intentionally rate close to the average rating of the resource, the average rating is never shown; only the summative rating of each resource is shown in the list of search results, as explained in the beginning of Sect. 3.

The expected number of resources contributed by the individual user (Q_I) is a fraction of Q_C. The users with higher resource-reputation C_I will get a larger Q_I, which means the users with a good resource reputation are expected to make more contributions while the expected total number of contributions from all users is fixed. If details are ignored, Formula (2) demonstrates how Q_C is distributed among users.

$$Q_I \approx Q_C \bullet \frac{C_I}{\sum C_I} \tag{2}$$

Q_I is the number of resources expected from the user.
Q_C is the expected sum of resources desired from all users.
C_I is the user's resource reputation.

The individual reward factor (F_I) defines the extent to which the resources contributed by the user are being rewarded. Users are not encouraged to make contributions after they have exceeded their expected number of contributions (Q_I). So F_I is a function that is a constant value as long as the number of the user's contributions is less than or equal to her Q_I. When the number exceeds the expectation, F_I drops to one fourth of the constant value instantaneously and keeps decreasing with the increment of the users' contributions (Fig. 4).

3.5 How the mechanism works

Before explaining the adaptive rewards mechanism, we need to introduce briefly the scoring model that we proposed in (Cheng and Vassileva 2005). It was used to measure the contributions made by users and to determine each user's membership level. Generally, the comprehensive evaluation of the user's contributions (V_{oe}) was

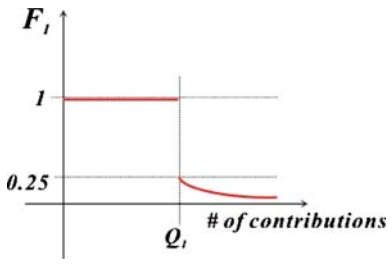


Fig. 4 The changing of the individual reward factor (F_I)

based on the number of times (T_i) the user made a contribution of a certain form (here we consider two forms of contribution: share a link to a paper or give a rating, but there can be more rewarded forms of contribution in general) and the weights introduced to denote the importance of each form of the contribution (W_i). V_{oe} was computed through formula (3). n is the number of the forms of contribution considered.

$$V_{\text{oe}} = \sum_{i=1}^{n} W_i \cdot T_i \tag{3}$$

According to their values of V_{oe}, users were classified into several levels of membership. Those with higher values of V_{oe} would obtain higher-level membership (e.g. silver or gold), and consequently gain more prestige and other rewards in the community (see Fig. 2).

In the adaptive reward mechanism, instead of fixed constants for the weights (W_i), varying weights $W_i(t)$ are used for the different forms of contribution (e.g. sharing resources, giving ratings). If we represent with $t = (1, 2, 3, \ldots, T_i)$ the sequence of the contributions of each form, the overall evaluation of a user's contributions (V_{oe}) is calculated through Formula (4).

$$V_{\text{oe}} = \sum_{i=1}^{n} \left[\sum_{t=1}^{T_i} W_i(t) \right] \tag{4}$$

We mainly consider two forms of contribution: sharing a resource and contributing a rating. The weights $W_i(t)$ for each form of contribution are modeled as discrete functions of time and depend on the states of the user's individual model and the community model at the current time. The current values of the weights are shown to the user at logon time in a personalized message, so that she can see what rewards she will receive for contributing a new resource and for rating resources contributed by others. Each personalized motivational message informs the user about the number of resources expected from her for the current topic (Q_{I}) and, if the user's C_{I} (the reputation for sharing) or R_{I} (the reputation for rating) is lower than a certain threshold, a reminder is displayed for her to pay attention to the quality of the resources or the ratings that she will contribute.

The adaptive weight for sharing resources (W_{S}) is calculated through Formula (5). Here W_{S0} is a constant, which is the initial value of the weight.

$$W_{\text{S}} = W_{\text{S0}} \bullet F_{\text{C}} \bullet F_{\text{I}} \tag{5}$$

F_{C} is the community reward factor.
F_{I} is the individual reward factor.

W_{S} is equal to W_{S0}, its initial constant value, when a new topic begins and the number of the user's contributions has not reached their expected value Q_{I}. After that, it decreases gently with time as F_{C} does (Fig. 3). Whenever the number of the user's contributions goes beyond her Q_{I}, W_{S} sharply decreases to 1/4 of its original value (as F_{I} does in Fig. 4) and continues to decrease with the accumulation of the user's contributions and time.

It can be seen that W_{S} inherits the features of both reward factors, F_{c} and F_{I}. In this way, a user who shares many resources but does not care about their quality gets

a low C_I (the user's reputation of contributing resources) and a small Q_I (the number of resources expected from her) and therefore, little reward for her subsequent contributions. Thus, the personalized message to the user would be to contribute less new resources and try to improve their quality. This situation continues until the user finally improves her reputation in sharing.

On the other hand, if a user tends to share a small number of good resources, she obtains a high C_I and a large Q_I. This user will earn more rewards by sharing new resources unless she starts compromising the quality. For both kinds of users, early contributed resources always earn more rewards. Hence, the adaptive weight W_S is able to restrict the quantity of user contributions, elicit the contributions from the users with a good reputation, discourage contributions from users who ignore quality and stimulate users to share early in the discussion period.

The adaptive weight for giving ratings (W_R) is proportional to the user's reputation of giving high-quality ratings (R_I) and is updated according to Formula (6), where K is a constant.

$$W_R = K \bullet R_I (K \text{ is a constant}) \tag{6}$$

The users who have gained a good rating reputation get higher weight for their subsequent ratings, which should stimulate them to rate more resources. However, those with low R_I will not get rewards for rating. They have to rate less and improve the quality of their ratings to win their reputation back and this would be the suggestion of the personalized message.

4 Implementation of the mechanism in the New Comtella System

The mechanism presented in the previous section was implemented in a version of the Comtella system, which was used as a test environment for the evaluation of the mechanism. Another version of the system, with very similar "look and feel", but using just the hierarchical memberships mechanism (Cheng and Vassileva 2005) was implemented too to serve as a control environment in the evaluation. Next, we will focus only on the design of the test system including the mechanism.

Figure 5 shows the "welcome page" (the front page) of the test system. On this page, the following information and features are provided for the user as follows:

- *The user's contribution levels in the previous week and in the current week*. The different forms of user contributions (e.g. sharing, rating) are evaluated separately, so that the user knows, which one to emphasize. To enhance the motivational effect, the contribution levels shown are updated immediately after the contributions are made.
- *The weights for different forms of contributions*. To limit the variable space so that it is possible to evaluate the proposed mechanism, in this version of Comtella, only two forms of contribution, *sharing articles* and *giving ratings*, are considered (we could have considered, for example, also *giving comments* or *discussing links in the discussion forum*). However, not only the number of times these two activities are performed, but also the quality of articles shared and of the ratings made are taken into account in the participation measure. The weights for different forms of contribution are listed at Table 1. *The user's current membership level*. It is updated weekly, based on the comprehensive evaluation of the user's contributions in the

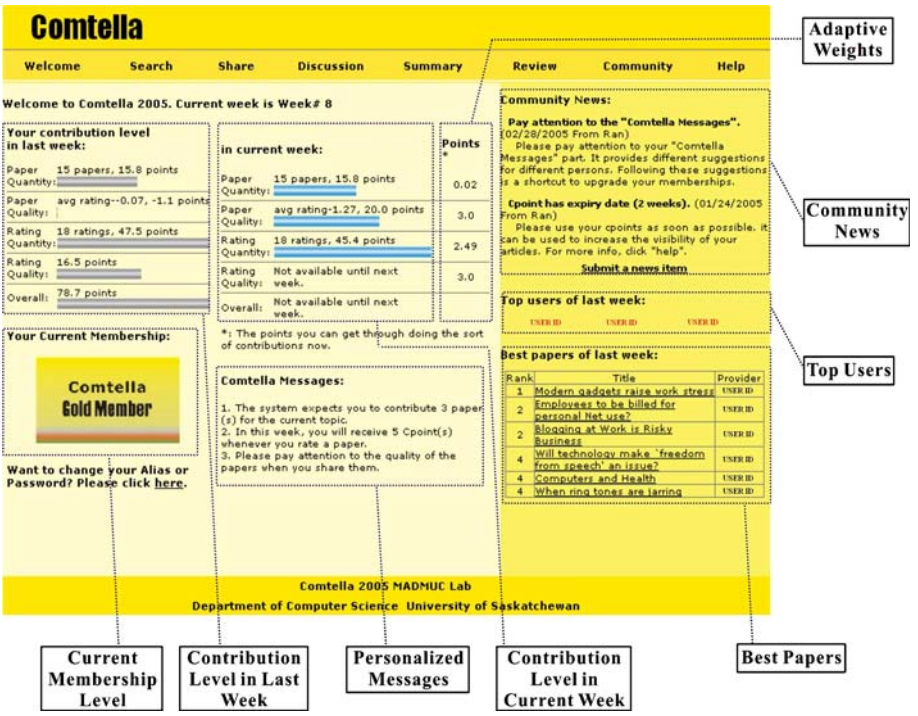


Fig. 5 The welcome page of Comtella

Table 1 Weights for different forms of contribution

Weights	Description	Value
W_{SN}	Weight for the quantity of resources contributed	Varying depending on the average quality of resources the user shared and the current need of the system
W_{SQ}	Weight for the quality of resources contributed	Constant
W_{RN}	Weight for the quantity of ratings made	Varying depending on average quality of ratings made by the user.
W_{RQ}	Weight for the quality of ratings made	Constant

previous week. The users' memberships are updated at the beginning of each week, based on their contribution levels in the previous week. The comprehensive evaluation of the user's contributions is computed through formula (3).

- *Personalized messages for the user*. The personalized messages inform the user of the number of articles she should share for the current topic and also remind the user to pay attention to the quality of her contributions and ratings when necessary.
- *Community news*. The community administrator (e.g. the instructor) can inform the users about events in community and provide course information such as deadlines for submitting assignments, time and place of examinations, etc. Besides, gold members are also allowed to release community news, as one of their privileges.

- *Top users and best papers of last week.* The top users and the best papers of the previous week are announced in the welcome page, as another stimulation to contribute. The top users are those three or four users with the highest comprehensive level of contributions in the system. Five – seven of the best papers (with the top-three summative ratings) are shown in the best-paper list together with the users who contributed them.

The rating interface is embedded in the "search page" of the system (Fig. 1), through which the user can rate the articles immediately after searching and viewing them. The current total of earned ratings and the times the article was viewed are shown for each article allowing the user to select for reading articles that have received highest number of ratings or that been viewed most often by others. By default, the articles are sorted in the list by the number of *c-points* and their sharing time. The users can re-sort them by article title, number of earned ratings or view-times.

Whenever the user rates an article, she earns a certain number of *c-points*. This number depends on the user's reputation of giving ratings, evaluated with formula (1). To stimulate users to consume their *c-points* sooner and rate more articles, the *c-points* are effective only within two weeks after they are obtained. When sharing a resource, the user has the option to invest a certain number of *c-points* in the resource (Fig. 6). The user can decide how many *c-points* to attach to the article, depending on how interesting, readable and relevant they think it is. However, to avoid the unbounded competition for the largest number of *c-points*, there is a limit (50 *c-points*) to the number of *c-points* that can be invested in one article.

There are four hierarchical membership levels in the new Comtella system—Gold, Silver, Bronze, and Plain (Plastic) member. The user interfaces provided for the members with different memberships are different in appearance and color, but same in terms of functionality. Although higher level members are not offered additional functions, they have certain privileges in the community. For example, they receive more ratings to give out, which makes them more influential in the community. Also, Gold members are enabled to publish news-items in the Community News.

5 Evaluation of the Proposed Mechanism

To test whether the mechanism discussed in the previous section can achieve the goals of regulating the quality and quantity of contributions and ensuring sustained

Fig. 6 The interface for sharing resources in Comtella

Table 2 Differences between the two systems for the two groups

Feature	System for Test Group	System for Control Group
Hierarchical Memberships	✓	✓
Showing Contribution Levels in Previous and Current Week	✓	✓
Interface for Rating Articles	✓	✓
Cpoints as Reward for Rating	✓	×
Adaptive Weights for Sharing and Rating	✓	×
Personalized Messages	✓	×

participation in online communities, an evaluation based on the Comtella system was carried out. It aimed to answer the following questions:

- Will the users in the test group rate articles more actively?
- How well will the summative ratings reflect the real quality of the articles?
- Will the users tend to share resources earlier in the week?
- Will the actual number of contributions be close to the desired one?
- Will the users share the number of articles that is expected from them?
- Will the users contribute a higher percentage of high-quality articles?
- Will there be information overload?

According to the evaluation plan, the users were divided into two groups: a test group and a control group. Two different Comtella systems were created to serve the two groups which were identical except for the extra features (see Table 2) provided by the new mechanism that were offered only in the system for the test group.

5.1 Evaluation Design

After the two systems were set up, a case study was launched in a class on Ethics and Information Technology offered in the 2004–2005 winter session to evaluate the effectiveness of the proposed mechanism. The participants were encouraged to use the system to share Web-articles related to the topic of each week. The course content involves extremely broad spectrum of issues, ranging from freedom of speech and the internet, to privacy, to intellectual property, to workplace issues, discrimination, outsourcing, etc. Many of these issues are discussed widely in the press in news stories, recent court case developments, relevant issues as reflected in political campaigns, etc. Requiring students to find and share articles in Comtella pursued several goals. The first goal was to share the onerous task of looking up all these articles so that the instructor is not overloaded. The second goal is to awake the students' awareness of ethical issues as they appear in the real world reflected in the media rather than viewing the course as purely "academic", abstract content and textbook cases. The most interesting stories found by the students were discussed in class from ethical perspective, the students took sides and argued for the different viewpoints or the different stakeholders involved in the story. Comtella provides a good platform for the students to share and receive extensively the newest information. The rating system helps to share the heavy workload of evaluation the relevance and quality of the stories among all the users, thereby making it feasible. According to the curriculum, a

different topic was introduced in each week except that the same topic was discussed in the fourth week and the fifth week. In total, eight topics were discussed during the study. Five percent (5%) of the final mark in the course was based on the number of shared articles and another 5% was based on the quality of the weekly summaries that the students had to write of an article chosen by them, from those shared in Comtella as follows:

- *Participants*

The participants were 31 fourth-year undergraduate students who took the course. They were assigned into two groups: a test group of 15 users and a control group of 16 users. To account for cultural and gender-based differences in the users' initial predisposition for participation, the assignment of students to groups was based on having equal proportion of Canadian to foreign and male to female students in each group. Besides the stratification for nationality and gender, the students were randomly assigned to the two groups. Before the study, all the participants voluntarily signed consent forms, which granted us permission to use their data.

- *Methods*

To evaluate the effect of the improved incentive mechanisms, we compared the behaviours of the test group, which used the system with all the features of the mechanism and the control group, which used the system where some functions were blocked, as shown in Table 1. To avoid the effects that the contribution patterns of one group can have impact on the behavior of the other group, the two groups formed completely separated online communities, and shared different articles. We tried to make the two systems visually as similar as possible; they shared the same entry (URL) and students were automatically redirected, so that it was hard to notice that there were different systems for different users. The system for the control group was not weaker one in terms of system functionality and usability. It just did not offer the features of the new incentive mechanism. Yet the students in both groups followed the same class schedule and shared lectures, classroom, and coursework throughout the study. We could not inhibit students talking with each other in the class, neither could we put them into two different classrooms. Therefore, we could not guarantee that there was no influence between the two groups, but we did what was possible to minimize the influence.

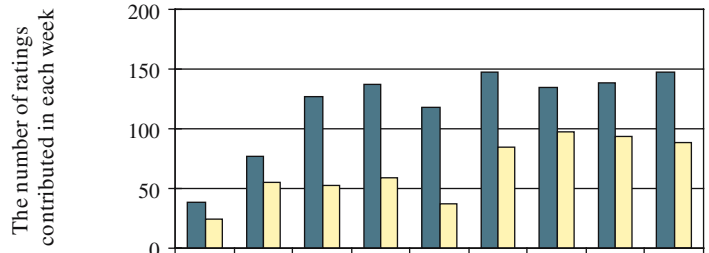
After the evaluation, a post-experiment questionnaire was distributed to the participants to collect feedback about their experiences in their online communities. Comtella was programmed to record the times and the types of various user actions and contributions in the online communities, including logging on the system, sharing, reading and rating articles, commenting on articles, uploading summaries, etc.

5.2 Results

The data from the system-logs and the post-experiment questionnaire showed the following results.

5.2.1 The users in the test group were more active in rating articles

The system logs showed that the number of ratings given in the test group was consistently higher than that in the control group in each week (Fig. 7).



	1	2	3	4	5	6	7	8	9
Test Group	39	77	127	137	118	147	135	138	147
Control Group	24	55	53	59	37	85	98	93	89

Fig. 7 The number of ratings contributed in the two groups in each week

Throughout the 9 weeks, the difference between the total number of ratings in both groups was significant, resulting in a total of 1,065 ratings in the test group and 593 ratings in the control group. The percentage of the articles rated was also higher in the test group than in the control group (78.3 vs. 53.3%). We also counted the ratings contributed by each individual user and applied the Mann–Whitney test (Lowry 1999) to verify the hypothesis that the users in the test group rated articles more actively than those in the control group. The results of the test showed that the z-ratio was equal to 1.727, which confirmed the hypothesis with statistical significance beyond the 0.05 level. This clearly shows that the proposed incentive mechanism with *c-points* and the associated rewards showed sustained effectiveness in stimulating users to rate articles during the experiment. Besides, the users' attitudes toward the *c-points* were mostly positive. Among the 15 users in the test group, nine users (60%) indicated in the questionnaires that the *c-points* were useful to make their articles more visible to others; eleven of them (73.3%) used more than 40% of all the *c-points* they earned to make their contributed articles more visible. Of all the *c-points* awarded to the users, 62.8% were attached to articles.

5.2.2 The articles with higher ratings were more likely to be chosen by users to summarize

This result was found by analyzing users' activity-logs in both systems. We observed that the articles with higher summative ratings were chosen by the students for summarization[1] more frequently. We can safely assume that the students tended to choose interesting or good-quality articles to summarize because they desired good marks for their summaries. The attention and the effort they paid to the articles they chose also indicated the articles were valuable. So the number of times an article was chosen can be used as an indicator of its quality. Then the correlation between articles' ratings and the number of the times they were chosen can be explained by two hypotheses:

(i) The articles' summative ratings reflect well the quality of the articles, therefore they showed a correlation with the number of times the articles were summarized.

[1] Writing summaries was a weekly-based assignment of the course; the students were free to choose to summarize any of the articles shared for the topic of the week.

(ii) The students tend to choose articles by looking at their summative ratings. If this hypothesis is true, the correlation does not mean that the ratings correctly reflect the quality of the articles.

In the exit questionnaire the users were asked whether they chose articles to summarize according to the summative ratings or the times the articles were viewed or none of these two indicators. Seven users answered that their selection was guided by the summative ratings; two users answered that their choice was guided by the number of times the articles were viewed; the remaining 22 users (71.0%) indicated that they chose articles based on none of these two indicators.

Since there is a connection between the article's summative rating and the number of times it was viewed, we removed from consideration the data from the seven users who looked at the ratings and the two users who looked at the number of view-times, and took into account only the data from the 22 users who stated that they did not choose articles based on any of these indicators. The results still showed a strong correlation between the articles' rating and the number of times they were summarized by the 22 users. Table 3 shows that the articles that earned higher ratings were on average chosen and summarized more often. Since these 22 users selected articles according to neither the summative ratings nor the number of view-times and the correlation between the articles' summative ratings and the number of times they were chosen by these users is statistically evident, we can conclude that the articles' summative ratings are able to reflect their quality.

An interesting result is that although there was no limit on the type or the length of the articles that could be shared, most of the articles shared by the users in the two groups were magazine articles, which were very popular and easy to read. Some academic papers were also shared during the study but their number was small. The articles with high summative ratings were usually the ones that were easy to read and understand and presented original viewpoints, so they were very interesting stories. Academic papers that were long and more profound were usually ignored by the users. So the quality reflected by summative ratings indicated the interestingness and readability of the articles rather than the technical or academic quality of the articles.

Table 3 The relation between the articles' summative ratings and the times they were summarized by the users

R	N	T	N/T
Greater than 4	7	5	0.714
4	26	15	0.577
3	89	31	0.348
2	167	34	0.204
1	303	39	0.129
0	47	7	0.149
Less than 0	154	6	0.039
Not rated	407	18	0.044

R: Summative rating
N: The number of the articles that had the rating shown in column R
T: The total times these articles were chosen for summarization (by the 22 users)
N/T: The number of times these articles were chosen on average

5.2.3 The users in the test group were more satisfied with the summative ratings received by their articles

We found that the users in the test group were more satisfied with the summative ratings that their contributions received from other users. The data from the questionnaire showed that 53% of the users (eight users) in the test group thought that the final ratings received by the articles *they shared* reflected fairly their quality, while in the control group only 31% (five users) thought so. This is partly because the users of the control group rated fewer of their contributions than users of the test group. The ratio of the articles to the ratings in the test group is 1: 1.74 and that in the control group is 1: 1.01. Around half (46.7%) of the articles shared in the control group were not rated at all while in the test group only 21.7% were unrated. More deserving articles did not receive ratings in the community of the control group than in the test group community. Apparently, the quality evaluation based on collaborative rating requires a critical number of user ratings. Before this number is reached, increasing the number of user ratings in the system through incentives can improve the accuracy of the quality evaluation of the shared resources.

5.2.4 The users in the test group tended to share resources earlier in the week

The system-logs showed that the users in both groups shared more articles in the first half of each week than in the second half. By "week" we mean a calendar week of 7 days, since students still could submit articles for credit until Sunday midnight and many did use the weekend for their Comtella activities. Yet, 66.1% of all the articles in both systems were contributed in the first 3 days of the week. The users in the test group shared a higher percentage of their contributions in the first 3 days (71.3%) than the users in the control group (60.6%) and the difference between the two groups was quite large in each week (ranging between 7 and 14%).

In Comtella, one topic was active in the system for only 1 week and the resources shared for the topic could be rated only in that week. So the resources posted earlier in the week gained more time and chance to attract attention and to collect ratings, which was realized by 17 users (54.8%) across the two groups (according to the data from the questionnaire). This explained why the users in both groups tended to contribute articles in the first half of the week.

The higher percentage of early contributions in the test group proved that the adaptive reward mechanism was effective in motivating users to share resources early. The reason for this conclusion is that the adaptive reward mechanism was applied only in the system used by the test group and it was the only difference between the two systems that was related to the timeliness of making contributions. For the users of test group, the adaptive weight for sharing resources deceased with time.

5.2.5 There was no big difference between the total numbers of shared articles across the two groups: In the test group, more than half of the users tried to share the number of articles that was expected from them. the overall number of articles in the test group was not excessive

The difference between the total numbers of articles shared over 9 weeks in the two groups was small. There were 613 articles shared in the test group vs. 587 in the control

group, i.e. the total number of submitted articles by the test and control groups during the experiment varied with only with ±2.16% from the average of 600 articles. In the control group, the quality of the contributions was not taken into account, so the users could have easily shared more articles to receive higher membership levels. However, even though this incentive existed, the users did not contribute excessively.

This may be due to the characters of the students in the control group. Some people are easy to be motivated by glory and recognition, and some are not. Even the presence of just one person with "greed" for higher status may have lead to an evolving competition, which could have tipped the system into a different state. In fact, this is exactly what happened in the previous version of the system, experimented in 2004 (Cheng and Vassileva 2005).

The number of articles contributed by each user ranged from 3 to 111 in the test group and from 0 to 124 in the control group. The standard deviation in the test group was slightly smaller than that in the control group (29.4 vs. 32.1).

The users of the test group were asked in the final questionnaire whether they followed the suggested number of contributions. Eight users (53% out of 15) responded positively. We calculated for each user the average difference between the actual shared number and the expected number over 9 weeks and found that for eight of the users (53%) the average differences were less than 2, which means these eight users almost contributed the number of articles that was expected from them. Interestingly, the two groups of eight users did not entirely overlap. Table 4 shows each user's answer to this question and the average of the differences between the actual number of articles she shared and the number expected from her in the 9 weeks. These results indicate that more than half of the users in the test group were persuaded to share resources in or close to the number that was expected from them.

Table 4 The difference between the users' real and expected contributions

User #	Average Difference	Answer to the question
04	0.75	No
07	0.88	Yes
12	0.88	Yes
13	1.00	Yes
03	1.13	Yes
10	1.13	Yes
01	1.38	Yes
15	1.63	More
14	2.00	No
11	2.25	Yes
08	2.75	Less
05	3.38	Yes
09	5.25	Less
06	8.38	No
02	9.00	More

The question

Did you pay attention to the number of articles the system expects you to contribute?

Four options:

Yes - Yes, I tried to share in the number the system expected.
More - I always shared more than the number.
Less - I always shared less than the number.
No - I did not care about that number; I shared as many as I wanted.

In the test group, the total number of contributions did not have a big discrepancy from the overall expected number for most of the topics since about half of the users tried to share the number expected from them and the extra contributions made by users who tended to share more were compensated roughly in the same number by the contributions of users who tended to share less. Table 5 shows the differences between the total number of shared articles and the overall expected number for the eight topics.

It can be seen that the differences for all the topics were less than 20% except for Topic 3, 5, and 8. The students over-contributed for Topic 3 possibly because the topic ("Wiretapping and encryption") happened to be of highest interest for them. We have yet to find an explanation for the large difference between the expected and actual number of contributions for Topic 5. The overall difference for the eight topics was equal to 12.3% of the total of the expected numbers. This shows that the number of articles in the test group was not excessive and the approach of setting specific goal for each user to contribute was helpful to control the overall quantity of resources in the system.

5.2.6 In both groups, the users' attitude towards the quality of the articles were generally neutral

As for the quality of the articles in both systems, the questionnaire asked the users in both the control and test group to give the rough estimate of the percentages of articles with high, medium and low quality in their respective systems. The data in Table 6 shows the averages of users' estimations, which indicates that their attitude toward the quality of the articles in their own communities was mostly neutral.

Table 5 The differences between the actual number and the expected number in the test group for the eight topics

Topic	1	2	3	4	5	6	7	8	Sum
Week	1	2	3	4 & 5	6	7	8	9	
E	60	60	61	121	62	62	60	60	546
A	57	66	98	121	83	72	69	47	613
D	−3	6	37	0	21	10	9	−13	67
−5.0%	10.0%	60.7%	0.0%	33.9%	16.1%	15.0%	−21.7%	12.3%	

The same topic was discussed in week 4 and 5. So there was only one expected number for both of the weeks.
E: The expected number of articles shared for the topic.
A: The actual number of articles shared.
D: The difference between A and E, equal to (A-E).
P: The percentage of the difference, equal to (A-E)/E.

Table 6 The users' estimations of the percentages of the articles with high, medium and low quality in both groups

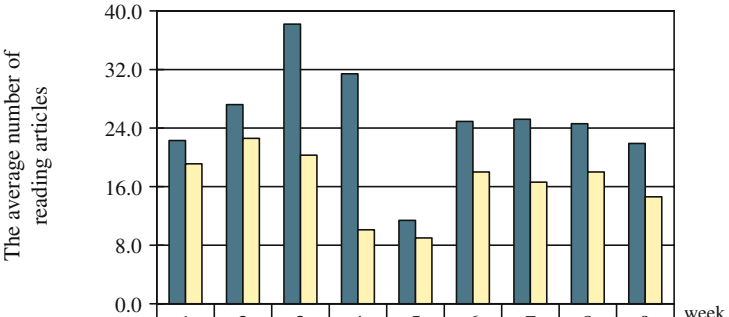
Group	High (%)	Medium (%)	Low (%)
Test Group	23.5	47.1	29.4
Control Group	27.3	41.5	31.1

However, it is hard to compare the quality of the articles in the two groups because the users in any group had experience only in one system, the articles shared in each community were different, and the users might have had different criteria of quality evaluation.

5.2.7 The users in the test group were more active in terms of logging on the system and reading articles

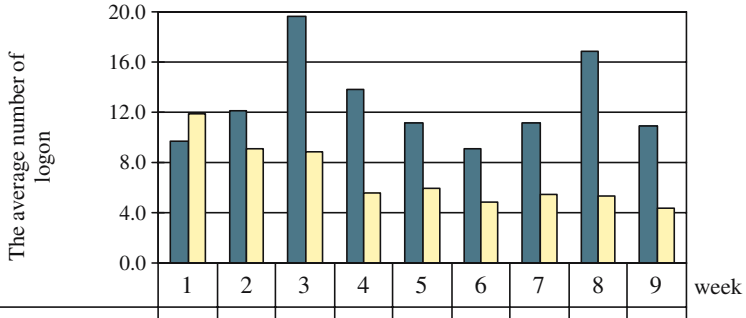
The analysis based on the system-logs showed that the users in the test group participated in the system more actively than the users in control group did. The number of times of reading articles and logging on the system were computed for both groups of users over 9 weeks. Figures 8 and 9 show the average number of times/actions of accessing articles (possibly reads) and logon actions in both groups over the 9 weeks.

In each week, on average, the users in the test group consistently read more articles and logged on the system more times than the users in the control group (except that in the first week the members of the control group logged on more times). Throughout



	1	2	3	4	5	6	7	8	9
Test Group	22.3	27.2	38.2	31.4	11.4	24.9	25.2	24.6	21.9
Control Group	19.1	22.6	20.3	10.1	9.0	18.0	16.6	18.0	14.6

Fig. 8 The average number of times of reading articles for the users in both groups



	1	2	3	4	5	6	7	8	9
Test Group	9.7	12.1	19.6	13.8	11.1	9.1	11.1	16.9	10.9
Control Group	11.9	9.1	8.8	5.6	5.9	4.9	5.4	5.3	4.4

Fig. 9 The average number of times of logon for the users in both groups

the study, the total number of times of reading in the test group and in the control group were 3,407 and 2,373, respectively; the total number of logon times were 1,714 and 982. This clearly shows that the activities of reading articles and logging on the system in the test group were more frequent than those in the control group, which means that the proposed mechanism could ensure more active and sustained user participation. A noteworthy fact is that these activities were not rewarded in either of the systems. Therefore, the more active user participation in the test group can not be directly attributed to a particular incentive in the mechanism.

One possible explanation could be that the test group users' more active participation in reading articles and logon is a by-product of stimulating them to rate articles. However, this is not the case. After analyzing the relevant data, we found no correlation between the number of ratings contributed and the times of reading articles or the times of logon in both systems during the 9 weeks. Table 7 shows the number of ratings contributed, the number of times the users read articles and the number of times they logged on the system in each of the 9 weeks for both groups.

Obviously, more ratings in a particular week are not related to more readings or more logons in that week (e.g. for the test group, Week 9 and for the control group, Week 7). We also computed for both systems the correlation coefficients of the number of ratings and the times of reading or logon over the 9 weeks (see Table 6). It can be seen that none of them is greater than 0.25, which clearly shows that the activities of reading and logon are not correlated with rating in either system. Actually, for all the users in both groups, the numbers of the articles they read are far more than the numbers of articles they rated. The system-logs showed that the users rated on average only about 30% of the articles they read. The users of the test group rated a higher percentage than those of the control group (34.6 vs. 25.7%). Obviously, the users could rate more articles that they had read without any extra effort of reading additional articles, since there were plenty of articles that they had read but not rated yet. Therefore, the incentive mechanism was able to increase the proportion of the articles the users rated from the articles they read, but did not stimulate users to read more articles. This explains why more ratings in particular weeks did not guarantee more times of reading articles in those weeks during the study. Similarly, Table 6 shows that there is no correlation between the users rating and logging in the system. The users logged on the system for various purposes and they could rate as many articles

Table 7 The number of ratings, the times of reading and number of logon actions in each week and the correlation coefficients for both groups

Week	Test Group			Control Group		
	Rating	Reading	Logon	Rating	Reading	Logon
1	39	335	145	24	306	191
2	77	408	181	55	361	146
3	127	573	294	53	325	140
4	137	471	207	59	162	89
5	118	171	166	37	144	94
6	147	374	137	85	288	79
7	135	378	167	98	265	87
8	138	369	253	93	288	85
9	147	328	164	89	234	71
Correlation coefficient with the no. of ratings	0.124	0.246		0.067	−0.757	

as they wanted during one logon. So the motivational mechanism's effect on encouraging users to rate articles had no influence on their activities of reading articles or logging on.

In addition to the difference in the numbers of reading and logon actions between the two groups, we also observed that in the control group the number of logons decreased during the 9 weeks and showed a negative correlation with the number of ratings because ratings increased during the study (see Table 6). However, in the test group no such decrease was observed. This gives some hint that the levels of information overload are different in the two groups. Jones' and Rafaeli's (1999) research indicated that information overload results in a decrease or end of the users' participation in online communities. The results from our previous study (Cheng and Vassileva 2005) also showed that the users' contributions and participation decreased after the system was flooded with shared articles. Evidently, user participation is roughly in inverse proportion to the level of information overload in the online community provided that there are no motivational factors working on the users. Since reading articles and logging on system are not among the activities that were rewarded by the proposed mechanism in any of the groups group, the difference between the users' levels of engaging in these activities in the two groups shows that the information overload in the test group was less serious than that in the control group.

6 Discussion

6.1 Confirmed hypotheses

The data from the evaluation showed that throughout the study, the users in *the test group consistently contributed more ratings* than the users in the control group, which supports the hypothesis that the collaborative rating mechanism can persistently stimulate the users to rate more articles. Besides, an evident correlation was found between the numbers of times the articles were chosen for summarization and the articles' summative ratings, which shows that *the summative ratings can reflect the quality of the articles*. The users in the test group seemed more satisfied with the ratings earned by their contributions because their contributions were rated more often than those in the control group. Apparently, a critical number of user ratings have to be reached for peer-evaluation based quality evaluation systems to work accurately.

The data showed that the proposed mechanism successfully *motivated the users in the test group to contribute their articles earlier* in each discussion period. The overall number of the contributions in the test group did not exceed the expected number in most of the weeks. More than half of the users in that group were persuaded to contribute the number of articles that was expected from them or close to this number. These results show that *the mechanism has a positive effect on controlling the overall number of resources in the system*. Besides, it was found that throughout the study, the users in the test group were consistently more active in reading articles and logging on the system, two activities that were not rewarded by the mechanisms in either of the two groups. This indicates that the proposed mechanism *is able to ensure more active and sustained user participation in the online community*. According to the results from the previous evaluation on Comtella and the research by others (Jones and Rafaeli 1999), the users' participation level decreases with the aggravation of information overload. Therefore, the users' different participation levels in the two groups indicate that *the*

level of information overload in the test group was lower than that in the control group. However, whether information overload was entirely avoided in the test group is still unknown.

6.2 How to choose the parameters for the mechanism?

There are several more general issues about the design of the mechanism that deserve discussion. First of all, the group model is "hand-crafted", rather than generated from real data from successful online communities, unlike for example, that proposed by (McLaren et al. 2006). There are many "magic numbers" in the mechanism, for example, the expected sum of contributions from the whole community on a given topic (week), the thresholds for reputation values, the exact shapes of the community reward function and the individual reward function. A possible criticism to our methodology is that the results may depend on the appropriate setting of the values while we give no guideline about how these values should be set. However, it is not our goal to prove that a certain set of values is better than another; in fact we believe, similar to (Alfonseca et al. 2006), that since the numbers are empirically derived, there is no optimal set of values, since it would be impossible to compare the effects of different numbers in exactly the same conditions. However, we believe, that there is a range of acceptable values, leaving a freedom of choice to the community administrator (e.g. instructor) depending on the goals of the community, the characteristics of community contributions (e.g. average quality, timeliness), and contribution dynamics. We believe that setting these parameters for particular community can be achieved by using simulations calibrated with empirical data from previous usage of the system in a similar type of community. Some of our current work uses system dynamics-based modeling and simulation to explore further how to select appropriate values for the parameters and appropriate shapes of reward functions to fit better a given community "character" and purpose.

Our choice of values for ratings (just $+1$ or -1 in Comtella) deserves some discussion, since other systems deploy a larger spectrum of values. Allowing for a wider scale of options generally increases the cognitive load of rating. Also it does not necessarily allow a better way to capture the quality of a reading, since it involves many aspects, such as readability, interestingness, popularity of the topic, originality, technical quality, etc. We considered allowing the students to rate separately each of these aspects. However, this would increase the complexity of rating action and finally discourage them from rating. In the current rating system, the users give an overall evaluation on the article based on their personal attitude to it (hopefully, a resulting form an implicit evaluation of the combination of all the abovementioned aspects), which is easy to do, similar to voting. From the analysis of the papers that ended up with high-ratings, we found that in such a rating system, the readability, originality and brevity of the paper were more important factors for the students than the paper's technical or academic quality. This was somewhat in line with the pedagogical goals of the class instructor, who encouraged the use of Comtella mostly to achieve a broad overview and ensure novely/freshenss of cases and stories, rather than depth and argumentation.

6.3 What is an appropriate "objective" measure for the quality of ratings?

An interesting issue to debate is whether it is appropriate to use the average rating that a resource has obtained so far as a measure for the quality of subsequent ratings.

Recall that the quality of a rating for a given resource is computed as reciprocal to the difference between the rating and the average of the ratings that the resource has received so far. Does this mean that the average taste should "rule"? There are arguments in both directions. On one side, the average taste of the community, for example in a system applied in educational context, like Comtella, may tend toward more superficial, easy to read articles, which was the case in our experiment. In communities with users that are trusted to have high quality standards, or who are expected to define the quality standards of the community (e.g. the instructor in an educational community), it is possible to take the ratings of these users as standard, even when they have not rated every single resource.

Unlike Slashdot, we did not weight differently the ratings coming from different users. This is possible to do, but it may put too much onus on measuring the quality of ratings, since users who tend to rate similarly to what later turns out to be the average rating will gain more "voting power" thus creating a circular self-reinforcing emphasis towards the "average". Since we do not have publicly known and trusted "moderators" to steer consciously through their ratings towards the values of the community, such a circular weighted scheme may create an additional undue bias in the system towards the average taste. Yet, we could incorporate easily weighted ratings in future, with dedicated evaluators.

Differential mechanisms for trust and reputation, which can be applied to propagate trust in the rating abilities of community members based on the similarity between their ratings can be used to provide personalized ratings rather than overall community ratings. Algorithms developed in the area of collaborative filtering systems (Herlocker et al., 2000) can be applied to find clusters of users with similar tastes based on commonly rated resources and use them to extrapolate anticipated ratings that these users would give to unrated resources and recommend articles that these users would likely like.

6.4 Design to prevent gaming the system

Many of the decisions guiding the design of the incentive mechanism and the corresponding user interface (what information is shown and what is not, what is shown on the same screen, etc.) were made to "outsmart" the users who we envisaged would attempt to game the system. For example, we developed the adaptive motivational mechanism to deal with the cognitive overload in the community, caused by excessive contributions by users trying to game the system, motivated by the hierarchical membership mechanism, i.e. the very need to develop the new adaptive incentive mechanism came from the presence of gamers.

Another example—we did not show in the table of search results who contributed each resource to prevent users from shilling, i.e. forming cliques of friends rating each others' contributions high to earn higher reputation for sharing resources. Showing the contributor of each resource would have been beneficial in creating a feeling of community, in discovering users with similar interests and in navigating the search results.

The average rating of a resource was also not shown in the list of search results (the only place where the users can rate resources), but instead the summative rating was shown. This made it difficult/impossible for the users to infer the average rating. The reason was that we wanted to prevent users from gaming the system by submitting ratings close to the current average. This would have been an easy way for them to increase their reputations in rating.

Another obvious way to game the system is to submit meaningless ratings without bothering to read the resource. In this way a user can quickly earn a lot of *c-points*, which can be invested to increase the visibility of user's own contributions and therefore, their chance to be rated. Of course, this does not necessarily bring benefit to the user, especially if the resource attracts negative ratings. Yet, to alleviate the danger of this form of gaming, the interface was designed to allow sorting the search results by any of the criteria shown on the first line of the table of results (see Fig. 5): by the total number of ratings received, by the number of view times, alphabetically by title, etc. However, most users did not use the sort options but relied on the default sorting (by *c-point*s and then by time), which shows prominently the "sponsored" links.

We do not allow users to modify the rating they have made on papers. As soon as the rating was given by the user, it could not be taken back and a certain amount of *c*-points was awarded immediately to the user. This decision was made to protect the system from users who would try to game the system and earn *c*-point through rating the same articles multiple times. Yet, it would have been perhaps good to have an option to reverse a rating given, if the user honestly changes her mind about the article.

To prevent users from stocking up too many *c-points*, which would give them too much power to promote their own contributions, users have a limited number of ratings that they can give away, defined by their membership level. Also to ensure rating activity on a regular basis, we introduced "expiry date" for *c-points*.

To prevent users from going into bidding wars in "sponsoring" their contributions, we allowed only a limited number of *c-points* to be invested in a contribution, and all invested *c-points* were rounded down, i.e. two resources with 32 or 37 *c-points*, even though they will be sorted correctly, are displayed as 30+ in the list of results, as shown in Fig. 1.

Yet, of course our current design is not unbeatable. An incentive mechanism evokes the "sleeping computer gamers" in many users by providing a challenge for their ingenuity. It seems that this cycle of "design to prevent gaming" is typical for all incentive-based systems. Optimizing utility to reap the maximum benefits by investing minimal efforts seems to be a basic feature in human economic behavior (Levitt and Dubner 2005). Instead of looking pessimistically at these findings, we prefer to look at the attempts to game as evidence that the mechanism has a strong motivational effect on users, even if not exactly as intended.

7 Conclusions and Future Work

To achieve a sustainable level of user participation in online communities, it is important to control the quality and the quantity of users' contributions and avoid information overload or degrade its level. Therefore, an incentive mechanism with adaptive rewards was designed and evaluated in a case study. The mechanism includes two parts. The collaborative rating mechanism ensures a decentralized way of measuring the quality of contributions by encouraging the users to rate each other's contributions. Based on this quality measurement, the adaptive rewards mechanism encourages users' contributions differently, taking into account the users' individual reputation and the current needs of the community.

The results from the case study showed that the proposed mechanism worked well in the online community, and was able to achieve most of the goals as desired.

Although the results of the case studies of the proposed mechanisms are quite positive and exciting, there are still some questions that have not been answered by the study and also some very interesting directions that deserve further research.

First, it is still unknown whether the proposed mechanism is able to improve the quality of the resources in online community. The results from the questionnaire about the quality of shared articles are unable to answer the question about which group as a whole produced higher-quality contributions. We suppose that the articles in the test group are of higher quality since they were read more frequently by users. However, more work is needed to prove this hypothesis. One possible solution is to invite a new group of students or an expert to evaluate the articles shared in both systems. For example, we may have them rate the articles on a one to five scale and then calculate and compare the average ratings for the articles in both groups. Because none of them have contributed any of the articles, their opinions would be more unbiased.

The performance of the collaborative rating system can be improved. It was found that a higher number of user ratings could make the quality evaluation more accurate. However, when the number of ratings reaches the critical level, the summative ratings will converge toward the community measure of quality of the resources, and more ratings will not improve the accuracy of the evaluation. It would be valuable to find out the critical number of user ratings for the peer-based rating systems to work properly. When this threshold is known, the rewards for rating can be adapted so that rating articles is rewarded more before the threshold is reached. Finding the threshold seems to be a statistical question based on the variance of the ratings. Also, ratings themselves are dependent on the experience, care, and skill of the rater. Finally, the rating of an article has a time function; the second article to address the same small detail of the weekly topic is not worth as much as the first one. A mathematical model with these properties can be build, but it will need to be calibrated, and for this a large amount of ratings will be necessary to ensure that they are converging. This is an interesting direction for future work.

In addition, the problem that some good articles may be never read by others than the original providers and may end up with no ratings remains unsolved. To encourage users to rate unrated articles, it is possible to offer the first rater of the article some extra benefit, especially if the article eventually ends up with a high-rating.

A personalized recommender system can be applied to help the users to overcome the problem of information overload. Although the adaptive reward mechanism has proved to be capable of lowering the level of information overload in the online community, the problem cannot be completely eliminated, especially for large-scale online systems with thousands of users. It would become much harder to control the overall quantity of the contributions in the system if the user population keeps increasing, since its diversity will also increase and the ratings will become less meaningful. Therefore, providing personalized recommendations for users based on the ratings given by users with a history of giving similar ratings will allow to reduce the search results list and may help them to filter out the information they do not want.

Some changes in the adaptive reward mechanism will be needed, if it is applied for different kind of online communities. The adaptive reward mechanism was designed in the context of a community for sharing class-related resources. The main characteristics of this type of community is that it is closed (a class with fixed number of students) and that the topic for sharing resources in the system is changed weekly and at any given time there is only one active topic being discussed. Therefore, the evaluation of users' contributions and reputation and the updating of the users' memberships follow

a weekly rhythm. However, in open interest-based online communities, like discussion forums, blog-systems, collaborative filtering systems, etc., usually contributions are made simultaneously in many forums/categories and there is no time limit to the participation in each forum/category. Despite these differences, it is still possible to motivate or control users' contributions based on the quality data of their previous contributions. We could build the user's reputation of making contributions on all the topics or on each topic separately, depending on the specifics of the community. The discussion on some categories is ephemeral. For these categories, it is necessary to encourage users to contribute early because later contributions may not get enough notice.

In a small team-like community, where the members are expected to collaborate to achieve a common goal, it may be possible to motivate participation without rewards. Just providing awareness of the common state of tasks or resources, through a shared coordinating representation, as in (Introne and Alterman 2006), showing the current levels of different kinds of contributions, may be sufficient to motivate users to participate. Every online community has its own characteristics. We do not believe that it would be possible to find universal mechanisms that are perfect for all systems. However, the ideas of rewarding desirable activities, assigning different status and service to users to arouse comparison and adapting users' reward to influence their contributions can be applied widely.

Finally, it would be very interesting to explore deeper the affective impact of the incentive mechanism and its influence on the users' motivation to participate. Masthoff and Gatt's work (2006) suggests a way of building affective models of users in terms of satisfaction and embarrassment in a group recommender system. Their approach may be applicable in our case, since the visualization of the users' status or any kind of participation metric is likely to cause exactly these emotions. How these emotions relate to motivation for participation is not clear now, but it might be another direction for future work.

Acknowledgements This research has been supported by the NSERC Discovery Grant of the second author. The authors are grateful to Lingling Sun and Weidong Han for implementing parts of the Comtella system and to the students in the CMPT408 (2004/05) class for participating in the experiment and granting consent for using their data. This paper (or a similar version) is not currently under review by a journal or conference, nor will it be submitted to such within the next three months.

References

Alfonseca, E., Carro, R., Martin, E., Ortigosa, A. and Paredes, P.: The impact of learning styles on student grouping for collaborative learning: a case study, 2006

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., Kraut, R. E.: Using social psychology to motivate contributions to online communities. Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, pp. 212–221, Chicago, Illinois (2004)

Burgahain, C., Agrawal, D., Suri, S.: A game theoretic framework for incentives in P2P system. In: Proceedings 3rd IEEE International Conference on P2P Computing, Linkoeping, Sweden, IEEE Press (2003)

Cheng, R., Vassileva, J.: User motivation and persuasion strategy for peer-to-peer communities. Proceedings of the 38th Hawaii International Conference on System Sciences (Mini-track "Online Communities in the Digital Economy"), pp. 193–202, IEEE Press

Golle, P., Leyton-Brown, K., Mironov, I.: Incentives for sharing in peer to peer networks. Proceedings Electronic Commerce 2001, pp. 264–267, Tampa, FL, ACM Press, 12–17 Oct 2001

Greer, J., McCalla, G., Vassileva, J., Deters, R., Bull, S., Kettel, L.: Lessons learned in deploying a multi-agent learning support system: the I-help experience. Proceedings of the 10th International Conference on Artificial Intelligence in Education, pp. 410–421, San Antonio (2001)

Herlocker, J. L., Konstan, A. J., Riedl, J.: Explaining collaborative filtering recommendations. Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, 241–250, (2000)
Hiltz, S. R., Turoff, M.: The Network Nation: Human Communication Via Computer. Addison-Wesley Publishing Company, Inc., London (1978)
Introne, J., Alterman, R.: Using Shared Representations to Improve Coordination and Intent Inference, in this issue
Jones, Q., Rafaeli, S.: User population and user contributions to virtual publics: a systems model. Proceedings of the international ACM SIGGROUP Conference on Supporting Group Work, pp. 239–248, Phoenix, Arizona (1999)
Lampe, C., Resnick, P.: Slash(dot) and burn: distributed moderation in a large online conversation space. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 543–550, Vienna, Austria (2004)
Levitt, S., Dubner, S.: Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. Harper Collins, New York (2005)
Lowry, R.: Subchapter 11a: the Mann-Whitney test & subchapter 12a: the Wilcoxon Signed-Rank test. Concepts and Applications of Inferential Statistics, 1999 Available online at http://faculty.vassar.edu/lowry/webtext.html. (last accessed on May 30, 2006)
Masthoff, J., Gatt, A.: In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems, 2006
McLaren, B., Walker, E., Harrer, A., Bollen, L., Sewall, J.: Creating cognitive tutors for collaborative learning: steps toward realization, 2006
Merton, R., Zuckerman, H. A.: The matthew effect in science: the reward and communication systems of science are considered. Science, **199**(3810), 55–63 (1968)
Preece, J.: Online Communities: Designing Usability, Supporting Sociability. Chichester, UK, John Wiley & Sons (2000)
Shenk, D.: Data Smog: Surviving the Information Glut. HarperCollins, New York (1997)
Shirky, C.: Fame vs. fortune: micro-payments and free content. First published Sept. 5, 2003 on the "*Network, Economics and Culture*" mailing list, available on line at: http://shirky.com/writings/fame_vs_fortune.html. (last accessed on May 30, 2006)
Vassileva, J.: Motivating participation in peer to peer communities. Proceedings of the Workshop on Emergent Societies in the Agent World, ESAW'02, pp. 141–155. Madrid, Springer Verlag LNAI 2577, 2002
Vassileva, J., Cheng, R., Sun, L., Han, W.: Stimulating user participation in a file-sharing P2P system supporting university classes. *P2P Journal*, July 2004 issue, 14–23. Available online at: http://bistrica.usask.ca/madmuc/Papers/P2PJ.pdf (last accessed on May 30, 2006)

Authors' vitae

Ran Cheng University of Saskatchewan, Department of Computer Science, 176 Thorvaldson Bldg., 110 Science Place, Saskatoon, Saskatchewan, S7N 5C9, Canada.

Ran Cheng received his B.S. in Computer Science from Beijing Information Technology Institute in 2002 and his M.S. degree in the same field from University of Saskatchewan in 2005. While pursuing his M.S. degree at University of Saskatchewan, Ran Cheng mainly worked in the area of user motivation and persuasion in peer-based resource-sharing systems. His paper summarizes most of the research work he did in this area during his graduate study.

Julita Vassileva University of Saskatchewan, Department of Computer Science, 176 Thorvaldson Bldg., 110 Science Place, Saskatoon, Saskatchewan, S7N 5C9, Canada.

Dr. Julita Vassileva is Associate Professor of Computer Science at the University of Saskatchewan. She received her Ph.D. degree in Mathematics and Computer Science from the University of Sofia, Bulgaria in 1992. Dr. Vassileva has worked in the areas of artificial intelligence in education, adaptive hypertext and hypermedia and multi-agent systems. More recently, her research has focused on ways to encourage participation in online communities, personal information management as well as recommendation systems using trust and reputation mechanisms. She has authored over hundred technical papers and has co-edited two books. Dr. Vassileva was program co-chair of the 8th international conference on User Modelling in Sonthofen, Germany, 2001.