Automatic Recommendation of Students' Answers for Error Mediation

Alexander Robert Kutzke[1], **Alexandre Direne**[1]

[1]Department of Informatics
Federal University of Paraná
Curitiba, PR, Brazil
Caixa Postal 19.081 – 81.531-980 – Curitiba – PR – Brazil

{**alexander,alexd**}**@inf.ufpr.br**

Abstract. *The problem of recommending answer records for error mediation in educational environments is introduced. Teachers face several difficulties in analysing student's incorrect answers due to the high workload it requires, even if these answers are digitally stored. In order to provide and facilitate the error mediation, this study describes an algorithm for answer recommendation. This algorithm generates automatic recommendations of relevant answers and questions to groups of similar students. Preliminary tests in a real application have indicated that the algorithm is capable of defining groups of students with actual similarities on their errors and of generating relevant recommendations. The main conclusions of the study are described and future works are pointed out.*

Resumo. *O problema da recomendação de registros de resposta para a mediação de erro em ambientes educacionais é introduzido. Professores enfrentam várias dificuldades em analisar as respostas incorretas de seus alunos, devido à alta carga de trabalho exigida, mesmo que essas respostas são armazenadas digitalmente. A fim de proporcionar e facilitar a mediação de erro, este estudo descreve um algoritmo para recomendação resposta. O algoritmo proposto gera recomendações automáticas de respostas e perguntas relevantes para grupos de estudantes considerados semelhantes. Testes preliminares em uma aplicação real indicaram que o algoritmo é capaz de definir grupos de estudantes com semelhanças reais sobre seus erros e de gerar recomendações pertinentes. As principais conclusões do estudo são descritas e trabalhos futuros são apontados.*

1. Introduction

Recent studies show the many contributions, to teachers and students, of the analysis of errors committed by students during the teaching and learning processes [Kutzke and Direne 2014, Isotani et al. 2011]. However, due to numerous difficulties (lack of time, many answers to analyse, etc.), students' errors end up being left out in educational environments. Thus, the error is not mediated.

Based on the cultural-historical psychology [Vygotsky 2012], mediating the error means making it the subject of an educational activity, i.e., problematizing it as part of a

development process [Kutzke and Direne 2014]. Thus, to promote the error mediation, it is necessary to provide access to relevant answer records for both teachers and students.

Recommender systems have gained space and attention in virtual education environments as they facilitate the access to different types of information [Manouselis et al. 2011]. However, the majority of the works is concentrated in retrieving learning resources. There are no studies about automatic recommendations of relevant answers for further teacher analysis or even for the student self-critique. In other words, there are no recommender systems aimed to promote error mediation.

This paper presents a new concept of answer recommendation as a support for students' errors mediation. Thus, teachers receive recommendations of potentially relevant answers for groups of similar students and, then, are able to analyse these responses with their students. An algorithm for automatic responses recommendation is described in detail.

The proposed algorithm was implemented in a web application to teach computer programming. Preliminary tests tend do indicate that the algorithm was capable of defining groups of students with real similarities on their errors and of generating relevant recommendations.

The remainder of this paper is organized as follows: Section 2 presents an overview of current work in the educational recommender systems field. The new conception of answer recommendation for error mediation is introduced in Section 3. Section 4 describes an algorithm for answer recommendations. Preliminary test results are presented in Section 5. Finally, Section 6 presents the main conclusions obtained with the study and poits out future work.

2. Related work

Most educational recommender systems only provide learning resource recommendations such as learning objects, contents to study, course enrollment suggestions etc. [Peña-Ayala 2014, Manouselis et al. 2011].

Vialardi et al. [Vialardi et al. 2011], for example, exhibit recommender systems to assist students in choosing courses for enrollment. Champaign and Cohen [Champaign and Cohen 2010], in turn, show recommendations of notes and comments made by the students during the interaction with learning objects for similar students.

A tool capable of providing pedagogical recommendations for teachers based on automatic educational data mining of online courses is presented by [Paiva et al. 2013]. The recommendations include resources available in the online environment, such as exercises and educational materials. Although the referred work presents an automatic form of evaluation of educational data to generate recommendations, it does not display an analysis of the students' errors. The work considers only simple data such as the number of unanswered questions, number of accesses to the system, student performance, etc.

Zapata-Gonzalez et al. [Zapata-Gonzalez et al. 2011] propose a hybrid recommender system to assist users in the search for learning objects. In the same direction, Durand et al. [Durand et al. 2011] present a recommender system of "learning paths" to help teachers to create courses in Learning Management Systems.

The authors of the present study are not aware of any recommender systems that provide answer recommendations aimed at error mediation for teachers or students. Few recommender system studies provide teacher support in students' error analysis. They usually ignore any analysis of relations among different error records. In other words, so far, it's been possible to verify an individualist error handling that lacks the study of error similarities among diferent students.

3. Answer recommendation and error mediation

When learning complex activities (like the ones taught in schools), the students are demanded to form *scientific concepts* [Vygotsky 2012]. These are mental structures that will regulate and reorganize the student's thinking and that will act reflexively. The development of scientific concepts is produced by the teaching that acts over the *Zone of Proximal Development* (ZPD) [Vygotsky 2012]. Thus, as the ZPD is in constant change, the formation of scientific concepts is always under improvement, which is constantly happening.

According to this concept, we can admit that, if there is no concept assimilation in its finished form, the error (incorrect answer of a student) is certainly part of the process of scientific concept formation. Thus, in order for the error not to become just one step of the empirical apprehension of reality, as in a trial-and-error learning, it must be *mediated*.

Students make mistakes during learning and several incorrect answers to different questions are presented by them. Therefore, assuming a system capable of storing all students' responses, in a short time, there will be many answer records to be analysed by the teacher. Certainly, many of these records are only slips and do not represent serious learning problems. However, other records can pose difficulties not only of one student, but also to groups of students with similar learning problems. With so many records, teachers face difficulties to find and select which of these answers are pedagogically relevant to mediate the students' errors.

The objective of the recommendation introduced by this work is to automatically find answer records and questions that may be of high pedagogical relevance to a specific group of students. Such recommendations allow the teacher to put more effort in error mediation and less in the search for relevant answers. In addition, the set of stored responses is supposedly large enough to prevent the teacher's observation of all answer records.

4. An algorithm for answer recommendation

This Section presents an algorithm for automatic recommendation of potentially relevant answers to student groups. This recommendation is made based only on the similarity relations among answers from different students. The framework presented in [Kutzke and Direne 2014] is used here to introduce the algorithm. The framework embodies a system in which all the answers submitted by students are digitally stored. The answer records are, then, compared in order to determine the similarities among them. Based on the comparisons, a similarity graph of answers is formed. In this graph, the vertices represent answer records and the edges connect similar answers. The weight of each edge represents the degree of similarity between two responses.

The algorithm consists of 3 steps: (A) groups of students with high similarity degree among their responses are found through the creation of a similarity graph of students, derived from the similarity graph of answers; (B) for each group, potentially relevant questions are defined (a question is said to be relevant for a group of students if the similarity among the group's answers to that question is greater than a given threshold); and (C) among the group's answers, the most representatives, i.e., those with the highest mean similarity among others responses, are selected.

Each of these three steps are detailed in the following sections.

4.1. Deriving the similarity graph of students

Let us assume the undirected similarity graph of answers A where each vertex $a_i \in V(A)$ represents an answer. An edge $(a_i, a_j) \in E(A)$ if and only if the answers a_i and a_j are similar[1]. The weight of each edge $(a_i, a_j) \in E(A)$ indicates the similarity degree between a_i and a_j, which is given by the similarity function[2] $sim_A(a_i, a_j)$. Let us also consider a set of students $S = \{s_1, s_2, s_3, \ldots, s_L\}$ and a set of questions $Q = \{q_1, q_2, q_3, \ldots, q_M\}$. Each answer a_i belongs to a student s_j and to a question q_k. The set of answers given by the student s_i to the question q_j is defined as $SQA_{s_i,q_j} \subset V(A)$.

From the graph A, it is possible to derive an undirected graph S, which represents the similarity between the set of students. An edge $(s_i, s_j) \in E(S)$ if and only if s_i and s_j are similar. It is important to note that the similarity among two students is considered only if it is greater than a given threshold Θ. The similarity sim_S between two students s_i and s_j is given by:

$$sim_S(s_i, s_j) = \sum_{q_k \in Q} \frac{rank_Q(s_i, s_j, q_k)}{M} \tag{1}$$

where M is the number of questions to which students s_i and s_j have access and $rank_Q(s_i, s_j, q_k)$ is the mean similarity of all answers given by s_i and s_j to question q_k. That is, given the induced subgraph $X = A[SQA_{s_i,q_k} + SQA_{s_j,q_k}]$, which is the graph of all answers given by the students s_i and s_j to question q_k, we have:

$$rank_Q(s_i, s_j, q_k) = \sum_{(a_m,a_n) \in E(X)} \frac{sim(a_m, a_n)}{|E(X)|} \tag{2}$$

In summary, the weight of an edge $(s_i, s_j) \in E(S)$ is the mean similarity between the answers given by s_i and s_j to all the questions they answered. Thus, the more similar the answers are, more similar the students will be.

4.2. Defining potentially relevant questions

Once the graph S is formed, it can be said that each connected component of S is a group of similar students. For each of these groups, based on their answers, it is possible

[1]The similarity relation among answers is considered to be a symmetrical relationship, i.e., if a_i is similar to a_j, then a_j is similar to a_i in the same degree.

[2]The similarity function implementation depends on the educational field (maths, physics, etc.) and on the format of the questions.

to indicate which questions are potentially relevant. Thus, we define the most relevant question to a group of students $S_i \subset V(S)$ as:

$$relevant\_question(S_i) = \max_{q_j \in Q}(rel(q_j, S_i)) \tag{3}$$

where $rel(q_j, S_i)$ represents the relevance of question q_j to the group of students S_i, which is given by:

$$rel(q_j, S_i) = \sum_{(s_m, s_n) \in E(S[S_i])} \frac{rank_Q(s_m, s_n, q_j)}{|E(S[S_i])|} \tag{4}$$

In Equation 4, $E(S[S_i])$ is the set of edges from the induced subgraph $S[S_i]$, i.e., the edges from the connected component formed by S_i. At this point, the system is already able to recommend potentially relevant questions for a group of students based only on the similarity of their responses. However, it is possible to recommend the answers that best represent the attempts of this group of students to the relevant question.

4.3. Defining the most representative answers

Considering the set of answers given by the students S_i to the question q_j as $SS = \sum_{s_m \in S_i} SQA_{s_m,q_j}$, and given the induced subgraph $Y = A[SS]$, which represents the graph of all S_i students' answers to the question q_j, the answer that best represents all the others in $V(Y)$ is given by:

$$rep\_answer(q_j, S_i) = \max_{a_k \in V(Y)}(rep(a_k, q_j, S_i)) \tag{5}$$

where $rep(a_k, q_j, S_i)$ is the representativeness degree of a_k to all the answers given by the students in S_i to the question q_j, and is defined as:

$$rep(a_k, q_j, S_i) = \sum_{a_n \in (V(Y) - \{a_k\})} \frac{sim(a_k, a_n)}{|V(Y) - \{a_k\}|} \tag{6}$$

Hence, the system is able to recommend the most representative answers given by a group of students for a potentially relevant question.

5. Preliminary results and discussion

The described algorithm was implemented in a web application to teach computer programming skills. In this application, the answers are source codes, which are compiled and automatically tested (matched) against input and output pairs. Two test sessions were carried out. The first aimed at analyzing how the algorithm would react when running on a real data set and determining, empirically, the quality of the recommendations generated. The second test session was held for a longer period of time and aimed to observe the algorithm's behavior for different sizes of data sets.

During the first test session, 53 students, organized into 3 virtual classes, C_1, C_2 and C_3, had registered to the system, and a set of 725 answers (138 correct and 587 incorrect) were collected. The classes had, respectively, 18, 15 and 20 students and access to 22, 22 and 3 questions. Only incorrect answers were considered. Three different similarity graphs of students were generated, one for each class. The algorithm was executed only one time and was able to generate, to class C_1 (237 incorrect answers), a total of 8 recommendations for 3 groups of 3 students each. Class C_2 (35 incorrect answers), in turn, had 2 recommendations for a single group of 7 students. Finally, class C_3 (315 incorrect answers), had 4 recommendations to 2 groups of 7 students each. Of all 14 recommendations, 10 were passed on by teachers to students. Table 1 sumarizes these data.

Table 1. Test session 1

	C_1	C_2	C_3
Students	18	15	20
Groups	3/3	1/7	2/7
Recommendations	8	2	4
Accepted recommendations	6	1	3
Questions	22	22	3
Wrong answers	237	35	315

The quality of generated recommendations strongly depends on the the function sim_A. However, even with a simple similarity function (source code similarity and output comparison), the recommendations were well received by the teachers. In most cases, the groups of students indicated by recommendations had real similar difficulties.

The relatively small data set produced impacts on the algorithm's results. The ideal algorithm's scenario is that all students have answered all the questions. However, this is not reflected in reality. Therefore, to define the degree of similarity between two students, only questions that both students had responded were considered. Thus, in the same group, there may be students who answered completely different questions, causing, in some cases, the impossibility to find relevant questions for certain groups. One approach to alleviate this problem and improve the recommendations' relevance is to consider only fully-connected components. However, this strategy could reduce dramatically the number of formed groups and hence the number of recommendations as well.

The second session of tests analysed the algorithm outcomes for a longer period of time. A set of 63 students, organized into a single class, interacted for 80 days with the system, and submitted 3211 answers (2415 incorrect and 796 correct). During this period, the algorithm was run multiple times a day, generating, at each time, different sets of recommendation. The data were stored during this process.

Two constraints were imposed to the algorithm during this test session. (1) for each response, only the 10 highest weighted edges were maintained in the graph of answers. That is, the maximum graph's degree was limited to 10. Only the most similar relations (edges) were kept; and, (2) let the set $A = a_1, a_2, .., a_n$ be the student's attempts to a given question. Only the responses a_1, $a_{\frac{n}{2}}$ and a_n from A are considered. In other words, for each student, only one set of, at most, 3 answers for each question were used

by the algorithm: the first attempt, an intermediate one and the last one. These measures were taken to reduce the algorithm execution time.

The analisys of the results obtained by the proposed algorithm were carried out in different experiment moments. It attempted to expose its functioning when used against different numbers of questions and answers. The data collected during the experiment are described in Table 2. Each line represents one moment of the data collection. Each column displays the data obtained for each of these moments. The data shown are respectively: total submitted answers, total incorrect submitted answers, number of available questions and the number of students who answered at least one question.

Table 2. Test Session 2

Answers	Incorrect Answers	Questions	Students
3201	2400	17	63
2961	2160	17	63
2721	1920	17	63
2481	1680	13	63
2241	1440	9	59
2001	1200	9	59
1761	960	9	57
1521	720	9	45
1281	480	7	37
1041	240	5	28
901	100	5	13

Figure 1 displays the graph of the recommendation data obtained for different amounts of incorrect answers. An increasing trend in the number of of produced recommendations should be noted as the amount of incorrect answers and of available questions growth. With a higher number of responses, there is a greater probability of forming edges in the graph of students. Thus, more recommendations end up being generated.

However, this is not a direct relationship. As it can be seen, there was a drop in the number of recommendations generated in the last three moments of the experiment. This decreasing coincides with the provision of new questions for students. In this case, two points should be taken into account: (1) the inclusion of new questions is a time where not all students answered these questions and (2) new answers to new questions may be so different that previously considered similar students could have their edges removed from the graph students. In other words, the augmenting of the number of incorrect answers can also lead to a scenario with less recommendations.

The graph depicted in Figure 1 points out, also, the number of connected components (CC) formed by the algorithm, the average number of students in each CC and the average number of recommendations generated for a CC. Note that, from 1000 incorrect answers, the algorithm was able to produce about 2 recommendations for each group of similar students.

As mentioned in Section 4.1, the edge between two students is added the graph of students only if its weight is higher than a given threshold Θ. Both experiments 1 and 2 were conducted considering $\Theta = 0.8$. However, the effect of the value of Θ was also

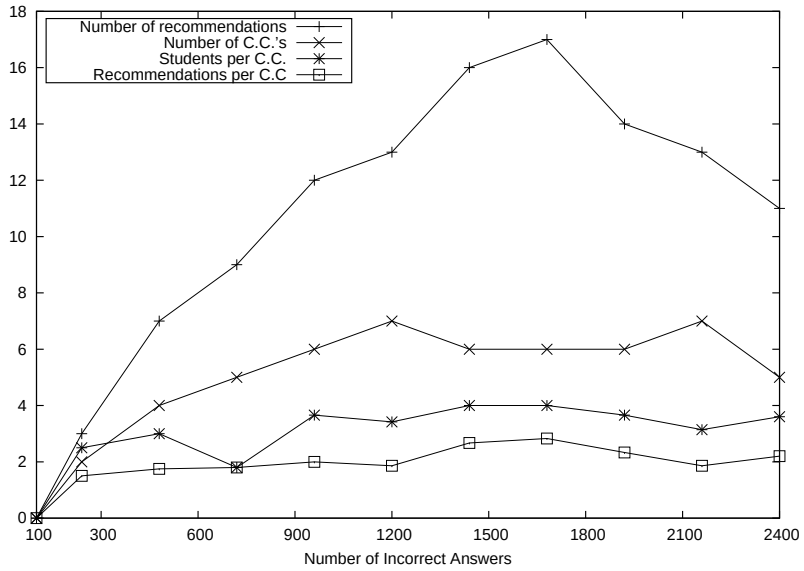

Figure 1. Recommendation data obtained for different amounts of incorrect answers

analysed using the data collected by second experiment. It took into account the results obtained for different values of Θ in environments with 240, 1440 and 2400 incorrect answers. The graphs shown in Figures 2 and 3 indicate the results obtained.

It should be noted that low values of Θ, which allow low weight edges to be added to the graph of students, make it difficult to create a higher number of connected components, because most students are considered similar to each other. In this case, the formed CC's are large and, thus, reduces the quality of recommendations (the similar characteristics among students tend to be very generic). In this sense, Figure 2 shows that with the increase of Θ, the number of obtained CC's, as expected, also increases. The removal of edges tend to cause a division of the CC's into smaller components. With the increase of Θ the algorithm is able to achieve a better students clustering in groups. Also, the quality of recommendations is refined, since the members of the groups have more specifics similarities.

Figure 3, in turn, indicates the number of recommendations obtained for different values of Θ. In this case, only values between 0.8 and 0.9 caused significant decreases in the number of recommendations. This is due to the fact that the sharp decrease of the number of edges in the graph students.

The tests performed indicate that the algorithm is able to generate relevant recommendations in different scenarios. The number of recommendations obtained showed acceptable at all times checked. However, there is still the need for further tests with more complete data sets that can lead to algorithm's improvements.

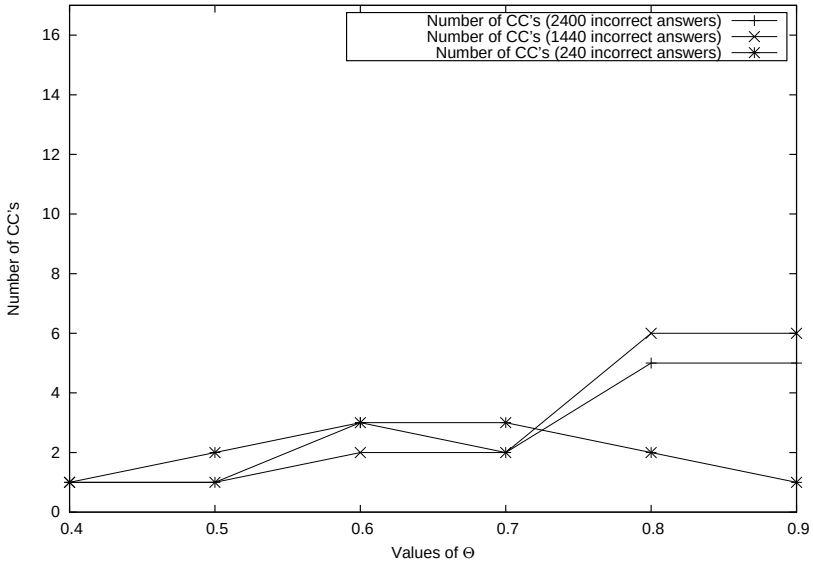

Figure 2. Number of CC's obtained for different values of Θ

6. Conclusion

This paper described the problem of recommending answer records for error mediation in educational environments. The difficulties faced by teachers and students in error mediation were pointed out. Based on this, a novel algorithm for automatic answer recommendation was introduced.

The described algorithm was implemented in a web application to teach computer programming skills. A set of preliminary tests have indicated promising results. The algorithm was capable of defining groups of students with real similarities on their errors and of recommending relevant answers for teacher's mediation. Further experiments are planned to verify the recommendations' quality on more complete data sets.

References

Champaign, J. and Cohen, R. (2010). An annotations approach to peer tutoring. In *The 3rd International Conference on Educational Data Mining (EDM 2010)*, pages 231–240. Citeseer.

Durand, G., Laplante, F., and Kop, R. (2011). A learning design recommendation system based on markov decision processes. In *KDD-2011: 17th ACM SIGKDD conference on knowledge discovery and data mining*.

Isotani, S., Adams, D., Mayer, R. E., Durkin, K., Rittle-Johnson, B., and McLaren, B. M. (2011). Can erroneous examples help middle-school students learn decimals? In *Towards Ubiquitous Learning*, pages 181–195. Springer.

Kutzke, A. R. and Direne, A. I. (2014). Mediação do erro na educação: um arcabouço de sistema para a instrumentalização de professores e alunos. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 25.

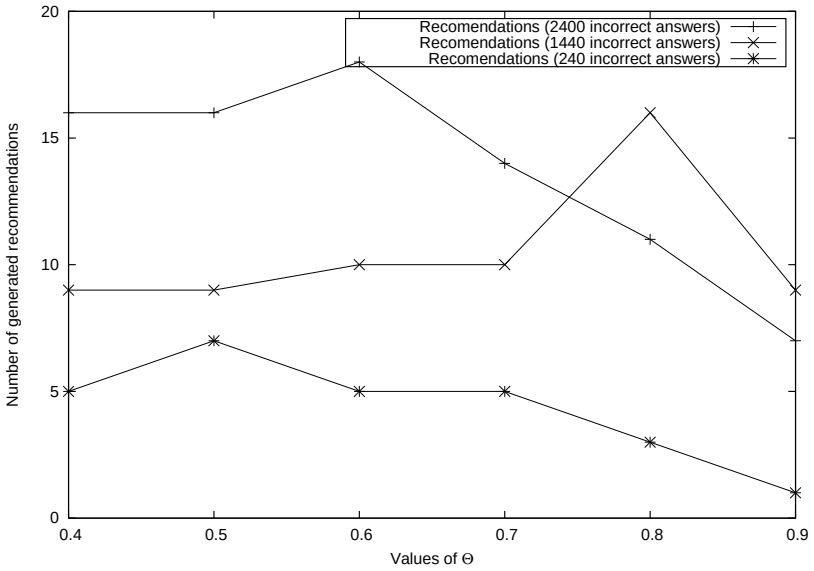

Figure 3. Number of recommendations obtained for different values of Θ

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., and Koper, R. (2011). Recommender systems in technology enhanced learning. In *Recommender systems handbook*, pages 387–415. Springer.

Paiva, R., Bittencourt, I. I., and da Silva, A. P. (2013). Uma ferramenta para recomendação pedagógica baseada em mineração de dados educacionais. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 1.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4, Part 1):1432 – 1462.

Vialardi, C., Chue, J., Peche, J., Alvarado, G., Vinatea, B., Estrella, J., and Ortigosa, Ã. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User-Adapted Interaction*, 21(1-2):217–248.

Vygotsky, L. S. (2012). *Thought and language*. MIT press.

Zapata-Gonzalez, A., Menendez, V., Prieto, M., and Romero, C. (2011). Using data mining in a recommender system to search for learning objects in repositories. In *The 4th International Conference on Educational Data Mining (EDM 2011)*.