The title for your project.

Yang Liu, Zhenge Zhao

Abstract—
We present a visualization tool for demonstrating the relations between different courses based on students' grades in these courses. The goal of this visualization tool is to show the hidden connections among courses. Having effective visualizations of course data is valuable for understanding how one course benefits another and will be useful for setting up the prerequisites in a university, building recommendation systems or academic advising. Using enrollment data from a research university in Canada, we implement a robust mathematical comparison model to calculate correlation coefficient between two records. We evaluate our design choices through engagements with professors and students.

1 Introduction

Navigating the curriculum of educational institution, fulfilling prerequisite and choosing between course options has been a feature of the educational environment dating back to Platos inscription "Let no one ignorant of geometry enter!" inscribed at the entrance of his academy [1]. In efforts to remain competitive, course options at institutes of higher learning have exploded, often offering hundreds of course possibilities that satisfy their general education requirements [4].

The real world scene is that coursework contents can be related, making one course as the prerequisite of the other will help students better learn knowledge and obtain better grades. Academic advisor, for example, deals with these issues a lot.

While some course dependency have already been explicitly annotated in practice, more remain unclear and potential. In this work, wed like to design a visualization view to group related courses in clusters based on student enrollment history, and another view to show correlations of two related courses based on the discrete grades of students who have enrolled both courses.

The aims of this research are:

- provide a tool to visualize coursework contents similarity based on student membership.
- study ordering of neighbours in the node-link diagram to highlight interesting neighbours.
- study simliarity metrics of two courses based on grade history of one student and multiple students.

2 Background

Collaborative-filtering is a recommendation approach that uses similarity between users and the benefit they have received from items in the past to make recommendations. Three main approaches to collaborative filtering are memory-based, model-based and hybrid methods [6] . Memory-based approaches, specifically item-based recommender 1, use an item-user rating matrix to compute pairwise similarities between items. In the contexts of courses, enrollment roster, previous courses students taken and grades course give, is a natural way to reason about course similarity.

However the recommendation approaches are not appropriate for our problems, since they are used for making predictions which means, the approach is designed to fill in the missing values in the matrix. For our case, however, we want to figure out how two courses are correlated with each other. We dont want to fill the matrix since it will blur the original data. What more, for each two courses we are only interested in the users who take both of them. Visualization is an effective and intuitive way to approach our goal.

	CS101	CS202	CS301	CS401
User1	80	99	87	??
user2	79	88	86	85
User 3	76	95	91	86
User 4		78		
User 5	98	80	80	70
User 6			78	

Compute the score of CS401 by user 1?

Fig. 1. Typical student course grade table

Therere many existing similarity measures to compare two entities. Well known metrics include Euclidean similarity, Jaccard(Tanimoto) similarity, Pearson Correlation Coefficient, cosine similarity and Log-likelihood similarity etc. Euclidean similarity measures the distance between courses grading vectors. Jaccard(Tanimoto) similarity is calculated by dividing the intersection of the sets by the union of those sets [3]. Cosine Similarity envision users ratings as points in space and measures the cosine of the angle between these lines drawn from origin to each point. The Log-likelihood similarity is a measure of how often items from 2 sets appear together versus how often they appear apart. Pearson Correlation Coefficient is a number between -1, 1. It measures the tendency of the rating vectors, paired one by one, and its typically used in early research papers. It formula is given as $pearson-correlation(u,w) = \frac{cov(R_u,R_w)}{\sigma x \sigma y}$ where cov stands for covariance and σx stands for standard deviation of x. We are interested in the strong positive correlation and the positive correlation as illustrated in 2.

For our work, Euclidean similarity is not suitable because courses taken by more students will be added more distances. Jaccard(Tanimoto) similarity and Log-likelihood dont count grades students get. We choose Pearson Correlation coefficient to indicate the similarity between two records.

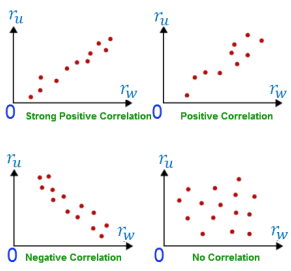

Fig. 2. Pearson Correlation Illustrated

- *Your Name is a graduate student at the University of Arizona. E-mail:[your NetID]@email.arizona.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2.1 Related Work

The problem of coursework similarity has been studied in the context of course recommendation system. Bendakir et al. [2] proposed a recommendation system based on decision tree of course history. Their approach, however, does not consider students' grades at all. Thus, their tool may wrongly correlate totally different courses simply due to historical mistakes. Sandvig et al. [3] did use the GPA information, but GPA, as an average metric, doesn't say much about each specific class.

When it comes to the visualization problem. Since our goal is to cluster similar classes together, a node-link diagram naturally jumps into our mind. D3 library has a force-directed graph that is close to our needs. But we are hesitant about its fisheye distortion and curved link variant because these variants make it hard to click on nodes or edges for further details. We are also aware that force directed drawing is criticized for local minima. A multilevel approach [7] might fix it but we are not focusing on algorithmic style improvement in this proposal.

Other well-known techniques include Rheingold-Tilford Tree, whose tree hierarchy is too constrained to express clustered nodes; arc diagram, whose purpose is to highlight existing cycles. We investigate but decide not to use them.

3 Proposed Work

Our ultimate goal, is to design a visualization tool for understanding courses interactions. For this, we design three components within the interface of the tool to support the analysis task: A) Courses global view: a map graph, illustrated in part A of 3 to deploy all the courses in a map, courses are connected with arrows. A high level course (like 4XX) will be drawn bigger compared to the prerequisite courses. Also, we use the thickness of arrows to show the similarity strength between two courses. Naturally the higher the similarity measures, the thicker the arrow. For This map graph we follow common "overview first, zoom and filter, details on demand" navigation parttern [5]. Each major is depicted with different colors. The view consist a zoomable navigation map and a thumbnail which show you the position and scaling of the current area. B) Course view: a node-link diagram, illustrated in part B of 3, which details the course you select or your selected arrow pointing at from courses global view. The specified course will be put in the center, with all the courses directing to it. The size of the circles and the thickness of the arrows will still follow the provision we set in part1.

C) Grades view: a parallel coordinates diagram, illustrated in part C of 3, to show the specific grades changes for students who have taken both courses. Each lines start point is the score that a student gets in the previous course, and the end point is the score the same students gets in the later course. Different colors are utilized to show the grade section for the first course. The view also contains a timeline slider which can be dragged to show all the lines in the padded area.

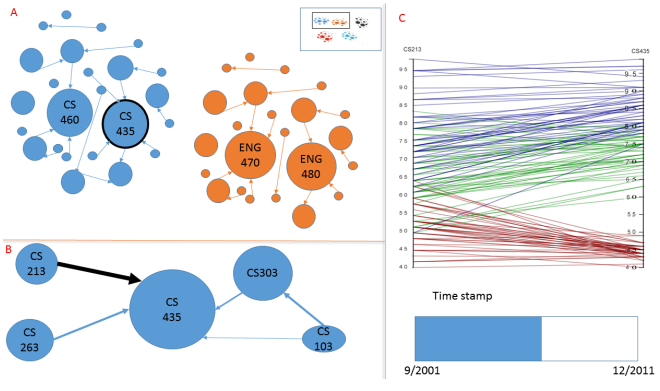

Fig. 3. Our Interface consists of three components: A) a global navigation map for all the course clusters ;B) For the selected Course, all the potential prerequisites for our focus course ;C)For the selected arrow, an interface to show all the transition of scores of students taking both two courses.

3.1 Data

Towards this end, we performed a pilot study using student data from a Canadian research university. This data included all students who had taken a computer science course at that university between September 2001 and December 2011.

The data was in a comma-separated file made up of rows with 6 fields: a unique, anonymous student identifier, the term (which could be Spring, Summer or Fall and the year), the subject ID (such as CS or ENGL), the course code (such as 101), the percentile grade received and the students major.

The data needs to be preprocessed in preparation for the similarity calculation and the visualization. A unique, sequential identifier was added to each line. There's a variety of identifiers for problems in a course, such as WD indicating withdrawal, DNR indicating the final exam was not written. These were replaced with 0 so that there would be a consistent, natural number domain for the grades. Some of the course codes had an optional letter suffix, which would indicate if it was offered online or at another campus. We ignore them because they don't affect course content a lot.

We removed courses and corresponding records with less than 10 students, on the basis that it was insufficient to measure the interactions between two courses. After these adjustments, the training set consisted of 37,392 students with data about 468,632 courses they took. There were 2,326 unique courses in the dataset.

3.2 Evaluation

It is lucky that we don't rely on explicit prerequisite information to train our model. So the first evaluation would be comparing our model's output with existing prerequisite information found online.

If that goes well, we can invite other groups in this class to evaluate it by agreeing to participate their evaluation.

If we go to the full-scale evaluation, the population must be carefully selected. I'd suggest including academic advisors both having experience with other tools and totally fresh to this kind of tool. Since academic advisors usually are well-knowledged about the course content by simply looking at the course name, it's also interesting to let students(other than freshman) in different years participate and evaluate the result without knowing the course content in great detail.

3.3 Timeline

Table 1. Project Milestones

Date	Milestone (%)
Oct 7	D3 library setup, familiar with node-link graph and parallel coordinates
Oct 15	Data preprocessing for D3 drawing
Oct 22	Prototype Drawing
Oct 30	Extract explicit prerequisite info from university website
Nov 7	Evaluate prototype with explicit prerequisite
Nov 14	Evaluate prototype with class groups

4 Impacts

Although this is not a pioneering work, we think it does add to an academic advisor's arsenal for it considers grades more significantly than state-of-the-art.

We believe our main contribution is that this tool not only gives clustering of courses, but tells why two courses are clustered together in the parallel coordinates diagram. Existing course recommender system based on decision tree or neural network may not persuade users in a straight forward way how the conclusion is made. We use a simple yet convincing heurisitic.

We use parallel coordinates to show the trend in students' grades to indicate coursework content. This may inspire other researchers to study item similarity with this diagram.

REFERENCES

[1] W. Anglin. *Mathematics: A Concise History and Philosophy: A Concise History and Philosophy*. Readings in Mathematics. Springer New York, 1994.

[2] N. Bendakir and E. Aïmeur. Using association rules for course recommendation. In *Proceedings of the AAAI Workshop on Educational Data Mining*, vol. 3, 2006.

[3] J. Sandvig and R. Burke. Aacorn: A cbr recommender for academic advising. Technical report, Technical Report TR05-015, DePaul University, 2005.

[4] B. Schwartz. *The Paradox of Choice: Why More Is Less, Revised Edition*. HarperCollins, 2009.

[5] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pp. 336–. IEEE Computer Society, Washington, DC, USA, 1996.

[6] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, Jan. 2009. doi: 10.1155/2009/421425

[7] C. Walshaw. A multilevel algorithm for force-directed graph drawing. In *International Symposium on Graph Drawing*, pp. 171–182. Springer, 2000.