Analyzing collaborative filtering for UNED first-year student enrolment Recommendation system

Adrián Clavero[1], Víctor Fresno[2], Fernando Latorre Torres[2] and Salvador Ros[3]

[1] *Centro Asociado de la UNED en Barbastro, Spain*
[2] *Computer Systems and Languages Department, ETSI Informática, UNED, Spain*
[3] *Communication and Control System Department, ETSI Informática, UNED, Spain*

Abstract

First-year students enrolment is an important situation that can influence their future results. An adequate decision about how many and which subjects are the most adequate for the personal characteristic of a first-year student contribute to decreasing the dropout and improving their results. This work presents the study of different algorithms to develop a collaborative filter recommendation system for UNED first-year students. Two algorithms have been evaluated and analyzed. The best algorithm for the recommendation system is based on cosine similarity improving in the best scenario up to 50% of the results.

Keywords

Recommendation Systems, learning analytics, dropout, education.

1. Introduction

National Distance Education University (UNED) is a distance university characterized by its special methodology and the special profile of its students. UNED students' profile is a student in his thirties combining his studies with professional activity or family reconciliation. This special profile determines the way these students enroll at the university. While the sophomores usually choose a few subjects and select them carefully according to a principle of effectiveness, the first-year student tends to enroll in the whole year, or if they decide on a partial enrollment, they haven´t enough information apart from the syllabus to make the most effective enrolment.

Analyzing the public data of the UNED´s statistical portal (https://app.uned.es/evacaldos/), it is remarkable to compare the evaluation rates between the first two years and the last two years, the latter being clearly higher. These data could be interpreted as the first-year student hasn't had time enough to study the subject or even selected subjects that needed previous content delivered in another subject. Anyway, the bad planning of the study is behind these data. On the contrary, the high rates among the sophomores suggest better planning of the study, selecting more carefully, and according to their needs, the subjects, obtaining a better optimization of their efforts and results.

This paper presents an enrollment recommendation system for the UNED´s first-year students, whose objective is to suggest the number of subjects to enroll based on the individual features of the students and conducted to get the best academic results and reduce the university dropout rate. This work has been carried out as a Final Degree Project in the Computer Engineering Degree at the UNED. The data used for the analysis were provided by the central university services, having been previously anonymized according to the RGPD.

Learning Analytics Summer Institute Spain (LASI Spain) 2022, June 20–21, 2022, Salamanca, Spain
EMAIL: aclavero@barbastro.uned.es (A. 1); flatorre@barbastro.uned.es (A. 2); vfresno@scc.uned.es (A. 3); sros@scc.uned.es (A. 4)

2. State of the art

Nowadays, we can state that recommender systems are a cornerstone of many successful new business models. Amazon, YouTube, Spotify, Netflix [1] and all the content streaming companies use recommenders to improve their offer and build customer loyalty. For this purpose, they have access to customer behavioral data and analyze them to recommend their products efficiently, increasing their profits. In the same way, it is possible to use these systems to improve different scenarios in the learning processes. According to [2], recommendation systems have been applied to other areas of the education field, being the academic election the area of more significant application. This area includes tasks such as selecting a university, course, or specific discipline. Other education areas would be related to academic performance, content or learning resources trying to predict the content preferences according to the student profiles [3].

The enrolment of students is a matter of interest for universities since a good selection of the subject for enrolling contributes to a good student performance and decrease the dropout rate.

The most common technique used in recommendation systems is the collaborative filter [4-7].

This technique uses the user information and their correlations to make a recommendation. Suppose this information is categorized by the users' preferences. In that case, we are talking of user-based filtering, which tends to group users with the same preferences hypothesizing that if they have, then they require similar products. On the other side, we consider an element-based filter if we consider the product rating patterns. The collaborative filter provides better results when there is a large amount of data and, on the contrary, has problems when starting with new elements [8]. Another disadvantage of this technique is the cold-start effect and the sparsity. The cold-start problem arises when a new user starts using the system and has very little information about him. The sparsity implies a lack of data or very irregular data. This problem is relevant in collaborative filtering systems since they are based on user data [9].

Other approaches are, on the one hand, content-based recommenders whose primary goal is to recommend products like those that the users have used and liked. This approach is like the element-based collaborative filter with the main difference that it only collects the data from the user to whom it recommends. Content-based filtering has the advantage of not depending on the preferences of other users since it is based on the comparison between elements. The disadvantages of this technique are similar to collaborative filtering [9]. On the other hand, we have knowledge-based recommenders. These recommenders make deterministic recommendations and are not affected by the introduction of new elements and use information obtained from the previous different actions of the user. The information needed for this technique is captured using different methods, and usually, this acquisition of knowledge has a high cost in terms of computation, time and resources [9].

Finally, the hybrid recommender combines different approaches to obtain more robust recommendation models using the main advantages of the used models and decreasing the negative effect of the chosen approaches. The approach used for the recommenders' applications is mostly hybrid, so different techniques are combined to obtain the result, being the collaborative filter the individual techniques most used. It should be noted that not the use of this approach depends on the available data, and it is not always possible to apply. As we highlighted before, these systems are time-consuming, so it could require a long execution time depending on the algorithm and the amount of data it processes. In these cases, it would be convenient to evaluate the results with different data volumes to optimize results and execution time.

Other considerations about recommender implementations are [9] a) the scalability, the larger the dataset, the efficiency could be decreased, so the application of big data architectures must be present, b) privacy protection, especially in the learning process. The data must be protected effectively, and the system doesn´t have to require more data than needed, c) Over-specialization.

This problem occurs when there is no diversity in the pattern of recommendations, and the chances of the user discovering something beneficial are practically nil d) Gray-Sheep, this effect refers to a user who does not show a clear preference or has inconsistent behavior. Therefore, the recommendation becomes difficult, and, thus, the effectiveness of the recommender system decreases.

3. UNED enrollment recommender system

The proposed recommender is focused on the task of academic enrollment of UNED first-year students. The recommender system will help first-year students choose the number of subjects that best suits their characteristics. We have selected this area of application because it is a crucial point in preventing dropouts, as can be seen in the data of the UNED statistical portal. These data reflect a significant difference between the number of enrolled and presented students in the first-year subjects.

The recommender system developed uses a collaborative filter technique as we rely on students' experiences from previous courses to make the recommendation. We discarded the content-based filter as we focused on indicating a range of subjects rather than specific subjects. In the case of knowledge-based systems, they do not suit our needs as we want to consider the experience of students from previous years, which provides excellent information.

Since this is a first approach to solve the problem, our recommender works without subject`s information regarding their complexity so It only can recommend a rank of number of subject that could guarantee a good performance.

3.1. Dataset

One of the principal problems when we face the implementation of a recommender in a university, is to obtain the data [10]. In this case, this project has been supported by the Vice-rectorate of technologies offering us total access to the data in an anonymized way. All the students' identifiers have been modified not to be able to identify a unique student.

Figure 1 shows the entity relation of the available information. This scheme shows we have available information about subjects, enrolment, students, evaluation performance, access to the university genre and UNED associate centers. These tables contain information from the last eleven years, and all the information used has been anonymized before processing it. As it can be shown and taking into account that the UNED is the highest university in Spain with almost 200.000 students enrolled per year, the volume of data is huge.

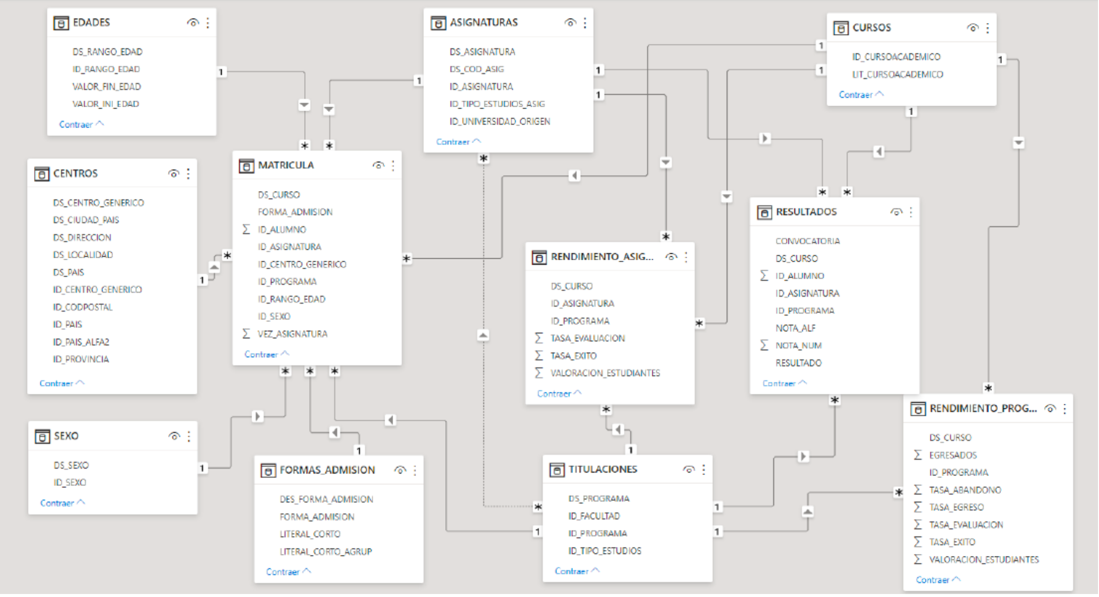


Figure 1: Entity-relation scheme of the available data. Labels are in Spanish. All the ID fields correspond to anonymized data.

However, although it may seem that we have a large amount of data, we must choose those that meet the requirements of our project. Thus, we should only select data from students in their first year and who have taken the exams at least once. Therefore, we must choose only the data related to the first year for the recommender implementation. Also, we have preprocessed the data to keep those that provide valuable information eliminating those that contained some inadequate or null value and those students who did not take any of the subjects enrolled because they may correspond to extraordinary situations, and we considered them as outliers.

3.2. Characterization of the UNED´s first-year student

Since our user is a new student and we have no information about his behavior (i.e., previous enrollment information), we must use partial information about our scheme since we are using collaborative filtering. The profile of a first-year student was defined by different features obtained from the data stored in the University central systems. We have defined five main features: a) Age range (i.e. this feature is split in different age ranges from 0 to 115) b) Access method to the university. We have identified twelve methods of access. C) Genre, for this property we only can identified 2 possible values, d) UNED Degree in which students want to enroll to limit the recommendation and finally e) Associate center where the student is making the enrollment information. The associate center is an essential field because it is related to the UNED geographical structure and allows us to segment information geographically. Therefore, we are looking for students like our first-year student in age, method of access to the university, genre and associate center. Once we have similar students identified, we check how many subjects they enrolled in and if they were successful or not, allowing us to recommend a range of numbers of subjects.

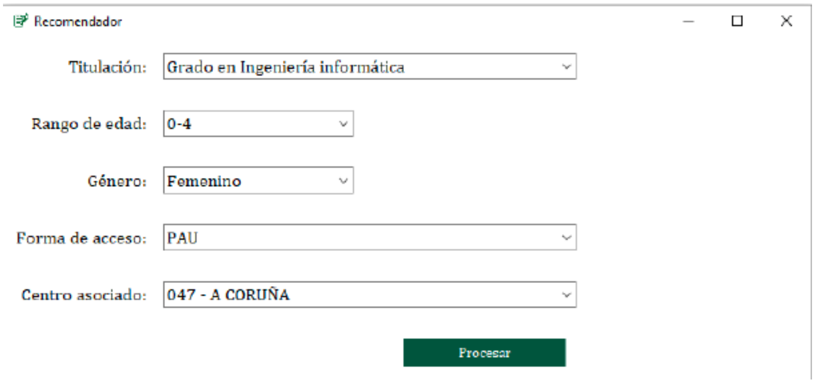
For computational purposes, this profile is coded as a one-shot vector whose fields code the information of the selected features, Table 1.

Table 1
First-year student One-shot vector

Features	UNED’s degree			Age range			Access			Campus			Genre	
Fields	7101	7102	…	20	25	…	1	2	…	1	2	…	M	F
Values	1	0	0	0	1	0	0	1	0	1	0	0	1	0

Thus, when the student meets one of these characteristics, we will put a 1 in that component while, if he/she does not meet it, we will put a 0. For example, a student whose age range corresponds to 20, will have a 1 in that component and a 0 in the rest of the components that refer to the age range.

The information is asked to the new first-year student using a simple web interface, Figure 2.



Titulación: Grado en Ingeniería informática

Rango de edad: 0-4

Género: Femenino

Forma de acceso: PAU

Centro asociado: 047 - A CORUÑA

Figure 2: User interface to collect first-year student information.

3.3. Applied Algorithms

In this work, were analyzed two algorithms for making recommendations: a) K-means b) cosine similarity. We have selected these two algorithms to compare two different approaches and obtain highlights about the best scenarios in which could be used.

3.3.1. K-means based recommender

It is a machine learning algorithm. This algorithm avoids relying on similarity measures between pairs of items, and it aims to divide all the input elements into K groups where their similarity is maximized.

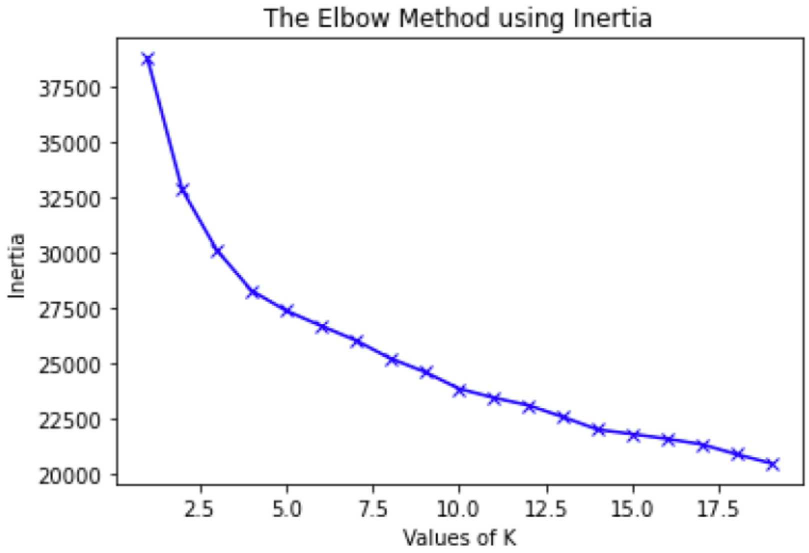


Figure 3: K-value calculated by the elbow method.

In a first step, the elements that will be the centroids of each group are defined (either by the user or by random initialization). In subsequent steps, these centroids and the elements belonging to each group are re-defined. Once the groups have been created, the algorithm will allow us to classify a new component of the most suitable group, i.e., the group whose centroid is most similar. This algorithm has a high time cost in learning, but then the classification of a new student is fast, which allows obtaining a fast recommendation.

3.3.2. Cosine similarity-based recommender

The similarity estimation offered by this method is based on calculating the cosine of the angle formed by the two vectors. If the angle between two vectors $ee1$ and $ee2$ is equal to 0º, the cosine value is 1; both vectors will have the same direction. On the contrary, if the angle is 180º, the cosine value is -1; the vectors will have opposite directions. Mathematically it is defined as follows:

$$\cos(e_1, e_2) = \frac{e_1 \cdot {e_2}^T}{\|e_1\| \cdot \|e_2\|}$$

3.4. Experimentation

To analyze the best algorithm that fits our problem, we have selected a small dataset built with data from five different grades. These grades have been selected for their different global success rates to determine if this parameter influences the selection of the algorithm. The degrees selected are in Table 2. These data have been obtained from the statistical portal of the UNED.

Table 2
Degrees selected to evaluate the recommender.

Code	Degree Denomination	Global Success Rate
6301	Social Education	58.12%
6502	Business Administration and Management	31.89%
6702	History of Art	48.64%
6801	Electrical Engineering	16.84%
7101	Computer Engineering	27.41%

For the experimentation, we have used all the data for the last eleven years except for the 2019-2020 academic year data due to COVID 2019 pandemic because it does not follow the trend of previous years. We have also eliminated the records of enrollments after the first year since we focused on the enrollment recommendation for the first year. Also, we have split our dataset in two following the rule 80:20, 80% for training and 20% for tests. The students used as reference were those that made their first enrollment in the academic year 2018-2019.

To calculate the adequacy of the recommendation, we used two error metrics. First, we define the Real Fail Rate, RFR, which refers to the rate of real fail subjects of a student. We defined the Real Fail Rate as:

$$RFR = \frac{(FG + NP)}{NS}$$

Where FG is the number of failing subjects, NP is the number of subjects that, in the end, the student didn´t take the exam, and NS is the number of subjects the student had enrolled.

$$RMFR = \frac{(RNS - RA)}{RNS}$$

Where RNS is the number of recommended subjects for enrolling, and RA is the number of student success subjects.

3.5. Results

To evaluate both algorithms, we built the one-shot vector of our test students and ran the algorithm against the dataset, and computed RFR for each degree. For RMFR, we calculated its value for the two extreme values of the recommended range. We defined RMFRL for the left value of the range and RMFRR for the right value of the range. Below, we show the test results for each algorithm, Table 3.

Table 3
RFR and RMFR results for all the grades in the dataset

Code	Degree Denomination	Real Data	Cosine Similarity		K-mean	
		RFR	$RMFR_L$	$RMFR_R$	$RMFR_L$	$RMFR_R$
6502	Social Education	0.51	0.31	0.45	0.42	0.45
6301	Business Administration and Management	0.35	0.33	0.39	0.43	0.42

6801	History of Art	0.68	0.18	0.34	0.16	0.36
6702	Electrical Engineering	0.32	0.37	0.50	0.78	0.78
7101	Computer Engineering	0.59	0.22	0.34	0.39	0.41

4. Discussion and Conclusions

With all the experimentation carried out and after the analysis of all the data, we can affirm that the K-Means algorithm is the one that offers the worst results, being notably better in the calculation of the recommendation with cosine similarity. Therefore, the classification algorithm cannot produce a set of sufficiently precise groups to improve the success rate or dropout. One of the reasons is the lack of information we have about first-year students. Our one-shot vector is simple, and it could be improved with more knowledge to try to define better the groups of students that share the same characteristics. The selection of K in the algorithm is another drawback of this system. To obtain the K parameter, we have applied the elbow algorithm dynamically to adjust the algorithm to the user profile. More work in this selection would improve the systems.

On the other hand, of these results, it is noteworthy that the recommender system does not improve the RFR of grades with a low RFR, such as History of Art, but it would help to significantly reduce the number of failed subjects in degrees in which the percentage of subjects they fail is currently very high. According to the results obtained, the system would make it possible to reduce the percentage of failed subjects by up to 50% in the best scenario.

In future lines of work, it would be especially useful to add information on subjects and degrees because although students are similar, the degrees they take may vary the results of the recommendation given to everyone. This would even make it possible to recommend certain combinations of subjects to create a more efficient and balanced sequence of study.

I would like to emphasize the need for meaningful learning in the students since it would facilitate understanding the contents of each subject. Although no specific order is established for the study of the subjects, a good structuring of them would help us correctly scaffold the concepts. This can be achieved by including additional information on the assignments.

5. References

[1] Gironacci, I. (2021). Literature Review of Recommendation Systems (pp. 119–129). https://doi.org/10.4018/978-1-7998-4339-9.ch009

[2] Rivera, A. C., Tapia-Leon, M., & Lujan-Mora, S. (2018). Recommendation Systems in Edu-cation: A Systematic Mapping Study. In Á. Rocha & T. Guarda (Eds.), Proceedings of the International Conference on Information Technology & Systems (ICITS 2018) (pp. 937–947). Springer International Publishing. https://doi.org/10.1007/978-3-319-73450-7_89

[3] Charnelli, M. E., Lanzarini, L. C., & Díaz, F. J. (2018). Sistemas recomendadores aplica-dos en educación. XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste). http://sedici.unlp.edu.ar/handle/10915/67261

[4] Lynn, N. D., & Emanuel, A. (2021). A review on Recommender Systems for course selec-tion in higher education. https://doi.org/10.1088/1757-899X/1098/3/032039

[5] Maphosa, M., Doorsamy, W., & Paul, B. (2020). A Review of Recommender Systems for Choosing Elective Courses. https://doi.org/10.14569/ijacsa.2020.0110933

[6] O'Mahony, M. P., & Smyth, B. (2007). A recommender system for on-line course enrol-ment: An initial study. Proceedings of the 2007 ACM Conference on Recommender Sys-tems, 133–136. https://doi.org/10.1145/1297231.1297254

[7] Ricci, F., Rokach, L., & Shapira, B. (2010). Recommender Systems Handbook. In Recom-mender Systems Handbook (Vols. 1–35, pp. 1–35). https://doi.org/10.1007/978-0-387-85820-3_1

[8] Gupta, S., & Dave, M. (2020). An Overview of Recommendation System: Methods and Techniques (pp. 231–237). https://doi.org/10.1007/978-981-15-0222-4_20

[9] Sharma, R., & Singh, R. (2016). Evolution of Recommender Systems from Ancient Times to Modern Era: A Survey. Indian Journal of Science and Technology, 9(20), 1–12. https://doi.org/10.17485/ijst/2016/v9i20/88005

[10] Fernández-García, A. J., Rodríguez-Echeverría, R., Preciado, J. C., Manzano, J. M. C., & Sánchez-Figueroa, F. (2020). Creating a Recommender System to Support Higher Educa-tion Students in the Subject Enrollment Decision. IEEE Access, 8, 189069–189088. https://doi.org/10.1109/ACCESS.2020.3031572