A College Major Recommendation System

Samuel A. Stein
Computer and Info. Science Department, Fordham University, NY, USA
sstein17@fordham.edu

Gary M. Weiss
Computer and Info. Science Department, Fordham University, NY, USA
sstein17@fordham.edu

Yiwen Chen
Computer and Info. Science Department, Fordham University, NY, USA
ychen638@fordham.edu

Daniel D. Leeds
Computer and Info. Science Department, Fordham University, NY, USA
dleeds@fordham.edu

ABSTRACT

College students are required to select a major but are often provided with only a modest amount of support in making this important decision. A poor decision is detrimental to the student, since it may result in the student later switching to a different major with a delay in graduation—or even result in the student leaving the university. This also impacts the university since time to graduation and retention rate are used to evaluate the quality of a university. There is a general lack of research on recommender systems for college majors, with the most relevant systems focusing on course-level recommendations. This study describes and evaluates a recommender system for selecting an undergraduate major, utilizing nine years of historical student data from a large university. The system bases its recommendations on the courses that the student takes in the first few years of college, and how well they performed in these courses. The system is designed to recommend majors that the student is likely to be interested in and will perform well in. Recommendations are evaluated based on the likelihood that the student's actual major was in the top five recommended majors, and whether the student performed above average in that major. The recommendation system dramatically outperforms the baseline strategy of randomly selecting a major, and when the recommendation is followed the student is 12% more likely to perform above average in the major.

CCS CONCEPTS

• **Information systems**; • **Information systems applications**; • **Data mining**; • **Collaborative filtering**; • **Decision support systems**; • **Data analytics**; • **Applied computing**; • **Education**;

KEYWORDS

Recommender systems, collaborative filtering, nearest neighbor, educational data mining

ACM Reference Format:
Samuel A. Stein, Gary M. Weiss, Yiwen Chen, and Daniel D. Leeds. 2020. A College Major Recommendation System. In *Fourteenth ACM Conference on Recommender Systems (RecSys '20), September 22–26, 2020, Virtual Event, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3383313.3418488

1 INTRODUCTION

Early in their college career, a student must make the life-changing decision of choosing a major. This decision may be made with minimal guidance. Students may infer basic strategies—e.g., if they are good at math they should consider a STEM (Science, Technology, Engineering, and Math) discipline—but these strategies may not leverage historical data maintained by the university. Schools that use historical data often do so in an ad-hoc manner. This paper describes a recommender system that uses historical data to recommend a major that the student is likely to be interested in and perform well in.

This paper describes and evaluates a recommendation system that utilizes collaborative filtering to recommend a set of undergraduate major disciplines. The system bases its recommendations on the courses that a student chooses to take in their first few years of college, and how well they perform in those courses. The system is designed to recommend majors consistent with the courses selected and for which the student performs well. Recommendations are evaluated based on the likelihood that the student's actual major is in the top five recommended majors, and whether the student performs above average in that major. The recommendation system is based on data for over 18,000 students, collected over a nine-year period from the undergraduate colleges of Fordham University, which is based in New York City.

This work is noteworthy because it applies recommender system technology to a relatively new and important application domain. The use of such an approach can greatly benefit students by guiding them in their choice of major. The results in this paper show that this recommendation system can lead to substantially improved student performance within their major. However, choosing a major is a significant life-choice and we do not advocate leaving such an important decision up to an automated system; rather we view this system as a resource to assist the student in making an informed choice.

2 RELATED WORK

Recommender systems have been used within the education domain for a variety of tasks, such as selecting courseware for specific courses [5], choosing specific content to improve the learning

RecSys '20, September 22–26, 2020, Virtual Event, Brazil
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7583-2/20/09.
https://doi.org/10.1145/3383313.3418488

process [6], selecting courses that will allow for certain skill development [13], and selecting programs to which a student should apply in order to maximize chances of acceptance and funding [14]. However, there is a notable lack of research for selecting a student major. The most relevant research involves recommending majors based on the student's interests, character and/or skills [16][17].

Our proposed system is an example of educational data mining, an emerging field that has its own society (educationaldatamining) and associated international conference and journal. Educational data mining includes a great deal of research on predicting student academic performance, but most of the research involves predicting performance in a single course. This prior work has used a variety of methods, including Bayesian networks [7], logistic regression [8], neural networks [11] and decision trees [9]. Methods from recommender systems have also been utilized to predict student performance [10].

3 METHODOLOGY

This section describes the education data set, the data preparation steps, the process used to identify a set of majors to recommend for a student, and the metrics used to evaluate the recommendations.

3.1 The Raw Education Data Set

The raw data contains 473,256 student-course records, where each record describes the performance of one student in one class. The data was obtained from Fordham University and covers a nine-year period. Due to privacy concerns, no non-academic information is contained within the dataset. The following fields were used in this study: student id, course identifier, college name (Fordham has multiple undergraduate colleges), graduation year, student major, and student course grade (using a 4-point scale). All undergraduate courses are numbered so the first digit indicates the intended year in which to take the course (i.e., 1000 level courses for the first year, 4000 level courses for the fourth/final year). Records not belonging to the largest undergraduate college were discarded to facilitate system evaluation, as some majors are college specific.

The data set contains 68 current majors (discontinued majors are ignored). The popularity of the majors varies widely as shown in Figure 1. The percentage of students in each major is specified by the values on the left y-axis, while the cumulative totals of the first n most common majors is shown using the red curve and values on the right y-axis. Only students that start and complete their degree within the nine-year period are included and that leaves 7,187 students. The popularity of majors is heavily skewed, with the first 28 majors covering 90% of the students. Foreign languages (e.g., “German”, “Arabic”) could be aggregated into a single major, but for this analysis are kept separate. This contributes to the exponential decay shown in Figure 1. Future work will examine such granularity issues.

Fordham university has a large required core curriculum, which includes diverse topics such as mathematics, social science, and performing arts. For this study all courses are labelled as “core” or “non-core”. We do not use non-core courses in our recommendation system since this would “bias” our system and lead it to making uninteresting recommendations (i.e., students taking “Computer Algorithms” would often be recommended to major in Computer Science). We believe that it is more interesting and useful to make recommendations based on courses that fulfill common requirements—but it should be understood that this does impact the performance of the system.

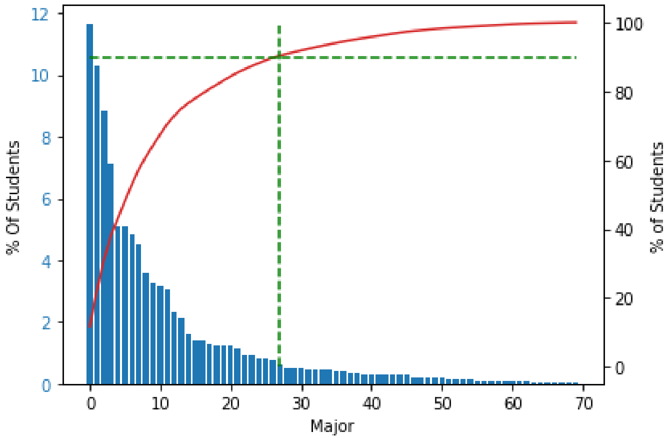

Figure 1: Distribution of student majors

3.2 Data Preprocessing and Transformation

As mentioned earlier, the popularity of majors varies widely, with several majors rarely selected. Including the most unpopular majors in the recommendation system degrades the performance of the system, as there is insufficient data to make informed decisions for these majors. Consequently, our initial system excludes the least common majors, which collectively covers 10% of students; this leaves the 28 most popular majors. Future work will relax this constraint.

The system only utilizes course information for the first two years of study to make recommendations, since most students declare a major near their third year. Furthermore, we only consider core courses for our primary analysis. Student performance within their major is considered in our evaluation process. This is measured by averaging the grades from all classes in the student's major, and then normalizing it to form an nGPA (normalized Grade Point Average) with a mean of 0 and a standard deviation of 1.

The recommender system requires the data to be at the student level, not the student-course level. This transformation is accomplished by aggregating the student-class records for each student. The relevant course records are determined based on the course level and type (core or non-core). The resulting records for our main analyses have 63 fields, which include all 1000 and 2000 level core courses, as well as the nGPA. This data set contains 4,562 student records; 10% of the students are dropped because they major in an uncommon discipline, and other records are dropped because some students had too few courses, suggesting that they transferred from another university.

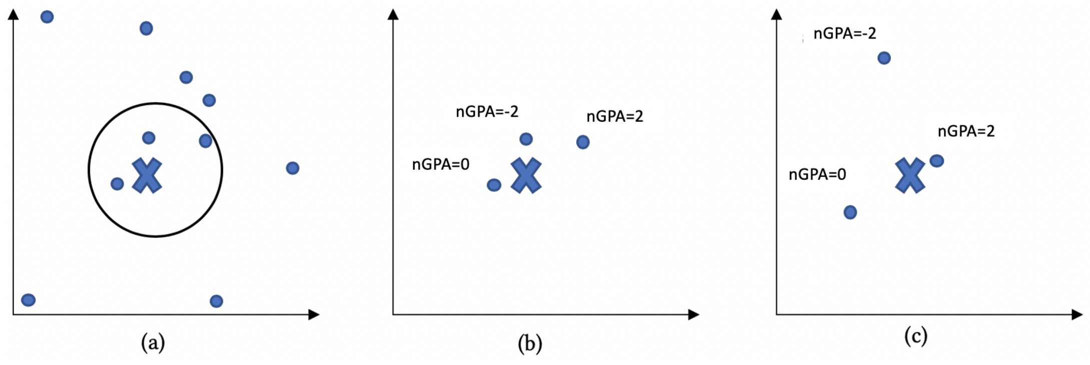

Figure 2: (a) initial choice of n nearest neighbors, (b) nGPA of neighbors, and (c) scaled distance

3.3 Methodology for Determining which Majors to Recommend

Our recommender system utilizes a nearest neighbor approach, where similarity is based on the transformed records described above. Previous research demonstrated that the distance metric has a large impact on performance [2]. For this reason, we explored a variety of distance metrics; due to space limitations, this paper focuses only on *CosD'*, adjusted cosine distance. The cosine distance computes the difference in grade ratios—a student who consistently performs poorly appears similar to a student who consistently performs well. Hence, the closest *n* students, with all other students discarded, have their distances rescaled using the scaling function in Equation 1.

$$CosD'(x, y) = CosD(x, y) \times \left(1 - \tanh\left(nGPA_y \times g\right)\right) \tag{1}$$

CosDx,y is the original cosine distance between students *x* and *y*, *nGPAy* CosD(x is the nGPA for student *y*, and *g* is a scaling factor that modulates the effect of tanh, with higher values increasing the effect of student performance. The effect of Equation 1 is visualized in Figure 2, where students who have a better normalized GPA are given more weight (smaller distance).

The goal of our system is to determine the distance to prospective major disciplines by looking at the majors of the closest students, and then combine those values. For major M_i, we compute $MajorDistance_i$ using the formula below, where $d_{M_i,j}$ represents the j$^{\text{th}}$ distance for a student in major i, n_i is the number of nearby students in major i, and *z* is the exponent factor to the number of students voting for the major. z is bound between values 1 and 2, with values closer to 2 leading to greater emphasis placed on the number of times a major is voted for instead of the vote distance

$$Major\ Distance_i = \frac{\sum d_{M_i,j}}{n_i^z} \tag{2}$$

3.4 System Evaluation Metrics

Evaluation should consider how the recommendation will be used. We want to evaluate (weighted) accuracy of predicting the selected major and likelihood of academic success in the predicted major. Selection prediction accuracy is a common measure for recommender systems. For movie recommendations, this could be measured by whether a recommendation was followed or was in the user's top *n* movies. Here, we measure the fraction of students whose actual majors are within the top *n* recommendations. We refer to this as major recommendation accuracy, or just accuracy.

Due to a large variance in the number of students within each major, accuracy is not necessarily the most informative metric, just as accuracy is usually not the primary evaluation metric for imbalanced classification problems. As in the class imbalance case, we focus on the F1 score. To calculate the F1 score, where there are *n* majors, and r_i and p_i are the recall and precision associated with each major, we use the following formula:

$$F1\ score = \frac{1}{n}\sum_{i=1}^{n}\frac{2r_i p_i}{r_i + p_i} \tag{3}$$

Finally, just because the major we recommend turns out to be the major selected by the student, does not necessarily mean that it is a suitable major. We assess the suitability of a major based on whether the student's performance in the major is at or above the average performance of students in that major (i.e., nGPA ≥ 0). We use the following two metrics to factor in academic performance, focusing on the first metric that represents the fraction of recommended majors in which the student performs better than average.

$$\text{Quality of Recommendation (QOR)} = \frac{\sum \text{Major Recommended nGPA} \geq 0}{\sum \text{Major Recommended}} \tag{4}$$

$$\text{Quality of Not Recommended (QONR)} = \frac{\sum \text{Major Not Recommended nGPA} \geq 0}{\sum \text{Major Not Recommended}} \tag{5}$$

4 RESULTS

This section presents the results for the recommender system experiments. Section 4.1 provides the main results with the "standard" set of parameter values. briefly explores the impact of varying the system parameters. All results are based on leave-one-out cross validation, which means each student recommendation is based on data from all other students.

Table 1: Recommendation Results

Evaluation Metric	Major Recommendation Strategy			
	Random	Most Common	Actual Major	Recommender System
Recommended & nGPA ≥ 0 (QOR)	55%	55%	55%	67%
Recommended & nGPA < 0	45%	45%	45%	33%
Not Recommended & nGPA ≥ 0 (QONR)	55%	55%	55%	44%
Not Recommended & nGPA < 0	45%	45%	45%	56%
Student Major Accuracy	18%	39%	100*	61%
Mean F1-Score	18%	18%	100*	42%

∗ These values are guaranteed to be 100% based on how accuracy was defined. As discussed earlier, these values are of limited utility since it is not known that the actual student major is the best major for the student.

4.1 Main Results

The results using our standard parameter settings are summarized in Table 1. Specifically, the adjusted cosine distance metric is used with g=0.5, 50 nearest neighbors are used, and accuracy is measured using the top 5 recommendations. As mentioned earlier, the 10% of students in the least common majors are excluded, and predictions are based only on core courses taken within the first two years.

The results for the recommender system are provided in the last column. The three other recommendation strategies provide useful baselines, in increasing order of sophistication: "Random" randomly selects a major, "Most Common" selects the most common major, "Actual Major" selects the major the student actually chose, and "Recommender System" uses the recommendations produced by our system. The results in Table 1 show that our recommender system performs quite well, and outperforms all of the baseline strategies on the QOR metric, which we view as the best evaluation metric—and on this metric our system even outperforms the student's actual choice (67% vs. 55%). That means that in the cases when our system recommends a major and the student chooses that major, then the student is more likely to perform well in the major. The metrics associated with what happens when we do not recommend a major are provided for completeness but are not as important. With respect to the student major accuracy and mean F1-Score, our system significantly outperforms the two simplest baseline strategies; however, it cannot outperform the strategy of choosing the actual major, since that will always be correct. As mentioned before, the accuracy and F1-Score metrics are of limited utility since we cannot say that the actual major that the student chooses is the best one. The system's ability to include the student's actual major in the top 5 recommendations 61% of the time suggests that the recommender system is making reasonable recommendations. Overall, we can say that if the students follow the system's recommendation, they will generally improve their academic performance.

4.2 Effect of Parameters

Experiments were executed to explore different parameter settings, but due to space limitations we can only present high level summaries of these experiments. The distance metric was found to be relatively insensitive—four other distance metrics were evaluated but the performance never varied by more than 3%. Increasing the number of nearest neighbors led to improvements in accuracy but a decrease in F1-score, as the higher values caused the system to focus more on the most common majors. Varying the number of years of core courses demonstrated that information necessary for good recommendations occurs within the student's first two years of college (more years did not help). However, including non-core courses allowed for significant gains on all measurable results; but as mentioned before this is not surprising since this information often encodes the students actual major and how they will perform in it. As for the parameters z, the exponent within the function describing the overall vote for a major, and g, tanh scaling coefficient, this was found to have a local minimum within the 1-2 range, which was used.

5 MAJORS WITH SIMILAR RECOMMENDATIONS

This paper focuses on describing and evaluating a system for recommending college majors. However, as a side-effect of its operation, the system can also generate some descriptive data mining results, by showing which majors are close together from a recommendation perspective. That is, if several majors are often recommended together, then in some sense they are related. This can be useful to college administrators and advisors and may provide insight into the relationships between disciplines. The results in Figure 3 visually depict the proximity of a group of majors from the "Communications" and "Neuroscience" majors. In each figure, the proximity to other majors from the designated major is based solely on the *radius* of the other major.

Figure 3Left shows that the communications major is closest to the psychology and English majors. The proximity to English is clear since they two majors do share much in common. The connection to psychology may not be quite as clear, but psychology is not considered a "hard science" and, like communications, leads to a Bachelor of Arts degree. With respect to Figure 3Right, we see that the four closest disciplines to the interdisciplinary neuroscience major are biology, chemistry, international political economy, and psychology. Biology and psychology are both participating department in the neuroscience major and play a central role in the major. Chemistry is not quite as central, but neuroscience students must study chemistry. Only the connection to the international political economy major is a bit odd, but that major is also interdisciplinary and perhaps attracts similar types of students.

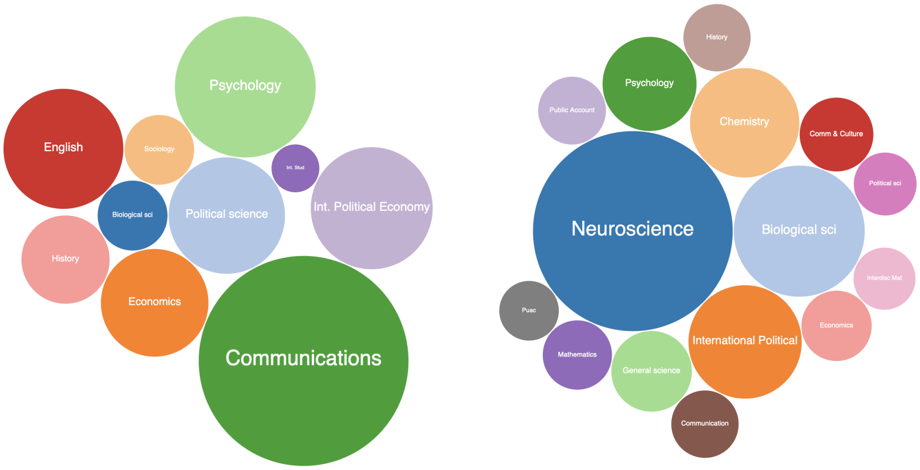

Figure 3: Majors recommended – Left: Communications and Right: Neuroscience Major

6 CONCLUSION

This paper describes and evaluates a system for recommending undergraduate majors. The evaluation used nine years of undergraduate student data from a large US university. Recommendations were based on students who took similar courses and performed similarly, as well as how the student was projected to perform in the major. The results demonstrate that it is possible to generate reasonable recommendations utilizing two years of core curriculum courses. These results show that 61% of the time the student's actual major was within the top-5 recommended majors, and that if the student follows the recommendation then 67% of the time they will outperform the average student in the major. Notably, when students follow the system's recommendation, they are 15% more likely to outperform the average student than when they choose their own major. This system is well-positioned to be used as a tool to assist students as they explore potential majors. The code that implements the recommender system is publicly available [15].

This is an early study and the work can be expanded in several ways. The methodology can be applied to data from other universities. Methods other than nearest neighbor can also be evaluated. Finally, we plan to more fully explore the use of different parameter settings and consider additional information in the decision-making process. Nonetheless, we do believe that our approach of basing decisions mainly on courses taken and grade performance is an interesting approach, which differs from most previous work.

7 REPRODUCIBILITY STATEMENT

The data used for this research study includes nine years of undergraduate student course and grade data. While student identifiers were anonymized, the low-level data is still too sensitive to share. However, we do provide a public copy of our software [15] so that our methodology can be reproduced and applied to other data.

REFERENCES

[1] Asad M Madni, Recommender Systems in E-Commerce, 2014, 2 014 World Automation Congress

[2] V. B. Surya Prasath, Haneen Arafat Abu Alfeilate , Ahmad B. A. Hassanat, Effects of Distance Measure Choice on KNN Classifier Performance - A Review, 2019

[3] J. Ben Schafer, Joseph Konstan, John Riedl, Recommender Systems in E-Commerce, University of Minnesota

[4] Mariela Tapia-Leon,Sergio Lujan-Mora, Recommendation Systems in Education: A Systematic Mapping Study,2018, Proceedings of the International Conference on Information Technology & Systems

[5] E. Garc ía, C. Romero, S. Ventura, C. D. Castro, An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering, User Modeling and User-Adapted Interaction 19 (1-2)

[6] N. Soonthornphisaj, E. Rojsattarat, S. Yim-ngam, Smart E-Learning Using Recommender System, in: International Conference on Intelligent Computing, 518–523, 2006

[7] R. Bekele, W. Menzel, A Bayesian Approach to Predict Performance of a Student (BAPPS): A Case with Ethiopian Students, in: Artificial Intelligence and Applications, Vienna, Austria, 189–194, 2005.

[8] H. Cen, K. Koedinger, B. Junker, Learning Factors Analysis A General Method for Cognitive Model Evaluation and Improvement, in:Intelligent Tutoring Systems, vol. 4053, Springer Berlin Heidelberg, ISBN 978-3-540-35159-7, 164–175, 2006

[9] N. Thai-Nghe, P. Janecek, P. Haddawy, A Comparative Analysis of Techniques for Predicting Academic Performance, in: Proceeding of 37th IEEE Frontiers in Education Conference (FIE'07), Milwaukee, USA, IEEE Xplore, T2G7–T2G12, 2007

[10] N. Thai-Nghe, D. Lucas, A. Krohn-Grimberghe, L Schmidt-Theme, Recommender Systems For Predicting Student Performance, in: Procedia Computer Science 1 (2010) 2811-2819; University of Hildesheim, Hildesheim, Germany

[11] C. Romero, S. Ventura, P. G. Espejo, C. Hervs, Data Mining Algorithms to Classify Students, in: 1st International Conference on EducationalData Mining (EDM'08), Montral, Canada, 8–17, 200

[12] N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, R. Koper, Recommender Systems in Technology Enhanced Learning, in: Kantor,P.B., Ricci, F., Rokach, L., Shapira, B. (eds.) 1st Recommender Systems Handbook, Springer-Berlin, 1–29, 2010.

[13] B Bakhshinategh, G Spanakis, O Zaiane, S Elatia, A Course Recommender System Based on Graduating Attributes, in: 9th International Conference on Computer Supported Education

[14] M Hasan, S Ahmed, D Abdullah, S Rahman, Graduate School Recommender System: Assisting admission seekers to apply for graduate studies in appropriate graduate schools, in: International Conference on Informatics, Electronics and Vision (ICIEV), 2010

[15] GitHub Repo of code available at the github repo of Guigetzu1224

[16] M. R. M. Arroyave, A. F. Estrada, A college degree recommendation model, Universidad Nacional De Colombia (2015); Universidad de Guayaquil

[17] C Obeid, I Lahoud, H.E. Khoury, P Champin, Ontology-based Recommender System in Higher Education, The Third Edition of Educational Knowledge Management Workshop (2018); Lyon, France