Recommending Remedial Learning Materials to the Students by Filling their Knowledge Gaps

Konstantin Bauman
Stern School of Business, NYU
kbauman@stern.nyu.edu

Alexander Tuzhilin
Stern School of Business, NYU
atuzhili@stern.nyu.edu

Abstract—**We present an approach of providing recommendations of remedial learning materials to the students that is based on the proposed "filling-the-gap" method. According to this method, we first identify gaps in student's mastery of various course topics. Then we identify those items from the library of assembled learning materials that help us to fill those gaps, and then we recommend these identified materials to the student. We show empirically through A/B testing that this approach leads to better performance results, as measured by student's improvement of average score on that exam in comparison to the previously taken courses.**

I. Introduction

Due to the recently increased interest in on-line educational technologies and educational delivery methods, the topic of recommendations in the educational domain has become increasingly important lately. In particular, it has been studied in various communities, including RecSys, UMAP, Advanced Learning Technologies, and the Technology-Enhanced Learning communities, and many approaches have been proposed on how to recommend learning materials to the students to improve their learning performance [1].

The TEL-based recommendations can be classified into two major types. The first type constitute "knowledge enhancing" recommendations that focus on the next learning activity expanding and broadening students knowledge of the subject matter. This type of recommendation is advocated by Khan Academy [2], Knewton [3] and some other companies and authors [4]. The second type constitutes "remedial" advice that identifies existing gaps in student's knowledge of the subject matter while the student progresses through the course and tries to "fill in" these gaps by recommending appropriate learning materials and activities.

In this paper, we present a methodology of identifying gaps in students' knowledge and propose specific algorithms to fill-in these gaps by providing recommendations of remedial learning materials to the students. Furthermore, we show empirically (through A/B testing) that this approach leads to better performance results for the "good" students who had average grades in all the previously taken courses between 70 and 90. They were useful in the sense that they lead to a significantly improved performance of these students on the final exams compared to their prior performance before they received personalized recommendations.

Although there exists prior work on recommending learning materials to the students in the TEL environments, there have been only few prior methods that identify the gaps in student's knowledge of the subject matter of the course and that also try to close those gaps [2], [3], [5], [6]. Most of the methods that propose specific algorithms of how to do it, take a more forward-looking proactive approach of recommending the "next learning activity" rather than taking a defensive filling-the-gap reactive approach advocated in this paper.

EdRecSys 2016. Copyright is held by the authors ©2016.

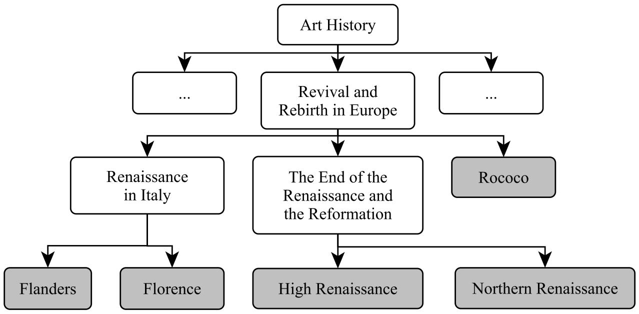

Fig. 1. Part of taxonomy for Art History course.

II. Filling-the-Gap Recommendation Method

In order to provide the "filling-the-gap" recommendations to the students taking a course, we first assemble the library of learning materials to be used for recommendation purposes. Then we construct the learning structure of the course and identify various topics covered in it. After these two steps, we identify the gaps in students' performance in the course and recommend remedial learning materials in order to fill-in students' knowledge gaps. In the rest of this section we describe the specifics of each of these steps.

(1) Building the knowledge structure of the course. For each course in a curriculum, we build taxonomy of the topics covered in that course. In this project, we automatically built the course taxonomy from the syllabus, the weekly learning objectives and the list of weekly reading materials provided by the instructor using information extraction and text mining methods. The on-line university with which we have worked has a very precise and well-developed course structure consisting of 8 weeks of learning sessions.

The taxonomies of the courses in our project consist of a tree of topics and subtopics, where topics correspond to weekly learning objectives of the course. For example, Figure 1 shows a part of the *Art History* course taxonomy where each node represents a topic covered in the course. Each topic node in the taxonomy has a set of obligatory reading materials that were selected by the instructor and assigned to that topic. The leaves of the tree constitute the smallest (atomic) topics and only one piece of reading material (e.g. a web-page, a chapter or a section of a book, an article, etc.) is associated with each leaf. The taxonomy tree does not have to be balanced. In practice, however, it usually has the depths of three or four layers, depending on the nature of a particular course.

(2) Building the library. In this step we build the library of the course related reading materials that go beyond the set of obligatory reading materials assigned by the course instructor. These additional reading materials include popular textbooks, on-line articles, web pages, on-line videos, and other on-line materials that are related to the course.

We build the library from the open on-line sources using the following steps: (a) identify key concepts for each topic in the course taxonomy using the TF-IDF measure computed from the text material(s) assigned to the topic and the corpus of Wikipedia articles; (b) for each concept (keyword), we automatically launch a search query with this concept on the Web and collect the first n results (we set $n = 10$) returned by the search engine, including the links and the actual documents; (c) select relevant documents, that contain more than one key concept from the list produced in Step (a). In [7], the authors followed a similar approach and showed that their method is able to automatically enrich a textbook with the suitable number of links to the related web content.

(3) Matching learning materials with the course topics. In this step for each topic in the course taxonomy, we identify the best matching set of "units of knowledge" from the library (e.g., a book chapter, a Web page or an article), thus establishing the links between the topic in the course taxonomy and the corresponding reading materials from the library. We do this by representing texts assigned to topics and units from the library as a vectors of the TF-IDF measures in the space of the key concepts for the course. We use cosine similarity measure to calculate the distance between topics and units, and identify the best matching materials.

(4) Matching the quizzes questions with the course topics. For each topic in the course taxonomy we determine the set of corresponding test questions from the course quizzes. We represent test question as a vector based on its text using the same TF-IDF methodology as in previous step and compute similarity measure ρ between topic vector and questions vectors. Finally, we assign question to the topic when the similarity measure between the two exceeds a certain threshold. Note that each question can be assigned to multiple course topics and each topic can get multiple corresponding questions. In our study, we used the cosine similarity measure for ρ, but it can also be any other similarity measure, such as Pearson correlation. For example, question "The Rococo style began in (Italy/Flanders/France/Spain), at the end of the reign of Louis XIV" may match topic Rococo because the question is "close" to that topic.

(5) Building students' learning profiles. For each student and a course offering we determine how well the student understood all the leaf topics specified in the course taxonomy by analyzing how well the student has done on the quiz questions corresponding to each of those topics. We calculate the performance score of a topic as the weighted fraction of the correct answers to the list of questions pertaining to this topic, where weights are the similarity measures between the topic and the questions.

For example, if there are 10 questions in the test corresponding to the topic Rococo in the Art History example with the same equal weight, and Joe answered 9 of them correctly, then Joe's score for this topic is 0.9, which means that Joe understood the topic Rococo well. In contrast, if John answered only 5 questions correctly, his score for the topic Rococo would only be 0.5, which means that he did not master this topic.

(6) Identification of students' knowledge gaps in the course. We presume that the student has a *knowledge gap* in a topic if either (a) the performance score of the student in this topic is low, i.e., below a certain threshold level (we use the median of the scores in this topic taken over all the students in the class); or (b) the student has knowledge gaps in a "sufficient" number of subtopics (e.g., more than 66%) of that topic, and therefore needs remedial actions for these subtopics. Referring to Figure 1, if Joe has a knowledge gap in topics "Rococo" and "Renaissance in Italy," we'll presume that Joe's knowledge in the whole topic "Revival and Rebirth in Europe" is poor and identify it as a *knowledge gap*.

(7) Preparing and providing recommendations. Given the structure of a course, the identified gaps in student knowledge in the class, and the links between the topics in the course and the supplemental reading materials from the library obtained in Step (2), we next provide recommendations of these supplementary reading materials to the students in order to close these knowledge gaps. In particular, for each knowledge gap topic node in the course taxonomy, we recommend to the student those supplementary reading materials that are linked to that node.

Note that these recommendations are personalized to the individual students since different students have different knowledge gaps. Still, in case of two students having the same set of knowledge gaps, they will receive the same set of recommended materials.

It is also important to note that if a student has a knowledge gap in some topic that has subtopics in the course taxonomy, we will provide recommendations for this topic but will not provide separate recommendations for any of its subtopics. For example, if Joe has knowledge gaps in topics Rococo and Renaissance in Italy and, therefore, in topic Revival and Rebirth in Europe (see Art History taxonomy in Figure 1), then he will receive recommendations only for the entire topic Revival and Rebirth in Europe.

III. Experimental Settings

Dataset. We conducted a field study with participation of the students from a major on-line university over a period of three semesters, where each semester consists of 9 weeks, 8 of which are dedicated to the studies and the last week to the final exams. In particular, we worked with 910 students taking one or more courses in Computers Science (CS), Business Administration (BA) and General Studies (GS), such as Mathematics, English, Psychology, Art History, etc. The unit of analysis in our study is the student/course pair specifying a course that a student takes during a particular semester. Overall, we collected data on 1512 student/course pairs throughout three semesters covered in our study.

To validate our method described in Section II, we conducted the following experiment (the so called A/B test). Within each course we randomly split the student/course pairs into the following three groups in order to provide different types of recommendations to them and compare their performance results:

1) The *control* group to whom we did not provide any recommendations in the course.
2) The *non-personalized* group consisting of the students who received "generic" (non-personalized) recommendations sent to everybody in that group. In particular, they received the same set of recommendations as the students in the personalized group who failed all their tests and thus needed help in all the topics of the course.
3) The *personalized* group of students received recommendations based on the method described in Section II.

Providing Recommendations. As the first step, we build taxonomy of the courses based on the syllabus and a set of learning materials selected by the instructor and assigned to the course. In particular, each week of studies is carefully structured in terms of the studying process at that university and contains a set of obligatory reading materials, assigned to all the students for the week. Most of the courses have taxonomies consisting of a root representing the entire course, eight nodes on the second level (one node for each week of the studies), and two or three children nodes at the third level specifying either the learning objective(s) for the week or the obligatorily assigned reading materials for the week. Next we used Google API in order to find web content that is relevant to the course materials, add it into the course library, and matched it to the course topics as discussed in Section II.

We provided recommendations to the students in *non-personalized* and *personalized* groups by sending them emails with the lists of links to the recommended materials up to three times per semester: (1) in preparation for the graded Quiz 1 during the third week of studies and addressing the knowledge gaps in student's performance on weekly quizzes during weeks 1 and 2; (2) in preparation for the graded Quiz 2 during the sixth week of studies addressing the knowledge gaps during weeks 3-5; (3) in preparation to the final exam addressing the knowledge gaps identified through out the course. The number of recommended materials for the *non-personalized* group of students is equal to the number of weeks covered by the graded quiz. For example, in preparation for the final exam they received 8 materials corresponding to each week of studies.

Performance Measures. The goal of our experiments is to show that personalized recommendations lead to better performance results in the course. Therefore, we compare the performance results of the students from the *control*, *non-personalized* and *personalized* groups on the final exams in terms of the measures defined below.

Each course in our experiment has the Final Exam in the form of a multiple choice quiz that includes questions covering most of the topics from the course. Since our methodology of filling-the-gap recommendations helps the students to master poorly performed topics and therefore should lead to better performance in the course, it is natural to use student's (absolute) performance on the final exam as a performance measure for our methodology. Moreover, we use the normalized grade on the final exam as another performance measure, where normalization is done by subtracting the average grade from the student's grade and dividing it by the standard deviation for all the students in the course.

Another good performance measure is the improvement of student's performance in the current course vis-á-vis his/her performance in the previous courses. The "previous courses" can be the last course, the last two courses and all the previous courses, either taken across all the subjects or within the same subject area as the current course, e.g., within CS, BA or MATH. Furthermore, we also use this same student improvement performance measure, but in the normalized (vs. absolute) form, as explained above.

Survey. At the end of each semester we also sent a survey to those students who have received at least one recommendation during the entire semester. The purpose of the survey is to see how well they perceived our recommendations and also to detect possible biases and problems with the experimentation.

As our survey results show, a significant majority of the students in the *non-personalized* and *personalized* groups who have completed the survey liked our recommendations, found them to be relevant and helpful for the course, and would like to recommend our recommendation tool to their friends.

IV. Results

First, we compared the *control*, *non-personalized* and *personalized* groups in terms of their average performance across the set of previously taken courses, including the last, the two previous and all the previously taken courses. The results show that there are no statistically significant differences between three experimental groups for *all* those performance measures.

We next compare the performance of the three experimental groups. In particular, Figure 2 shows the comparison in grade improvements across the three groups for the students having average performances on all the previously taken courses ranging from 55 to 100 points (as shown on the X axis). As Figure 2 shows, performance improvements for personalized recommendations are higher across all the types of the good students (ranges 70-75, 75-80, 80-85 and 85-90) in comparison with the *control* and the *non-personalized* groups. Furthermore, for all the "good" students (ranges 70-90), t-test shows that the performance improvement for the *personalized* group

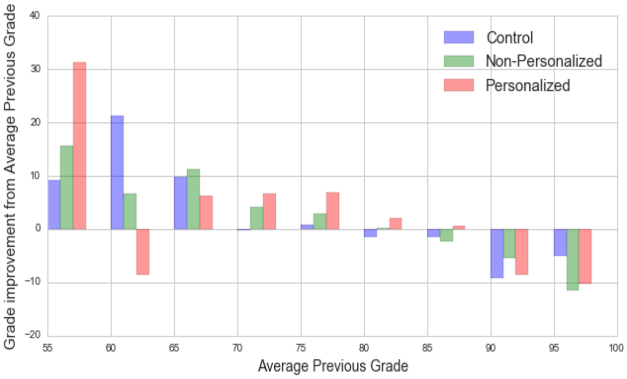

Fig. 2. Histogram of grade improvement by average previous grade.

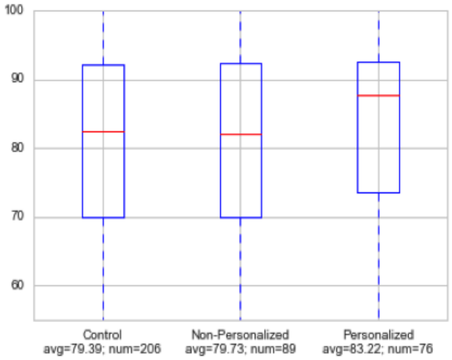

Fig. 4. Performance of “good” students on final exam.

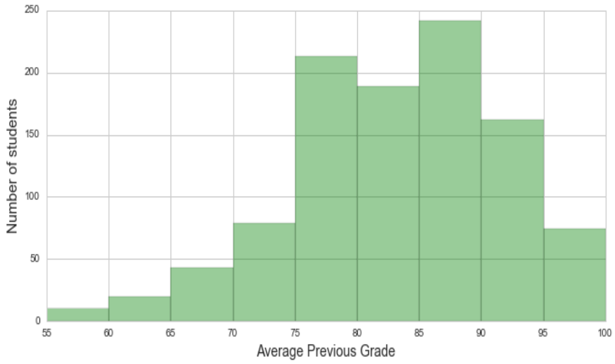

Fig. 3. Histogram of the number of students by average previous grade.

is significantly better than for the *control* group. Note that such students constitute about 70% of the total number of students, as is shown in Figure 3. Moreover, for the “falling-behind” ($GPA < 70$) and the “excellent” ($GPA > 90$) students the comparison between the *personalized* and the other groups are not statistically significant. The “falling-behind” group is quite small: as Figure 3 shows, it constitutes 7% of the students. The reason why personalized recommendations have no effects on “excellent” students is because our system seldomly sends recommendations to them (since they are already doing fine in their studies). Similar results are observed for the “good” students in the case of measuring not only absolute but also normalized performance improvements.

In addition to the average performance improvement over all the previously taken courses, we also considered the improvement taken over the last previously taken course. We also observe similar results in this case: the group of good students receiving personalized recommendations has significantly higher performance improvements than the control group of good students. Furthermore, similar results hold for the case of normalized version of this metric.

Furthermore, we also consider those “good” students who followed our recommendations, i.e. clicked on the provided links, in *non-personalized* and *personalized* groups ($\sim$ 40% of students). Figure 4 presents the comparison of their performance results on the final exam with the “good” students from the *control* group. The difference between *personalized* (83.22) and *control* (79.39) groups is statistically significant, whereas the difference between *non-personalized* and *control* is not statistically significant. Moreover, similar results hold for the “good” students for most of the other measures described in Section III, such as improvement of the student’s performance vis-á-vis her performance in previous courses, in both absolute and normalized cases.

V. Conclusion

In this paper we presented a methodology for providing automatic personalized filling-the-gap recommendations to the students that gather data on students’ progress through the course, identifies “knowledge gaps” in their learning of the course materials, and provides remedial recommendations of learning materials to the students with the purpose of filling these gaps. Furthermore, we empirically tested our filling-the-gap method in the setting of an on-line university using the A/B testing methodology and showed that is actually worked, i.e. the “good” students (whose average final exam scores across the previously taken courses was between 70 and 90) improved their performance on the final exams significantly more (in comparison to their prior performance before they received personalized recommendations) than the students from the *control* group.

References

[1] N. Manouselis, H. Drachsler, V. Katrien, and D. Erik, *Recommender Systems for Learning*. Springer, 2013.

[2] R. Murphy, L. Gallagher, A. Krumm, J. Mislevy, and A. Hafter, “Research on the use of khan academy in schools.” in *SRI Education report*, 2014.

[3] K. Wilson and Z. Nichols, “The knewton platform: A general-purpose adaptive learning infrastructure.” in *A Knewton white paper*, 2015.

[4] H. Drachsler, K. Verbert, O. C. Santos, and N. Manouselis, *Recommender Systems Handbook*. Springer US, 2015, ch. Panorama of Recommender Systems to Support Learning, pp. 421–451.

[5] J. S. Underwood, “Metis: A content map-based recommender system for digital learning activities,” in *Ed. Rec. Sys. and Technologies: Practices and Challenges*, O. C. Santos and J. G. Boticario, Eds., 2012.

[6] A. Klanja-Milievi, B. Vesin, M. Ivanovi, and Z. Budimac, “E-learning personalization based on hybrid recommendation strategy and learning style identification,” *Computers & Education*, vol. 56, no. 3, 2011.

[7] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu, “Enriching textbooks through data mining,” in *ACM Symposium on Computing for Development*, ser. ACM DEV ’10, 2010.