Classsourcing: Crowd-Based Validation of Question-Answer Learning Objects

Jakub Šimko, Marián Šimko, Mária Bieliková, Jakub Ševcech, Roman Burger

Institute of Informatics and Software Engineering,
Slovak University of Technology in Bratislava,
Ilkovičova, 842 16 Bratislava, Slovakia

{name.surname}@fiit.stuba.sk

Abstract. The Web 2.0 principles reflect into learning domain and provide means for interactivity and collaboration. Student activities during learning in this environment can be utilized to gather data usable for learning corpora enrichment. It is now a research issue to examine, to what extent the student crowd is reliable in delivering useful artifacts and to bring in suitable tools to enable this. In this paper we present a method for crowd-based validation of question-answer learning objects involving interactive exercise for learners. The method utilizes students' correctness estimations of answers provided by other students during learning. We show that aggregate student crowd estimations are to big extent comparable to teacher's evaluations of provided answers.

Keywords: crowdsourcing, education, question, answer evaluation, technology enhanced learning

1 Introduction

In this paper, we deal with a phenomenon of *crowdsourcing within learning environments*. Web 2.0-induced paradigm shift, reflected in learning, triggered collaboration and availability of learning content to the masses. Besides taking advantage of an educational system, by consuming provided content, learners produce a significant number of user-generated data that constitute a tremendous potential for further improvement of learning process: not only they leave footprints in system usage logs, they also actively contribute by adding, annotating or modifying learning materials. Some user activities in educational systems may therefore be managed to produce useful artifacts (as by-products of learning), which can be utilized to supplement learning content (e.g., student-created explanatory notes attached to original learning material). Such artifacts may either be *learning objects*[1] themselves or can be useful otherwise, e.g. as metadata describing the learning objects. Research possibilities are open here, calling for devising learning activities which (besides educational effects)

[1] We adopt a broader definition by IEEE, which defines a learning object as "any entity, digital or non-digital, that may be used for learning, education or training" [5].

lead to creation of useful artifacts. One of the key aspects such approaches must consider is ensuring the quality of created artifacts, even if we involve relatively inexperienced students in the process.

In this context, a particular issue we focus on is the *acquisition of correctness information on student-created answers to exercise- and exam-like questions*. During learning exercises, students often interact with questions and answers (related to course domain). If a student answers a question during a learning course, he expects feedback on his answer correctness. Usually, this feedback is provided by the teacher (e.g. in-class learning sessions). However, the teacher's responses are not always available during *online* sessions. Instead, the online learning application comes to place, giving an automated feedback. This, however, can be done only for certain types of exercises (questions), such as multiple choice answers. The free text answers (which are sometimes the only suitable option from didactical perspective) cannot be evaluated sufficiently by today's automated means. The evaluation of free text answers can only be done by a human (and this burdens the teacher).

We present a method, which comprises learning from existing questions and answers (either correct or wrong) and at the same time, involves students into evaluation of correctness of answers for their peers during learning sessions. It is based on an interactive exercise, in which student is confronted with a question and existing student-created answer (further referred to as *question-answer learning object* or QALO). For example, in a software engineering course the question "What is the purpose of the feasibility study" may be answered with "To determine, if a problem is worth solving". The task for the student is to evaluate correctness of this answer. After the student does so, he receives feedback based on the aggregate crowd correctness evaluation based on previous evaluations provided by other students. By deploying this method into online learning environment, following effects are achieved:

1. The students are able to exercise their course domain knowledge autonomously.
2. The correctness estimations of individual students (to the same QALO) constitute the *crowd correctness estimation*, which, as we show in our experiments, approximates the true answer correctness (according to teachers).

The outcoming crowd correctness estimation is used in the exercise itself, but may also be used for other purposes, for example to give feedback to the original author of the answer (lifting some burden from the teacher). This enables the use of *question answering* exercise, where the question answerers are fed back by their colleagues over time. However, as the input for our method, any QALOs may be used, for example results of exams or homeworks, which may be reused this way.

We devised our method and deployed it within our Adaptive Learning Framework (ALEF) [12] in the Principles of Software Engineering course. Through it, we collected usage data for the exercise, computed crowd answer correctness estimations and evaluated their accuracy against grand truth provided by teachers. We show that student-generated data obtained collectively with no prior pedagogical knowledge can to big extent substitute evaluations provided by a teacher. As a result, students can support each other during their learning sessions and teacher's efforts in learning corpora creation and feedback provision is reduced.

2 Related work

We utilize principles of crowdsourcing, a paradigm which uses human computation for substituting machines in performing tasks hard or impossible to automate. Crowdsourcing often involves lay users, which raises the question of quality of the solutions they provide. This issue is in general solved by redundant task solving, collaborative filtering, consensus, peer-reviews etc. [10]. Our method bears a similarity with principles of community question answering systems, such as Yahoo! Answers2, which are a subgroup of crowdsourcing approaches. In these systems, answers to questions are acquired from some crowd members and are secondarily evaluated by other crowd members. The best answers eventually emerge. To reach quality answers, some works use automated analysis of the answer texts [2], other focus more on voting and filtering [3]. Focus is also given to predicting answerer level of expertise [1]. Community question-answering systems often collect answers as solutions of problems in specific domains. With our work, we aim to explore, whether their principles could be used in a didactical scenario, where users-reviewers do not seek answers to their problems but learn and test their knowledge instead.

The crowdsourcing principles are in general, relevant for the learning domain. With the emergence of technology enhanced learning, we witness paradigm shift in learning, especially when considering web-based learning environments. Benefiting from concepts introduced by Web 2.0 and moving towards genuine Read-Write Web [9], a student becomes more autonomous and less dependent on the teacher. She is provided with more competences since she can tag, rate, share and collaborate during learning. She becomes an active contributor rather than a passive consumer of learning content [4]. The activity of a learner is “boosted” not only in relation with an educational system, but also when considering collaboration during learning [11,13].

The distributed nature of the Web allows to connect and virtually gather various learners in a convenient way – anytime, anywhere. The students can therefore be viewed as potentially useful crowd force. To participate in the creative process, they can be motivated internally (by their own will to learn) or externally (by course points or gamification). Student activities may often result into new learning materials created intentionally [8, 14], or are utilized to promote existing educational content [6]. Crowd activities are implicitly connected with collaboration or collective intelligence – in either explicit or implicit manner [7].

3 Crowd Validation of Question-Answer Learning Objects

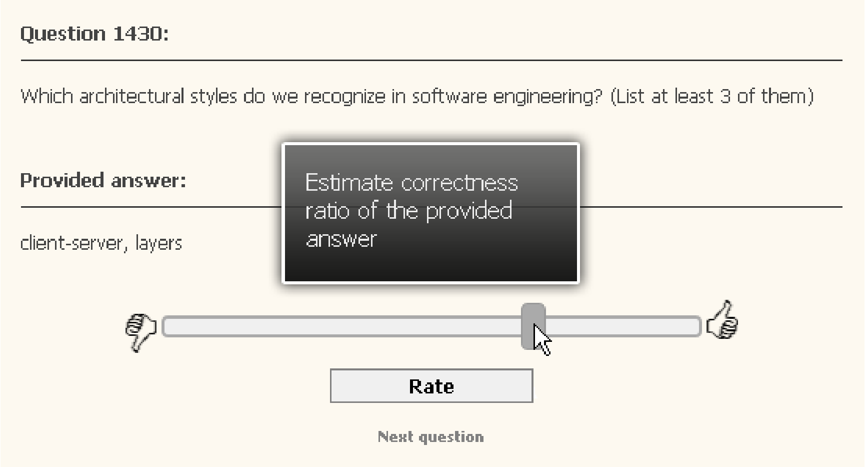
For retrieval of information on correctness of answers within *question-answer learning objects* (QALOs), we present a method consisting of interactive student exercise and a subsequent automated interpretation of student activity within this exercise. Our method retrieves the correctness information via crowdsourcing of the group of students attending a learning course. During her learning sessions, the student pulls,

[2] answers.yahoo.com

reviews and rates QALOs. By rating we mean *estimating correctness* of an answer provided in QALO. After that, she retrieves feedback in form of global QALO correctness estimation computed from estimations provided by other peers (a global "crowd truth"). The QALOs used by our method can be of any origin, provided they are relevant to the student by topic and difficulty. In a common use case, these would be questions and answers used and created during exercises in the learning course.

A typical session with QALO exercise can be described by a repetitive scenario:

1. A student makes a request for QALO (within an educational system). Usually, she does so during her "home" online learning session, either as a "starting" activity (when she wants to discover what to learn next) or "finishing" activity (when she wants to reaffirm her newly gained knowledge). It is generally expected that this scenario especially occurs prior to exams or seminaries, where students might be tested. Thus, more than one student is usually working with the system at a time.
2. The QALO selector picks a suitable QALO for the student. Its aim is to maintain an effective allocation of crowd power and to avoid the situation when many QALOs are rated by insufficient number of students. The QALO selector is described in more detail below.



Question 1430:

Which architectural styles do we recognize in software engineering? (List at least 3 of them)

Provided answer:

client-server, layers

Estimate correctness ratio of the provided answer

Fig. 1. A screenshot of the QALO interface. Using the slider, a student expresses her estimation of the correctness of the answer. After clicking the "Rate" button the estimation is stored and feedback information (the "crowd truth" correctness of the answer) is displayed.

3. The QALO is displayed to the user (see Fig. 1), consisting of a question (e.g., "Which architectural styles do we recognize in software engineering? List at least 3.") and a provided answer (e.g., "client-server, layers"). The student reviews the QALO and decides to what extent the provided answer is correct, i.e., she provides QALO correctness estimation. She expresses this by moving the marker on a slider between two extremes: incorrect (internally represented by 0) and correct (internally represented by 1). In our example, if the correct answer to the question is a list of three items and the provided answer lists only two correct options, user may

move the marker to two thirds (from left) of the slider's width, indicating the correctness estimation of the provided answer.

4. After submitting her estimation, the student is presented with the current global correctness estimation for the answer (the "crowd truth").
5. The global correctness estimation for the QALO gets updated to include student's rating. It is defined as the *average of all individual correctness estimations of this QALO* (a value between 0 and 1).
6. If the QALO received *sufficient feedback* from students, it is excluded from further processing. For the sake of experimentation (in order to acquire uniform data set) we defined this by a constant number of rating actions needed for one QALO. For the practical use though, we would rather use a dynamic metric to determine, whether the current crowd answer can be considered close to definitive. Such metric could take into account, for example, the variance of the crowd answers and exclude a QALO from the process earlier, if the variance drops under certain margin.

Computation of the global correctness for particular answer as estimated by the crowd presents core of our method. As a real values between 0 and 1, the crowd answers may be used "as is" (e.g., for feedback to students), but to give them nominal interpretation, we discretize them further into three possible values: *correct*, *incorrect* or *unknown* using two parameters: t and ε. The t (threshold) splits the correctness interval into two areas designating two possible values: *correct* and *incorrect*. The crowd answer is then determined according to which interval its real value falls. The second parameter (ε) adds a third possible value: *unknown* by inserting an "uncertainty interval" around the t value (rendering the values that fall into it as unknown), resulting in intervals $\langle 0, t-\varepsilon\rangle$, $(t-\varepsilon, t+\varepsilon)$ and $\langle t+\varepsilon, 1\rangle$ for incorrect, unknown and correct estimations, respectively. We have experimented with different values of t and ε to yield the best results.

Our method requires that a certain minimum number of students is attending the course and participates in QALO correctness estimation. In order to acquire valid global correctness estimations, the number of participating students must be equal to number needed in worst-case scenario from the *sufficient feedback per QALO* point of view (in our experiments, defined by a constant). To provide feedback to students for their estimations, the requirement is even lower: units of previous feedback actions are required – the student is always informed, how many of his peers evaluated the QALO before her and can take the feedback with adequate seriousness. Practically, in a scenario when students review potential exam questions, the motivation to interact with the QALO content is high (which was also shown in our experiments) and is therefore no problem to provide feedback in most cases.

It is important to note that our method is independent of the semantics of the QALOs and it is "portable" to any educational course where a sufficient number of QALOs is available. The questions and student answers should also be simple and test small pieces of knowledge for smoother and more controllable process.

Considering the computation of crowd answer as a valid estimation of correct answer, the process of assigning QALOs to students could be completely random if an unlimited crowd power is at hand. However, we expect that in many cases, the number of student ratings required for whole dataset exceeds the available force that the student crowd is willing to offer.

Therefore we devised the *QALO Selector* – a routine for QALO picking, executed upon each QALO request. It aims to complete the evaluation of a particular QALO in a relatively short time by assigning it to students frequently so it receives the sufficient student feedback faster. The basic heuristic to do this is to assign the QALO that has not yet been validated sufficiently but has most of the validations already done. This way, the crowd force is used effectively, leaving only a minimum number of partially validated (i.e., unusable) QALOs. Keeping this "working frame" (a set of partially validated QALOs) narrow is, however, contradictory to other requirements:

1. A QALO must be validated by different students; and
2. A single student has a need for topical diversity or adaptation to her knowledge within QALO she rates.

A student motivation to participate might drop, if she encounters the same question or even QALO. Occasional repeating of the same QALO to the same student in a short time is exploitable in many ways (speeding up the evaluation, testing the student's consistency) but due to the possible loss of motivation, we avoided it, so the student encounters each QALO only once. The same question can be encountered more times, but only after student passes a certain number of other questions.

4 Evaluation: A Real-world Experiment in Class

In order to evaluate our method, we have conducted a real-world experiment in a setting of software engineering course lectured at the Slovak University of Technology in Bratislava. Over the period of two weeks, students were free to pull, read, consider and validate QALOs, which were assembled from the questions from last term's tests and respective answers provided by last term's students. Based on obtained correctness estimations we computed average correctness estimation for each QALO. We compared the results with a gold standard – correctness estimations of QALOs provided by teachers.

Hypothesis. The correctness estimations of answers in question-answer learning objects (QALOs), obtained by student crowd using our method, are the same as teacher's correctness assignments to these QALOs. This, we measure through *accuracy* (i.e., ratio of correct crowd answers to all its answers) and *dropout* (i.e., ratio of cases where crowd reached the *unknown* answer to all cases). We conducted the experiment for multiple settings of parameters t and ε. We expected the optimal t value to be 0.5.

Environment and context. The data collection spanned over two weeks around the mid-term of the software engineering course. The course consists of weekly lectures and exercises and also comprises supplementary online learning materials within the educational system ALEF [12]. The same system provided the platform for our method for this experiment. During the course, students also undertake weekly mini-exams. These exams comprise similar or identical questions as those used in this experiment and therefore we expected a natural interest from the student side to participate in the QALO validation.

Participants. Overall, 142 students (of *Principles of software engineering* course) participated in the experiment (out of 162 students who enrolled in this course). Participation was completely voluntary. We considered no prior knowledge about the domain expertise of the participating students.

Data. We have used 200 questions, each with 20 answers to construct the initial QALO set (thus comprising 4,000 QALOs). The answers in QALOs used were taken from real exams (commenced year earlier), so we could utilize the existing teacher correctness evaluations in the experiment. According to course syllabus, each QALO has been assigned with week, when its respective topics were discussed and students were asked only those already covered by lectures. Overall, 9,939 QALO correctness estimations were collected. 479 QALOs were provided with equal or more than 16 correctness estimations (our threshold for sufficient student feedback; we used threshold equal to 16 at which further estimation would change the crowd answer only marginally in worst case).

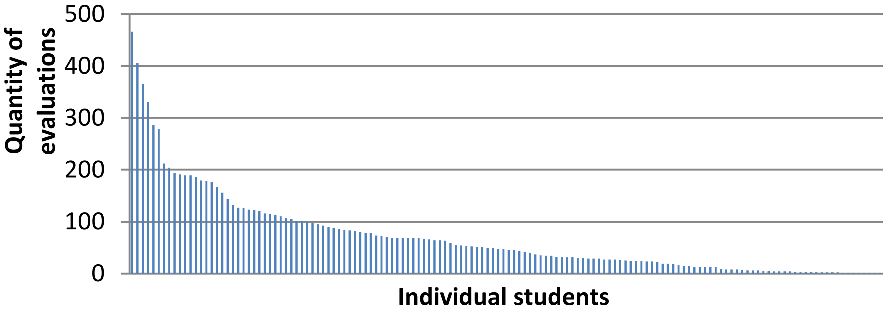


Fig 2. Quantities of evaluations provided by individual students

In average 70 QALO evaluations were collected per student, however, students greatly differed by quantity of estimations they delivered: while few of them evaluated hundreds of QALOs (top user delivered 466), many solved only units. Such distribution follows power law (see Fig. 2). Especially, the "best performers" can be accounted to different motivations for participants: besides the educational motivation (to learn and to test one's knowledge), students were also motivated by few extra points to their course assessments and also by *gamification* mechanisms (e.g., ranking among other users) present in the used educational system.

Yet more interesting "non-uniformity" we observed in the collected data was the indication of a *tendency of the students to consider incorrect question answers as correct answers*. From all QALOs considered in the experiment, 65 % had answers marked as correct by the teacher. However, 79 % of all student evaluations (those with some tendency, i.e., those not equal to 0.5, which was also the default value) were positive about the correctness (see the histogram in the Fig. 3 which illustrates this phenomenon). This suggests, that students tend to trust the answers created by other students. This also corresponds with our pedagogical experience: when students are unsure or wrong when answering questions, they at least *try* to make their answers *look correct*. Such answers then easily confuse other students (who validate QALOs) and "trick" them into belief, that they are correct.

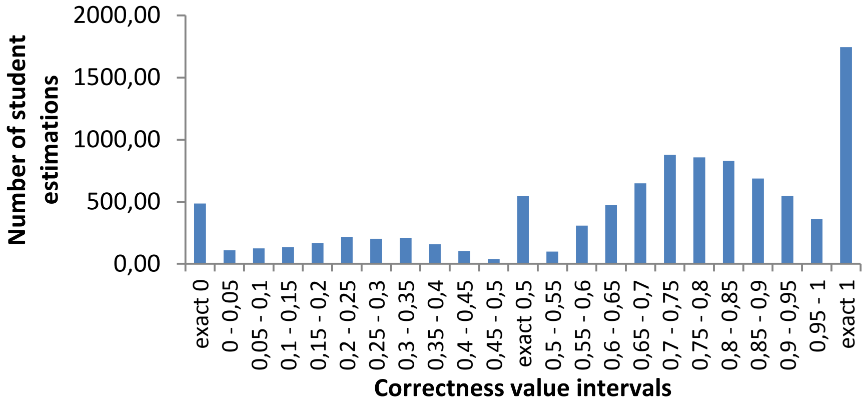


Fig 3. Distribution of all QALO student correctness estimations according to their values. Many students state extreme estimations and also tend to consider many answers correct.

Results. The overall real correctness estimations were computed for each QALO that received exactly or more than 16 estimations. After discretization (to three possible values, according to parameters t and ε), the estimations were compared to the reference set – answer correctness information assigned by teachers in the last term: correct or incorrect. The Table 1 shows the resulting accuracy of the method (the proportion of QALOs with correctly estimated correctness – both *correct* and *incorrect* – of provided answers in all QALOs) along with the percentage of *unknown* value cases.

Table 1. Method's *accuracy* and *unknown cases percentage* (in parentheses) for different parameter setups (t – correctness threshold, ε – uncertainty factor).

t	ε = 0.0	ε = 0.05	ε = 0.10
0.55	79.60 (0.0)	83.52 (12.44)	86.88 (20.40)
0.60	82.59 (0.0)	86.44 (11.94)	88.97 (27.86)
0.65	**84.58 (0.0)**	**87.06 (15.42)**	**91.55 (29.35)**
0.70	80.10 (0.0)	88.55 (17.41)	88.89 (37.31)
0.75	79.10 (0.0)	79.62 (21.89)	86.92 (46.77)

On the contrary to the initially expected correctness threshold $t = 0.5$, as best parameter configuration, the correctness threshold $t = 0.65$ has emerged. With no uncertainty interval ($\varepsilon = 0$), the method was rendered promisingly 84.58 % accurate. With introduction of the uncertainty interval we even see an increased accuracy in crowd's decision, though dropout (unknown cases) percentages are significant too.

We consider the results very promising as reasonably high accuracy of student answer correctness can be obtained via validation performed by students themselves. The accuracy increases over 90 % if approximately 30 % of QALOs are omitted. For our purpose even a higher "loss" is affordable as our primary goal is to support learning corpora enrichment while reducing teacher's efforts and not necessarily to get correct validations for all provided answers.

One could naturally expect the threshold t = 0.5 to be optimal. Instead for student crowd a higher value (t = 0.65) was observed as better. We account this to the significantly often occurring "trusting student phenomenon" described above, where students validate incorrect answers as correct ones. The ratio of false positive and false negative crowd correctness estimations supports this assumption. With t = 0.5 (i.e., the expected "normal conditions"), 91 % of false crowd answers were false positives (i.e., cases when students wrongly stated that an answer is correct).

5 Discussion and Conclusions

We have presented a method for student-crowd-based acquisition of correctness information of answers to questions in the context of learning course. It benefits from collective "wisdom" of a group of lay students. Functioning also as a didactical tool, our method enables to re-use existing question-answer learning objects and to give feedback to answer creators. Since our method is not constrained in terms of the domain of the course, it is portable and applicable to any course, where question-answer learning objects are available. In our experiments, our findings were as follows:

1. In an *implicit collaboration* scenario, while undertaking their learning sessions with our method, students as a crowd are able to validate learning objects with quality comparable to their teachers.
2. An interesting effect that unfavorably skewed the student crowd answers was the "trusting student" phenomenon, where students in significant numbers evaluated incorrect question answers as correct.
3. The interactive character of the QALO correctness evaluation exercise, combined with gamification incentives successfully motivated students to participate – one student evaluated averagely 70 QALOs.

There are also several possible improvements of the base method (focusing solely on estimating correct answer without learning goals that in real case scenario are always present) that are a subject of our future work. First, it is an introduction of smarter identification of sufficient student feedback on the QALO – if the correctness estimates show only a little variance from the start, the QALO might be excluded from the process earlier and spare some student actions for other QALOs. Secondly, to further increase the speed and accuracy of our method for validating answers by student crowd we want to introduce a model of individual influence of students in average correctness estimation. The influence would source from student's level of knowledge in the course domain and would be acquired by means common in academic courses such as previous exam results.

Acknowledgements. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grants No. VG1/0675/11 and No. VG1/0971/11, the Slovak Research and Development Agency under the contract No. APVV-0208-10 and is the partial result of the Research & Development Operational Programme for the project SMART II, ITMS 26240120029, co-funded by the ERDF.

6 References

1. Adamic, L. A., Zhang, J., Bakshy, E., Ackerman, M. S.: Knowledge sharing and yahoo answers: everyone knows something. In: Proc. of the 17th int. conf. on World Wide Web (WWW '08), pp. 665-674. ACM, New York (2008)
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proc. of the int. conf. on Web search and web data mining - WSDM ’08, pp. 183-194. ACM, New York (2008)
3. Chen, B. C., Dasgupta, A., Wang, X., Yang, J.: Vote calibration in community question-answering systems. In Proc. of the 35th int. ACM SIGIR conf. on Research and development in information retrieval (SIGIR '12), pp. 781-790. ACM, New York (2012)
4. Downes, S.: E-learning 2.0. eLearn magazine 2005, 10 (1). ACM, New York (2005)
5. IEEE LTS: Draft Standard for Learning Object Metadata. IEEE Standard 1484.12.1. IEEE, (2002) [retrieved March 2013]
6. Ghauth, K. I., Abdullah, N. A.: The Effect of Incorporating Good Learners' Ratings in e-Learning Content-based Recommender System. Educational Technology & Society, 14 (2), 248–257 (2011)
7. Golovchinsky, G., Qvarfordt, P., Pickens, J.: Collaborative information seeking. Information Seeking Support Systems. (2008)
8. Kidd, J., O'Shea, P., Baker, P., Kaufman, J. Allen, D.: Student-authored Wikibooks: Textbooks of the Future?. In: McFerrin, K. et al. (eds.) Proc. of Society for Information Technology & Teacher Education Int. Conf. 2008, pp. 2644-2647. Chesapeake, VA: AACE (2008)
9. Lawson, M.: Berners-Lee on the read/write web. BBC, Technology (2005) http://news.bbc.co.uk/1/hi/technology/4132752.stm [accessed 31/03/2013].
10. Quinn, A. J., Bederson B. B.: Human computation: a survey and taxonomy of a growing field. In: Proc. of the 2011 annual conf. on Human factors in computing systems (CHI '11), pp. 1403–1412. ACM, New York (2011)
11. Stahl, G., Koschmann, T., Suthers, D.: Computer-supported collaborative learning: An historical perspective. In: Sawyer, R. K. (eds.) Cambridge handbook of the learning sciences, pp. 409–426. Cambridge, UK: Cambridge University Press (2006)
12. Šimko, M., Barla, M., Bieliková, M.: ALEF: A Framework for Adaptive Web-based Learning 2.0. In: Reynolds, N., Turcsányi S. M. (eds.) KCKS 2010, IFIP Advances in Information and Communication Technology, Vol. 324, pp. 367–378. Springer (2010)
13. Šimko, M., Barla, M., Mihál, V., Unčík, M., Bieliková, M.: Supporting Collaborative Web-Based Education via Annotations. In: Proc. of W. Conf. on Educational Multimedia, Hypermedia & Telecommunications, ED-MEDIA 2011, pp. 2576–2585. AACE (2011)
14. Wheeler, S., Yeomans, P., Wheeler, D.: The good, the bad and the wiki: Evaluating student-generated content for collaborative learning. British Journal of Educational Technology, 39(6), 987-995 (2008)