В. А. Бородина, С. А. Коровин, М. С. Воробьева, М. В. Аврискин

Тюменский государственный университет, г. Тюмень

УДК 004.9

ПОДХОД К ОПРЕДЕЛЕНИЮ ИНФОРМАЦИОННОЙ ПЛОТНОСТИ ОБРАЗОВАТЕЛЬНОГО КУРСА НА ОСНОВЕ ТЕРМИНОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

Аннотация. В статье рассмотрен комбинированный подход к определению информационной плотности лекционных материалов образовательных курсов. Предложены и апробированы метрики информационной плотности, демонстрирующие распределение материала по курсу. Описаны и протестированы алгоритмы для выделения содержательных частей материалов курса.

Ключевые слова: информационная плотность, оценка образовательных курсов, обработка естественного языка, извлечение информации, выделение терминов.

Введение. В настоящее время в вузах, в том числе в ФГАОУ ВО «Тюменский государственный университет», где внедрены образовательные модели ИОТ и «2+2+2» [1, 2], часть дисциплин успешно реализуются в формате онлайн-курсов. Частота выбора курсов, реализуемых в таком формате, обучающимися увеличивается [3, 4]. Для того чтобы создавать и рекомендовать студентам образовательный контент, доступный для их понимания, полезно оценивать соответствие его содержания уровню знаний целевой аудитории.

Образовательный курс включает в себя лекционные и практические блоки, материалы для самостоятельного изучения и тестовые задания. В данном исследовании рассматривается лекционный контент, представленный в виде аудиозаписей, извлеченных из видеоматериалов курса. Для оценки содержания этого лекционного контента используется полученный из аудио текст, сложность восприятия и понимания которого зависит от содержащегося в нем количества информации.

В качестве показателя, влияющего на восприятие текста конкретной аудиторией, Шелестюк Е. В. в своей работе [5] рассматривает информационную плотность, выражающую долю новых смысловых блоков относительно общего объема информации. При высоком значении этого показателя затрудняется понимание текста, а при низком — наблюдается высокая информационная избыточность, которая может быть вызвана повторами и многочисленными разъяснениями, что увеличит объем материала и сделает текст менее привлекательным для целевой аудитории.

Существуют различные подходы к измерению информационных показателей текстов, например, предложенный в статье [6], где информационная избыточность текста вычисляется на основе теории информации Шеннона, однако не учитывается содержание и смысл текста напрямую. Исследователями в работе [7] рассматривается иной подход: использование пропозиционального моделирования для вычисления информационной насыщенности текста, при этом подход был реализован наполовину в ручном режиме, что является его недостатком.

Оценке дидактической сложности учебных материалов посвящена публикация Майера Р. В. [8], где предложен учитывающий структурную и семантическую сложность учебного текста метод, в котором плотность информации рассчитывается с учетом коэффициента свернутости информации каждого слова, значение данного коэффициента у терминов, ранее неизвестных аудитории, выше. Влияние использования терминологии на сложность восприятия материала аудиторией с различным уровнем знаний также отмечается в статье [9].

В работе [10] проведен обзор и корреляционный анализ критериев информационной и лексической насыщенности на примере выпускных квалификационных работ студентов, описана реализация вычисления этих критериев, в том числе информативности, показывающей долю введенных определений, что может быть полезно при работе с терминологией в текстах. Исследование [11] посвящено разработке системы оценивания материалов образовательных курсов. Среди автоматических критериев оценивания качества электронного учебного контента была выделена информационная насыщенность, вычисляемая как доля новых понятий к общему числу слов в тексте, также приведена интерпретация данного показателя.

В рамках данного исследования будет рассмотрен комбинированный подход, основанный на определении информационной плотности как доли новых смысловых блоков среди всех высказываний текста. Под смысловыми блоками подразумеваются разделенные на «упомянутые» (названо только само понятие) и «объясненные» (дано определение и/или объяснение) термины, так как они отражают содержание курса и играют ключевую роль в формировании знаний, при этом будут учитываться только термины определенной предметной области.

Цель исследования — разработка комбинированного подхода к определению информационной плотности текста, полученного из аудиозаписей лекций, составляющих образовательный курс.

Проблема исследования. Дана последовательность лекций одного курса $L = \{l_i\}_{i=1}^{a}$, каждый элемент которой представлен $T_i = \left\{t_g^i\right\}_{g=1}^{b}$ — последовательностью слов или словосочетаний текстового содержания лекции.

Пусть $TBG = \{tbg_m\}_{m=1}^{e}$ — множество терминов, составляющих словарь студента, и $TSA = \left\{tsa_j\right\}_{j=1}^{c}$ — множество терминов, составляющих словарь предметной области. Термин tsa_j представлен кортежем $[name_j, \{definition_j^z\}, z \in \mathrm{N}]$, где $name_j$ — понятие предметной области, $definition_j^z$ — множество определений. При этом $\forall T_i \exists \left\{term_p^i\right\}_{p=1}^{y} = Term_i \subset T_i$ и $Term_i \subset TSA$.

Пусть $TermPrev = \bigcup_{n=1}^{i-1} Term_n$ — множество терминов, содержащихся в предыдущий лекциях. Тогда термин $term\_new$ называется «новым» в T_i, если $term\_new \in Term_i, term\_new \notin TBG, term\_new \notin TermPrev$; а термин $term\_old$ — «уже изученным», если $term\_old \in Term_i, term\_old \in TBG$ или $term\_old \in TermPrev$.

Требуется найти:

- $Terms\_new_i = \left\{term\_new_u^i\right\}_{u=1}^{h}$ — последовательность «новых» терминов, $term\_new_u^i = \left[name_u^i, definition_u^i, quantity_u^i\right]$, где $name_u^i$ — понятие, $definition_u^i$ — выделенное в T_i определение (подмножество слов $\left\{t_g^i\right\}$ или $\emptyset$), $quantity_u^i$ — количество упоминаний термина в тексте;
- $Terms\_old_i$ — последовательность «уже изученных» терминов, $term\_old_r^i = \left[name_r^i, quantity_r^i\right]$, где $name_r^i$ — самое понятие, $quantity_r^i$ — количество упоминаний термина в тексте.

В качестве оценок информационной плотности каждой лекции l_i выбраны четыре метрики:

1) *NTP-LWC (New Terms Proportion per Lecture Word Count)* — доля $Terms\_new_i$ относительно всех терминов лекции из расчета на N слов текста лекции;

2) *CNTR-LD (Course-wide New Terms Ratio Adjusted for Lecture Duration)* —доля $Terms\_new_i$ относительно всех терминов курса с учетом длины лекции в словах относительно всего курса, т. е. показатель обратно пропорционален доле длительности лекции в общей длительности курса;

3) *NTEP-LWC (New Terms with Explanation Proportion per Lecture Word Count)* — доля $Terms\_new_i$ с определениями или объяснениями относительно всех терминов лекции из расчета на N слов текста лекции;

4) *ECNT-L (Explanation Coverage of New Terms per Lecture)* — доля предложений с определениями или объяснениями новых терминов от общего числа предложений лекции.

Материалы и методы. В качестве основного материала для работы был использован курс «Теория вероятностей», автором которого является профессор ТюмГУ А. Н. Шевляков (далее — курс А). Для сравнения был рассмотрен курс «Добрая теория вероятностей для ЕГЭ», автор курса — С. М. Балакирев (далее — курс В). Курс А представлен 32 видеороликами и сгруппирован по темам в семь лекций, их средняя продолжительность составляет 69 минут. Курс В состоит из 21 видеоролика средней продолжительностью 9 минут. С помощью модели для автоматического распознавания речи в текст "Whisper Medium" были получены транскрипты извлеченной из видеофайлов аудиодорожки в формате TSV с разбиением на фрагменты с указанием времени его начала и окончания, которые были скорректированы, чтобы каждый фрагмент содержал законченное предложение.

Словарь предметной области для анализа курсов по теории вероятностей, состоящий из 310 терминов, сформирован из предметного указателя пособия Н. И. Черновой «Теория вероятностей»[1] с последующим получением определений терминов из «Википедии» и генерацией двух вариантов «эталонных» определений с помощью LLM-модели "Sber GigaChat". Список из 182 терминов, определяющий словарь студента-первокурсника, был составлен на основе спецификации контрольно-измерительных материалов для проведения ЕГЭ по математике профильного уровня в 2024 году[2].

Для определения информационной плотности текста лекций курса с помощью разрабатываемого подхода необходимо в тексте лекции T_i найти $Term_i$ — множество упоминания терминов из словаря предметной области TSA, учитывая, что сложные слова для автоматического перевода речи в текст (например, фамилии ученых) могут быть искажены. Для нахождения $Term_i$ используется следующий алгоритм:

1) предобработка T_i и для каждого понятия из TSA: лемматизация, удаление стоп-слов и небуквенных символов;

2) для каждой n-граммы t_g^i, где $n = 1, \dots, 5$ вычисление коэффициента сходства $sim\_term(t_g^i, tsa_j)$ с каждым термином, где функция sim_term проверяет выполнение следующих условий: совпадающие части речи, коэффициент сходства строк и косинусное расстояние между векторными представлениями выше заданных порогов;

[1] Чернова Н. И. Теория вероятностей: пособие. URL: https://tvims.nsu.ru/chernova/sibguti/tv-sibguti.pdf.

[2] Спецификации КИМ ЕГЭ по математике. URL: https://fipi.ru/ege/demoversii-specifikacii-kodifikatory#!/tab/151883967-2.

3) для каждой n-граммы выбор пары (t_g^i, tsa_j) с наибольшим коэффициентом сходства больше заданного порога.

Выделение определений, извлеченных на предыдущем этапе терминов $Term_i = \{term_p^i\}_{p=1}^{y}$, в тексте, разбитом на предложения, $T_i = \{sentence_{ss}^i\}_{ss=1}^{rr}$ осуществляется с помощью следующих шагов:

1) предобработка T_i, каждого понятия и всех вариантов эталонного определения из TSA аналогично шагу № 1 в предыдущем алгоритме;

2) для каждого $term_p^i$ выделяются кандидаты $CDN = \{cdn_q\}_{q=1}^{v}$ — где каждый кандидат cdn_q представляет собой набор последовательных предложений $[sentence_{ss-\varepsilon_1}^i, \dots, sentence_{ss+\varepsilon_2}^i]$, где $sentence_{ss}^i$ — предложение, содержащее $term_p^i$, а ε_1 и ε_2 — границы кандидата (0 или 2). Для каждого $sentence_{ss}^i$ рассматриваются все сочетания ε_1 и ε_2;

3) вычисление косинусного расстояния между векторными представлениями всех кандидатов CDN и каждого определения из TSA;

4) для каждого эталонного определения $definition_z^p$ термина $term_p^i$ формируется рейтинг на основе косинусного сходства и частоты упоминания термина $term_p^i$ в лекции l_i, первые несколько кандидатов с оценкой выше заданного порога рассматриваются далее;

5) итоговый список значимых фрагментов формируется из кандидатов, семантически близких наибольшему числу эталонных определений.

Сходство строк вычислялось на основе расстояния Левенштейна. Векторные представления предложений были получены с помощью предобученной модели BERT[1]. Пороговые значения для сходства терминов и определений подбирались эмпирически. На основе извлеченных из текстов лекций терминов и их определений вычисляются показатели информационной плотности по метрикам: *NTP-LWC, CNTR-LD, NTEP-LWC* и *ECNT-L*.

Результаты. Для реализации комбинированного подхода для определения информационной плотности лекций образовательных курсов были разработаны алгоритмы извлечения терминов и выделения определений, для оценки работы которых были размечены транскрипты видеороликов (около 900 предложений). Качество извлечения терминов оценивалось по метрикам: $accuracy$, $precision$, $recall$ и f-мера, при этом корректными примерами для всех метрик, кроме $accuracy$, считались частично совпадающие эталонные и полученные списки терминов. Результаты для извлечения терминов представлены в табл. 1.

Таблица 1

Оценка извлечения терминов на размеченном датасете

Рассматриваемые термины	*accuracy*	*precision*	*recall*	*f-мера*
Все термины, выделенные вручную	0.739	0.980	0.742	0.845
Термины только из словаря	0.949	0.974	0.974	0.974

[1] DeepPavlov/rubert-base-cased-sentence. URL: https://huggingface.co/DeepPavlov/rubert-base-cased-sentence.

Метрика $recall$ показывает, что большая часть содержащихся в словаре терминов, в том числе с некорректным написанием из-за перевода аудиозаписи в текст, были выделены корректно, за исключением случаев употребления неполного названия термина или замены слова однокоренным другой части речи. Было выявлено, что слова, разные по смыслу, но схожие по написанию, определяются как одинаковые, чего можно избежать, определяя сходство векторных представлений не только самих фраз, но их контекста. Более низкие показатели по метрикам $accuracy$ и $recall$ для всех терминов, выделенных вручную, объясняются тем, что в тексте лекции присутствовали термины, отсутствующие в словаре.

Все предложения, выделенные в качестве определений терминов, оценивались вручную, так как при разметке отмечались только те предложения, которые наиболее похожи на определения, тогда как в результате работы алгоритма было выделено значительно большее количество групп предложений, близких по содержанию к эталонным определениям, при этом некоторые предложения были выделены как определения сразу нескольких терминов.

В результате 26% предложений можно считать определением или объяснением, т. е. предложением, содержащим значимую информацию о термине. Остальные 74% предложений можно условно поделить на три группы: соседние с объяснением; содержащие значимую информацию о другом термине; фрагменты примеров и рассмотренных задач, где употребляются термины. В результате корректно обработано не менее 80% предложений, в эту долю входят корректно распознанные предложения, содержащие значимую информацию о терминах, и предложения, не содержащие существенной информации предложения, не выделенные в качестве определений.

Значения, полученные при расчете метрик информационной плотности для курса А, с использованием $N = 100$ как значения среднего темпа речи, представлены в табл. 2. Первые три метрики в разных вариациях показывают долю новой информации, измеряемой употребляемыми терминами, с учетом длины текста лекции, на рис. 1 представлен график изменения каждого из этих показателей внутри курса. По значениям метрик *NTP-LWC* и *NTEP-LWC* можно сделать вывод о том, что доля новой информации от лекции к лекции постепенно уменьшалась (за исключением пятой лекции, где наблюдается скачок). Разница между метриками *NTP-LWC* и *NTEP-LWC* показывает долю новых необъясненных терминов и может использоваться для корректировки материалов курса в виде добавления объяснений некоторых терминов.

Таблица 2

Параметры метрик информационной плотности по курсу А

Метрика	*MIN*	*MAX*	*AVG*	*10%*	*25%*	*50%*	*75%*	*90%*
NTP-LWC	0.0022	0.0246	0.0126	0.0058	0.0083	0.0136	0.0157	0.0202
CNTR-LD	0.0018	0.0171	0.0084	0.0020	0.0050	0.0088	0.0107	0.0135
NTEP-LWC	0.0011	0.0191	0.0081	0.0023	0.0045	0.0073	0.0100	0.0146
ECNT-L	0.0208	0.2725	0.1358	0.0470	0.0759	0.0927	0.2064	0.2611

Метрика *CNTR-LD* показывает распределение новой информации между всеми лекциями и напрямую зависит от того, насколько равные по длительности лекции составляют курс,

в данном случае разброс значений может быть объяснен влиянием существенных отличий в длительности лекций. Кроме того, метрика *CNTR-LD* показывает, что несмотря на то, что в первой лекции почти все термины являются новыми, она не является самой информационно плотной. Метрика *ECNT-L* показывает долю предложений лекции, объясняющих новые термины или содержащих иную информацию, связанную с их употреблением. По значениям метрики *ECNT-L* можно делать выводы о концентрации новой информации в каждой лекции курса.

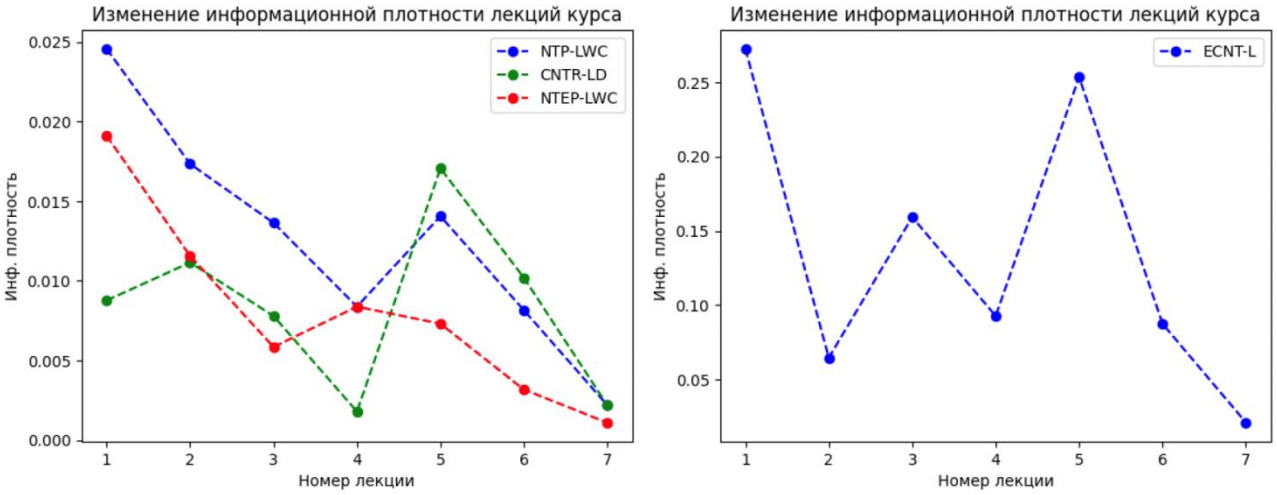

Рис. 1. Графики изменения метрик информационной плотности на курсе А

С помощью предложенных метрик можно сравнивать различные курсы, связанные одной предметной областью. Значения метрик для курса В представлены в таблице 3. Сравнивая полученные для двух курсов значения, можно заметить, что в курсе В есть лекции, не содержащие новой информации (значения метрик 0), причем таких лекций по метрике *ECNT-L* 5 из 21. При этом максимальные значения всех метрик у курса В выше, чем у курса А, а также выше значения метрик, начиная с 25-го и 50-го процентиля. В совокупности это показывает, что распределение материала между лекциями курса В менее сбалансированно.

Таблица 3

Параметры метрик информационной плотности по курсу В

Метрика	*MIN*	*MAX*	*AVG*	*10%*	*25%*	*50%*	*75%*	*90%*
NTP-LWC	0.0000	0.1293	0.0455	0.0000	0.0204	0.0442	0.0590	0.1115
CNTR-LD	0.0000	0.0311	0.0092	0.0000	0.0036	0.0063	0.0141	0.0211
NTEP-LWC	0.0000	0.1293	0.0337	0.0000	0.0000	0.0246	0.0454	0.0892
ECNT-L	0.0000	0.4691	0.1158	0.0000	0.0000	0.0723	0.2113	0.2697

Заключение. В рамках проведенного исследования был предложен и разработан комбинированный подход к определению информационной плотности составляющих курс лекций по предложенным метрикам. Реализованный подход был апробирован на курсах по теории вероятностей авторов А. Н. Шевлякова и С. М. Балакирева. Для выделения содержательных

частей материалов курса были разработаны алгоритмы извлечения терминов и выделения определений из текстового представления лекций и оценено качество их работы. Для разработки алгоритмов использовался язык программирования Python версии 3.10 и библиотеки: Pandas, spaCy, Gensim, Scikit-Learn, Transformers и Torch.

Дальнейшая работа может быть направлена на модификацию предложенных алгоритмов путем более точного выделения границ значимых фрагментов текста и выбора одного наиболее подходящего термина, определением которого может являться предложение. Для распознавания предложений со значимой информацией можно применить иные подходы, например, решать задачу классификации, выделяя определения терминов, а затем сопоставляя их с терминами словаря путем сравнения с эталонными определениями. Возможно, имеет смысл проведение дополнительных исследований для выделения определений терминов и примеров для их объяснения отдельно, так как эти содержательные части выполняют разные роли.

Получаемые значения метрик информационной плотности образовательного контента могут быть полезны не только для анализа распределения материала по курсу и рекомендаций доступных для понимания потенциальными слушателями, но и для сравнения образовательного контента. В перспективе возможно масштабировать полученные результаты путем создания программного продукта для анализа множества образовательных курсов и получения данных для проверки исследовательских гипотез в сфере образования.

СПИСОК ЛИТЕРАТУРЫ

1. Федорова Н.К. Индивидуализация образования: модель Тюменского государственного университета / Н.К. Федорова. — Текст: электронный // EdCrunch Томск: материалы международной конференции по новым образовательным технологиям, Томск, 29-31 мая 2019 года. — Томск: Издательский Дом Томского государственного университета, 2019. — С. 301-305. — URL: https://elibrary.ru/download/elibrary_42266584_84397714.pdf (дата обращения: 03.03.2024).
2. Гаврилюк Т.В. Переход к обучению по индивидуальным образовательным траекториям в оценках студентов и преподавателей (на примере Тюменского государственного университета) / Т.В. Гаврилюк, Т.В. Погодаева. — Текст: электронный // Социологический журнал. — 2023. — Т. 29, № 2. — С. 51-73. — URL: https://cyberleninka.ru/article/n/perehod-k-obucheniyu-po-individualnym-obrazovatelnym-traektoriyam-v-otsenkah-studentov-i-prepodavateley-na-primere-tyumenskogo (дата обращения: 03.03.2024).
3. Рощина Я.М. Спрос на массовые открытые онлайн-курсы (МООС) опыт российского образования / Я.М. Рощина, С.Ю. Рощин, В.Н. Рудаков. — Текст: электронный // Вопросы образования. — 2018. — № 1. — С. 174-199. — URL: https://cyberleninka.ru/article/n/spros-na-massovye-otkrytye-onlayn-kursy-mooc-opyt-rossiyskogo-obrazovaniya (дата обращения: 12.03.2024).
4. Кичикова Д.В. Внедрение онлайн-курсов в образовательный процесс: опыт ТюмГУ / Д.В. Кичикова, И.А. Тимофеева. — Текст: электронный // Инновационные технологии в образовательной деятельности: Материалы XXIV Международной научно методической конференции, Нижний Новгород, 02 марта 2022 года. — Нижний Новгород: Нижегородский государственный технический университет им. Р. Е. Алексеева, 2022. — С. 183-188. — URL: https://elibrary.ru/download/elibrary_48276288_90085885.pdf (дата обращения: 13.03.2024).
5. Шелестюк Е.В. Методика выявления количественных показателей истинности, информативности и информационной плотности текстов / Е В. Шелестюк — Текст: электронный // Система языка: синхрония и диахрония: Межвузовский сборник научных статей. — Уфа: РИЦ БашГУ, 2009. —

С. 151-156. — URL: https://www.academia.edu/6467473/Методики_выявления_количественных_показателей_истинности_информативности_и_информационной_плотности_текста (дата обращения: 04.04.2024).
6. Рогожникова Т.М. Построение математической модели для оценки информационной избыточности текста / Т.М. Рогожникова, Н. Н. Воронов. — Текст: электронный // Теория и практика языковой коммуникации: материалы VIII международной научно-методической конференции, Уфа, 23–24 июня 2016 года. — Уфа: ГОУ ВПО "Уфимский государственный авиационный технический университет", 2016. — С. 216-232. — URL: https://elibrary.ru/download/elibrary_28335966_35840890.pdf (дата обращения: 04.04.2024).
7. Солнышкина М.И. Пропозициональное моделирование для оценки информативности текста / М.И. Солнышкина, Е.В. Мартынова, М.И. Андреева. — Текст: электронный // Ученые записки национального общества прикладной лингвистики. — 2020. — № 3 (31). — С. 47-57. — URL: https://elibrary.ru/download/elibrary_43896198_85937173.pdf (дата обращения: 4.04.2024).
8. Майер Р.В. Проблема оценки сложности дидактических объектов и ее решение / Р.В. Майер — Текст: электронный // АНИ: педагогика и психология. — 2019. — № 4 (29). — С. 126-129. — URL: https://cyberleninka.ru/article/n/problema-otsenki-slozhnosti-didakticheskih-obektov-i-ee-reshenie (дата обращения: 14.04.2024).
9. Матвеева О.В. «Информационная насыщенность» как характерная особенность специальных текстов / О.В. Матвеева. — Текст: электронный // Актуальные проблемы гуманитарных и естественных наук. — 2014. — № 4-1. — С. 363-368. — URL: https://cyberleninka.ru/article/n/informatsionnaya-nasyschennost-kak-harakternaya-osobennost-spetsialnyh-tekstov (дата обращения: 24.04.2024).
10. Стрельников А.И. Исследование методов анализа информационной и лексической насыщенности научных текстов / А.И. Стрельников, М.С. Воробьева — Текст: электронный // Математическое и информационное моделирование: материалы Всероссийской конференции молодых ученых, Тюмень, 18-23 мая 2022 года / Министерство науки и высшего образования Российской Федерации, Тюменский государственный университет, Институт математики и компьютерных наук. — Вып. 20. — Тюмень: ТюмГУ-Press, 2022. — С. 221-229. — URL: https://elibrary.ru/download/elibrary_49518827_29997537.pdf (дата обращения: 24.04.2024).
11. Городович А.В. Модели, алгоритмы и инструментальная система оценивания и модернизации учебного контента: дис. канд. техн. наук / А.В. Городович. — Томск, 2022. — 166 с. — Текст: электронный. — URL: https://postgraduate.tusur.ru/system/file_copies/files/000/003/204/ original/Диссертация_Городович.pdf (дата обращения: 14.04.2024).