Next-Term Student Grade Prediction

Mack Sweeney, Jaime Lester, Huzefa Rangwala
George Mason University
Fairfax, VA, United States
{*msweene2, jlester2*}@*gmu.edu, rangwala@cs.gmu.edu*

***Abstract*—An enduring issue in higher education is student retention to successful graduation. To further this goal, we develop a system for the task of predicting students' course grades for the next enrollment term in a traditional university setting. Each term, students enroll in a limited number of courses and earn grades in the range A-F for each course. Given historical grade data, our task is to predict the grades for each student in the courses they will enroll in during the next term. With this problem formulation, the next-term student grade prediction problem becomes quite similar to a rating prediction or next-basket recommendation problem. The factorization machine (FM), a general-purpose matrix factorization (MF) algorithm suitable for this task, is leveraged as the state-of-the-art method and compared to a variety of other methods.**

Our experiments show that FMs achieve the lowest prediction error. Results for both cold-start and non-cold-start prediction demonstrate that FMs can be used to accurately predict in both settings. Finally, we identify limitations observed in FMs and the other models tested and discuss directions for future work. To our knowledge, this is the first study that applies state-of-the-art collaborative filtering algorithms to solve the next-term student grade prediction problem.

***Keywords*-matrix factorization; factorization machines; grade prediction; cold-start; recommender system; educational data mining; regression**

I. Introduction

An enduring issue in higher education is student retention to successful graduation [1]. The 2001 National Research Council report [2] identified the *critical need* to develop innovative approaches to enable higher-education institutions to retain students, ensure their timely graduation, and ensure they are well-trained and workforce-ready in their fields of study. The ability to predict student grades in future enrollment terms provides valuable information to help students, advisers, and educators achieve these goals. One can attempt to predict grades for the next term or for several terms ahead. This information can be used to help students choose the most suitable majors, properly blend challenging and easy courses in a semester's schedule, and indicate to advisors and educators when students need additional attention. This task is complicated by the ever-increasing volume of data streaming in from increasing student enrollments and by the continually shifting characteristics of the overall student populations.

As the volume and variety of information being collected in traditional university settings continue to expand, new opportunities to apply big data analytics arise. In this paper, we develop a system to accomplish one such task: predicting students' course grades for the next enrollment term in a traditional university setting [1]. Students take courses over a sequence of academic terms. During each academic term, students enroll in one or more courses and earn grades in each course in the range A-F. Given historical grade data, our task is to predict the grades students will obtain in the courses they will take in the next term.

With this problem formulation, the next-term student grade prediction problem becomes quite similar to a rating prediction problem or a next-basket recommendation problem. These problems have been well-researched in the e-commerce domain. Perhaps the most notable examples arose during the Netflix prize challenge [3], though recommender systems research was still quite active before and has continued to be active after the challenge ended in 2009. Throughout the years, a variety of methods have been used to address this problem, including nearest neighbor approaches [4] [5] [6], collaborative filtering techniques [7] [8], restricted boltzmann machines [9], and even topic modeling methods [10], among others. Despite their ever-growing popularity, there has been relatively little application of these methods to the field of educational data mining. The few notable examples will be discussed further in the literature review. Many of these studies look exclusively at online data, and none of them explore the problem of next-term student grade prediction in as much detail as we do in this study.

In this paper, we first formalize the problem of next-term student grade prediction. We then compare several baseline methods to the factorization machine (FM) model. We restrict this experiment to the collaborative filtering (CF) setting, meaning that we leverage only student grades for our predictions. Our experiments show that FMs outperform all other methods, making it the choice for our next-term prediction system.

Traditionally, recommender systems have issues predicting for unseen users and items (cold-start dyads). However, our experiments show that the use of course and student bias terms alleviate this weakness sufficiently to produce reasonable predictions for cold-start records. While making a poor product recommendation is not desirable, it is much

[1]Source code is available at: https://github.com/macks22/ntsgp

less desirable to make a particularly poor course grade prediction. Doing so could potentially lead a student to avoid taking a course he or she would have excelled in or to take a course too advanced for his or her current skills. With this in mind, we perform a detailed analysis of the errors made in prediction and characterize the most significant shortcomings of the models tested in predicting student grades.

The rest of the paper proceeds as follows. Section II presents a literature review. Section III formulates the problem and Section IV presents the dataset. Section V lays out the methods we tested for grade prediction, and Section VI presents results and discussion for predictions on cold- and non-cold-start predictions. The section concludes with a detailed exploration of the errors being made by the FM model. Concluding remarks, and directions for future work are contained in Section VII.

II. Literature Review

Much of the research pertaining to MF for recommendation tasks has been done in the domain of e-commerce. In particular, the task of movie recommendation has been widely studied during the Netflix Prize Challenge [3]. Nearest-neighbor (NN) approaches and the use of Restricted Boltzmann Machines (RBMs) have also shown good results. Several recommender systems have been developed for reducing information overload, improving product recommendation, and targeting advertisements in industrial settings [4] [11]. Naturally, these systems use related techniques but on a different problem domain.

The application of recommender system technology, and particularly MF techniques, to student grade prediction is largely a new area of research. Similar work has mostly been done in online learning communities, such as Massive Open Online Courses (MOOCs) [12] [13] [14]. In contrast, the goal in the present study is to predict grades in a traditional university learning environment.

Perhaps the most similar work attempts to predict grades in a traditional university setting at the individual activity level. This includes predicting grades for assignments, quizzes, and tests. The authors employed a custom mixed-membership multi-linear regression model to predict student grades using a variety of data, including grades and learning management system (LMS) features [15]. The present study does not use data from or predict grades for individual activities. Instead, grades are predicted only at the course level using only past student grades.

In another study by Sorour et al., comments from learning reflection assignments in a course are used as features for grade prediction [16]. Latent Semantic Analysis (LSA) is used to extract topics and reduce dimensionality of the comments. Then k-means is used to cluster the comments into per-grade clusters. The features of this study consist only of bag-of-words term vectors and the problem is cast as a classification task. The present study does not leverage any term features and casts the grade prediction task as a regression problem.

In another vein of closely related research, Thai-Nge et al. [17] map student success outcomes to binary ratings and try to predict first successful attempts on course activities. The results on this task are compared to regression techniques to demonstrate improved performance when mapping items to knowledge componenents (skills required for a task). This mapping provides much denser data than the items by themselves. The authors disregard cold-start, leaving that for another paper. In comparison, our task is predicting student grades, regardless of pass/fail, and we predict for both cold-start and non-cold-start records.

III. Problem Formulation

Given a database of (student, course) dyads, our goal is to predict grades for each student for the next enrollment term. More formally, we have n students and m courses, analogous to users and items in the traditional recommender system setting. The (student, course) dyads comprise an $n \times m$ sparse grade matrix G. This is the primary source of information leveraged by matrix factorization techniques, such as Singular Value Decomposition (SVD). For those methods, the task is often cast as a matrix completion problem. However, we do not wish to complete the entire matrix. Our efforts here assume that domain knowledge will be available to filter the list of empty cells to those which are relevant to a particuar student. We are further restricted in our testing efforts to only those courses students select themselves, rather than all those which they could select from.

So we are predicting for a subset of cells in the matrix G. We consider each cell to be a (student, course) dyad and represent it as a feature vector $x_i \in \mathbb{R}^{n+m}$ and a grade $g_i \in [0, 4]$. The first n variables represent the one-hot encoded student IDs and the next m represent the one-hot encoded course IDs. This is simply the traditional data representation for regression problems. We train our models on all feature vectors x_i preceding the current term. Given the trained model, the task is to predict grades $\hat{g}_i$ for all feature vectors x_i in the current term. So each term we train a model on the cumulative grade matrix, then predict for that term. This is repeated for each term in sequence until we have made predictions for all grades.

IV. Dataset Description

The data used for this study comes from a public university (George Mason University) with an enrollment of 33,000 students as of Fall 2014 [18]. Observations begin in Summer 2009 and continue until Spring 2014, for a total of 15 terms. All students whose cohorts pre-date this time period and all non-transfer students were excluded from the data. In total, there are 13,845 students declared in one of

Table I
COLD START PROPORTION BY ACADEMIC TERM

Term	Term #	Dyads	NCS	CS	% CS
'09 Summer	0	59	0	59	100.00
'09 Fall	1	11,467	35	11,432	99.69
'10 Spring	2	11,171	8,962	2,209	19.77
'10 Summer	3	659	579	80	12.14
'10 Fall	4	21,459	9,289	12,170	56.71
'11 Spring	5	20,906	19,800	1,106	5.29
'11 Summer	6	1,531	1,378	153	9.99
'11 Fall	7	31,333	17,027	14,306	45.66
'12 Spring	8	30,680	28,899	1,781	5.81
'12 Summer	9	2,494	2,365	129	5.17
'12 Fall	10	40,896	27,626	13,270	32.45
'13 Spring	11	39,985	38,445	1,540	3.85
'13 Summer	12	3,595	3,450	145	4.03
'13 Fall	13	47,613	33,256	14,357	30.15
'14 Spring	14	46,709	45,698	1,011	2.16

CS: Cold-Start, NCS: Non Cold-Start, Term # denotes chronological ordering of academic terms in dataset.

105 majors, each of which belongs to one of 13 colleges. During this time period, these students have taken 3,060 unique courses, each of which is classified as one of 138 disciplines and taught by one of 4,287 instructors. After discarding records with no grades or grades which do not translate to the A-F scale (such as withdrawals and audits), we have 310,498 dyads. All this data was collected and anonymized in accordance with Institutional Review Board (IRB) policies.

The A-F letter grades have nominal equivalents in the range 0-4, so the target space is actually discrete. As a result, this problem could be cast as either classification or regression. While some classification methods were explored in preliminary experiments, these all fail to capture the ordinal nature of the data. In particular, an F (0.0) is lower than a D (1.0), and a D is much lower than an A (4.0). In order to properly capture this ordering information, we cast the problem as a regression task.

A. *Cold Start Predictions*

In the context of the next-term prediction task, cold-start records are (student, course) dyads for which either or both the student and course occur in a term and do not occur in any previous term. For our dataset, there are 73,748 (23.75%) cold-start records and 236,809 (76.25%) non-cold-start records. 9,998 (13.56%) of these are dyads for which both the student and the course are cold-start. 50,523 (68.51%) are student-only cold-start, and 13,227 (17.94%) are course-only cold-start. Table I breaks down the proportion of cold-start records by academic term.

V. METHODS

We explore two classes of methods for the next-term student grade prediction task. These are (1) simple baselines and (2) MF-based methods.

A. *Simple Baselines*

We devised three simple baselines to better understand the effect of leveraging three types of central tendencies in the data.

- *Uniform Random*: Randomly predict grades from a uniform distribution over the range [0, 4].
- *Global Mean*: Predict grades using the mean of all previously observed grades.
- *Mean of Means*: Average the global mean, the per-student mean, and the per-course mean.

The Uniform Random method illustrates the result of making predictions by randomly guessing. The Global Mean method illustrates the informative value of the overall central tendency of the data. The Mean of Means method takes this strategy one step further, also incorporating the per-student and per-course (row and column) averages. When compared to the Global Mean method, this illustrates the added benefit of a small level of personalization and historical knowledge about the course difficulty. In the cold-start setting, we allow the Mean of Means method to use whatever information is available. For a particular dyad, either the student or course may be present, but not both. If neither is present, it reduces to the Global Mean method.

B. *MF-based Methods*

We look at three MF-based methods:

- *SVD*: Standard singular value decomposition of rank k.
- *SVD-kNN*: SVD post-processed with kNN, as described in [19].
- *FM*: Factorization machine model, as described in [20].

Pure MF methods such as SVD are unable to make predictions for cold-start records. Completing a matrix that does not contain a particular user row or item column will not yield any information for that user or item. FMs can incorporate arbitrary side information while also leveraging the sparse user-item matrix. However, when applied to collaborative filtering problems, they suffer from the same cold-start issues. Each of these models attempt to capture the pairwise interaction of variables by decomposing the feature space into a k-rank reduced subspace. This results in two sets of latent vectors. One for the students ($v_s \in \mathbb{R}^k$), and one for the courses ($v_c \in \mathbb{R}^k$). We first look at SVD:

$$g_i = \sum_{f=1}^{k} v_{s,f} v_{c,f} = v_s^T v_c. \tag{1}$$

Each grade is simply predicted as the dot product of the latent student and course feature vectors. These latent feature vectors can be considered a less noisy, condensed representation of the student and course information. Paterek [19] showed that post-processing SVD with kNN yields improved predictive performance. The predicted grade for a course is replaced with the predicted grade for the course with the highest cosine similarity in the latent feature space.

We compare SVD and SVD-kNN to the FM model. With this data, the model reduces to the sum of global, student, and course bias terms and the factorized interaction of the student with the course. This last term is the same dot product seen in the SVD model. We first show the general equation and then the reduced equation for training on G:

$$\begin{aligned} g_i &= w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j'=j+1}^{p} x_j x_j' \sum_{f=1}^{k} v_{s,f} v_{c,f} \\ &= w_0 + w_s + w_c + v_s^T v_c. \end{aligned} \tag{2}$$

We observe that this model is able to capture all of the information captured by the simple baselines as well as the information captured by SVD. In general, it captures the global central tendency, 1-way linear relationships between the predictors and the grade (bias terms), and 2-way factorized interactions between each predictor and the grade. We use the libFM library [20] for a fast implementation of the FM algorithm. Markov Chain Monte Carlo (MCMC) is used for learning the FM model. We set all parameters using a grid search.

VI. Experimental Results and Discussion

In order to evaluate the performance of the various methods at the next-term course grade prediction task, we perform two seperate experiments. First we predict only for non-cold-start records. Then we predict for all records, including cold-start.

A. Metrics

Evaluations are performed in terms of two common regression metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{g}_i - g_i)^2}{N}} \qquad MAE = \frac{\sum_{i=1}^{N} |\hat{g}_i - g_i|}{N}$$

RMSE penalizes severe prediction errors more heavily than small ones. Given our task, we would prefer not to declare a method the best if it performs very well for half the students and very poorly for the other half. Hence RMSE is the metric we use to compare methods. MAE allows us to understand the range of grades we might actually be predicting. If a student is actually going to get a B, and the MAE is 0.33, we expect our model to predict either a B-, B, or B+. All MAE measurements are accompanied by the standard deviation of the absolute errors. Note that the absolute errors are not quite Gaussian, so it is not correct to assume approximate 95% confidence intervals between $\pm$ two standard deviations. All tables below list the results in order of best to worst RMSE.

Table II
Non-Cold-start Prediction

Method (Settings)	RMSE	MAE
FM (i=200, s=0.2)	**0.7758**	**0.5658 $\pm$ 0.5308**
Mean of Means	0.8594	0.6460 $\pm$ 0.5668
SVD-kNN (k=5)	0.9340	0.6911 $\pm$ 0.6283
SVD (k=5)	0.9585	0.6999 $\pm$ 0.6548
Global Mean	0.9587	0.7386 $\pm$ 0.6111
Uniform Random	1.8670	1.5388 $\pm$ 1.0574

Table III
Including Cold-start Prediction

Method (Settings)	RMSE	MAE
FM (d=8, i=200, s=0.1)	**0.8041**	**0.5892 $\pm$ 0.5472**
SVD-kNN (k=5)	0.9379	0.6927 $\pm$ 0.6323
Global Mean	0.9532	0.7332 $\pm$ 0.6091
SVD (k=5)	0.9560	0.6991 $\pm$ 0.6521
Mean of Means	1.0524	0.8127 $\pm$ 0.6687
Uniform Random	1.8459	1.5145 $\pm$ 1.0553

B. Non-Cold-start

The results for non-cold-start prediction are shown in Table II. We see the Uniform Random and Global Mean methods perform worst, as expected. It is somewhat surprising to see SVD perform almost the same as the Global Mean method. This is likely due to overfitting, which is a common issue with SVD [19]. While the post-processing is able to improve the SVD results, they have considerably more error than those produced by the Mean of Means method. This indicates the per-student and per-course biases are quite informative. The FM model outperforms all the others by a wide margin. So a combination of all the other methods proves to be most suitable for this task.

C. Cold-start

We first evaluate how well each model is able to predict all grades, including cold-start predictions. Then we pick out the best model (the FM) and determine how well it predicts for cold-start records vs. non-cold-start records.

For cold-start prediction (Table III), we see roughly the same trend as for non-cold-start predictions. SVD is now outperformed by the Global Mean method, but they are still close. In this setting, the Mean of Means method performs worse than the Global Mean method. Given that 86.44% of the records are predicted through a sum of the global mean and either the student or course bias, we can conclude that including only the student or only the course bias can lead to overly confident predictions about a new course or student, respectively. In this setting, the FM model is relying on essentially the same information as the Mean of Means method, but it is able to avoid overconfident learning patterns through the use of Bayesian complexity control (integrating out the regularization hyperparameters).

We next look at cold-start error vs. non-cold-start error for the FM method. Table IV lays out the results. The FM

Table IV
COLD-START (CS) VS. NON-COLD-START ERROR

Group	Dyad %	RMSE	MAE
Non-CS	76.27	0.7758	0.5659 ± 0.5306
CS student-only	16.27	0.9106	0.6929 ± 0.5909
CS course-only	4.26	0.7623	0.5730 ± 0.5028
CS both	3.20	0.9535	0.7095 ± 0.6370

results for course-only cold start are better than those for non-cold-start records. This indicates that the student trends are much more important for predicting next-term grades than the course trends. However, the discrepancy between the student-only cold start and the cold-start records with no prior information is larger than the discrepancy between non-cold-start and course-only cold-start. This indicates that the course trends are still quite useful in the absence of any prior observations for a student. The worst predictions are those made on student cold-start records. This is interesting evidence that the student cold-start prediction problem will be the most challenging subproblem of the next-term student grade prediction problem.

D. Error Analysis

While making a poor product rating prediction is not desirable, it is much less desirable to make a particularly poor course grade prediction. Doing so could potentially lead a student to avoid taking a course he or she would have excelled in or to take a course too advanced for his or her current skills. With this in mind, we seek to better understand the distribution of error in our predictions. Fig. 1a lays out the distribution of error in terms of RMSE in a cohort by term number matrix. The cohort is the term the student was admitted to the university, while the term number is the objective chronological ordering of the academic terms in the dataset. Each cell represents the aggregate RMSE for all grade predictions for the students admitted in a particular cohort taking courses in a particular academic term. The bar chart on top of the heatmap shows the per-term RMSE, and the bar chart on the right shows the per-cohort RMSE.

Note that we have excluded Summer terms, which account for only 2.67% of the predictions. This clarifies the trends observed. It is also important to note that Spring cohorts account for only 1.10% of the total predictions. Hence they have much more variability than the Fall terms. Fig. 1b shows a heatmap of the dyad counts for each of the cohort/term cells.

Looking down the diagonal, we see that predictions for each cohort for the first term are generally poorer than predictions for subsequent terms. This reflects the increased difficulty of cold-start predictions, since these are all cold-start students and may also be cold-start courses. Following each row from left to right, we expect the predictions to decrease in error as we accumulate more information about the students in this cohort. With the exception of the Fall

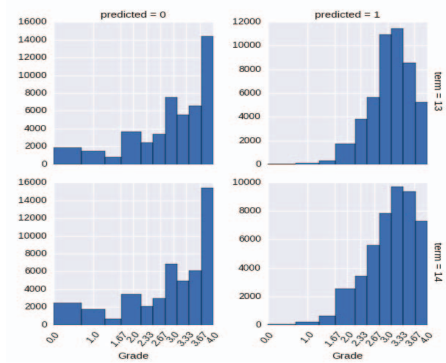


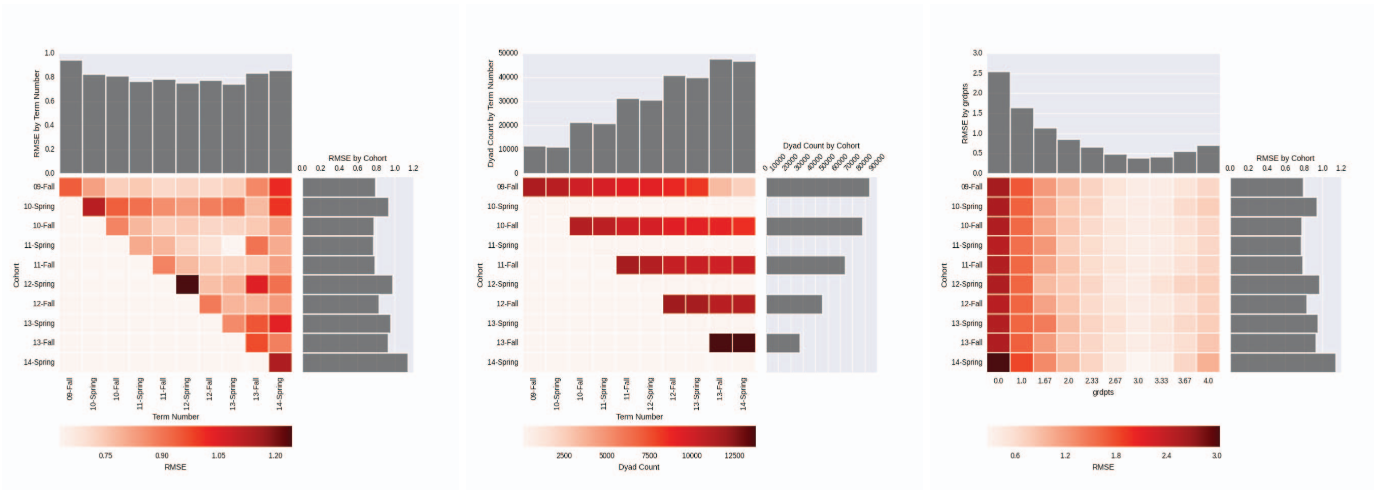
Figure 2. Actual vs. Predicted Grade Distribution For Last Two Terms

of 2013 and Spring of 2014, we do see this trend. So in general, our predictions improve as we accumulate more historical grade data. By these last two terms, many students are beginning to graduate from the original cohorts and enrollments are higher than ever before. As a result, the characteristics of the students are beginning to shift. In particular, we see many more new students (<30 credit hours) in these terms than in previous terms. Predictions for these students are generally worse.

While the FM model is able to learn effectively as it observes more data and gradually increase predictive performance, it has one serious shortcoming which is shared by all the other models tested. Briefly, none of them predict well for failing grades. Fig. 1c shows a cohort by grade points heatmap of RMSE. Each distinct grade value a student can earn is listed in order on the x-axis. So each cell of this heatmap represents the RMSE obtained on predictions for students in a particular cohort who earned a particular grade. Fig. 2 shows the actual vs. predicted grade distribution for the last two terms (Fall 13 and Spring 14). All other terms, excluding the very first, which involved only cold-start predictions, show the same trend. All of the models we experimented with produce a grade distribution which is centered on 3.0 and tapers off smoothly in both directions. This is quite clear in Fig. 1c, which shows almost no error around 3.0 grades and increases noticeably in either direction from there. The issue here is that the grades themselves do not taper off smoothly. We see an almost exponential trend from 2.0 to 4.0, with dips and jumps in between. 4.0 grades are far and above the most common grade, and predictions for these are also somewhat poor.

VII. CONCLUSIONS AND FUTURE WORK

Motivated by the need for institutions to retain students, ensure timely graduation, and ensure students are well-trained and workforce-ready in their field of study, we have developed a system for next-term course grade prediction. After experimenting with a variety of CF models, we determined that the MF-based FM model is best-suited to this task. Using this model, we can predict grades for both



(a) (b) (c)

Figure 1. Cohort by term number error visualization. (a) Error breakdown for FM predictions. (b) Dyad counts, excluding summer/spring terms. (c) Cohort by grade points error breakdown for FM predictions

new and returning students and for both new and existing courses. While this system significantly outperforms random guess predictions, and noticeably outperforms the CF models tested, it still has limitations we seek to address in future works.

In general, our predictions improve as we acquire additional grade data. However, the system seems slow to adapt to shifting characteristics in the student population as a whole. Learning global, 1-way, and 2-way interaction trends on separate time windows of the data may help compensate for this. Additionally, we have found that our system predicts poorly for failing grades, which is a critical component of effective grade prediction. The first step in addressing this problem is to incorporate side information and reassess from there. Afterwards, we plan to experiment with pass vs. fail grade prediction in the future. If we can accurately classify failing dyads, we expect the present system to predict very well for the remaining passing dyads. Once we have improved predictive performance for failing grades, we plan to deploy this system for live use and perform A/B testing to better understand its performance and the effect it has on student decision-making.

ACKNOWLEDGMENT

This research was funded by NSF IIS grant 1447489.

REFERENCES

[1] S. Aud and S. Wilkinson-Flicker, *The Condition of Education 2013*. U.S. Department of Education, National Center for Education Statistics., 2013, no. NCES 2013-037.

[2] N. R. Council, *Building a Workforce for the Information Economy*. National Academies Press, 2001.

[3] R. M. Bell, Y. Koren, and C. Volinsky, "The BellKor solution to the netflix prize," *AT&T Labs Research*, 2007.

[4] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80, 2003.

[5] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Physical Review E*, vol. 64, no. 4, 2001.

[6] J. F. Huete, J. M. Fernndez-Luna, L. M. de Campos, and M. A. Rueda-Morales, "Using past-prediction accuracy in recommender systems," *Information Sciences*, vol. 199, pp. 78–92, 2012.

[7] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 145–186.

[8] H. Shan and A. Banerjee, "Generalized probabilistic matrix factorizations for collaborative filtering," in *ICDM '10*, 2010, pp. 1025–1030.

[9] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proc. of the 24th ICML*. ACM, 2007, pp. 791–798.

[10] L. Alsumait, P. Wang, C. Domeniconi, and D. Barbar, "Embedding semantics in LDA topic models," in *Text Mining*, M. W. Berry and J. Kogan, Eds. John Wiley & Sons, Ltd, 2010, pp. 183–204.

[11] B. Kanagal, A. Ahmed, S. Pandey, V. Josifovski, J. Yuan, and L. Garcia-Pueyo, "Supercharging recommender systems using taxonomies for learning user purchase behavior," *Proc. VLDB Endow.*, vol. 5, no. 10, pp. 956–967, 2012.

[12] C. Romero, S. Ventura, and E. Garca, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education*, vol. 51, no. 1, pp. 368–384, 2008.

[13] A. Pea-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[14] C. Romero, M.-I. Lpez, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Computers & Education*, vol. 68, pp. 458–472, 2013.

[15] A. Elbadrawy, R. S. Studham, and George Karypis, "Personalized multi-regression models for predicting students' performance in course activities," *LAK, '15*, 2015.

[16] S. E. Sorour, T. Mine, K. Goda, and S. Hirokawa, "A predictive model to evaluate student performance," *Journal of Information Processing*, vol. 23, no. 2, pp. 192–201, 2015.

[17] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 1, no. 2, pp. 2811–2819, 2010.

[18] "Gmu facts 2014-2015 facts and figures," 2014. [Online]. Available: https://irr.gmu.edu/FastFacts/#stud

[19] Arkadiusz Paterek, "Improving regularized singular value decomposition for collaborative filtering," *KDD*, 2007.

[20] S. Rendle, "Factorization machines with libFM," *ACM TIST*, vol. 3, no. 3, 2012.