СЕКЦИЯ 4
СИСТЕМЫ ДЛЯ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ: ПРОГРАММНЫЕ РЕШЕНИЯ

С. Д. Бессонова, И. В. Кондратьев, М. С. Воробьева
Тюменский государственный университет, г. Тюмень
УДК 004.8

КОМБИНИРОВАННЫЙ ПОДХОД К РЕКОМЕНДАЦИИ УЧЕБНЫХ ДИСЦИПЛИН НА ОСНОВЕ АНАЛИЗА ДАННЫХ СТУДЕНТОВ

Аннотация. Исследование направлено на изучение и реализацию методов рекомендаций элективных курсов на основе данных о студентах и их интересах. Комбинированный подход к рекомендации сочетает в себе контентную и коллаборативную фильтрации. Коллаборативная фильтрация основывается на алгоритме AgglomerativeClustering, группирующем студентов по их сходству, а контентная использует языковую модель paraphrase-multilingual-mpnet-base-v2 для векторизации описаний дисциплин.

Ключевые слова: индивидуальная образовательная траектория, машинное обучение, кластеризация, векторизация текста, рекомендательная система, персонализированные рекомендации, контентная фильтрация, коллаборативная фильтрация.

Введение. Построение индивидуальной образовательной траектории (ИОТ) — распространенная практика в современных вузах, позволяющая студентам комбинировать дисциплины из разных областей, ориентируясь на личные цели. Например, в Тюменском Государственном университете (ТюмГУ) с 2017 года студенты формируют ИОТ с помощью выбора элективов — дисциплин по выбору, которые студент выбирает самостоятельно на 2-4 семестры.

Актуальность технологических решений для помощи в построении ИОТ подтверждается современными исследованиями. В работе [1] предложен метод автоматизированного ранжирования учебных материалов на основе анализа семантической близости аннотаций с использованием моделей семейства BERT (например, SciBERT, Multilingual Universal Sentence Encoder). Авторы демонстрируют, что оценка косинусного сходства векторных представлений текстов позволяет выявлять публикации, наиболее релевантные для построения индивидуальных траекторий, что решает проблему субъективности при подборе источников. Эксперименты на коллекциях научных статей продемонстрировали высокую степень согласованности результатов модели с экспертными оценками, что подтверждает ее пригодность для задач персонализации обучения. Это особенно важно в контексте растущего интереса к ИОТ как инструменту развития профессиональных компетенций студентов.

Студенты испытывают сложности при выборе элективных, что подтверждается в статье, освящающей опыт формирования ИОТ в ТюмГУ [2]. Согласно результатам анкетирования, 26% опрошенных первокурсников и 82,1% второкурсников испытывали трудности при выборе элективов.

Также в исследовании, проведенном среди студентов Саратовской государственной юридической академии [3], выяснили, что большинство студентов (60%) полагаются на рекомендации старших студентов при выборе элективных дисциплин, что делает новые элективы менее привлекательными.

Существуют примеры сервисов, помогающих студентам в выборе элективных дисциплин. Так, команда студентов ТюмГУ разработали систему рекомендаций элективных дисциплин на основе подписок в социальной сети «Вконтакте» [4]. Студент получает персонализированные рекомендации, но не может скорректировать свои предпочтения, так как изменить интересы можно только с помощью подписок на сообщества.

В исследовании [5] представлен инструмент для рекомендаций дисциплин по выбору, соответствующих образовательной программе студентов. Инструмент имеет функционал для самостоятельного подбора элективов, но не предоставляет персонализированных рекомендаций.

Проблема исследования. Студенты сталкиваются с трудностями при выборе элективных дисциплин из-за большого количества постоянно меняющихся курсов и отсутствия персонализированных рекомендаций. Существующие системы пока не в полной мере учитывают семантику курсов и индивидуальные предпочтения, что может ограничивать точность рекомендаций.

Таким образом, цель исследования состоит в сравнении и реализации методов рекомендаций элективных курсов на основе данных о студентах или их интересах.

Материалы и методы. В исследовании рассмотрены два подхода — контентная и коллаборативная фильтрация — для повышения точности рекомендаций. Контентная фильтрация компенсирует недостатки коллаборативной, позволяя анализировать новые курсы. В свою очередь, коллаборативная фильтрация учитывает данные студентов и выявляет скрытые связи. Контентная фильтрация основана на анализе содержания элективных дисциплин. Подход заключается в создании векторного представления описания каждого электива и их последующего сравнения с интересами пользователя, представленных в векторном виде.

Формулируем математические постановки решаемых задач: 1) поиск элективов, удовлетворяющих заданному запрос; 2) кластеризация студентов по мотивационному профилю.

Задача 1. поиск элективов, удовлетворяющих заданному запрос

Дано:

$El = \{el_1, el_2, \ldots, el_n\}$ — множество элективных курсов.

q — запрос студента

Найти: подмножество элективов $El^* \in El$, состоящее из k курсов, таких что:

$$El^* = arg \max_{El' \subset E, |El'|=k} \sum_{e_i \in El'} cos(\theta_{Q, e_i}), \text{ где}$$

$E = \{e_1, e_2, \ldots, e_n\}$ — множество векторных представлений элективов;

$e_i = v(el_i)$ — векторное представление описания электива;

$Q = v(q)$ — векторное представление текстового запроса студента;

$\cos(\theta_{Q, e_i}) = \frac{Q * e_i}{\|Q\| * \|e_i\|}$ — косинусное сходство между вектором запроса Q и вектором электива e_i.

Коллаборативная фильтрация основывается на анализе данных студентов: если несколько студентов имеют схожие данные, то курсы, которые нравятся одному из них, могут быть рекомендованы другим.

Задача 2. кластеризации студентов по мотивационному профилю

Дано:

$S = \{S_1, S_2, \ldots, S_n\}$ — множество студентов;

$S_i = \{c_1, c_2, \ldots, c_m\}$ — мотивационный профиль студента, представленный множеством характеристик, где:

m — количество мотивационных факторов студента;

c_j — степень развития мотивационного фактора j.

Необходимо:

1. Определить оптимальное количество кластеров k:

$$k^* = \arg\max_k S(k),$$

где $S(k)$ — средний коэффициент силуэта s_i кластера:

$$S(k) = \frac{1}{|S|}\sum_{i \in S} s_i \,; s_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

2. Разделить студентов на k кластеров $C = \{C_1, C_2, \ldots, C_k\}$.

Контентная фильтрация использует языковую модель для векторизации текста. Для выбора наиболее подходящей модели сравнились следующие решения:

- SentenceTransformer представляет собой технологию модификации предварительно обученной сети BERT. SentenceTransformer имеет множество предобученных моделей для различных языков, в том числе и русского.
- FastText. Библиотека, содержащая предобученные векторные представления слов и классификатор — алгоритм, разбивающий слова на классы.
- TF-IDF. Статистическая мера, используемая для оценки важности слова в контексте документа или набора документов.

Данные об элективах были получены с помощью парсинга сайтов информационных систем «Отзывус» и «Modeus». Список актуальных для выбора дисциплин был получен от менеджеров Управления индивидуальных образовательных траекторий (УИОТ) ТюмГУ.

В ходе работы был проведен сравнительный анализ моделей для оценки соответствия предложенных элективов запросам студентов. Для этого сопоставлялись показатели косинусного сходства между векторными представлениями запросов и описаний элективов, которые были предложены каждой из перечисленных выше моделей. Тестирование проходило на трех видах запросов: коротком (одно слово), развернутом (одно предложение) и сложном (несколько предложений со словами из различных сфер).

На всех запросах лучше всего себя показала модель SentenceTransformer с paraphrase-multilingual-mpnet-base-v2: вне зависимости от сложности запроса она находила элективы с наивысшим косинусным сходством (короткий: 0,5-0,6, сложный: 0,6-0,7), при этом элективы соответствовали запросу. С увеличением длины и сложности запроса paraphrase-multilingual-mpnet-base-v2 продолжает подбирать релевантные элективы, как и другие модели SentenceTransformer. FastText с усложнением запроса подбирает элективы с высоким косинусным сходством (0,7 и выше), но сами элективы не всегда релевантны.

В основе коллаборативной фильтрации лежит кластеризация студентов по их мотивационным профилям. Для выбора метода кластеризации сравнивались несколько алгоритмов: K-Means, DBSCAN, HDBSCAN, GMM, AgglomerativeClustering.

Данные студентов, прошедших диагностику в рамках проекта «Центры компетенций», включая ФИО, пройденные элективы, а также результаты диагностики в виде мотивационных профилей были предоставлены УИОТ ТюмГУ. Персонализированные данные были обезличены.

Данные о мотивационных профилях содержат показатели 2 554 студентов. Результаты представлены в числовом виде (от 200 до 800 баллов по каждому показателю). Обработка профилей происходила с помощью языка программирования Python и включала в себя следующие этапы:

1. Пропущенные значения были заменены нулями, так как считается, что пропущенное значение — отсутствие мотивационного признака.

2. Признаки с высоким коэффициентом корреляции были заменены на одно агрегированное значение — стандартное отклонение. Для каждого студента считается, насколько сильно различаются его значения в 16 столбцах, включая «Альтруизм», «Креативность», «Саморазвитие» и другие личностные характеристики.

3. Для дальнейшей кластеризации данные были стандартизированы.

Число кластеров подбиралось заранее с помощью метода локтя и коэффициента силуэта. Для оценки качества кластеризации использовались индекс Девиса-Болдина и коэффициент силуэта. Индекс Девиса-Болдина показывает среднюю степень сходства кластера с его наиболее похожими соседними кластерами. Коэффициент силуэта оценивает сплоченность кластера, измеряя среднее расстояние между всеми точками внутри кластера, и степень разделения, измеряя расстояние до ближайшего соседнего кластера [6].

На основании сравнительного анализа выбран метод AgglomerativeClustering — алгоритм иерархической кластеризации, который последовательно объединяет точки в кластеры, основываясь на их мере сходства, пока не будет достигнут желаемый уровень кластеризации. Алгоритм имеет наименьшее значение индекса Дэвиса-Болдина при среднем значении коэффициента силуэта среди всех алгоритмов.

Гиперпараметр linkage модели AgglomerativeClustering отвечает за определение расстояния между кластерами. На каждом шаге агломеративной кластеризации два ближайших кластера сливаются в один. Какие два кластера сливаются, зависит от выбранного linkage. Для используемого набора данных лучше всего подошло значение complete гиперпараметра linkage (максимальное расстояние между всеми наблюдениями двух кластеров).

Для визуализации кластеров размерность данных была уменьшена с помощью t-SNE и PCA. Визуализация кластеров представлена на рис. 1. Хотя кластеры компактны и разделимы, в некоторых из них сосредоточено значительно больше студентов, чем в остальных.

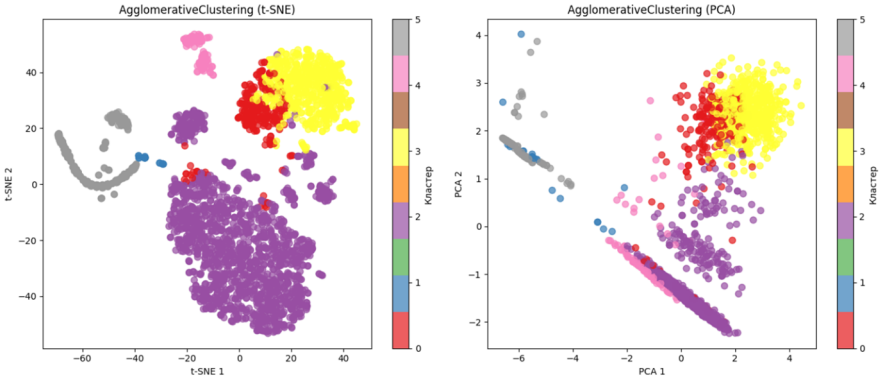

Рис. 1. Визуализация кластеров с помощью t-SNE и PCA

Результаты. В ходе работы было проведено сравнение методов рекомендаций элективных дисциплин и разработан подход к формированию списка рекомендуемых курсов на основе имеющихся данных о студентах ТюмГУ. Он совмещает в себе два метода: коллаборативная фильтрация на основе кластеризации студентов по их мотивационным профилям и контентная фильтрация, анализирующая текстовые описания дисциплин и запрос пользователя.

Для реализации модели коллаборативной фильтрации был проведен сравнительный анализ различных методов кластеризации. Разбиение оценивалось с помощью коэффициента силуэта и индекса Девиса-Болдина. В ходе исследования AgglomerativeClustering показал наилучшие результаты, достигнув 0.3602 по коэффициенту силуэта и 1.2101 по индексу Дэвиса-Болдина. Результаты исследования кластеризации представлены в табл. 1.

Таблица 1

Результаты исследования кластеризации

Метод	*Гиперпараметры*	*Кол-во кластеров*	*Кластеры*	*Коэф. силуэта*	*Индекс Дэвиса-Болдина*
K-Means	n_clusters=7	7	(761), (724), (447), (289), (281), (102), (78)	0.2515	1.4795
DBSCAN	eps=2, min_samples=40	4	(1438), (681), (306), (257)	0.4102	2.0695
HDBSCAN	min_cluster_size=20	4	(2389), (192), (59), (42)	0.4825	1.5810
GMM	n_components=6	6	(1411), (634), (288), (157), (98), (94)	0.3262	2.1469
Agglomerative Clustering	n_clusters=6, linkage='complete'	6	(1548), (519), (275), (218), (102), (20)	0.3602	1.2101

На основе кластеризации методом AgglomerativeClustering все студенты были разделены на 6 групп, что позволило создать внутри каждого кластера ранжированный перечень элективов согласно частоте их выбора участниками группы.

Для реализации модели контентной фильтрации описания курсов были предобработаны с помощью библиотеки Natasha. Для дальнейшей работы описания были преобразованы в векторные представления текстов с помощью модели, предобученной для русского языка, — paraphrase-multilingual-mpnet-base-v2 из библиотеки Sentence-Transformers. Сравнение методов SentenceTransformer, fastText и tf-idf выявило, что модели SentenceTransformer рекомендовали наиболее релевантные элективы с высоким показателем косинусного сходства. Результаты сравнения методов векторизации представлены в табл. 2.

Разработанный алгоритм успешно определяет студента в один из имеющихся кластеров для дальнейшей рекомендации элективов, а также находит дисциплины, соответствующие запросу пользователя с косинусным сходством не менее 0.4, что позволяет рекомендовать студентам подходящие курсы.

Проведенные эксперименты показали, что комбинированное использование кластеризации и анализа текста позволяет повысить точность рекомендаций и адаптировать их под индивидуальные запросы студентов.

Таблица 2

Сравнение методов векторизации с помощью среднего значения косинусного сходства для 10 подобранных элективов

Модель		*Простой запрос*	*Средний запрос*	*Сложный запрос*
Sentence-Transformers	paraphrase-multilingual-mpnet-base-v2	0,59	0,51	0,71
	paraphrase-multilingual-MiniLM-L12-v2	0,57	0,53	0,67
	paraphrase-xlm-r-multilingual-v1	0,49	0,48	0,62
fastText		0,39	0,69	0,9
TF-IDF		0,1	0,07	0,07

Заключение. В ходе исследования разработан алгоритм рекомендации элективных дисциплин на основе данных о студентах и их интересах. Ключевым результатом работы стало комбинирование методов контентной и коллаборативной фильтрации, что позволило преодолеть нивелировать характерные недостатки каждого из подходов. Использование алгоритма AgglomerativeClustering позволило выделить среди студентов шесть групп с различными предпочтениями. Применение модели paraphrase-multilingual-mpnet-base-v2 помогло повысить точность подбора элективов (косинусное сходство достигло 0.71 для сложных запросов).

Практическая применимость исследования подтверждена запуском чат-бота в Telegram, который предоставляет студентам персонализированные рекомендации и ссылки на образовательные ресурсы.

Перспективы работы основаны на расширении функционала за счет разработки административной панели, а также использования дополнительных данных для повышения точности рекомендаций.

Исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания (FEWZ-2024-0052).

СПИСОК ЛИТЕРАТУРЫ

1. Михайлов Д.В., Емельянов Г.М. Трансформерные модели BERT, взаимное сходство смыслов коротких текстов и их ранжирование по близости эталону // Математические методы распознавания образов: тезисы докладов 21-й Всероссийской конференции с международным участием. — Москва, 2023. — С. 159-161
2. Донкова И.А., Якубовский Ю.Е., Карякин И.Ю., Карякин Ю.Е. Опыт формирования индивидуальных образовательных траекторий // Инженерное образование. — 2024. — № 35. — С. 119-130.
3. Вьюшкина Е.Г., Щербакова О.В. Индивидуальная образовательная траектория: основания для выбора элективных дисциплин // Известия саратовского университета. Новая серия. Серия: Акмеология образования. Психология развития. — 2022. — № 2 (42). — С. 112-119.

4. Коровин С.А., Бородина В.А., Кочеткова К.В., Слободяник К.А., Аврискин М.В. Выбор элективных дисциплин в ТюмГУ на основе цифровых портретов студентов, сформированных на базе анализа подписок пользователя в VK // Математическое и информационное моделирование: материалы Всероссийской конференции молодых ученых. Вып. 20. — Тюмень, 2022. — С. 451-460.
5. Воробьева М.С., Перевалова М.Н. Диагностика предпочтений студентов при проектировании индивидуальных образовательных траекторий // Информатизация образования и методика электронного обучения: цифровые технологии в образовании: материалы v международной научной конференции. Часть 2. — Красноярск, 2021. — С. 60-63.
6. Хечми Шили. Кластеризация в аналитике больших данных: системный обзор и сравнительный анализ (обзорная статья) / Хечми Шили // Научно-технический вестник информационных технологий, механики и оптики. — 2023. Т. 23. — № 5. — С. 967–979.