Proceedings from the First International Workshop on Educational Knowledge Management (EKM 2014),

Linköping, November 24, 2014

Edited by

Inaya Lahoud and Lars Ahrenberg

Linköping Electronic Conference Proceedings, No. 104
ISBN: 978-91-7519-218-5
ISSN: 1650-3686
eISSN: 1650-3740
URL: http://www.ep.liu.se/ecp_home/index.en.aspx?issue=104
Linköping University Electronic Press
Linköping, Sweden, 2014

Preface

The First International Workshop on Educational Knowledge Management (EKM 2014) was organized as a satellite event to the 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW), held in Linköping, Sweden, November 24th to 28th, 2014. The workshop was held in the afternoon of the 24th of November.

The interest in Knowledge Engineering and Knowledge Management for the educational domain has been growing in recent years. This can be seen in the series of conferences organized by the International Educational Data Mining Society and in papers discussing the role of knowledge management in higher education. As education is increasingly occurring online or in educational software, resulting in an explosion of data, new techniques are being developed and tested, aiming for instance to improve educational effectiveness, determine the key factors to the success of educational training, support basic research on learning, or manage educational training by satisfying the needs of a community, local industry, or professional development.

The event aimed to bring together researchers, academic and professional leaders, consultants, and practitioners, from the domain of semantic web, data mining, machine learning, linked data, and natural language processing to discuss and share experiences in the educational area.

In the call of papers, we invited submissions reporting original research related to any problem of managing and exploring information in the educational area in schools, colleges, universities, and other academic or professional learning institutions.

A non-exhaustive list of topics for the workshop included the following:

- Educational knowledge management and ontology
- Educational knowledge acquisition, extraction, reuse
- Natural language processing to improve educational effectiveness
- Providing feedback to teachers and other stakeholders generated from EKM methods
- Generic frameworks, methods and approaches for EKM
- Mining the results of educational research
- Educational process mining

Two members of the following program committee have reviewed our submissions:

Halil Ibrahim Bulbul, Gazi University, Turkey
Stefan Dietze, Leibniz University Hanover, Germany
Catherine Faron Zucker, University of Nice Sophia Antipolis, France
Davide Fossati, Carnegie Mellon University, US
Ayako Hoshino, NEC Knowledge Discovery Research Laboratories, NEC Corp, Japan
Roger Nkambou, Université du Québec, Montréal
David Monticolo, University of Lorraine, France
Nuno Pombo, University of Beira Interior, Portugal

Following the reviewers' recommendations, three full papers were accepted for presentation at the workshop and inclusion into the workshop proceedings volume; subject to revisions as recommended by the reviewers. We invite participants and readers to enjoy the workshop program.

Workshop co-chairs

Inaya Lahoud
Lars Ahrenberg

Website: http://www.ida.liu.se/conferences/IWEKM14/

Acknowledgements:

Financial support for the organization of the workshop has come in part from the EKAW sponsors, and in part from the University of Galatasaray (http://gsu.edu.tr/)

Table of Contents

Discovering Educational Potential Embedded in Community Question Answering

Ivan Srba and Mária Bieliková

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia

{name.surname}@stuba.sk

Abstract. Community Question Answering (CQA) systems, such as Yahoo! Answers or Quora, are mostly perceived and studied from their primary knowledge sharing perspective. In spite of that, CQA systems have also a potential to become an effective means for people to acquire new knowledge. In the present, we can witness initial efforts on taking advance of this secondary perspective as CQA concepts have been recently applied in several educational applications. However, these educational systems take quite different approaches to the transition of CQA concepts from the open web to organizational and educational environments. One of possible reasons is that it is not clear which aspects of the question answering process have the best potential for knowledge acquisition. Therefore, we performed a study with a dataset obtained from CQA system Stack Overflow, in which we identified three scenarios that positively leads to improvement of users' expertise. The results of the study provide us with a better understanding how CQA systems can be applied as nontraditional learning environments.

Keywords: Community Question Answering, Informal Learning, Knowledge Sharing, Stack Overflow

1 Introduction

Information retrieval systems (e.g. web search engines or digital libraries) provide users with powerful tools how to identify valuable information and knowledge in the great information space of the current web. However, these systems are effective only if required information has been already codified and made publicly available. Moreover, standard information retrieval systems are not always successful to answer subjective, non-factual and context-aware queries, such as recommendations, advices or complex domain-specific problems. Current possibilities of the web allow us to employ supplementary sources of information to overcome these problems. These non-traditional sources of knowledge are often based on collective intelligence. Concept of collective intelligence [4] refers to shared knowledge which emerges from common collaboration of a community of users that share common practice, interests or goals. Collective intelligence is present in many popular web systems, such as forums, social networking sites or wikis (particularly Wikipedia is considered as one of the best examples of collective intelligence). In recent years, the new forms of web systems based on collective intelligence have appeared. One of them is Community Question Answering.

Community Question Answering (CQA) is a service where people can seek information by asking questions and share knowledge by providing answers on questions asked by other users. The well-known examples of CQA systems include Yahoo! Answers, Quora or domain-specific Stack Overflow where users concern with questions related to computer science and programming.

Millions of answered questions have already proved the successful concepts of CQA. Therefore, CQA systems became the interesting subject of many research studies. However, in spite of the increasing research effort in recent years, the beneficial effects of CQA systems has not been fully determined for intra-organizational environments yet, such as for educational or business organizations. Especially the educational domain, where students are quite often struggling with various problems related to a learning process, can benefit from the positive effects of the question answering process. Nevertheless, the full potential of CQA systems in the educational domain has not been discovered yet. In particular, the current state of research does not provide a clear answer which aspects of the question answering process contribute to successful knowledge acquisition.

The main goal of this paper is to describe an educational potential of CQA systems, specifically how CQA concepts can be successfully applied in educational and organization-wide environments, such as at universities. We determine this potential by means of an exploratory study with a dataset from CQA system Stack Overflow. The results from the study consist of an identification of three knowledge acquisition scenarios that lead to a positive improvement of users' expertise.

The paper is structured as follows: we describe a related work on CQA systems in Section 2; Section 3 contains a study aimed to discover a learning potential of CQA systems; finally, conclusions and future work are proposed in Section 4.

2 CQA Systems in an Educational Context

A typical process of question answering in CQA systems consists of several steps. At first, an asker posts a question by formulating a description of his or her problem. In addition, it is usually necessary to select an appropriate question's topic (a category or a set of related tags). Afterwards, answerers can provide their answer-candidates on the posted question, vote for the most appropriate answers and thus help all users, who are involved in the question answering process, to identify answers with the highest quality. The asker can finish the answering process by selecting the best answer, which satisfies his or her information needs best, and consequently the question is marked as answered and moved to the archive of solved questions.

The significant part of state-of-the-art research on CQA systems studies the question answering process from a perspective of knowledge sharing. In this perspective, the goal of a CQA system is to harness knowledge of a whole community to provide the most suitable answers on recently posted questions in the shortest possible time. Besides this primary view, we stress that there is also another interesting perspective how CQA systems can be perceived. People can gain new knowledge by reading, asking and also by answering questions. In addition, they are able to perceive different perspectives on a problem by discussions attached to a question or related answers. Thereby, it is very natural to speak about this knowledge acquisition as a special kind of informal learning in CQA systems.

One of possible ways, how to provide students with innovative learning environments, is to adopt concepts of Web 2.0 knowledge sharing applications, such as wikis, forums, social networking sites or content creation tools [8]. Due to the learning potential of CQA systems, it is natural that they have been also already applied as a model in proposals of educational systems. However, we are aware of only a few studies concerned with an employment of CQA systems within the educational domain.

At first, OpenStudy [5] is a large-scale open social learning environment which promotes knowledge sharing through Web 2.0 technologies. It adapts concepts of many social applications, such as CQA systems, online forums, real-time chats and social networking sites. In the present, OpenStudy (http://openstudy.com) has more than 1 million users that come from 160 countries.

While OpenStudy involves a great open community of students, remaining educational CQA systems are focused on smaller groups of students who enrolled for the same class. Piazza (http://piazza.com) is a learning system that is directly inspired by CQA. It is an online platform which offers a refined question answering process along with key features for effective course collaboration. It supports student-to-student collaboration as well as student-to-teacher discussions.

Green Dolphin is another social question answering board designed to support collaborative learning of programming [1]. Green Dolphin automatically identifies students who are experts on a particular topic. Afterwards, students can ask other students or directly experts identified by the system. This recommendation ensures high-quality answers while minimizing teachers overload. Students are awarded by points for asking and answering questions. Afterwards, they can use the earned points to direct their own questions to the recommended experts or teachers. The important concept of Green Dolphin is that new questions are postponed and hidden from teachers for some time, so students have enough time to provide answers by themselves. Only if a question cannot be answered in the given time, a teacher is notified and asked to take a participation on students' collaboration.

Authors in [2] investigated how to meet the needs of students and instructors while providing them with possibilities of social tools. Classroom Salon was proposed in which collaboration takes place in small groups termed Salons. Each Salon can be open to the entire community or only to a particular group. Students can use these Salons to post various documents, such as a piece of text, a program or a series of questions. Additionally, it is possible to annotate or vote on these documents. Authors in the set of experiments confirmed that their system based on principles of social tools, CQA being one of them, can successfully replace and outperform traditional online forums.

On the basis of performed analyzes, we identified an open problem that directs our further research. The analyzed educational systems are quite diverse in an application of CQA concepts for the purpose of learning. One of the main reasons is that their potential to elicit students' participation and collaborative learning is only to be discovered and thus it has not been well-documented yet [1]. The existing systems assume that an active participation on the question answering process (e.g. asking a question, providing an answer or a comment, searching for solved questions) leads somehow to an improvement of users' knowledge about particular topics. However, it is not very clear which specific situations in the question answering process have a potential for participating users to acquire new knowledge.

We suppose that especially collaborative scenarios in which users participate on question answering with other more experienced users can lead to improvement in their knowledge. With tangible and well-described identification of this kind of scenarios, it will be possible to optimally utilize the learning potential embedded in the question answering process. Naturally, the identification of these scenarios plays even more important role in the educational domain. It will suggest how to propose more effective learning environments as well as methods for adaptive collaboration support that will guide students towards these scenarios. There are many ways how to achieve it, e.g. we can provide students with a recommendation to take a participation on questions in which some of the identified collaborative scenarios will occur with a high probability.

3 Determination of CQA Learning Potential

In order to answer the identified open problem, we conducted a study aimed to discover in which scenarios and how users improve their knowledge in CQA systems. For this purpose, we analyzed a dataset from Stack Overflow and investigated relations between users' interactions and their level of expertise. Following the standard process in CQA systems, we identified three scenarios which occur during the question answering process and which can lead to improvement in expertise of participating users. Consequently, we stated a hypothesis associated with each of these scenarios. More specifically, we suppose that users improve their expertise:

- When an asker receives a high quality answer (Scenario 1).

  H1: The number of asked questions containing high quality answers (with a quality that exceeds asker's current expertise) is positively associated with improvement in asker's expertise.

- When an answerer utilizes additional external sources, which supplement his or her own expertise, to provide better answers (Scenario 2).

  H2: The number of provided high quality answers (with a quality that exceeds current answer's expertise) is positively associated with answerer's expertise.

- When a user participates (i.e. provides an additional answer-candidate while the best answer has not been selected yet) on a question answering in which other high quality answers has been previously posted (Scenario 3).

  H3: The number of provided answers to questions that has previously received other high quality answers (with a quality that exceeds current answerer expertise) is positively associated with answerer's expertise.

3.1 Dataset Description

We employed a dataset from Stack Overflow in our analyses. Stack Overflow is a domain-specific open CQA system in which users concern with various questions about programming. Stack Overflow is not applied directly in the educational domain, nevertheless our goal is to determine the learning potential embedded in the question answering process, which appears independently on the particular domain where CQA system is applied in. In addition, its large dataset is publicly available and thus it represents the best option for the purpose of our study.

Stack Overflow was founded in July 2008 and so far it contains more than 8.1M of questions and more than 14M of answers. It is considered as the fastest CQA system with median time to the first answer lower than 10 minutes. Several voting mechanisms are available: each question and answer can be voted up or down; each comment can be marked as a useful one. In addition, a question can be starred as a favorite one.

The anonymized dataset from Stack Overflow as well as from all other CQA systems built on the top of Stack Exchange infrastructure is published regularly under Creative Commons license and contains all publicly available data (http://blog.stackexchange.com/category/cc-wiki-dump/). The main part of dataset consists of users' posts (i.e. questions, answers and comments), their revision history and metadata (i.e. tags, votes, received badges).

At first, we analyzed the evolution of Stack Overflow during its history. More specifically, we focused on the number of new questions. Figure 1 shows that the amount of new questions is growing from the very beginning and is quite stable recently. Therefore, we decided to limit our further analyses at the content posted between January and December 2013. During this one year-long interval, users asked about 2.3M of questions and provided more than 3.4M of answers. About 91.9% of all questions received at least one answer and in addition, asker selected the best answer in the case of 48.7% of all questions. Quite impressive is also the speed of the question answering process as 67.5% of questions received the first answer in 10 minutes after being posted.

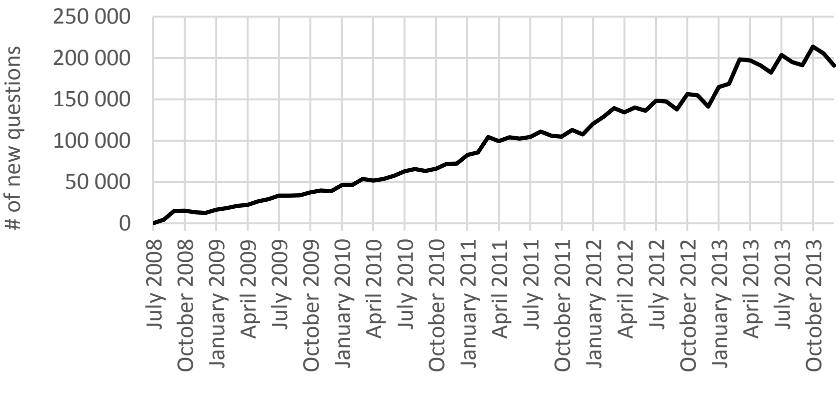


Fig. 1. Number of new questions posted each month during Stack Overflow history.

3.2 Estimation of Answer Quality and User Topical Expertise

Stack Overflow dataset provides information only about low-level interactions (e.g. creation of a new question). In order to confirm the stated hypotheses, it was necessary to process users' actions to more abstract variables.

Topic. To model question and answer topic, we decided to employ tags assigned by users at the time of the question creation. Corresponding answers inherit topics from the question they are related to. Some other studies supplement user-assigned tags by latent topics identified by methods such as Latent Dirichlet Allocation (LDA), e.g. [7]. However, this approach is important especially when questions are assigned only to one category (e.g. in Yahoo! Answer) and thus a little information is known about a question's topic. On the other side, in Stack Overflow a question can have unlimited number of tags (each question has 2.95 tags in average in the selected dataset).

Answer Quality. Determination of answer quality in CQA systems is quite a difficult and challenging task because most of questions can be subjectively oriented and the context of each question and corresponding asker is unique and sometimes not well known. Several different techniques have been used to determine answer quality in CQA systems so far. Some authors relied on an assumption than an answer is of high quality if it was selected by an asker as the best answer. Authors in [6] pointed out a problem of this discrete taxonomy. The best answer selected by an asker can be chosen subjectively and thus it can be biased while there can be also other high-quality answers. Another possibility how to achieve more precise answers' quality estimation is manual evaluation what is, however, really time consuming and thus it is not possible to apply it on great datasets. Finally, Stack Overflow, as well as other CQA systems, provides users with a voting mechanism which determines answers' quality by means of score. Score refers to a sum of positive and negative votes on questions as well as answers. Positive votes on an answer determine its correctness and giving a negative vote means that the answer is incomplete, incorrect or biased. More than 70% of all votes are created within 48 hours after the corresponding post was created. Therefore, we can consider score as a quite precise estimation of answer quality which is mostly independent on time when the answer was posted.

By analyses of Stack Overflow dataset, we found out that values of answer score follow a typical long-tail distribution and thus any calculation with them can be significantly skewed. For this reason, we performed two transformations on the score:

1. A logarithmic transformation to eliminate the undesired power law distribution.
2. A min-max normalization which transforms positive values to interval $\langle 0,1 \rangle$ and negative values to interval $\langle -1,0 \rangle$. This normalization is important because negative votes in the selected dataset represent only 11% of all votes and thus we emphasized their weight.

Consequently, we propose to calculate quality of an answer A_i as a normalized score (S') increased in the case when the answer was selected as the best answer (BA). We empirically set the influence of asker's best answer vote/selection as a value 0.1 that corresponds approximately to 5 standard up-votes by users other than the asker.

$$quality(A_i) = S'(A_i) + BA(A_i) \quad (1)$$

$$BA(A_i) = \begin{cases} 0.1 & if\ A_i\ was\ selected\ as\ the\ best\ answer \\ 0 & otherwise \end{cases} \quad (2)$$

User topical expertise. In the most of research works, an estimation of users' expertise is based on their previous contributions. Zhang et al. [10] proposed a very simple metric named z-score which describes how many answers and questions a user previously posted in the CQA system: $Z\_score = (a - q)/\sqrt{(a + q)}$, where a represents a number of posted answers and q is a number of asked questions. The assumption is that true experts only provide answers and do not ask any questions. Bouguessa et al. [3] proposed a probabilistic model that is based on another simple metric named InDegree. This metric represents only a number of best answers provided by the particular user.

In our approach, we similarly utilize the previous user activity. However in order to derive user expertise more precisely, we estimate it from quality of previously provided answers. More specifically, user expertise is calculated as a cumulative average of answers' quality. In addition, it is calculated separately for each topic (tag). It means that users can have different values of expertise for various topics and thus we are able to model real user expertise even more precisely. Similarly as all approaches based on previous user activities, also our approach has a drawback that we are not able to calculate user expertise for users with no or minimal activity in the CQA system. It means, that we have an estimation of user expertise with a high degree of uncertainty at the beginning and with the following answers, we are able to refine the expertise level and estimate it with a significantly higher confidence.

3.3 Evaluation of Knowledge Acquisition Scenarios

We employed the estimation of answer quality and user topical expertise to evaluate the stated hypotheses. At first, we identified all occurrences of three analyzed scenarios for each user and separately for each tag he or she provided answers on. Consequently, we calculated two numbers each time when user expertise has changed (i.e. when the particular user provided a new answer):

1. Number of scenario occurrences in the time interval from the analyzed point in time until user's last activity in the selected dataset.
2. Relative change of user's expertise between the analyzed point in time and user's last activity in the dataset.

Time intervals with a low-confidence estimation of user expertise were omitted because they can bring an undesired distortion to the evaluation. Finally, the calculated relative change of expertise was averaged across all users and tags with the respect to the number of scenario occurrences (see Figure 2).

The obtained results pointed out that all three scenarios are positively related with user expertise. We can see a logarithmic distribution for scenario H2 (when an answerer provides a high quality answer) and H3 (when an answerer provides an answer besides other high quality answers). It means that these two scenarios provide the best potential to boost knowledge acquisition. On the other hand, influence of scenario H1 (when an asker receives a high quality answer) follows rather a linear trend.

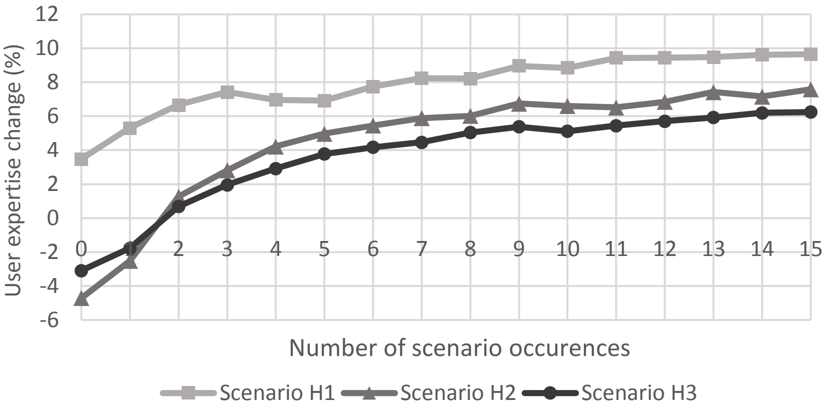


Fig. 2. Relation between average change of user topical expertise and the number of scenario occurrences in the observed time intervals.

To evaluate the relation between variables numerically, we used Kendall tau rank correlation coefficient as an evaluation metric. Kendall tau correlation was selected due to the non-linear character of evaluated relations as well as its better robustness in comparison with other standard correlation coefficients (i.e. Spearman rank correlation and Pearson correlation). All scenarios achieved strong correlations with a high significance (see Table 1).

Table 1. Overview of the achieved Kendall tau correlations for the evaluated hypotheses.

Hypothesis	*Correlation*	*P-value*
H1 (when an asker receives a high quality answer)	0.9167	< 0.001
H2 (when an answerer provides a high quality answer)	0.9333	< 0.001
H3 (when an answerer provides an answer besides other high quality answers)	0.9830	< 0.001

4 Conclusion and Future Work

Without any doubt, fast and reliable availability of information is the crucial step to success in business as well as in academia. We consider Community Question Answering (CQA) systems as promising knowledge management systems that appeared only recently. While the primary goal of CQA systems is to provide high quality answers on new questions in the shortest possible time, we recognized their potential to become novel learning systems that supplement a formal educational process.

On the basis of the dataset from Stack Overflow, which is one of the most popular CQA systems, we identified three scenarios that positively contribute to development of users' knowledge. The results of our study give us a deeper insight into the learning potential embedded in the question answering process. More specifically, the identified knowledge acquisitions scenarios can be utilized by researchers to propose more effective educational environments or methods for collaboration support. For example, during a personalized recommendation of questions to students, we can promote those recommendations that will direct students into the identified scenarios.

In our additional work, we took into consideration the knowledge acquisitions scenarios together with organizational and educational specifics to propose an educational organization-wide CQA system named Askalot. Askalot demonstrates the learning potential of CQA system as a complementary tool dedicated to knowledge sharing besides other educational systems ALEF [9] and PopCorm [8], which were developed and are used at our faculty.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11 and by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

1. Aritajati, C., Narayanan, N.H.: Facilitating students' collaboration and learning in a question and answer system. Proc. of the 2013 conf. on Computer supported cooperative work companion - CSCW '13. pp. 101–106. ACM Press, New York, USA (2013).
2. Barr, J., Gunawardena, A.: Classroom salon: a tool for social collaboration. Proc. of the 43rd ACM technical symposium on Computer Science Education - SIGCSE '12. pp. 197–202. ACM Press, New York, USA (2012).
3. Bouguessa, M., Dumoulin, B., Wang, S.: Identifying authoritative actors in question-answering forums. Proc. of the 14th ACM SIGKDD international conf. on Knowledge discovery and data mining - KDD 08. pp. 866–874. ACM Press, New York, USA (2008).
4. Lévy, P.: Collective Intelligence: Mankind's Emerging World in Cyberspace. Perseus Books, Cambridge, MA, USA (1997).
5. Ram, A., Ai, H., Ram, P., Sahay, S.: Open social learning communities. Proc. of the International Conf. on Web Intelligence, Mining and Semantics - WIMS'11. ACM Press, New York, USA (2011).
6. Sakai, T., Ishikawa, D., Kando, N., Seki, Y., Kuriyama, K., Lin, C.-Y.: Using graded-relevance metrics for evaluating community QA answer selection. Proc. of the fourth ACM international conf. on Web search and data mining - WSDM '11. pp. 187–196. ACM Press, New York, USA (2011).
7. Szpektor, I., Maarek, Y., Pelleg, D.: When Relevance is not Enough : Promoting Diversity and Freshness in Personalized Question Recommendation. Proc. of the 22nd international conf. on World Wide Web. pp. 1249–1259 (2013).
8. Srba I., Bieliková, M.: Designing Learning Environments Based on Collaborative Content Creation. In Proceeding of Workshop on Collaborative Technologies for Working and Learning, pp. 49–53, CEUR, (2013).
9. Šimko, M., Barla, M., Bieliková, M.: ALEF: A Framework for Adaptive Web-Based Learning 2.0. In Proc. of IFIP Advances in Information and Communication Technology, Vol. 324, pp. 367-378. Springer, Heidelberg (2010).
10. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities. Proc. of the 16th international conf. on World Wide Web - WWW '07. pp. 221–230. ACM Press, New York, USA (2007).

An agent based approach to annotate ideas during creativity challenges in an engineering school of innovation

Davy Monticolo[1], Chang T. K. [2], Inaya Lahoud[3]
[1]Lorraine University, Innovative Processes Research Team (ERPI Laboratory), 8 rue Bastien Lepage, 54 000 Nancy France {davy.monticolo@univ-lorraine.fr}
[2] Department of Information System, National University of Singapore, Singapore {changtk@nus.edu.sg}
[3] Department of Computer Engineering, University of Galatasaray, Istanbul, Turkey {inaya.lahoud@gmail.com}

Abstract. This paper presents a multi agent system architecture used to store, annotate and reused knowledge from the ideas created by the students of the University of Lorraine during the creativity workshop called "48 hours to generate ideas".

Keywords. Multi Agent System, Collaborative Creativity Process, Knowledge Annotation

1 Introduction

The engineering school of innovation (ENSGSI) of the University of Lorraine organizes every year a creativity workshop called "48 hours to generate ideas[1]" for the students. This challenge is international since there are twelve other universities all over the word, which participate to this event. During forty-eight hours the students will apply creativity method to generate hundreds of ideas in order to solve an industrial problem. The aim of creativity workshops is to develop creativity supported by a collaborative process where students groups generate an eco-system of ideas, evaluate them and make them evolve. The creativity collaborative process involves creativity participants, creativity experts and stakeholders. Creativity experts who are the professors of the engineering school of innovation, animate and lead the creativity process taking into account the skills and evolution of the groups of creativity participants, combining different creativity methods and installing a sequence of divergence/convergence phases helping the growth of the ideas eco-system. However,

[1] http://www.48h-innovation-maker.com/
[2] http://curbcreativepracticebootcamp.eventbrite.com/

during these workshops or challenges, ideas are usually written on post-its and then enriched and structured by means of paper forms that are difficult to exploit by both participants and creativity experts.

With the development of information and communication technologies, different innovation platforms are proposed. Several research studies on how to choose a modern interactive tool (or even a collaboration ecology), which simulates the creativity collaborations, have been recently presented in [1][2][3]. Another educational creativity workshop is the annual Creative Practice Bootcamp2 held in Nashville, TN, where students learn how to apply methods like Brainstorming [4], Brainpurge [5] or Brainwriting [6]. All these systems provide supports for distant and asynchronous innovation. However, even if their main aim is to favor the innovation process, they limit their contribution to feeding participants with information (from other participants, from different content providers, etc.) based on crowdsourcing principles. They are neither installing nor supporting the creativity process itself as creativity experts do during the creativity workshops. Others research works use a multi agent system approach to support the creativity process. In these works we observe two categories; the multi agent systems used to simulate the cognitive mechanism of the creative people like in [7], [8] or the multi agent systems which aim to manage a creativity support system like [9], [10], [11]. The multi agent system approach allows to realize complex tasks like annotate, evaluate ideas and also to take into account of the social features of the creative people like the way they cooperate and the information that they need to fulfill an activity [12].

In this paper we present an agents based approach called CIMAS (Creativity Ideas Managed by Agent System) to support the creativity process all along the challenge “48 hours to generate ideas”. In the next section, we will explain the interest to use a multi agent system approach to support the creativity process. In the following section, we will describe the architecture of CIMAS and the annotation process used by the agents to manage the knowledge inside ideas.

2 Overview of CIMAS

In this section we describe the concepts and the architecture of the CIMAS system. There are three types of users; the stakeholders the creativity quizmasters and the creative participants. The aim objectives of the CIMAS system are:

- To help creative participant to annotate and evaluate their ideas and to research others similar ideas;
- To assist stakeholders to search relevant ideas by using different points of views;
- To assist creativity quizmaster by providing indicators and ideas trends from all the creative participant groups.

[2] http://curbcreativepracticebootcamp.eventbrite.com/

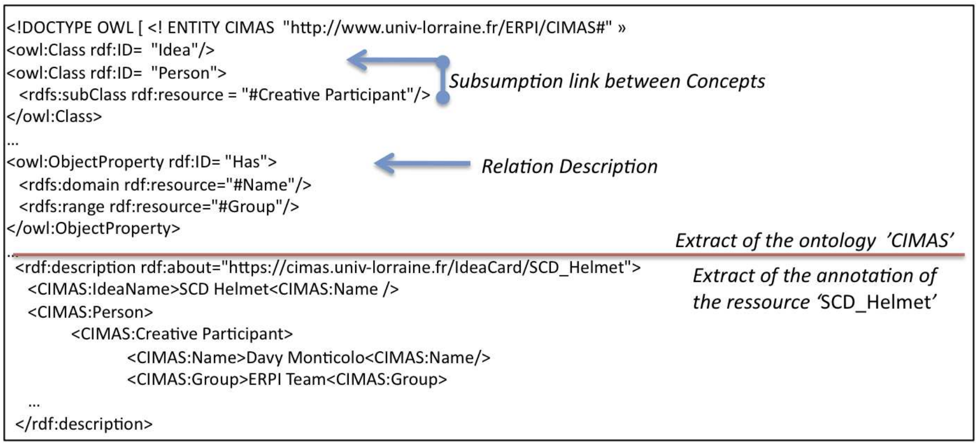
Agents dedicated to the ideas annotation

Doyle explains in [13] that “annotated environments containing explanations of the purpose and uses of spaces and activities allow agents to quickly become intelligent actors in those spaces”. The ideas landscape represents the annotated environment built by the agents. Indeed the CIMAS agents have to annotate each idea, sketch or post-it in order to handle and exploit this information.

The Semantic Web [14] represents a set of languages which facilitate the annotation of web resources. By using RDF language of the Semantic Web, we can describe the context and the content of an idea even if the idea is a text, a sketch or a video. Compared to the Web, the ideas have more delimited context. We can easily define who the creators are, the type of content, when the idea was created. Thus an ontological approach is conceivable to describe ideas. There are already several ontologies aimed at the annotation of ideas, such as ideas “ideaontology” [15], or “Idea Management ontology” [16]. Ideaontology is dedicated to the evaluation of the idea and use mono criteria methods to evaluate an idea. The second ontology is based on four groups of concepts; the concepts related to describe the origin of the idea, the concepts relative to describe the idea, the concepts which describe the innovative part of the idea (impact of the idea, target, feasibility, etc.), and the object (evolution of the idea, the process to develop it, etc.). In CIMAS we have built an ontology of concepts relating to description (types, use cases, etc.) and to contexts (creator, trust, evaluation, related project, etc.). The CIMAS ontology is formalized with OWL lite [17] which is related to provide a conceptual model to describe ideas and which the resources are defined separately.

The Figure 1 shows an extract of the CIMAS ontology and an example of annotation with literal and conceptual properties.

The CIMAS system does not lead directly with the web resources but with their annotation to support the ideas information management. Thus the CIMAS ontology represents a conceptual structure used by the agents to annotate ideas, to organize and research them.



```
<!DOCTYPE OWL [ <! ENTITY CIMAS  "http://www.univ-lorraine.fr/ERPI/CIMAS#" »
<owl:Class rdf:ID=  "Idea"/>
<owl:Class rdf:ID=  "Person">
  <rdfs:subClass rdf:resource = "#Creative Participant"/>
</owl:Class>
…
<owl:ObjectProperty rdf:ID= "Has">
  <rdfs:domain rdf:resource="#Name"/>
  <rdfs:range rdf:resource="#Group"/>
</owl:ObjectProperty>
…
 <rdf:description rdf:about="https://cimas.univ-lorraine.fr/IdeaCard/SCD_Helmet">
   <CIMAS:IdeaName>SCD Helmet<CIMAS:Name />
   <CIMAS:Person>
          <CIMAS:Creative Participant>
                    <CIMAS:Name>Davy Monticolo<CIMAS:Name/>
                    <CIMAS:Group>ERPI Team<CIMAS:Group>
   …
 </rdf:description>
```

Subsumption link between Concepts

Relation Description

Extract of the ontology ’CIMAS’

*Extract of the annotation of the ressource ‘*SCD_Helmet*’*

Fig. 1. Extract of the CIMAS ontology and annotation example.

Architecture

A Multi Agent System is a network of agents that work together in a cooperative way to solve problems that would be generally difficult to solve for any individual agent. Information Agents are a part of intelligent agents [18],[19]. Klusch made a list of the services that a multi-agent system can offer in a information management approach [20]:

- Search, acquire, analyse and classify information coming from various information sources;
- Give information to human and computing networks once usable knowledge is ready to be consulted;
- Negotiate on information integration or exclusion into the system;
- Give explanation to the quality and reliability related to the integrated information;
- Learn progressively all along the information management process;

The proposed approach to design a MAS is based on an organizational approach like the A.G.R model used in AALAADIN [21], OperettA [22] and methodologies like TROPOS [23] or RIOCC [24]. Thus the CIMAS architecture is viewed as a human society in term or role, skill and relationships.

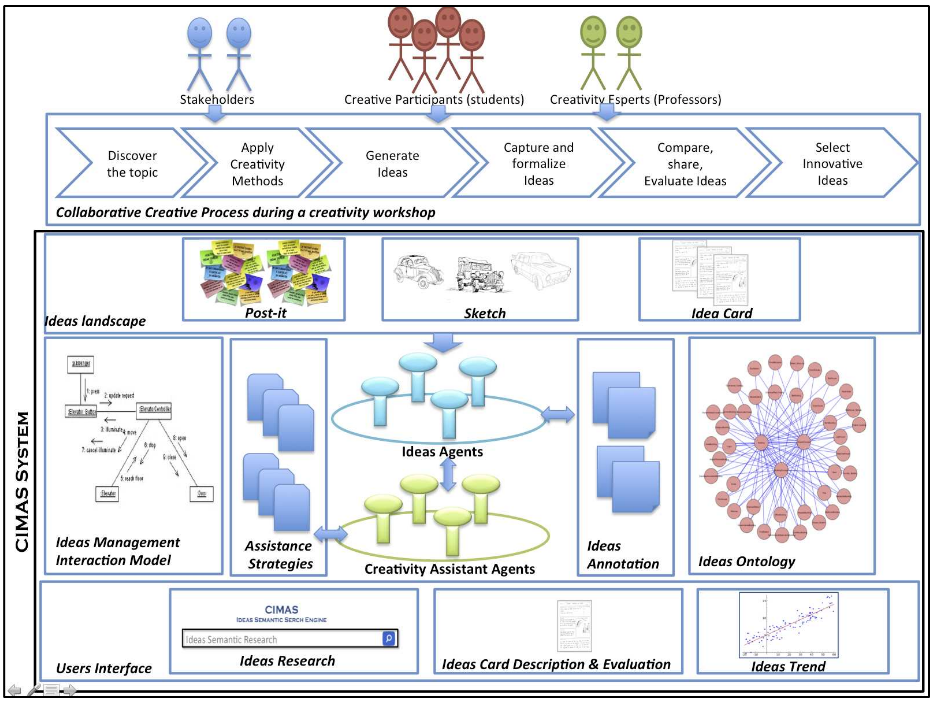

Stakeholders
Creative Participants (students)
Creativity Esperts (Professors)
Discover the topic
Apply Creativity Methods
Generate Ideas
Capture and formalize Ideas
Compare, share, Evaluate Ideas
Select Innovative Ideas
Collaborative Creative Process during a creativity workshop
CIMAS SYSTEM
Ideas landscape
Post-it
Sketch
Idea Card
Ideas Management Interaction Model
Assistance Strategies
Ideas Agents
Creativity Assistant Agents
Ideas Annotation
Ideas Ontology
Users Interface
CIMAS
Ideas Semantic Serch Engine
Ideas Semantic Research
Ideas Research
Ideas Card Description & Evaluation
Ideas Trend

Fig. 2. CIMAS architecture

The main objective of the CIMAS system is to manage ideas coming from different information sources (post-its, texts, sketches). The CIMAS system is based on three layers (Figure 2):

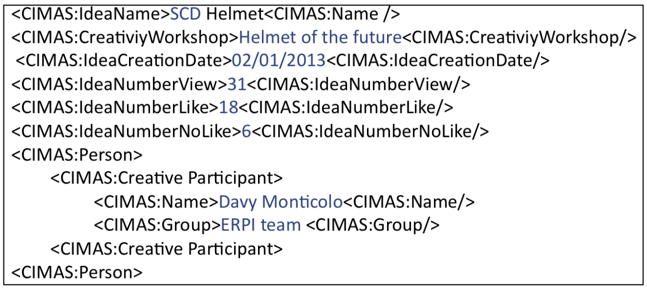
- The ideas landscape where users insert their ideas (text, sketch or post-its) in the system by using forms;
- The agents layer where the management of ideas is executed;
- The Interface layer where users can research an idea, consult an idea card or display the ideas trend inside a creativity workshop.

In this paper we describe the agents layer.

Ideas Agents Society in CIMAS

From the architecture analysis we can specify the two different agents' societies, the interactions between agents and the mechanisms they use to handle the annotations of ideas.

The Ideas Agents Society is dedicated to annotate the idea according to the ideas ontology. More explicitly, the agents use the structure of the ontology to annotate ideas. There are three Ideas agents, one by type of creativity workshop. There is one agent according to each type of content (post-it, sketch and idea card). The annotation of an idea is composed of a list of tags (Figure 3), which describe its creation (creator, creation date, team, creativity workshop) and its evolution (Number of views, number of "likes", etc.).

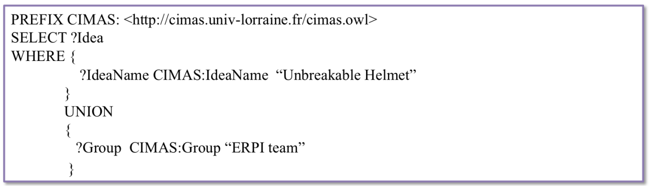


```
<CIMAS:IdeaName>SCD Helmet<CIMAS:Name />
<CIMAS:CreativiyWorkshop>Helmet of the future<CIMAS:CreativiyWorkshop/>
 <CIMAS:IdeaCreationDate>02/01/2013<CIMAS:IdeaCreationDate/>
<CIMAS:IdeaNumberView>31<CIMAS:IdeaNumberView/>
<CIMAS:IdeaNumberLike>18<CIMAS:IdeaNumberLike/>
<CIMAS:IdeaNumberNoLike>6<CIMAS:IdeaNumberNoLike/>
<CIMAS:Person>
      <CIMAS:Creative Participant>
            <CIMAS:Name>Davy Monticolo<CIMAS:Name/>
            <CIMAS:Group>ERPI team <CIMAS:Group/>
      <CIMAS:Creative Participant>
<CIMAS:Person>
```

Fig. 3. Annotation of an idea

The Ideas agents also build the result of the semantic researches when users enter keywords in the CIMAS search engine (Figure 2). They use two mechanisms; the first one is to built SPARQL requests [25] and the second is to calculate the semantic similarity between two ideas.

The first method is used to research the exact name of an idea, a creator or a group. Figure 4 shows the SPARQL request to search an idea which is called "Unbreakable helmet" created by the group "ERPI team".



```
PREFIX CIMAS: <http://cimas.univ-lorraine.fr/cimas.owl>
SELECT ?Idea
WHERE {
           ?IdeaName CIMAS:IdeaName  “Unbreakable Helmet”
         }
         UNION
         {
          ?Group  CIMAS:Group “ERPI team”
          }
```

Fig. 4. SPARQL request built by the Ideas Agents

The second method consists of calculate the similarity between two concepts in order to identify the similar concepts or the closed concepts. The method is based on the calculus of the semantic distances between two concepts in the RDF models embedded annotations. For example in the CIMAS ontology we have the following property:

$$[Post-it] \rightarrow (Creator) \rightarrow [CreativeParticipant]$$

The method will provide results such as:

$$[Sketch] \rightarrow (Creator) \rightarrow [Student]$$
$$[IdeaCard] \rightarrow (Creator) \rightarrow [\mathrm{Pr}\,ofessor]$$

To research the concepts, which are close semantically, the agents use the distance of Rada [26] counting the number of arcs on the shorter path between two terms (t1 and t2) (formula 1). By using this distance we can define the distance between two RDF triplets as the sum of the distances between: two relations, two concepts in first argument (domain) and two concepts in second argument (range) (formula 2).

$$dist(t1,t2) = length(t1,lest(t1,t2)) + length(t2,lest(t1,t2)) \quad (1)$$

$$\begin{aligned} dist(triple1,triple2) &= dist(domain(triple1),domain(triple2)) \\ &+dist(predicate(triple1),predicate(triple2)) \\ &+dist(range(triple1),range(triple2)) \end{aligned} \quad (2)$$

The algorithm gives a number between 0 and 1. The closer the number is to 1; the closer are the concepts semantically. A semantic research on all the RDF annotations on the word “Helmet” provides the results shown in Table 1:

Concepts	*Similarity index*
Headdress	0,632
Hard Hat	0,452
Crash Helmet	0,678
Bandore	0,128
Crown	0,321
Hat Head Protector	0,521
Safety Helmet	0,862

Table1. Research for the word “Helmet” in the Ideas Annotations

The Ideas agents will propose the three best results of the research i.e. the results are “Headdress”, “Had Hat” and “Crash Helmet” for the previous example.

The Society of Creative Assistant Agents in CIMAS

The Creativity Assistant Agents (CAA) interact with the users through the three following interfaces:

- The semantic research engine where they will send the elements of the request to the Ideas agent;
- The Ideas Card visualization and Evaluation. With this interface the CA presents the different idea cards and allow the users to add a comment or a mention “like” or “not like”;
- The Ideas Trend interface. This interface is a scatter chart showing the different ideas themes which are emergent in the workshop.

There are three different CA agents by creativity workshop. Each agent manages one type of interface; the ideas research interface, the ideas cards description & evaluation interface and the ideas trend interface.

3 Conclusion

This paper presents the architecture of a multi agent system dedicated to the ideas management during the creativity workshop “48 hours to generate ideas” organized by the engineering school of innovation of the University of Lorraine. The system uses the semantic web language and an idea ontology to build, research and exploit ideas annotations. The next work of this project will be to make the agents pro active, i.e. to allow the agents to inform the different type of users (Creative participants/students, creativity experts/professors and stakeholders/industrial partners) all along the workshop about the trend of ideas, or if a new idea is similar to another. Another perspective will be to evaluate the whole ideas generated during the creativity workshops and to propose actions to reuse them.

References

1. Smeaton A. F., Lee H., Foley C., Mc Givney S., “Collaborative Video Searching on a Tabletop”, *Multimedia Systems Journal*, vol. 12, n° 4, p. 375-391, 2006.
2. Morris M., Lombardo J., Wigdor D., “WeSearch : supporting collaborative search and sensemaking on a tabletop display”, *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW’10, ACM, New York, NY, USA, p. 401-410
3. Loke S., Ling S., “Analyzing Observable Behaviours of Device Ecology Workflows”, *In Proceedings of the 6th International Conference on Enterprise Information Systems*, p. 78-83, 2004.
4. Dugosh K. L., Paulus P. B., “Cognitive and social comparison processes in brainstorming”, *Journal of Experimental Social Psychology*, vol. 41, n° 3, p. 313 - 320, 2005.
5. VanGundy, A. B. (2008). 101 activities for teaching creativity and problem solving. John Wiley & Sons.
6. Isaksen S., Dorval K., Treffin D., *The Act of Creation, 2012*
7. Siddhartha B., Stellan O, Social creativity as a function of agent cognition and network properties: A computer model, in the international journal of social network, Volume 32, Issue 4, October 2010, Pages 263 – 278
8. Edmonds, E. A., Candy, L., Jones, R., & Soufi, B. (1994). Support for collaborative design: Agents and emergence. *Communications of the ACM,37*(7), 41-47.

9. Saunders, R. (2002). Curious design agents and artificial creativity.
10. López-Ortega, O. (2013). Computer-assisted creativity: Emulation of cognitive processes on a multi-agent system. *Expert Systems with Applications*, *40*(9), 3459-3470.
11. Lawson, B. (2005). Oracles, draughtsmen, and agents: the nature of knowledge and creativity in design and the role of IT. *Automation in construction*, *14*(3), 383-391.
12. Monticolo D., Mihaita S., Darwich H., Hilaire V., "A meta-model for knowledge configuration management to support collaborative engineering", in Computers in Industry, Vol5, P222-256, 2014
13. Doyle, Hayes-Roth, Agents in Annotated Worlds, Proc. Autonomous Agents, ACM, 1998, 173-180
14. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web. ",*Scientific american* 284.5 (2001): 28-37.
15. C. Riedl, S. Wagner, J. Leimeister (2010). Exploring Large Collections of Ideas in Collaborative Settings through Visualization, p0-6.
16. Westerski, A., & Iglesias, C. A. (2011). "The road from community ideas to organisational innovation: a life cycle survey of idea management systems Adam Westerski* and Carlos A. Iglesias, 7(4), 493-506.
17. McGuinness, Deborah L., and Frank Van Harmelen. "OWL web ontology language overview. " *W3C recommendation* 10.2004-03 (2004): 10.
18. S. Savarimuthu, M. Purvis, M. Purvis, "Tag-based Model for Knowledge Sharing in Agent Society, in the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), 10–15 May, 2009, Budapest, Hungary
19. Gandon, Fabien, Laurent Berthelot, and Rose Dieng-Kuntz. "A multi-agent platform for a corporate semantic web. " Proceedings of the first international joint conference on Autonomous agents and multiagent systems, AAMAS 2002.
20. M. Klusch, (Ed.): "Intelligent Information Agents: Agent-based Information Discovery and Management in the Internet", Springer, 1999.
21. Ferber, Gutknecht, "A meta-model for the analysis and design of organizations in multi-agent systems". IEEE Computer Society, Proc. 3rd ICMAS, 128-135, 1998.
22. Okouya, Daniel, and Virginia Dignum. "OperettA: a prototype tool for the design, analysis and development of multi-agent organizations". *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: demo papers*. International Foundation for Autonomous Agents and Multiagent Systems, 2008.
23. Bresciani, Paolo, et al. "Tropos: An agent-oriented software development methodology." *Autonomous Agents and Multi-Agent Systems* 8.3 (2004): 203-236.
24. D. Monticolo, I. Lahoud, and E. Bonjour. "Distributed knowledge extracted by a MAS using ontology alignment methods. " *Computer & Information Science (ICCIS), 2012 International Conference on*. Vol. 1. IEEE, 2012.
25. Prud'Hommeaux, Eric, and Andy Seaborne. "SPARQL query language for RDF " *W3C recommendation* 15 (2008).
26. Rada, R., Mili, H., Bicknell, E., Blettner, M., (1989) Development and Application of a Metric on Semantic Nets, IEEE Transaction on Systems, Man, and Cybernetics, vol. 19(1), pp. 17-30.

Generating Multiple Choice Questions From Ontologies: How Far Can We Go?

Tahani Alsubait, Bijan Parsia, and Uli Sattler

School of Computer Science, The University of Manchester, United Kingdom
{alsubait,bparsia,sattler}@cs.man.ac.uk

Abstract. Ontology-based Multiple Choice Question (MCQ) generation has a relatively short history. Many attempts have been carried out to develop methods to generate MCQs from ontologies. However, there is a still a need to understand the applicability of these methods in real educational settings. In this paper, we present an empirical evaluation of ontology-based MCQ generation. We examine the feasibility of applying ontology-based MCQ generation methods by educators with no prior experience in ontology building. The findings of this study show that this is feasible and can result in generating a reasonable number of educationally useful questions with good predictions about their difficulty levels.

1 Introduction

Automatic question generation is a relatively new field and dimension within the broad concept of technology-aided assessment. It potentially offers educators some help to ease the burden and reduce the cost of manual assessment construction. In terms of time, it is reported that assessment development requires considerable time [9, 18, 20]. In terms of cost, it is estimated that the cost of developing one question for a high-stake test can range from $1,500 to $2,000 [19]. More importantly, in terms of quality, as many as 40% of manually constructed questions can fail to perform as intended when used in assessments [11]. This has motivated many researchers to develop automated methods to generate assessment questions. Many of these methods have focused on the generation of Multiple Choice Questions (MCQs) which are typically used in high-stake testing.

Ontologies are machine-processable artefacts that can formally describe the main notions of a specific domain. Recent advancements in ontology languages and ontology tools have created an interest in ontology-based MCQ generation. Various attempts have been made to generate MCQs from ontologies [3, 17, 23, 24]. However, little is known about how useful these MCQs are when used in real educational settings.

We present a new case study for using ontology-based MCQ generation in real educational settings. The purpose of the study is to evaluate the feasibility of using ontology-based MCQ generators by instructors who have no prior experience in ontology development. Rather than using an existing ontology, we examine the case where a new ontology is required to be build from scratch. We estimate the cost of question generation including the cost of building a new ontology by a novice ontology developer (e.g., the instructor in this case). We also evaluate the quality of the generated questions and the accuracy of the generator's predictions about the difficulty of the generated questions.

2 Background

An MCQ item is an assessment tool which is made up of the following parts: 1) A stem, 2) A key and 3) Some distractors. The stem is a statement that introduces a problem to the student. The key is simply the correct answer. A number of incorrect, yet plausible, answers are called the distractors. The number of optimal distractors for MCQs remains debatable [12].

An ontology is a set of axioms which can be either terminological or assertional. Terminological axioms describe relationships between concepts. Assertional axioms describe relationships between individuals and concepts or between individuals and roles. Description Logics (DL) ontologies have formal semantics [7]. In this sense, an ontology is a logical theory which implies that implicit knowledge can be inferred from the explicitly stated knowledge. For a detailed overview of ontologies, the reader is referred to [7].

3 Related work

Prior to exploring the large body of research related to ontology-based question generation methods, we need to understand the basic/optional components associated with those methods. These components are:

3.1 Source preparation

Before being able to generate questions, a suitable source ontology must be prepared. Gavrilova et al. [10] present a 5-step strategy aimed at developing teaching ontologies. The stages are: 1) Glossary development, 2) Laddering, 3) Disintegration, 4) Categorisation and 5) Refinement. Sosnovsky et al. [21] present a case study for utilising the above 5-step strategy to develop an ontology for the domain of C programming.

3.2 Item generation

The next step of generation is to generate an assessment item or part of it (e.g., distractors) from the developed ontology. For example, Mitkov et al. [16] have developed an approach to automatically generate MCQs using NLP methods. They also make use of ontologies to generate distractors that are similar to the correct answer.

Zitko et al. [24] proposed templates and algorithms for automatic generation of MCQs from ontologies. They generate a random set of distractors for each questions. Papasalouros et al. [17] constrain the set of distractors to some neighbours of the correct answer in the ontology.

Williams [23] presents a prototype system for generating mathematical word problems from ontologies based on predefined logical patterns. The proposed method makes use of data properties in general ontologies. The data properties are used to replace certain place holders in the predefined patterns.

3.3 Characterisation

Some methods of ontology-based question generation vary the characteristics of the generated questions (e.g., their difficulty). For instance, Williams [23] proposes to vary the difficulty of mathematical problems by introducing/removing distractor numerical values and varying sentence complexity and length. Alsubait et al. [3, 2, 1] propose to vary the difficulty of MCQs by varying the similarity between the key and distractors.

3.4 Presentation

To enhance the readability of automatically generated questions, Williams [23] extends the use of SWAT[1] natural language tools to verbalise ontology terms which are used in the generated questions. For example, "has a height of" can be derived from the data property "has_height". Similarly, Papasalouros et al. [17] use simple natural language generation techniques to transform the generated questions into English sentences.

3.5 Post evaluation

Mitkov et al. [16] present an evaluation study of automatically generated MCQs in real testing settings. Item response theory (IRT) [15] has been used for the statistical analysis of students results. In particular, they study the following properties: (i) item difficulty, (ii) discrimination between good and poor students and (iii) usefulness of distractors. They also compare manual and automatic methods of MCQ generation and report that that automated methods perform better than manual methods of test items in terms of time without compromising quality.

In an earlier study [3], the authors evaluated a large set of multiple-choice questions which have been generated from three different ontologies. The evaluation was carried out using an automated solver which simulates a student trying to answer these questions. The use of the automated solver facilitated the evaluation of the large number of questions. The findings of this study show that it is feasible to control the difficulty of questions by varying the similarity between the key and distractors. A more recent study [5] in which the authors recruit a group of students in real testing settings confirms the results of the study carried earlier using the automated solver.

[1] http://swat.open.ac.uk/tools/

4 Ontology-based MCQ generation in practice

Introduction to Software Development in Java is a self-study course run by the School of Computer Science at the University of Manchester. It aims to ensure that students enrolled in Masters programs in the school have a thorough grasp of fundamental programming concepts in Java. Topics covered in this course include: object-oriented basics, imperative programming, classes, inheritance, exception handling, collections, stream and file I/O. The course material is delivered online via Moodle. As with any self-study course, students enrolled in this course need a series of self-assessments to guide them through their learning journey.

4.1 Materials and methods

Equipment description the following machine was used for the experiments in this paper: Intel Quad-core i7 2.4GHz processor, 4 GB 1333 MHz DDR3 RAM, running Mac OS X 10.7.5. In addition to the following software: OWL API v3.4.4 [14] and FaCT++ [22].

Building the ontology An ontology that covers the contents of the course has been built by an instructor who has an experience in Java but with no huge familiarity with materials of this course. In this case, the instructor had no prior experience in building ontologies. The online course material covers both fundamental concepts (i.e., terminological knowledge) and practical Java examples (i.e., procedural knowledge). Only terminological knowledge is suitable to be modelled in ontologies. This type of knowledge is typically a vital part of education in general and of assessment in particular. It is regarded as the basic level in Bloom's taxonomy of learning objectives [8]. The development of the ontology has gone through the following steps:

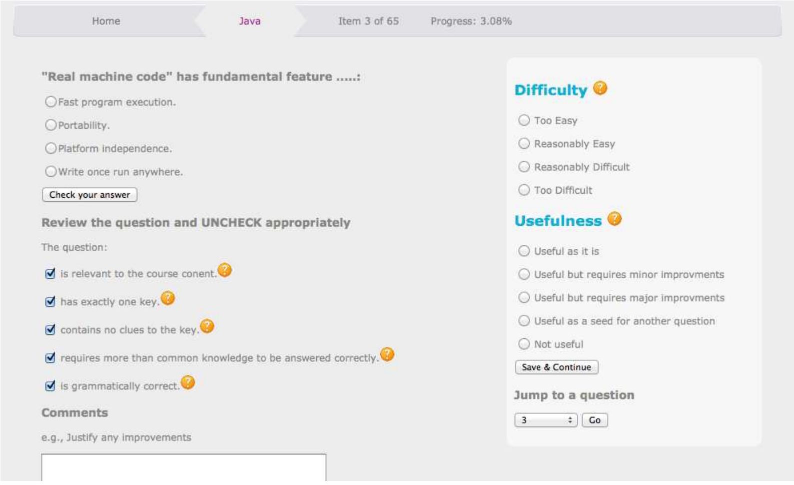
- The instructor has been introduced to basics of ontology development in an initial meeting which lasts for 2 hours. This included a brief hands-on tutorial on using *Protégé* 4 ontology editor. Further online materials [13] were forwarded to the instructor to familiarise herself on building and dealing with ontologies.
- The instructor built an initial version of the ontology. She went through the first 6 modules of the course, extracted and added to the ontology any encountered concepts and finally established links between the added concepts. This took a total of 10 hours and 15 minutes spread over 6 days. This has resulted in a total of 91 classes, 44 object properties and 315 axioms.
- A two-hours feedback session took place to highlight weak points in this version of the ontology. The instructor reported that, as the number of classes and relations increased, it got very hard to maintain the same level of understanding of the current state of the ontology.

- The second version of the ontology took 5.5 hours to build. The resulting ontology has a total of 91 classes, 38 object properties and 331 axioms. The main task was to restructure the ontology according to the received feedback. The decrease in the number of object properties is due to merging those object properties which had very similar meaning but different names. The increase in the number of axioms can be partially explained by the fact that the instructor was advised to assert negative facts in the ontology whenever and wherever possible. In addition, some concepts were re-categorised (e.g., declared as a subclass of another exiting class).
- To ensure that the ontology covers the main concepts of the domain, the instructor was advised to consult a glossary of Java-related terms which is part of the online course material. Adding new terms from the glossary in suitable positions in the ontology took a total of 10 hours over 4 days. The resulting ontology has a total of 319 classes, 107 object properties, 213 annotation assertion axioms and 513 logical axioms. The DL expressivity of the resulting ontology is $\mathcal{ALCHQ}$ which allows conjunctions, disjunctions, complements, universal restrictions, existential restrictions, qualified number restrictions and role hierarchies. For more information and examples, the reader is referred to [7].

Generating questions We follow the same question generation strategies described in [5] to generate multiple choice questions from ontologies. The first step of question generation is to compute the pairwise similarity for all the classes in the ontology using the similarity measures described in [6]. These similarity measures have been shown to be highly correlated with human similarity measurements [6]. The intuition behind using similarity measures as part of question generation is that varying the similarity between the key and distractors can make it possible to vary the difficulty of the generated questions [4]. In other words, increasing the difficulty to distinguish the correct answer among the given answers makes the question harder.

A total of 428 questions have been generated from the Java ontology. Then questions with less than 3 distractors have been excluded (resulting in 344 questions). Questions in which there is an overlap between the stem and the key have been filtered out (resulting in 264 questions). This step was necessary to ensure that there are no word clues in the stem that could make the correct answer too obivous. Previous attempts to generate MCQs from ontologies have identified this as a possible problem [5]. In this study, we filter out questions in which there is a shared word of more than three characters between the stem and key. This does not apply to questions in which the shared word is also present in the distractors. Finally, questions which can be described as redundant and that are not expected/recommended to appear in a single exam were manually excluded (e.g., two questions which have slightly different stems but the the same set of answers or vice versa). This step was carried out only to get a reasonable number of questions that can be reviewed in a limited time. The resulting questions are 65 questions in total. Among these are 25 easy questions and 40 difficult questions.

Reviewing questions Three reviewers have been asked to evaluate the 65 questions using the web interface shown in Figure 1. All the reviewers have experience in both the subject matter (i.e., programming in Java) and assessment construction. The reviewers have been randomly numbered as Reviewer 1, Reviewer 2 and Reviewer 3 with Reviewer 2 being the ontology developer. For each question, the reviewer is asked to first attempt to answer the question. Next, the reviewer is asked to rate the difficulty of the question by choosing one of the options 1) Too easy, 2) Reasonably easy, 3) Reasonably difficult and 4) Too difficult. Then the reviewer is asked to rate the usefulness of the question by choosing one of the options: (0) not useful at all, (1) useful as a seed for another question, (2) useful but requires major improvements, (3) useful but requires minor improvements or (4) useful as it is. In addition, the reviewer is asked to check whether the question adhere to 5 rules for constructing good MCQs. These rules were gathered from the qualitative analysis of previous reviewer comments in a previous evaluation study [5]. The rules are: R1) The question is relevant to the course content, R2) The question has exactly one key, R3) The question contains no clues to the key, R4) The question requires more than common knowledge to be answered correctly, and R5) The question is grammatically correct.



"Real machine code" has fundamental feature .....:

○ Fast program execution.
○ Portability.
○ Platform independence.
○ Write once run anywhere.

Review the question and UNCHECK appropriately

The question:

☑ is relevant to the course conent.
☑ has exactly one key.
☑ contains no clues to the key.
☑ requires more than common knowledge to be answered correctly.
☑ is grammatically correct.

Comments

e.g., Justify any improvements

Difficulty

○ Too Easy
○ Reasonably Easy
○ Reasonably Difficult
○ Too Difficult

Usefulness

○ Useful as it is
○ Useful but requires minor improvments
○ Useful but requires major improvments
○ Useful as a seed for another question
○ Not useful

Jump to a question

3

Fig. 1: The reviewing web-interface

4.2 Results and discussion

Total cost We report on the cost, in terms of time, of the three phases: 1) ontology building, 2) question generation and 3) question review. The ontology took around 25 hours to be built by an instructor who has no prior experience on ontology building and no huge familiarity with the course material used in this study. This cost could have been reduced with an appropriate experience in building ontologies and/or higher familiarity with course content. The generation of a total of 428 questions using the machine described above took around 8 hours including the time required to compute pair-wise similarities. Finally, Reviewers 1, 2 and 3 spent around 43 minutes, 141 minutes, and 56 minutes respectively. We exclude any question for which more than 15 minutes were spent. This indicates that the reviewer was interrupted during the review of that question. In addition, Reviewer 2 reported that she was taking side notes while reviewing each question. For this reason and for other reasons that could interrupt the reviewer, the cost of the reviewing phase should be regarded as a general indicator only.

In terms of cost, it is interesting to compare between two possible scenarios to generate MCQs. The first scenario is where the questions are manually constructed whereas the second scenario utilises ontology-based question generation strategies. The cost of manual generation is expected to be lower than the cost of developing a new ontology added to the cost of question generation and review. However, a few points should be taken into account here. First, in the second scenario, the ontology is expected to be re-used multiple times to generate different sets of questions. Second, the aim is to generate questions with highly accurate predictions about their pedagogical characteristics which has been shown to be possible in the second scenario. Third, no particular skills/creativity for MCQ construction are required when utilising ontology-based question generation strategies.

Usefulness of questions Figure 2 shows the number of questions rated by each reviewer as: not useful at all, useful as a seed for another question, useful but requires major improvements, useful but requires minor improvements, or useful as it is. As the figure indicates, a reasonable number of questions have been rated as useful by at least one reviewer. More precisely, 63 out of the 65 questions have been rated as either useful as it is or useful with minor improvements by at least one reviewer. And 50 questions have been rated as either useful as it is or useful with minor improvements by at least two reviewers. Finally, 24 questions belong to the same category as rated by all three reviewers. As a concrete example of a question that was rated useful by all 3 reviewers, we present the following question:

Q: ..... refers to ”A non-static member variable of a class.”:
(A) Loop variable
(B) Instance variable (Key)
(C) Reference variable
(D) Primitive variable

Quality of questions Adherence to the 5 rules for constructing goof MCQs indicates the quality of the generated questions. Figure 3 shows the number

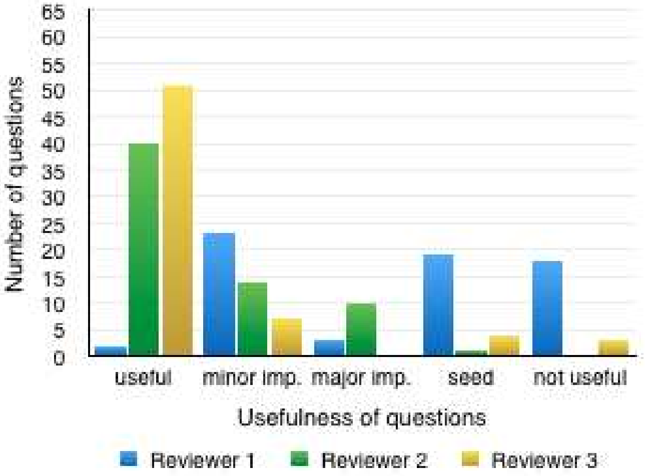


Fig. 2: Usefulness of questions according to reviewers evaluations

of questions adhering to each rule as evaluated by each reviewer. In general, a large number of questions have been found to adhere to Rules R1, R2 and R4. It can be noticed that only a few questions violate Rule R4 (i.e., no clues rule). Recall that a lexical filter has been applied to the generated questions to filter out questions with obvious word clues. This has resulted in filtering out 80 questions. This means that the lexical filter is needed to enhance the quality of the generated questions. The grammatical correctness rule (R5) was the only rule which got low ratings. According to reviewers' comments, this is mainly due to the lack of appropriate articles (i.e., the, a, an). Dealing with this issue and other presentation/verbalisation issues is part of future work.

Difficulty of questions according to reviewers' ratings Part of the objectives of this study is evaluate the accuracy of predictions made by the questions generation tool about the difficulty of each generated question. To do this, we compare difficulty estimations by each reviewer with tool's predictions. Recall that each reviewer was allowed to select from four options of different difficulty levels (too easy, reasonably easy, too difficult, reasonably difficult). This is to distinguish between acceptable and extreme levels of difficulty/easiness. However, tool's predictions can take only two values (easy or difficult). To study tool-to-reviewers agreements, we only consider the two general categories of difficulty. That is, the four categories of difficulty estimations by reviewers are collapsed into two categories only (easy and difficult). Figure 4 shows the number of questions for which there is an agreement between the tool and at least one, two or three reviewers. As the Figure shows, for a large number of questions (51 out of 65 questions) there has been an agreement between the tool and at least one reviewer. To understand the causes of disagreements, we further

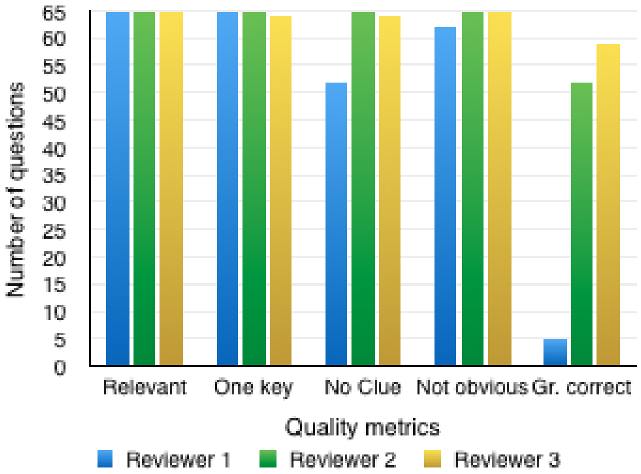


Fig. 3: Quality of questions according to reviewers' evaluations

categorise the agreements according to the difficulty of questions. Table 1 indicates that the degree of agreement is much higher with easy questions reaching 100% agreements with at least one reviewer. This could mean that the generated distractors for difficult questions were not plausible enough. This has been discussed with the ontology developer because we believe that better distractors could be generated by enriching the ontology. In particular, the ontology developer has indicated that many classes in the ontology have been assigned to a single superclass while they could possibly be assigned to multiple superclasses. Restructuring and enriching the ontology is expected to increase the ability of the tool to generate questions at certain levels of difficulty.

Table 1: Accuracy of difficulty predictions for easy and difficult questions

	At least 1 reviewer	At least 2 reviewers	At least 3 reviewers
Easy questions	100%	88%	52%
Difficult questions	65%	35%	2.5%
All questions	78.5%	55.4%	21.6%

Difficulty of questions according to reviewers' performance Each reviewer has attempted to solve each question as part of the reviewing process. Interestingly, non of the reviewers has answered all the questions correctly, including the ontology builder who answered 60 questions correctly. The first and third reviewers have correctly answered 55 and 59 questions respectively. This can have different possible explanations. For example, it could be possible that

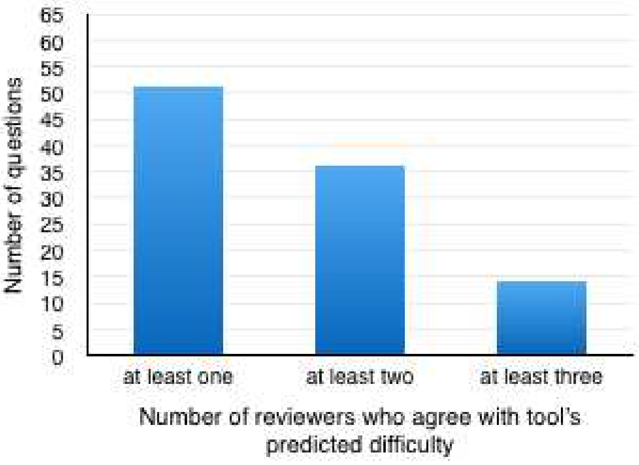


Fig. 4: Difficulty of questions according to reviewers' evaluations

the reviewer have picked a wrong answer by mistake while trying to pick the key. This has actually happened with the first reviewer who has reported this by leaving a comment on one question. Note also that the third reviewer has reported that in exactly one question there was more than one possible correct answer, see Figure 3. This means that if a reviewer picks an answer other than the one identified by the tool as the correct answer then his/her answer will not be recognised as correct. Figure 5 shows the number of questions answered correctly by at least one, two and three reviewers.

In exactly one question, none of the reviewers answered the question correctly, raising a question about the validity of this question as an assessment tool. The stem part of this question was "Which is the odd one out?". To required task to answer the question is to distinguish between the answers which have a common link (the distractors) and the answer which cannot be linked to the other answers (the key). Although all the reviewers have rated this question as "useful", we believe that it is too difficult and not necessarily very useful as an assessment item.

5 Conclusion and future research directions

We believe that ontology-based MCQ generation has proved to be a useful method for generating quality MCQs. The cost of generation is still considered to be high but is expected to be reduced over continuous uses of the same ontology.

As future work, we aim to administer the generated questions to a group of students in order to see how useful the questions are for the purposes of self-assessments. We also aim to add a verbaliser to the MCQ generator to enhance language accuracy. Finally, we believe that there is a potential in using the

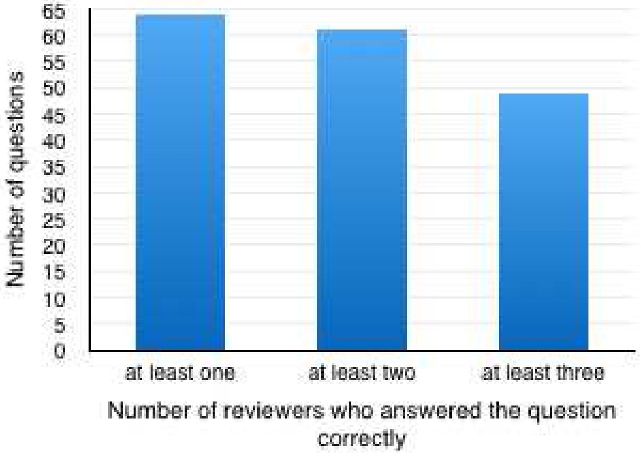


Fig. 5: Difficulty of questions according to reviewers performance

developed MCQ generation methods in application other than assessments. For example, we are interested in exploring the applicability of these methods for ontology evaluation and comprehension purposes.

References

1. T. Alsubait, B. Parsia, and U. Sattler. Automatic generation of analogy questions for student assessment: an ontology-based approach. In *ALT-C 2012 Conference Proceedings*, 2012.
2. T. Alsubait, B. Parsia, and U. Sattler. Mining ontologies for analogy questions: A similarity-based approach. In *OWLED*, 2012.
3. T. Alsubait, B. Parsia, and U. Sattler. Next generation of e-assessment: automatic generation of questions. *International Journal of Technology Enhanced Learning*, 4(3/4):156–171, 2012.
4. T. Alsubait, B. Parsia, and U. Sattler. A similarity-based theory of controlling mcq difficulty. In *Second International Conference on e-Learning and e-Technologies in Education (ICEEE)*, pages 283–288, 2013.
5. T. Alsubait, B. Parsia, and U. Sattler. Generating multiple choice questions from ontologies: Lessons learnt. In *The 11th OWL: Experiences and Directions Workshop (OWLED2014)*, 2014.
6. T. Alsubait, B. Parsia, and U. Sattler. Measuring similarity in ontologies: How bad is a cheap measure? In *27th Inernational Workshop on Description Logics (DL-2014)*, 2014.
7. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. (eds.) Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, second edition, 2007.
8. B. S. Bloom and D. R. Krathwohl. *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners. Handbook 1. Cognitive domain.* New York: Addison-Wesley, 1956.

9. B. G. Davis. *Tools for Teaching.* San Francisco, CA: Jossey-Bass, 2001.
10. T. Gavrilova, R. Farzan, and P. Brusilovsky. One practical algorithm of creating teaching ontologies. In *12th International Network-Based Education Conference NBE*, pages 29–37, 2005.
11. T. M. Haladyna. Developing and validating multiple-choice test items. *Hillsdale: Lawrence Erlbaum*, 1994.
12. T.M. Haladyna and S.M. Downing. How many options is enough for a multiple choice test item? *Educational & Psychological Measurement*, 53(4):999–1010, 1993.
13. M. Horridge. A practical guide to building OWL ontologies using Protégé 4 and CO-ODE tools, edition 1.3. http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/ [accessed: 18-04-2014], 2011.
14. M. Horridge and S. Bechhofer. The OWL API: A Java API for working with OWL 2 ontologies. In *In Proceedings of the 6th International Workshop on OWL: Experiences and Directions (OWLED)*, 2009.
15. M. Miller, R. Linn, and N. Gronlund. *Measurement and Assessment in Teaching, Tenth Edition.* Pearson, 2008.
16. R. Mitkov, L. An Ha, and N. Karamani. A computer-aided environment for generating multiple-choice test items.cambridge university press. *Natural Language Engineering*, 12(2):177–194, 2006.
17. A. Papasalouros, K. Kotis, and K. Kanaris. Automatic generation of multiple-choice questions from domain ontologies. In *IADIS e-Learning 2008 conference*, Amsterdam, 2008.
18. M. Paxton. A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education*, 25(2):109–119, 2001.
19. L. Rudner. *Elements of adaptive testing*, chapter Implementing the Graduate Management Admission Test computerized adaptive test, pages 151–165. New York, NY: Springer, 2010.
20. J. T. Sidick, G. V. Barrett, and D. Doverspike. Three-alternative multiple-choice tests: An attractive option. *Personnel Psychology*, 47:829–835, 1994.
21. S. Sosnovsky and T. Gavrilova. Development of educational ontology for C-Programming. In *Proceedings of the XI-th International Conference Knowledge-Dialogue-Solution, vol. 1, pp. 127132. FOI ITHEA*, 2006.
22. D. Tsarkov and I. Horrocks. FaCT++ description logic reasoner: System description. In *Proceedings of the 3rd International Joint Conference on Automated Reasoning (IJCAR)*, 2006.
23. S. Williams. Generating mathematical word problems. In *2011 AAAI Fall Symposium Series*, 2011.
24. B. Zitko, S. Stankov, M. Rosic, and A. Grubisic. Dynamic test generation over ontology-based knowledge representation in authoring shell. *Expert Systems with Applications: An International Journal*, 36(4):8185–8196, 2008.