Educational Data Mining by Correlation Analysis

Rainer Knauf[1], Kinshuk[2], Yoshitaka Sakurai[3], Setsuo Tsuruta[3]
[1]Faculty of Computer Science and Automation, Ilmenau University of Technology, Ilmenau
[2]Athabasca University, Edmonton
[3]School of Information Environment, Tokyo Denki University, Tokyo
rainer.knauf@tu-ilmenau.de

Abstract: *The paper introduces a "lazy" Data Mining technology, which models students' learning characteristics by considering real data instead of deriving ("guessing") their characteristics explicitly. Here, we introduce a technology to mine course characteristics similarities of former students' study traces and utilize them to optimize curricula of current students based to their performance traits revealed by their study achievements so far. This technology generates suggestions of personalized curricula. Furthermore, it is supplemented by an adaptation mechanism, which compares recent data with historical data to ensure that the similarity of mined characteristics follow the dynamic changes affecting curriculum (e.g., revision of course contents and materials, and changes in teachers, etc.).*

Keywords: personalized curriculum mining, adaptive learning technologies

1. Introduction

University students may not know exactly in advance, which performance skills are challenged in which course. Therefore, they cannot know whether or not they can perform really well in it. Furthermore, students may be not consciously aware of their own performance traits and consider them to compose their curriculum to optimally make use of it. Since each particular course challenges certain particular performance skills, the GP distribution on the various courses may characterize a student's performance traits. This information may be helpful to suggest best possible curricula complements to the courses taken so far, i.e. to a student's educational history. Proposals to characterize the skills challenged in university courses are usually explicit models to match with the teacher's style and material. Despite of their successful empirical evaluation such models cannot really been proven. However, they are derived from educational data. So, why not using the data itself as the model instead of deriving ("guessing") explicit assumptions and guesses from it? The approach introduced here refrains from explicit models, but uses a database of cases as the model and computes (positive and negative) correlations to a current case to derive suggestions for optimal curricula.

2. Performance Correlation

The basic idea behind this approach is that people have (a) certain intelligence traits and (b) certain learning style profiles, which may influence their chance to receive good results in subjects, which challenge these traits (a) and which (b) are taught and examined by the teacher in a way that fits to this style. The particular GP achieved in particular courses may aindicate, (a) up to which degree a student meets the intellectual skills challenged in a potential new course and (b) up to which degree a student's learning style matches with the teacher's way of teaching and examining and the material he issues in a potential new course. We believe that such individual profile parameters imply correlations between these degrees in different courses and thus, there are positive or negative correlations between the degrees of success in the courses. These correlations characterize implicitly a student's profile and can be derived by Educational Data Mining on (former) students' educational history. In case of perfect positive correlation, this student will receive about the same number of GP in course *y* than he/she achieved in course *x*. In case of perfect negative correlation, this student will receive a number of grade points in course *y* which is the higher (lower), the lower (higher) the number of GP in course *x* was. If there is no correlation at all between both courses, the challenged skills and the preferable learning style and learning material preferences are totally independent from each other. Statistically it means that whatever number of GP a student receives in course *x*, any number of GP may be the result in course *y*. To also include the circumstance that a correlation can by influenced by learning a particular course content or experiencing a particular teacher's style or material of teaching, we consider only correlations *corr(x,y)* of courses *y* taken at a later time than courses *x*. Having samples of *k* students *S = {s₁, ..., s_k}*, who took a course *y* after a course *x*, and each student s_i achieved g_i^x GPs in course *x* and g_i^y GPs in course *y*, the linear correlation coefficient is $corr(x,y) = \frac{\sum_{i=1}^{k}(g_i^x - \bar{g}^x)(g_i^y - \bar{g}^y)}{\sqrt{\sum_{i=1}^{k}(g_i^x - \bar{g}^x)^2}\sqrt{\sum_{i=1}^{k}(g_i^y - \bar{g}^y)^2}}$ with $\bar{g}^z$ being the average number of GPs achieved by the *k* students *{s₁, ..., s_k}* in course *z*. These correlations can be computed based of known educational histories of (former) students and applied to courses of the educational history of a current student to suggest optimal courses for upcoming semesters. Additionally, courses with no correlation with courses taken so far should be considered for upcoming semesters to involve other performance challenges and teaching styles and materials, which have not experienced so far.

3. Correlation Fusion

A potential upcoming course *y* correlates with the courses $X = \{x_1, ..., x_m\}$ taken so far in various ways, i.e. with some of the courses in *X* in a strong positive way, with others in a strong negative way and maybe, with others not at all. In practice, courses are weighted with their related number of units (in Japan) or credits (in USA, Canada and Europe). Since we apply our approach in a Japanese university, we use the term units here. Therefore, it is reasonable to weight the different correlations $corr(x_j, y)$ of a potential upcoming course *y* of the different courses taken so far $X = \{x_1, ..., x_m\}$ with the number of units u_j for the courses $x_j \in X$ and to compute a weighted average correlation of the course *y* to the courses in *X* as $corr(\{x_1,\dots,x_m\},y) = \frac{\sum_{i=1}^{m} u_i * corr(x_i,y)}{\sum_{i=1}^{m} u_i} = \frac{\sum_{i=1}^{m} u_i * \frac{\sum_{j=1}^{k_i}(g_j^{x_i} - \bar{g}_j^{x_i})(g_j^y - \bar{g}_j^y)}{\sqrt{\sum_{j=1}^{k_i}(g_j^{x_i} - \bar{g}^{x_i})^2}\sqrt{\sum_{j=1}^{k_i}(g_j^y - \bar{g}^y)^2}}}{\sum_{i=1}^{m} u_i}$. Here, (1) $\{x_1, ..., x_m\}$ is the set of courses the student, who looks for appropriate courses for his/her next semester, took so far, and *m* is the cardinality of this set, i.e. the number of these courses, (2) *y* is a candidate course for the next semester, for which the students met all prerequisites and for which the correlation is subject of computing based on the students' samples in the database, (3) k_j is the number of students in the database, who took course *y* after taking the course x_j ($x_j \in \{x_1, ..., x_m\}$), (4) $g_i^{x_j}$ is the number of grade points the *i*-th ($1 \le i \le k_j$) student in $\{s_1,\dots,s_{k_j}\}$ of the data base received in the course x_j ($1 \le j \le m$), (5) $\bar{g}^{x_j}$ is the average number of grade points achieved by the k_j students $\{s_1, s_s, \dots, s_{k_j}\}$ in the database, who took the course x_j ($x_j \in \{x_1, ..., x_m\}$), and (6) u_j is the number of units of course x_j ($1 \le j \le m$). In this formula, we consider all courses taken by the student so far equally and independent from in which of the former semesters he/she took it for selecting courses for the upcoming semesters.

4. Composing the Curriculum of an upcoming Semester

When a curriculum for an upcoming semester is composed, a list of all courses, which are possible to take in the next semester according to whether or not their prerequisites are met by the educational history $X = \{x_1, ..., x_m\}$, needs to be computed. Let $Y = \{y_1, ... y_n\}$ be the set of possible courses. Next, a reasonable fraction *F* between (a) the total number of units for courses that correlate very heavy with the educational history and (b) the total number *U* of units for courses, which require traits and learning preferences, which do not correlate at all in the upcoming semester has to be determined.

After that, the educational history $X = \{x_1, ..., x_m\}$ has to be divided into (a) a sub-list $X^+ = \{x_1, x_2, \ldots x_{m^+}\}$ of courses, in which the considered student performed best and (b) a sub-list $X^- = \{x_1, x_2, \ldots x_{m^-}\}$ of courses, in which the student performed worst. Then, the set of candidate courses $Y = \{y_1, ..., y_n\}$ should be sorted towards a list $\vec{Y} = [y^1,...,y^n]$ [1] according to decreasing absolute values of (a) $corr(X^+,y)$, if $corr(X,y) > 0$, i.e. if *y* correlates positive to the (complete) educational history respectively (b) $corr(X^-,y)$, if $corr(X,y) < 0$, i.e. if *y* correlates negative to the (complete) educational history of the considered student. Finally, the next semester is composed by subjects from the candidate set $Y = \{y_1, ..., y_n\}$ by (1) taking the courses beginning from the (strongly correlated) front end of $\vec{Y}$ until their total number of units reaches a value, which is as close as possible to $F * U$ (i.e. the fraction of correlated courses is reached) and (2) taking the rest of the units are from the (rather non correlated) rear end of $\vec{Y}$ until the total number of units reaches a value as close as possible to *U*.

In case a course teacher changes, the old data from the related former course need to be deleted from the database, which serves for computing the correlation, because at least the style of teaching the course and the kind of material changes by changing the teacher. In some cases, also the course content changes, too, and thus, the course profile changes even more, because the challenged traits may change additionally. Therefore, the correlation computing for between this course and any consecutive course has to begin from scratch. Also, if a teacher revises a course and changes his/her way of teaching or the teaching material significantly, this course has to be handled in the same was as in the above case, i.e. deleted from the database. Deleting a course from the database does not mean, that the complete students' traces, which contain this course have to be deleted, of course. In fact, this would be a waste of the other useful information within these student's samples. It only means that this course does not count any more for correlation computation, i.e. it is excluded from that. In case the same course is offered by multiple teachers, these courses have to be treated as different courses in the database, because different teachers may teach the same content in different ways.

5. Conclusions and Outlook

This paper introduced a technology to mine course characteristics similarities of former students' study traces and utilize them to optimize curricula of current students based to their performance traits revealed by their study achievements so far. Correlation analysis on a database of students' traces is some kind of "lazy" model (in the slang of Data Miners) is (1) easier to acquire and (2) completely trustable (100 % true) compared to "guessing" parameters (intellectual traits, learning styles, material preferences, …) of an "eager" modeling technology that derives a real model. This way, our technology generates suggestions of personalized curricula. Furthermore, this technology is supplemented by an adaptation mechanism, which compares recent data with historical data to ensure that the similarity of mined characteristics follow the dynamic changes affecting curriculum (e.g., revision of course contents and materials, and changes in teachers, etc.). Meanwhile, we collected sufficient data to mine such correlations (namely 186 students' traces) and start advising students optimal curricula based on this technology. Furthermore, we investigate more refined correlation measures (than linear correlation), which are said to be more robust against outliers in the data base.

Footnotes

[1] We chose top indices on purpose to clear up that each y^j normally origins from a y_i with $i \neq j$.

References

Knauf, R, Sakurai, Y., Takada, K., & Dohi, S. (2009) Personalized Curriculum Composition by Learner Profile Driven Data Mining. Proc. of the 2009 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2009), ISBN 978-1-4244- 2794-9, San Antonio, TX, USA, pp.\ 2137-2142.

Knauf, R., Sakurai, Y., Tsuruta, S., & Jantke, K. P. (2010) Modeling Didactic Knowledge by Storyboarding, Journal of Educational Computing Research. (42)4, ISSN: 0735-6331 (Paper) 1541-4140 (Online), Baywood Publishing Company Inc, pp. 355-383.

Tsuruta, S., Knauf, R., Dohi, S., Kawabe, T., Sakurai, Y. (2013) An Intelligent System for Modeling and Supporting Academic Educational Processes, book chapter (chapter 19) in Aljeandro Penã Alaya (Ed.): Intelligent and Adaptive Educational-Learning Systems: Achievements and Trends, KES-Springer Verlag Book Series, to appear in 2013.