УДК 004.852, 51-77
doi:10.18720/SPBPU/2/id24-46

Daria D. Danilova[1],
Bachelor;
Vadim G. Pak [2],
Associate Professor, Candidate of Physical and Mathematical Sciences

METHODS OF DATA ANALYSIS IN THE PROBLEM OF IMPROVING ACADEMIC PERFORMANCE WITH DISTANT LEARNING TECHNOLOGY

[1, 2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia; [1] danilova.dd@edu.spbstu.ru; [2] pak_vg@spbstu.ru

Abstract. The paper investigates the results of student grade history reports from the Moodle platform. Students are clustered according to their academic performance; based on the largest cluster of students, tasks are clustered according to difficulty level. For each course assignment, a search is made for professor's similar feedbacks. Based on their analysis, recommendations are given for further improvement of students' academic performance.

Keywords: data analysis, distant learning, academic performance, clustering, Gaussian mixture model, TF-IDF, cosine similarity.

Данилова Дарья Дмитриевна[1],
бакалавр;
Пак Вадим Геннадьевич[2],
доцент, канд. физ.-мат. наук

МЕТОДЫ АНАЛИЗА ДАННЫХ В ЗАДАЧЕ ПОВЫШЕНИЯ УСПЕВАЕМОСТИ НА КУРСАХ СИСТЕМЫ ДИСТАНЦИОННОГО ОБУЧЕНИЯ

[1, 2] Россия, Санкт-Петербург, Санкт-Петербургский политехнический университет Петра Великого;
[1] danilova.dd@edu.spbstu.ru, [2] pak_vg@spbstu.ru

Аннотация. В статье исследуются результаты отчетов по истории успеваемости учащихся с платформы Moodle. Проводится кластеризация студентов в соответствии с их успеваемостью; на основе самой большой группы учащихся, выполняется кластеризация заданий в соответствии с уровнем сложности. Для каждого задания курса выполняется поиск похожих отзывов преподавателя. На основе их анализа даются рекомендации по дальнейшему улучшению успеваемости студентов.

Ключевые слова: анализ данных, дистанционное обучение, академическая успеваемость, кластеризация, модель гауссовой смеси, TF-IDF, косинусное сходство.

Introduction

Most of the courses at Peter the Great St. Petersburg Polytechnic University have been taught for many years using the Moodle platform for mixed and distance learning. The platform provides a space for professors and students to work together. There are various features available in Moodle to track student progress.

The accumulated database of professors' feedback on the solutions of students' tasks can be used to analyze the improvement of students' academic performance. Improving academic performance is beneficial to both students and professors. Students will be able to better assimilate the teaching material, and professors will improve the quality of teaching and education.

The goal is to investigate the feedbacks to the tasks on the courses of the distance learning system to improve student academic performance.

The relevance lies in the fact that the analyzed courses belong to the mandatory part of the educational programmes. One of the courses belongs to the basic part of the bachelor's degree program and the second highlights the basic knowledge necessary for mastering the subsequent courses of the professional retraining program. Therefore, it is important for students to raise their academic performance in these subjects.

Tasks were set to research feedback texts of two courses over several years and give recommendations for each course to improve academic performance.

1. Analysis of courses

1.1. Analyzed courses

Two courses were chosen to analyze students' academic performance: "Mathematical logic for part-time students" and "DEV-DB. Database Basics for programmers (PostgreSQL)" on the Moodle platform.

The course "Mathematical Logic for part-time students" is taught for bachelors of the "Applied Information Science" at the Higher School of Intelligent Systems and Supercomputer Technologies and belongs to the basic part of the professional cycle of general mathematical and natural science disciplines of the curriculum. The course lasts a semester, there is data for analysis for the last 3 years.

The course "DEV-DB. Database Basics for Programmers (PostgreSQL)" is taught at a professional retraining course at the Higher Engineering School and lasts 5 weeks. There are data for analysis for the last year.

In these courses, students are not limited in the number of attempts to pass assignments and can improve their grades. Therefore, in this case, it is possible to track progress by the number of attempts.

1.2. Dataset preprocessing

Each course in Moodle has grade history report, which contains the professor's feedback history. And these reports will be used in this paper for data analysis.

To start with, the datasets should be preprocessed. Rows that do not have grader should be deleted. Only tasks will be analyzed, there is no need to include tests.

1.3. Clustering of students

To begin with, the students of each course should be divided into clusters according to academic performance.

The Gaussian mixture model (GMM) was chosen for clustering, since the traditional K-Means algorithm has some limitations [3]. For example, K-Means creates only circular clusters, does not provide probabilistic estimates of whether points belong to clusters, does not take into account cluster variance, relies on distance and ignores the distribution of each cluster. And the Gaussian mixture model studies the distribution and provides better clustering.

The Figure 1 shows the clustering results for each course.

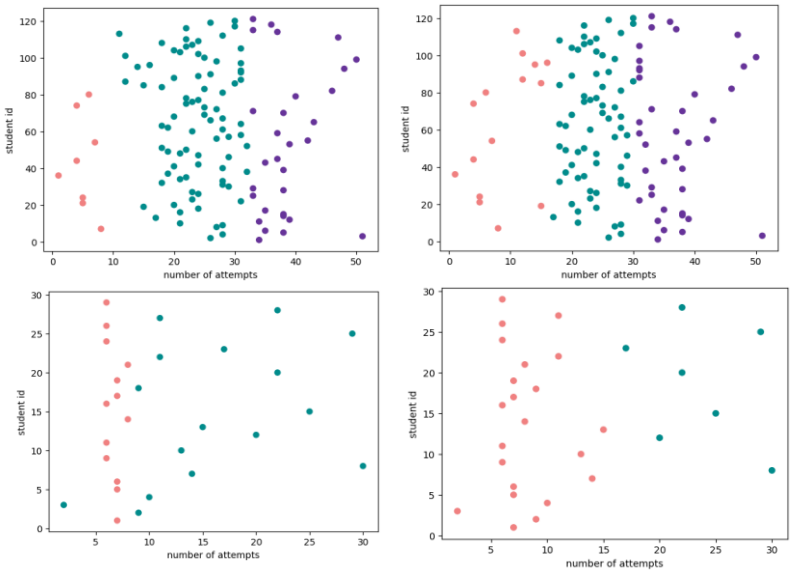

Fig. 1. Clustering of students by total number of attempts: upper quadrants are databases course, bottom quadrants are mathematical logic course, K-Means is used in left quadrants, GMM is used in right quadrants

In a database course, students can be divided into three clusters: successful, average and lagging behind. There are two clusters in the mathematical logic course.

1.1. Clustering tasks

Since the courses are aimed at the majority, it is needed to choose the largest cluster of students and within this cluster perform clustering of

tasks by the number of attempts per task — finding simple and complex tasks. To do this, first it is needed to calculate how many attempts each student took to solve each task and then determine the median of attempts for each task. The results of clustering are presented in Figure 2.

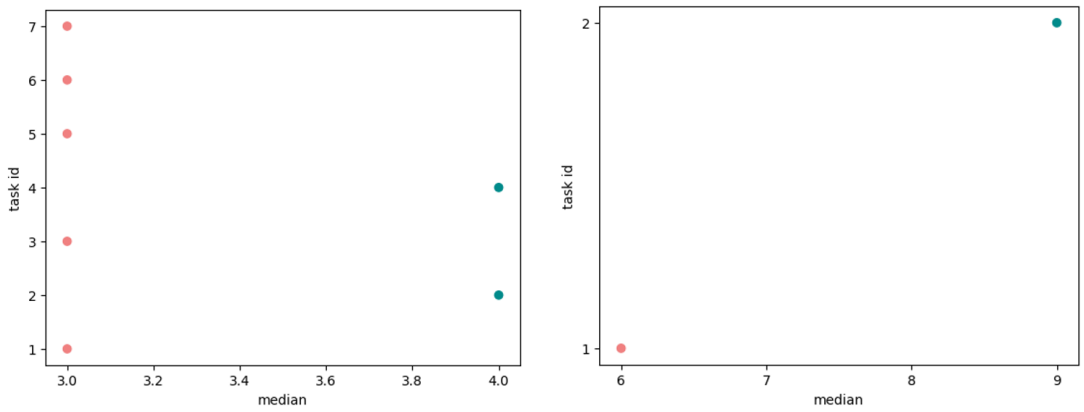

Fig. 2. Attempts of students from the largest cluster: on the left side for databases course, on the right side for mathematical logic course

It can be seen that the assignments in the mathematical logic course are difficult for students, while in the database course the assignments are of average difficulty level.

1.5. Reviews preprocessing

As a preprocessing it is necessary to convert the reviews to lower case, remove punctuation marks, numbers and other non-letter characters from them. Then delete duplicate reviews.

Now it is needed to remove the stop words, because they create noise. Stop words include prepositions, conjunctions, adverbs, etc. To stop words should be also included words such as “passed”, “failed”, “task” and their other forms, if they exist in other languages.

Since the form of the word matters in the analysis, all the words should be brought to the initial form, that is, lemmatization will be performed.

1.6. Search for similar reviews

The feedback professors gave to students for each task in each course needs to be analyzed. To do this, the most similar reviews can be found using cosine similarity [1, 4], which is calculated for feature vectors using the formula (1):

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

where A and B are feature vectors.

But first the reviews need to be vectorized. The reviews will be vectors of the weights of the words’ significance. The reviews vectorization will be done using TF-IDF (term frequency — inverse document frequency) [1]. This method allows determine not only the importance of the word in a particular

text, but also the importance of the word taking into account all texts. If a word is often found in all documents, then it is unlikely that it is of great importance (for example, stop-words) [2]. Conversely, if a word is uncommon, it probably determines document's content to a greater extent.

The frequency of a word relative to all words in a document can be found using the formula (2):

$$tf(t,d) = \frac{count\ of\ t\ in\ d}{number\ of\ words\ in\ d}, \quad (2)$$

where *t* is a term, *d* is a document.

The significance of a word in all documents can be calculated using the formula (3):

$$idf(t,D) = \log\left(\frac{number\ of\ documents\ D}{number\ of\ documents\ containing\ the\ term\ t}\right), \quad (3)$$

where *D* is number of documents, *t* is a term.

The weight is calculated by the formula (4):

$$tf - idf(t,d,D) = tf(t,d) \times idf(t), \quad (4)$$

where *t* is a term, *d* is a document, *D* is number of documents.

Now it is necessary to carry out vectorization — convert TF-IDF representations into vectors. Each review becomes a vector in a multidimensional space, where the dimensions represent the terms in the corpus — the collection of reviews under consideration. The reviews with the three highest cosine similarity values were selected for each task. The results of cosine similarities of reviews for database fundamentals course are in the Table 1.

Table 1

The results of cosine similarities for database course

Task	Cosine similarity
1	0.59
	0.59
	0.581
2	0.799
	0.773
	0.747
3	0.902
	0.853
	0.832
4	0.187
	0.116
	0.11
5	0.611
	0.367
	0.226
6	0.758
	0.741
	0.638
7	0.833
	0.693
	0.606

The results of cosine similarities for mathematical logic course are in the Table 2.

Table 2

The results of cosine similarities for mathematical logic course

Task	Cosine similarity
1	0.392
	0.331
	0.311
2	0.442
	0.384
	0.309

2. Processing of results, recommendations

After analyzing similar reviews, the following recommendations were created, presented in Table 3, to improve academic performance.

Table 3

Recommendations for courses

Course	Task	Recommendations
"DEV-DB. Database Basics for Programmers (PostgreSQL)"	1	Pay attention to null values
	2	Clarify that the query should be concise, this may affect its complexity and performance
	3	Pay more attention to the description of the subject area
	4	Pay attention in which cases it is inappropriate to use restrictions
	5	Add an explanation to the task
	6	Add the condition to the task description that restricts the use of non-required
	7	Focus on the keys — primary and surrogate
"Mathematical logic for part-time students"	1	Pay attention to the importance of the sequence of formulas. Pay attention to forbidden substitutions in proof trees. In the reviews, there is a similarity in the indications of such mistakes
	2	Focus more on Bernays predicate calculus

Conclusions

In this paper, reviews of two different courses over several years have been analyzed. Namely, the students of each course were clustered according to the level of academic performance. Within the largest clusters, tasks were clustered according to the difficulty level. For each assignment, similar reviews were found, based on which recommendations were made that could help improve students' academic performance.

The results and methods presented in this paper can be used to analyze other courses on distant learning platforms.

References

1. Chiny M., Chihab M., Bencharef O., Chihab Y. Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms. In: Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning, 2021. – Pp. 15–20.

2. Jalilifard A., Caridá V.F., Mansano A.F., Cristo R.S., da Fonseca F.P.C. Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. In: Thampi, S.M., Gelenbe, E., Atiquzzaman, M., Chaudhary, V., Li, KC. (eds) Advances in Computing and Network Communications. Lecture Notes in Electrical Engineering, vol. 736. Springer, 2021.

3. Patel E., Kushwaha D.S. Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. Elsevier, 2020. – Pp. 158–167.

4. Singh R., Maurya S., Tripathi T., Narula T., Srivastav G. Movie Recommendation System using Cosine Similarity and KNN. International Journal of Engineering and Advanced Technology, 2020. – Pp. 556–559.

УДК 004.822
doi:10.18720/SPBPU/2/id24-47

Пономарёв Василий Васильевич [1],
директор по исследованиям и развитию, канд. филол. наук;
Туманов Владимир Евгеньевич [2],
преподаватель, канд. хим. наук;
Пономарёв Всеволод Васильевич [3],
студент

МОДЕЛИРОВАНИЕ УСПЕШНОГО РЕШЕНИЯ ЗАДАЧИ ИМПОРТОЗАМЕЩЕНИЯ В СИСТЕМЕ ОБРАЗОВАНИЯ В ПОНЯТИЙНОЙ ФОРМАЛИЗАЦИИ ТЕРМИНОЛОГИИ ЦИФРОВОГО ИНТЕРАКТИВНОГО ДОКУМЕНТА

[1] Россия, Московская область, Ногинск, ООО Научно-Производственное Предприятие «РУМБ», vasily.ponomarev@gmail.com;
[2] Россия, Московская область, Ногинск, Московский областной медицинский колледж № 3, tve9000@gmail.com;
[3] Россия, Москва, Московский институт тонких химических технологий имени М.В. Ломоносова федерального государственного бюджетного образовательного учреждения высшего образования «МИРЭА» — Российского технологического университета, moshimik@mail.ru

Аннотация. Репрезентирован концептуальный инструментарий в балансе с понятийным аппаратом формализации терминологии цифрового интерактивного документа применительно к моделированию успешного решения задачи импортозамещения в системе образования.