МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Тольяттинский государственный университет»

Институт математики, физики и информационных технологий
(наименование института полностью)

Кафедра «Прикладная математика и информатика»
(наименование кафедры)
01.04.02 Прикладная математика и информатика
(код и наименование направления подготовки)

Математическое моделирование
(направленность (профиль)

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему «Методы и алгоритмы интеллектуального анализа данных для оценки поведения студента в системе электронного обучения»

Студент Е.Ю. Лесных _____
(И.О. Фамилия) (личная подпись)

Научный руководитель О.М. Гущина _____
(И.О. Фамилия) (личная подпись)

Руководитель программы д.ф.-м.н., доцент, С.В. Талалов _____
(ученая степень, звание, И.О. Фамилия) (личная подпись)
«_____»_____20_____г.

Допустить к защите
Заведующий кафедрой к.т.н., доцент, А.В. Очеповский _____
(ученая степень, звание, И.О. Фамилия) (личная подпись)
«_____» _____20_____г.

Тольятти 2019

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В настоящее время электронное обучение - важный элемент учебной и преподавательской деятельности в подавляющем количестве колледжей и университетов по всему миру. Во многих высших учебных заведениях основные библиотечные службы, ИТ и службы поддержки поддерживают услуги по электронному обучению. Из отчета UCISA за июль 2016 г. видно, что основные информационные службы предоставляют услуги ИОС и организуют его поддержку в 93% высших учебных заведений.

Электронное образование содержит в себе электронные учебники, образовательные услуги и технологии. К.В. Буваков [3] рассказал в докладе о том, что современные студенты и школьники — это поколение, для которых электронная информация и способы ее получения является повседневной составляющей жизни. В общем, современные технологии в образовании воспринимаются студентами положительно, - знания, умения и навыки, которые они получают, используются в дальнейшем не только в карьерном росте, но и в повседневной жизни.

В обзоре российского и мирового рынка электронного обучения [25] выделяется то, что применение технологий в обучении дает безграничные возможности для развития и получения полезных знаний, навыков. Получать знания в XXI веке стало быстро, увлекательно и доступно. С точки зрения интересов государства, электронное образование предоставляет возможность получения одинакового уровня образования людям, которые живут в удаленных точках страны, т. е. электронное образование позволяет устранить образовательное неравенство.

Актуальность исследования обусловлена необходимостью применить методы интеллектуального анализа данных для оценки поведения студента, что позволит с большей вероятностью прогнозировать возможные ошибки и подмену данных в системе электронного обучения.

В условиях формирования системы непрерывного образования и повышения квалификации, дистанционное обучение стало оптимальной

формой опережающего обучения и профессиональной переподготовки. Развитие дистанционного образования требует специальных информационных возможностей для обеспечения бесперебойной работы сети, а также для облегчения работы преподавателей и студентов.

Такие ученые, как А.А. Андреев, Ж.Н. Зайцева, В.А. Каймин, Э.В. Кинелев, Д.Э. Колосов, Ю.Г. Круглов, И.А. Липский, В.И. Овсянников, Е.С. Полат, А.А. Тихомиров, А.Н. Тихонов, В.В. Ярных и др., принесли большую пользу в разработке методик дистанционного обучения, однако, вопрос качества остается открытым.

В 1989 году Гриорий Пятецкий-Шапиро представил область интеллектуального анализа данных на семинаре. В то время, когда он был работником компании GTE Labs, заинтересовался возможностью автоматического нахождения определённых правил, для ускорения некоторых запросов к объемным базам данных. Тогда были введены два термина — Data Mining и Knowledge Discovery In Data. В 1993 году была представлена первая рассылка «Knowledge Discovery Nuggets», на год позднее, в 1994 году Гриорий Пятецкий-Шапиро создал один из первых сайтов по Data Mining [29].

Задачи, которые решаются при помощи методов интеллектуального анализа данных, можно разделить на описательные (англ. descriptive) и предсказательные (англ. predictive) [29].

В описательных задачах дается наглядное описание имеющихся скрытых закономерностей, тогда как в предсказательных решается вопрос о предсказании тех случаев, для которых данных ещё нет. Описательные задачи:

- поиск правил или паттернов;
- кластерный анализ;
- нахождение регрессионной модели.

Предсказательные задачи:

- классификация объектов;
- регрессионный анализ.

В настоящее время, интеллектуальный анализ данных является развивающимся направлением, с помощью которого можно улучшить качество электронного образования [28].

Целью работы является применение методов и алгоритмов интеллектуального анализа данных для исследования поведения студента в системе электронного образования.

Объект исследования – анализ поведения студента при прохождении курса электронного обучения.

Предмет исследования – методы и алгоритмы интеллектуального анализа данных в системе электронного обучения.

Гипотеза исследования состоит в том, что можно применять методы и алгоритмы интеллектуального анализа для принятия решения о корректности поведения студента в системе электронного обучения, если:

- обозначены методы оценки поведения студента;
- реализована технология, которая обеспечивает достоверный анализ поведения студента при прохождении курса электронного обучения;
- проведено исследование поведение студента при помощи методов и алгоритмов интеллектуального анализа данных.

Исходя из цели исследования и для проверки выдвинутой гипотезы, необходимо решить следующие **задачи**:

- провести сравнительный анализ существующих способов и средств анализа поведения студента при прохождении курса электронного обучения;
- исследовать возможные критерии сравнения поведения студентов;
- исследовать методы и алгоритмы интеллектуального анализа данных;
- разработать математическую модель оценивания студента по данным его поведения;

- спроектировать и реализовать математическую модель интеллектуального анализа поведения студента при прохождении курса электронного обучения;
- проверить эффективность разработанной модели и определить результативность ее внедрения.

Научная новизна исследования состоит в использовании методов и алгоритмов интеллектуального анализа данных для оценки поведения студента в системе электронного образования.

Практическая значимость заключается в анализе поведения студента при прохождении курса электронного обучения с использованием методов и алгоритмов интеллектуального анализа данных, который позволит с большей вероятностью прогнозировать возможные ошибки и подмену данных в системе дистанционного обучения.

Положения, выносимые на защиту:

1. Алгоритмы интеллектуального анализа данных для оценки проведения студента в системе электронного образования.
2. Компьютерная и математическая модели алгоритмов интеллектуального анализа данных для оценки проведения студента в системе электронного образования.

Работа представляет собой результат теоретической и практической деятельности в области оценки поведения студента при прохождении курса электронного образования, используемой в образовательном процессе для повышения качества дистанционного обучения.

Определив цель, объект и предмет исследования перейдем к анализу научных работ, описывающих способы и средства анализа поведения студента при прохождении курса электронного обучения, а также проанализируем, какие для этого используются критерии сравнения поведения студента.

Объем и структура диссертации: диссертационное исследование состоит из введения, четырех глав, заключения, библиографии (66 наименований). Работа изложена на 70 страницах, содержит 11 рисунков и 1 таблицы.

Глава 1 АНАЛИЗ ПОВЕДЕНИЯ СТУДЕНТОВ В СИСТЕМЕ ЭЛЕКТРОННОГО ОБУЧЕНИЯ

1.1 Системы электронного обучения

С развитием информационных технологий, электронное образование получило широкое применение. В настоящее время учебная и преподавательская деятельность включает в себя весовой элемент – электронное образование. Согласно отчету Ассоциации информационных систем университетов и колледжей (UCISA) за июль 2016 г., основные ИТ-службы предоставляют услуги информационной обучающей среды (далее ИОС) и организуют ее сопровождение в 93% высших учебных заведений.

Электронное образование представляет собой такой вид образовательного процесса, где обучение происходит при помощи цифровых технологий, с использование электронных учебников, образовательных услуг и технологий. К.В. Буваков [1] рассматривает в своем докладе то, что нынешние студенты и школьники — это поколение электроники и гаджетов, для которых электронный способ получения как учебной, так и развлекательной информации является неотъемлемой частью жизни. Большинство студентов приветствует современные технологии в образовании — знания, умения и навыки, приобретенные во время такого обучения, используются в дальнейшем в самосовершенствовании и карьерном росте.

В обзоре мирового и российского рынка электронного обучения [9] отмечается, что использование технологий в обучении открывает безграничные возможности для развития и получения новых знаний и навыков, снимая различные барьеры. Учиться в XXI веке стало быстро, увлекательно и доступно. С точки зрения интересов государства, электронное образование – это возможность получения одинакового уровня образования людям, проживающим в разных точках страны, т. е. способ устранения образовательного неравенства.

В последние годы количество исследований, изучающих эффективность электронного обучения, возросло. Это в первую очередь связано с возросшими

возможностями для ИТ и обучения, а также повышенным вниманием со стороны политиков и организаций к тому, «что работает» в обучении.

В диссертации [8] автор пишет, что активное внедрение электронного образования требует наличия как психологической и профессиональной подготовленности всех участников данного процесса, так и его теоретического освоения и методического обеспечения. Низкий уровень востребованности надежного профессионального образования на современном российском рынке труда, обусловленный недостаточными темпами развития экономики, влияет на интерес абитуриентов, студентов, их родителей, работодателей, самих высших учебных заведений в качестве образовательных услуг и желание части обучающихся иметь диплом при минимальных умственных и финансовых затратах.

Наличие нескольких способов понимания эффективности электронного обучения позволяет профессионалам и исследователям гибко измерять и определять эффективность решения электронного обучения. Невыполнение концепции электронного обучения могут привести к недоразумениям, а аспекты эффективности, которые имеют наибольшую ценность для участников и заинтересованных сторон, не могут рассматриваться. Освещение многих определений эффективности может привести к размышлениям и вдохновению для надлежащего использования концепции эффективности, что позволит специалистам по обучению лучше согласовывать свои ожидания и направлять свои измерительные усилия на то, что важно для них и для заинтересованных сторон.

Наличие нескольких способов понимания эффективности электронного обучения позволяет профессионалам и исследователям гибко измерять и определять эффективность решения электронного обучения. Невыполнение концепции электронного обучения могут привести к недоразумениям, а аспекты эффективности, которые имеют наибольшую ценность для участников и заинтересованных сторон, не могут рассматриваться. Освещение многих определений эффективности может привести к размышлениям и вдохновению

для надлежащего использования концепции эффективности, что позволит специалистам по обучению лучше согласовывать свои ожидания и направлять свои измерительные усилия на то, что важно для них и для заинтересованных сторон.

Необходимо повысить уровень образования посредствам анализа поведения студента с помощью методов и алгоритмов интеллектуального анализа данных в системе электронного обучения для выявления подмены или использования студентом справочной литературы во время итоговых работ.

Дистанционное обучение - это вид обучения, базирующийся на использовании разнообразных традиционных, новых информационных и телекоммуникационных технологий, с помощью которых создаются условия для свободного выбора образовательных дисциплин, согласно стандартам, виртуального общения с преподавателем, при этом обучение не зависит от расположения студента в пространстве.

В условиях постоянного развивающихся сфер деятельности человека, постоянно создаются системы непрерывного образования и повышения квалификации. Электронное обучение – это очень удобная форма обучения и профессиональной переподготовки.

Системы электронного обучения получили широкое распространение, что дало исследователям больший объем информации по сравнению с традиционным образовательным процессом. Это связано с активным использованием в электронном обучении различных технологий сбора данных, и с большим масштабом аудитории электронных образовательных сред (ЭОС). В начале 21 века появилось новое направление в области интеллектуального анализа данных – анализа образовательных данных, чему способствовал рост объема данных в системах электронного образования [1].

Различные методы кластеризации были применены в недавних исследованиях для прогнозирования успеваемости студентов. Алана М. и соавт. в [25] были исследованы методы поддержки процесса принятия решений

учителем путем группирования учащихся и планирования задач соответственно, был достигнут положительный результат.

В [27] автор исследовал различные исследования и наборы данных, вращающихся вокруг области EDM. Автор пришел к выводу, что EDM может использоваться как часть широкого диапазона зон, включая распознавание студентов, подвергающихся риску, выявление потребностей для адаптации потребностей различных групп студентов, увеличение показателей выпускников.

Еще одно научное исследование [29], согласующееся с [27], проведено Амджадом Абу Саа, которое изучает и прогнозирует успеваемость учащихся в различных сценариях с использованием методов кластеризации. В аналогичном исследовании [30] Томмазо и Алекс Бауэрс проанализировали различные аналитические методы: интеллектуальный анализ данных, обучение и академическую аналитику, и пришли к выводу, что применение методов интеллектуального анализа данных дает положительные результаты. , K-means - это современный алгоритм кластеризации на основе разделов, который применяется в EDM. Например, специальный выбор места учащегося в лаборатории или классе и его влияние на оценку ученика были оценены Иванцевичем, Целиковичем и Луковичем [31]. Другое исследование, представленное Ying, et al. [32] использовал K- means для понимания поведения студентов на основе описания 40 студентов. В исследовании, проведенном Eranki & Moudgalya [33], K-means был применен для изучения влияния человеческих характеристик на успеваемость ученика во время прослушивания музыки, давая очевидные классификации.

Анализом образовательных данных называют направление исследований, которое направлено на применение методов интеллектуального анализа данных, машинного обучения и статистики к информации, производимой образовательными учреждениями.

АОД пытается выявить закономерности из данных, получаемых в процессе обучения. Эти данные могут быть весьма обширны и содержать

большое число подробностей. Например, некоторые системы управления обучением (learning management system) отслеживают информацию о том, в какое время студент получил доступ к тому или иному учебному объекту, как часто студенты обращались к этому объекту и как много минут объект отображался на экране компьютера студента, какие тестовые задания выполнялись первыми и т.п. Уровень подробности этих данных таков, что даже короткий сеанс работы с ЭОС может произвести большой объем данных для анализа.

Другие данные могут включать гораздо меньше подробностей. Например, электронная зачетная книжка студента состоит из упорядоченных во времени списков курсов, которые прошел данный студент, оценки, полученные за тот или иной курс и т.п. АОД использует оба выше описанных типа данных, чтобы обнаружить закономерности в учебе студентов. Изучая данные ЭОС, можно заметить связь между темами, которым обучающийся уделял внимание во время прохождения курса, и итоговой оценкой этого студента.

АОД тесно связан с аналитикой обучения. У этих дисциплин во многом совпадают цели и задачи. Некоторые специалисты [1, 2] видят различие между данными направлениями в том, что АОД нацелен на автоматизации выявления закономерностей в образовательных данных, в то время как аналитика обучения большее внимание уделяет подготовке данных в виде, пригодном для анализа человеком.

Исследуемые данные являются одной из особенностей АОД. Часто они имеют сложную структуру, например, иерархическую, семантическую, представляя собой трудности для анализа традиционными методами. Важно и то, что это могут оказаться данные, с которыми работниками сферы образования обычно не сталкиваются. Анализ журналов использования сайта или базы данных образовательной системы может предоставить следующую информацию [5]:

- часто посещаемые страницы;
- используемые браузеры;

- количество посещенных страниц;
- место, откуда студенты входят в систему;
- количество посещений, их длительность для каждого обучающегося;
- популярные ключевые слова;
- количество кликов, просмотров, скачиваний материалов;
- количество страниц, которые были просмотрены обучающимся за сеанс;
- список электронных ресурсов, которые использует обучающийся;
- объём учебного материала, изучаемого обучающимся перед выполнением задания.

Такие данные может предоставить система Moodle. В работах [9, 10] был описан процесс сбора данных из Moodle, их анализ и некоторые результаты, полученные с помощью интеллектуального анализа данных.

Аналогом Moodle является система iSpring и в ней можно увидеть, кто, сколько раз и какие материалы просматривал. Соответственно, можно провести анализ активности студентов (сотрудников), популярности материалов и так далее.

Нужно отметить, что сервис iSpring Online LMS не является полноценной системой управления обучением, и с его помощью нельзя автоматизировать такие процессы как контроль знаний или управление программами обучения. Однако при использовании совместно с редактором iSpring Suite предоставляемых возможностей достаточно для создания и доставки учебного контента в рамках несложного учебного процесса.

GetCourse, Coursmos, EduTerra.PRO, PushtoLearn, Emdesell, StarStaff, Edulance и многие другие так же предоставляют возможности онлайн обучения и ведут статистику активности пользователей.

1.2 Модели анализа процесса онлайн-обучения и постановка задачи

В настоящее время изучение поведения в онлайн-обучении в среде больших данных в основном сосредоточено на трех аспектах, как показано на Рисунок 1.

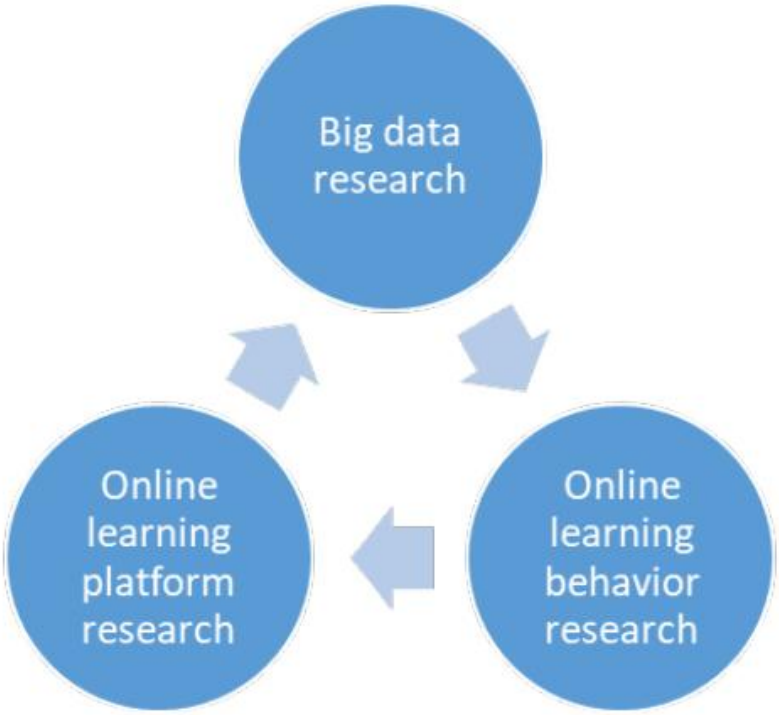

Рисунок 1 – Процесс изучения поведения в онлайн обучении

Big data research - исследование больших данных. На данный момент основные технологии анализа больших данных включают в себя визуальный анализ, алгоритмы интеллектуального анализа данных, семантический анализ. Например, модель Google Mapreduce ориентирована на большие наборы данных. Университет Пердью использует технологии больших данных для создания механизмов раннего предсказания результатов, собирая данные о студентов курса. Основное внимание современных исследователей к исследованию больших данных - извлечение ценности из них.

Online learning platform research - исследование онлайн-платформы обучения. Платформы онлайн-обучения можно разделить на две категории: коммерческие и некоммерческие. Например, Moodle является самой широко используемой в мире бесплатной платформой для онлайн-обучения, которая не

только используется во многих университетах мира, но также предлагает разнообразные инструменты онлайн-общения, всевозможные учебники для обучения. В последние годы типичным представителем платформы онлайн-обучения является массовый открытый онлайн-курс, также известный как МООС.

Online learning behavior research - исследование поведения в онлайн-обучении. Много исследований связано с поведением, объект исследования и выбор образца имеют определенные ограничения, в то время как некоторые исследования, касающиеся поведения в процессе онлайн-обучения, проводились при участии исследователей.

Кеннет изучал влияние учебного поведения и рефлексивного обучения на онлайн-бизнес-курсах. Приор разделил влияние онлайн-обучения на три аспекта: отношение к обучению, информационная грамотность и самоэффективность, Бучер изучил различные уровни начальных знаний и соотнес результаты обучения. Это показало, что более высокий уровень начальных знаний может привести к тому, что учащиеся приобретут более глубокий уровень в обучении.

Следуя модели анализа и модели процесса онлайн-обучения, модель анализа делится на три части: кластерный анализ, анализ рекомендаций и корреляционный анализ. В соответствии с процессом решения проблем, процесс модели анализа поведения в режиме онлайн обучения подразделяется на обработку данных, выбор метода и процесс анализа, вывод результатов и так далее, как показано на Рисунок 2.

На первом этапе собирается информация о студентах, которая хранится в системе электронного обучения, далее проводятся горизонтальный и логический анализы и осуществляется вывод результатов.

Также применение интеллектуального анализа данных в системах электронного обучения можно охарактеризовать как итеративный цикл, в котором приложения интеллектуального анализа данных вносят вклад в улучшение обучения, а также используют полезные знания для принятия

решений. Процесс интеллектуального анализа данных представлен на Рисунок 3.

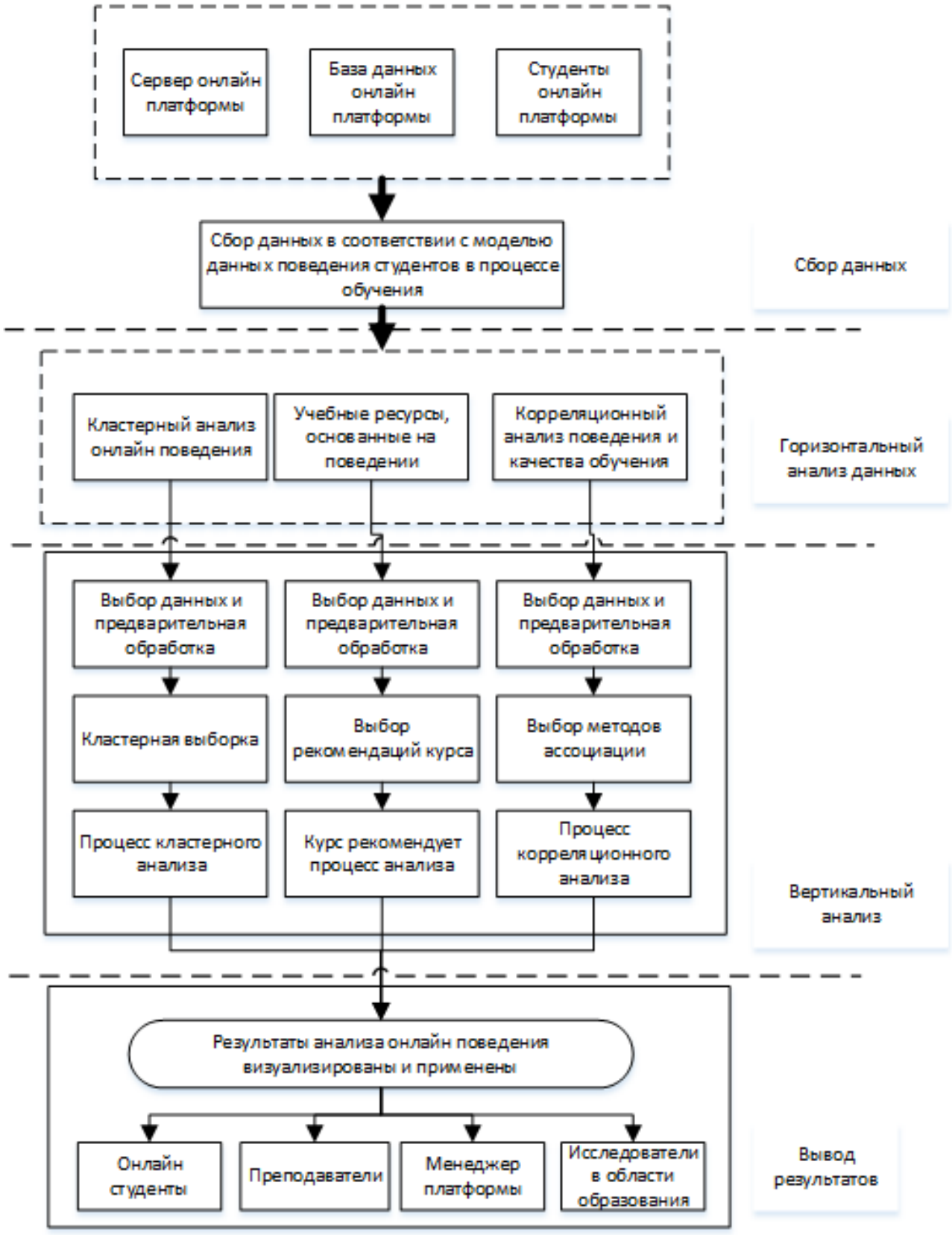

Рисунок 2 – Модель процесса анализа данных в онлайн-обучении

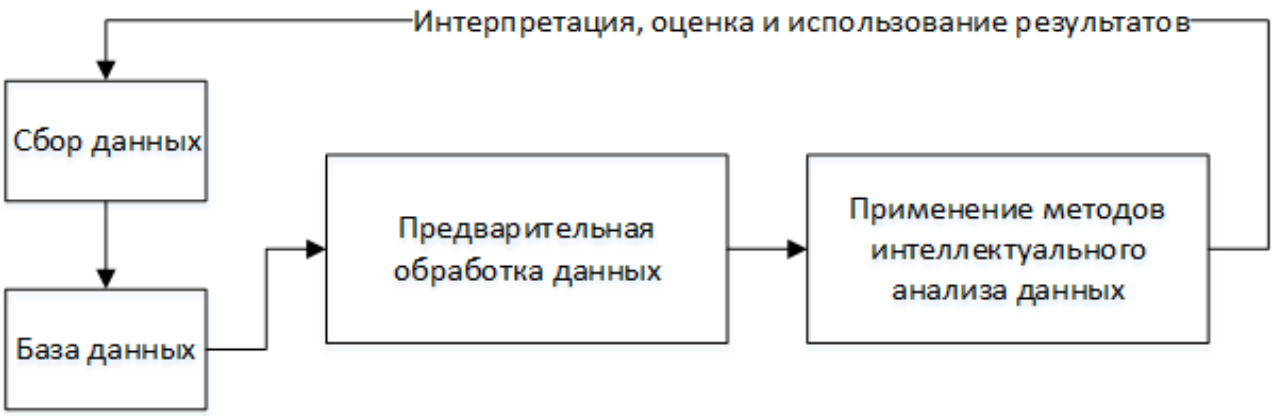

Рисунок 3 – Процесс интеллектуального анализа данных

Рассмотрим этапы более подробно.

1. Сбор данных. Информация об использовании и взаимодействии хранится в базе данных. Информация о взаимодействии хранится в базе данных.
2. Предварительная обработка данных. Данные очищаются и преобразуются в соответствующий формат для анализа. Для предварительной обработки данных можно использовать инструмент администратора базы данных или какой-то конкретный инструмент предварительной обработки. Данные преобразуются в соответствующий формат.
3. Применение интеллектуального анализа данных. Алгоритмы интеллектуального анализа данных применяются для построения и выполнения модели, которая обнаруживает и обобщает знания, представляющие интерес для пользователя (учителя, ученика, администратора и т.д.). Для достижения этой цели используется инструмент интеллектуального анализа данных.
4. Интерпретация, оценка и использование результатов. Полученные результаты или модель интерпретируются и используются учителем для дальнейших действий. Преподаватель может

использовать информацию для принятия решений о студентах с целью улучшения обучения студентов. Полученная модель интерпретируется и используется преподавателем для дальнейшего анализа.

В 2018 году ЮНЕСКО опубликовала отчет, в котором подробно рассказывается, как учебная аналитика обещает «превратить образовательные исследования в науку, основанную на данных, а образовательные учреждения - в организации, принимающие решения на основе фактических данных». За последние несколько лет многие из этих мер уже начали применяться.

Поведение в режиме онлайн обучения как многомерная сложная система, предусматривающая элементы, которые влияют на детальный анализ, для определения характеристик обучения, мотивации и стиля обучения, чтобы помочь учителям и менеджерам разработать разумную структуру обучения и стратегии обучения.

При анализе факторов влияния также учитываются не только внутренние факторы учащегося, но и внешние факторы, включая среду обучения, систему поддержки и режим обучения. Внутренние факторы включают в себя информационную грамотность учащегося, мотивацию обучения, первоначальные знания предмета учащегося, стиль обучения и самоэффективность учащегося. К внешним факторам относятся модели онлайн-обучения, ресурсы для онлайн-обучения, системы поддержки онлайн-обучения и навыки преподавания.

Исследование поведения в системе онлайн обучения включает в себя персональные требования, технические требования и работа с данными. Персональные требования включают в себя требования учащихся, организаторов курсов, менеджеров платформ и исследователей в области образования. Технические требования включают в себя анализ данных об образовании, анализ обучения и так далее. Работа с данными включает в себя сбор данных и хранение данных.

Анализ поведения, связанного с обучением в режиме онлайн в основном делится на явное и неявное. Среди них, большинство вещей, которые могут быть записаны платформой – это явное поведение при обучении, такое как просмотр, поиск, сохранение и т.д. Неявное – мотивация студента к обучению, старание, заучивание, тренировки, изучение дополнительной информации. Поведение в режиме онлайн обучения как сложная система, мы можем классифицировать ее по нескольким измерениям, а затем сформировать полную и всеобъемлющую систему классификации поведения студентов в системах онлайн обучения.

Например, имеется подход, в котором создается профиль студента с помощью технологии обработки больших данных. Во-первых, анализируются характеристики учащихся и факторы, влияющие на поведение. Затем рассчитывается сходство поведения студентов и используем алгоритм коэффициента Жакара для классификации студентов. Наконец, создается профиль студента, а также визуальный анализ.

Курс электронного обучения требует определенной цели, внутреннего мотива, синхронной обратной связи и независимости учащихся. Профиль студента помогает студенту понять его ситуацию, найти свои проблемы и улучшить уровень прохождения онлайн-курсов. С непрерывным накоплением данных об образовании и всесторонним развитием, профиль студента обязан способствовать здоровому развитию электронного обучения.

Совершенствование программных инструментов, предназначенных для анализа данных, позволило привлекать к исследованиям специалистов, не имеющий опыт в программировании, что явилось главным фактором, который поспособствовал развитию АОД.

По данным [1] значительная часть статей, которые были опубликованы в журналах Intelligent Tutoring Systems, и User Modeling and Adaptive Personalization, доклады на конференциях по АОД и аналитике обучения применяют такие свободно распространяемые пакеты, как: RapidMiner, R, Weka, KEEL и SNAPP. Данные пакеты содержат алгоритмы, которые

реализуют описанные выше методы анализа данных, так же обеспечивают импорт и поддержку предварительной обработки данных для применения в рамках этих методов. Обеспечивается поддержка проведения статистической проверки адекватности модели и визуализация данных.

Немногие из этих пакетов имеют открытый исходный код, но несмотря на это, исследователи разрабатывают модули, которые расширяют базовые возможности, добавляют все новые и новые функции, в которых нуждаются работники образовательных учреждений. Распространенный инструмент для статистического анализа данных R сейчас имеет около 10000 пакетов-расширений. Среди них имеются интерфейсы к другим развитым пакетам анализа данных. Так же, Weka и RapidMiner имеют расширения, которые позволяют использовать возможности R. Открытый программный код дает возможность глубоко изучить методы и алгоритмы, которые используются другими исследователями. Рассмотрим некоторые пакеты более подробно.

RapidMiner - инструмент, который был создан для интеллектуального анализа данных, преследуя цель, что аналитик не должен программировать. В RapidMiner включен большой набор операторов, с помощью которых решаются широкие спектры задач получения и обработки информации из различных источников.

R - программная среда вычислений с открытым исходным кодом, которая была создана в рамках проекта GNU. В начале R был реализован сотрудниками Оклендского университета Россом Айхэкой и Робертом Джентлменом. На данный момент язык и среда поддерживаются и развиваются организацией R Foundation[4].

Weka – свободно распространяемое программное обеспечение для анализа данных и машинного обучения, написанное на языке программирования Java в университете Уайкато. Weka - это уникальный набор средств для визуализации и алгоритмов интеллектуального анализа данных, а так же решения задач прогнозирования, со включенной графической пользовательской оболочкой для доступа к ним.

Orange - кросс-платформенный инструмент предназначенный для машинного обучения и интеллектуального анализа данных. Он объединяет в мощные рабочие процессы визуальное программирование, как интуитивное средство объединения анализа данных, и интерактивные методы визуализации.

Knime (Konstanz Information Miner) - Java-приложение с открытым исходным кодом, кросс-платформенное. Широко используется для интеллектуального анализа данных и их оптимизации. Его можно загрузить как основное приложение (Knime Desktop), так и весь SDK, основанный на Eclipse Helios.

Neural Designer – программный инструмент для расширенной аналитики. Он включает в себя инструменты для описательной, диагностической, предсказательной и предписывающей аналитики. Целью является получение доступа к действиям, что приводит к более умным решениям и лучшим результатам в бизнесе.

Однако, существующие программные разработки могут быть доработаны для того, чтобы отвечать требования для использования в АОД сотрудниками, не имеющими опыта в программировании.

Для того, чтобы лучше понять механизм анализа поведения студента в системе электронного образования, разберемся во всех его этапах:

1. В каком виде информация о студентах собирается и что в ней содержится.
2. Как выгружаются данные: какие методы и алгоритмы используются.
3. Как происходит анализ: какие методы и алгоритмы используются.

Исходя из изложенного выше, **задачей данной работы** будет являться – нахождение оптимальных методов и алгоритмов для анализа данных в системе электронного образования. Оптимальность методов и алгоритмов будем оценивать по схожести прогнозируемых и реальных результатов.

Глава 2 BIG DATA В ЭЛЕКТРОННОМ ОБУЧЕНИИ

2.1 BIG DATA в современных системах электронного обучения

Типичные исследования больших данных в образовании - это анализ данных в образовании и аналитика обучения студентов. Интеллектуальный анализ данных определяется как «новая дисциплина, связанная с разработкой методов исследования уникальных типов данных, которые поступают из учебных заведений, и использованием этих методов для лучшего понимания учащихся и условий, в которых они учатся». Аналитика обучающихся имеет разные определения из-за различных доступных инструментов аналитики с акцентом на использование возможностей для моделирования, где успех учащихся является одной из главных проблем.

Большие данные могут приносить пользу студентам, предоставляя им современную и динамичную систему образования. В исследовании [34] Атанасиос и Панагиотис проанализировали цели, задачи и преимущества больших данных и открытых данных в образовании. Авторы пришли к выводу, что система образования может быть улучшена путем использования новых подходов к обучению, чтобы сделать ее более эффективной и сфокусированной на ней. Более того, Annapoorna et al. [35], поддерживают ту же идею и ожидали, что большие данные могут быть эффективно использованы для прогнозирования результатов учащихся, а также улучшения как преподавания, так и обучения. Исследование, проведенное Туласи [36] и Беном Даниэлем [37], было направлено на высшее образование и изучало решения, предлагаемые системами больших данных, для решения проблем, стоящих перед высшим образованием. Крис Деде [38] продолжил продвижение темы, изучив «следующие шаги», которые могут быть предприняты с использованием больших данных в образовании, и пришел к выводу, что эта область имеет большой потенциал в улучшении индивидуального опыта обучения.

Многочисленные исследователи утверждают, что персонализация в академической среде позволяет использовать более эффективные и жизнеспособные формы обучения. Различные работы пытаются улучшить

качество и жизнеспособность электронного обучения, используя стандарты других исследовательских зон. Эта модель продвижения персонализации дополнительно проявляется в электронном обучении. Matteo et al., [39] представили новый инструмент: Интеллектуальный веб-учитель (IWT) для поддержки персонализированного электронного обучения в их исследовании персонализированного электронного обучения. Сравнение традиционных методов с IWT позволяет сделать вывод, что персонализация позволяет использовать более эффективные и мощные формы электронного обучения, демонстрируя растущий уровень выполнения как преподавателем, так и студентами. Система, разработанная Prawira et al. [40] и использовавшая Moodle оказалось способным улучшить учебный процесс и сотрудничество между преподавателем и студентом в высшем образовании.

Марьям Яранди и др. [41] использовали индивидуальные возможности обучения студентов, чтобы представить методику, основанную на онтологии, для улучшения адаптивной схемы электронного обучения. Предлагаемая система электронного обучения создает контент, позволяющий учащемуся узнать персонализированную систему электронного обучения, и утверждают, что система поощряет обучение благодаря реализации индивидуальных потребностей. Результаты согласуются с исследованием Chun-Hui Wu et al. [43], которое представляет теоретические основы адаптивного электронного обучения, самооценки и теории динамических лесов. Система предоставляет адаптированные учебные материалы для студентов на основе способностей студентов. Модель сеточного агента была предложена Жэнь и Юином в их исследовании [44] для эффективной адаптации систем электронного обучения, использующих искусственную психологию, к отдельным студентам, которым будет полезна эта персонализация. Кроме того, Синь Ли и Ши-Куо Чанг [45] предложили еще одну персонализированную систему электронного обучения, которая представляет собой экстрактор обратной связи с возможностью объединения пользовательских настроек.

Значение вышеупомянутой литературы заключается в том, что персонализированные схемы электронного обучения являются эффективными инструментами индивидуального обучения. В электронном обучении постоянно генерируется значительный объем данных, и они доступны в Интернете. Таким образом, требуются более сложные и пограничные методы кластеризации для сравнения данных EDM для получения внутренней информации. Чтобы справиться с вышеупомянутыми проблемами, рекомендуется кластеризовать подход к интеллектуальному анализу данных и интегрировать его с персонализированной системой электронного обучения. Интеграция подходов интеллектуального анализа данных делает систему обучения более интересной.

2.2 Форма представления данных о студентах

Все данные о студентах занимают значительный объем, для их обработки нужен конкретный подход. Под большими данными понимается широкое разнообразие массивов данных, которые не могут быть надлежащим образом обработаны традиционными приложениями из-за своего огромного объема или сложного состава.

Для начала опишем, в каком виде хранятся данные о студентах. Данные о студентах в системах электронного образования чаще всего хранятся в виде журналов сервера (log файлы), что позволяет сократить объем затрачиваемой памяти.

Студент взаимодействует и управляет своим профилем через интерфейс, развернутый на ноутбуке или смартфоне. Профиль пользователя и другая информация редко изменяются через Интернет. Согласно [53,54], профиль студента относится к типичной группе студентов. Функция профиля заключается в автоматическом определении потребностей и предпочтений студента. Данные, связанные со студентами, работают как источник для персонализации студенческих запросов и интеллектуального ответа на запросы.

Профилирование учащихся - это непрерывный процесс, который содержит как статические, так и динамические данные.

Данные, собранные статическим способом [54], включают личные, личностные, когнитивные, педагогические данные и данные о предпочтениях. Личные данные определяют биографическую информацию о студентах. Личные данные способствуют развитию у студентов внимания, навыков общения. Профиль студента отражает общий интерес и поведение студента. Когнитивные данные информируют о познании учащихся, а педагогические данные описывают различные стили и методы обучения. Если система поддержки профиля обнаруживает какое-либо необычное поведение в деятельности учащихся, она соответствующим образом обновляет профиль.

Типичный профиль студента может включать в себя сведения о его имени, возрасте, адресе, социально-экономическом статусе, школе, о результатах обучения, пройденном курсе, изученных модулях, результатах экзаменов, присвоенной степени.

Со стороны системы электронного образования профиль студента:

- обрабатывает информацию о студенте;
- собирает данные для получения подходящей информации об уровне знаний студента;
- обрабатывает отзывы от студентов, чтобы уточнить процесс рекомендации учебных материалов;
- обеспечивает динамическую оценку;
- рекомендует учебный материал для каждого учащегося на основе его знаний и профиля.

Со стороны студента электронного образования профиль студента:

- отвечает за все транзакции, которые могут происходить между сетью и компьютером;
- отвечает за наблюдение за поведением ученика и сохраняет его в модуле профилирования ученика.

Результаты анализа хранятся в виде набора шаблонов. Можно использовать описания наиболее распространенных последовательностей в данных для прогноза следующего наиболее вероятного шага в новой последовательности. Прогнозирующие запросы можно настраивать для того, чтобы они возвращали переменное число прогнозов или описательные статистические данные.

В базе хранятся все личные данные студента, включая имя, возраст, пол, адрес, почтовый индекс, а также данные, относящиеся к образованию, такие как квалификация. Кроме того, наличие такой информации, как опыт работы, карьерные цели, диапазон дохода, предыдущие курсы и интересующие курсы, были бы очень полезны для предсказания будущего поведения различных классов тредоустроенных людей. Кроме того, другая информация, такая как личные интересы и хобби, была бы очень полезна для инструмента интеллектуального анализа данных, чтобы обнаруживать скрытые шаблоны, создавая интеллектуальные модели, основанные на огромном количестве данных[4].

Рассмотрим пример на основе Moodle. Информация из файлов позволяет наставникам видеть, какие ресурсы используются и когда. Например, преподаватель может проверить, что отдельный студент прочитал лекцию и посмотреть, сколько времени он на это затратил.

Встроенное выпадающее меню используется, чтобы установить фильтры (Рисунок 4). Учитель может рассмотреть отчеты на уровне курса, уровне пользователя и/или уровне активности в определенный день или все дни.

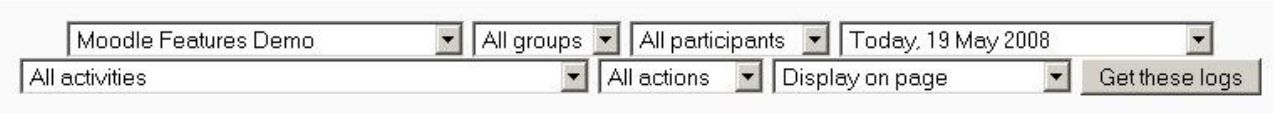

Рисунок 4 – Встроенное меню для составления отчетов

Фильтр можно настроить по нескольким параметрам и просмотреть результаты, а затем выгрузить информацию в виде текстового файла. Существующие фильтры:

1. Фильтр курса. По умолчанию используется текущий курс. Учитель должен иметь права на запрашиваемые курсы.
2. Фильтр студентов. По умолчанию все пользователи. Выпадающий список пользователей позволяет выбрать конкретного пользователя.
3. Дневной фильтр. По умолчанию все дни. Выпадающий список дней позволяет выбрать конкретную дату.
4. Фильтр активности. По умолчанию все действия. Выпадающий список может сузить фильтр до одного действия.
5. Фильтр действий. По умолчанию все действия. Можно ограничить действие просмотром, обновлением, удалением или всеми изменениями.
6. Фильтр отображения или загрузки. По умолчанию отображаются результаты фильтра. Журнал может быть загружен в виде текста, ODS или Excel.

Отчеты могут помочь определить, какие типы ресурса являются самыми популярными, а какие нет; это может помочь забронировать ресурсы на будущие курсы. Просмотр отдельных отчетов о проделанной работе может помочь контролировать студенческую деятельность (Рисунок 5).

Рисунок 5 – Пример отчета, показывающего активность курса в определенный день

Страница «Отчет» позволяет учителю или администратору просматривать журналы курса или сайта и отчеты об активности пользователей. Отчет можно найти в блоке администрирования. Отчеты сайта доступны пользователям, которым назначена административная роль.

Отображаемые журналы показывают активные ссылки на другие части курса. К ним относятся профиль пользователя, конкретная страница или ссылка на ресурс (Рисунок 6).

Рисунок 6 – Страница «Отчет»

Ранее мы обозначили, что все действия студентов регистрируются в журналах, но как обрабатывать данные, которые не имеют структуры? Первым шагом для аналитики электронного обучения является объединение всех данных, которые были собраны собрали и их «отчистка» от информации, не несущей ценности.

Хранилище данных является лучшим решением для хранения и обработки данных. Собирая всю информацию в удобной форме и одном месте, становится возможным анализировать и получать информацию о показателях студентов.

Для решения этой задачи имеются модели распределенной обработки данных, рассмотрим наиболее популярные из них.

2.3 Алгоритмы выгрузки больших данных

В большинстве случаев, при загрузке данных из источников возникает необходимость хранения этих данных в одной системе и их передачи в другую систему для дальнейшей обработки и анализа. С целью осуществления этого процесса создаются хранилища данных (ХД или DWH — Data Warehouse). Хранилище данных представляет собой базу данных для сбора и обработки информации, поступающей из различных источников. Спецификой хранилищ данных является ориентированность на подготовку отчётов с целью поддержки принятия решений в организациях [1]. Поскольку в данное время существует большое количество технологий и систем, различных по своей по структуре данных и архитектуре, данные при передаче из одной системы в другую, нуждаются в преобразовании. Основываясь на вышесказанном, можно выделить три последовательных этапа процесса работы с данными (Рисунок 7):

1. Извлечение (Extract)
2. Преобразование (Transform)
3. Загрузка (Load)

Рисунок 7 – ETL процесс

Эти этапы принято обозначать аббревиатурой ETL, которая описывает один из основных процессов управления данными при их извлечении из источников и последующей загрузки в хранилище данных с целью получения

достоверной информации. Остановимся на более подробном рассмотрении вышеуказанных этапов.

Извлечение данных. Этап заключается в извлечении данных из источников, при этом данные не изменяются независимо от их качества и загружаются в промежуточную область.

Преобразование данных. В процессе преобразования, данные подергаются группировке, а также преобразуются в нужный формат, согласно структуре хранилища данных. Также выполняется очистка данных, проверка на полноту, и формируются отчёты для дальнейшего исправления ошибок.

Загрузка данных. На данном этапе осуществляется загрузка трансформированных данных из промежуточной области в хранилище данных.

Стоит отметить, что загрузке подвергаются не все данные, а только те, которые являются новыми или были изменены. В процессе загрузки поддерживается версионность. Это является необходимым условием получения актуальной версии записи на произвольную дату. Довольно часто ETL становится промежуточным слоем между системами класса OLTP и хранилищем данных или OLAP-системой.

Online Transaction Processing (OLTP) — относительные небольшие транзакционные системы, обрабатывающие большие потоки данных в реальном времени.

Online analytical processing (OLAP) — системы динамического построения отчетов и документов, составления сложных запросов к базе данных для глубокого анализа.

Для осуществления ETL-процесса допустимо использовать почти любой современный язык программирования. Однако, если требуется не разовая конвертация, а постоянно выполнять интеграцию данных, то целесообразно рассмотреть специализированное ПО. При этом стоит учитывать скорость, расширяемость и масштабируемость выбранного инструмента. Среди лидеров на рынке ETL-инструментов выделяются Oracle, Informatica и IBM. Обычно системы, созданные указанными компаниями, перекрывают потребности

большинства компаний в области ETL. Исходя их этого, целесообразно выбирать ETL-инструмент основываясь на поставленных задачах, а также существующей платформе компании.

В случае преобладания продуктов IBM, стоит обратить внимание на решение Data Stage и Data Manager. Отличительной особенностью является наличие OLAP, что позволяет не строить сложные хранилища данных.

При преобладании Unix операционных систем, целесообразны решения PowerCenter и PowerMart от Informatica. Стоит отметить возможность разработки на языке Java [2,3].

При преобладании продуктов Microsoft, целесообразно применение SSIS от Microsoft. Указанный продукт располагает развитым пользовательским интерфейсом, однако не является кроссплатформенным [4].

Сейчас системы, которые относятся к ETL, являются не только решением проблем, возникающих при вводе данных, их переносе, возникающих при появлении системных ошибок или различиях между информационными системами, которые являются источниками и потребителями данных. Стоит отметить, что указанные задачи очень важны, поскольку в случае наличия большого количества различных ИС вероятно возникновение ошибок при передаче данных между системами, что приводит к снижению качества и достоверности данных. ETL-системы со временем стали включать в себя более широкий смысл, поскольку решаемые такими системами задачи получили высоки приоритет в компаниях. В конечном итоге бизнес пришёл к выводу о важности внедрения подобных информационных систем с целью повышения скорости принятия решений и ведения успешного конкурентного противостояния. Резюмируя вышесказанное, можно сделать вывод о том, что разработчики описанных систем должны стремиться к совершенствованию разрабатываемых систем и повышению распространённости таких систем на рынке.

2.4 База данных в системе электронного образования

После преобразования журналов сервера в базу данных, имеем дело со структурированными данными. Рассмотрим базу данных системы электронного образования (Рисунок 8).

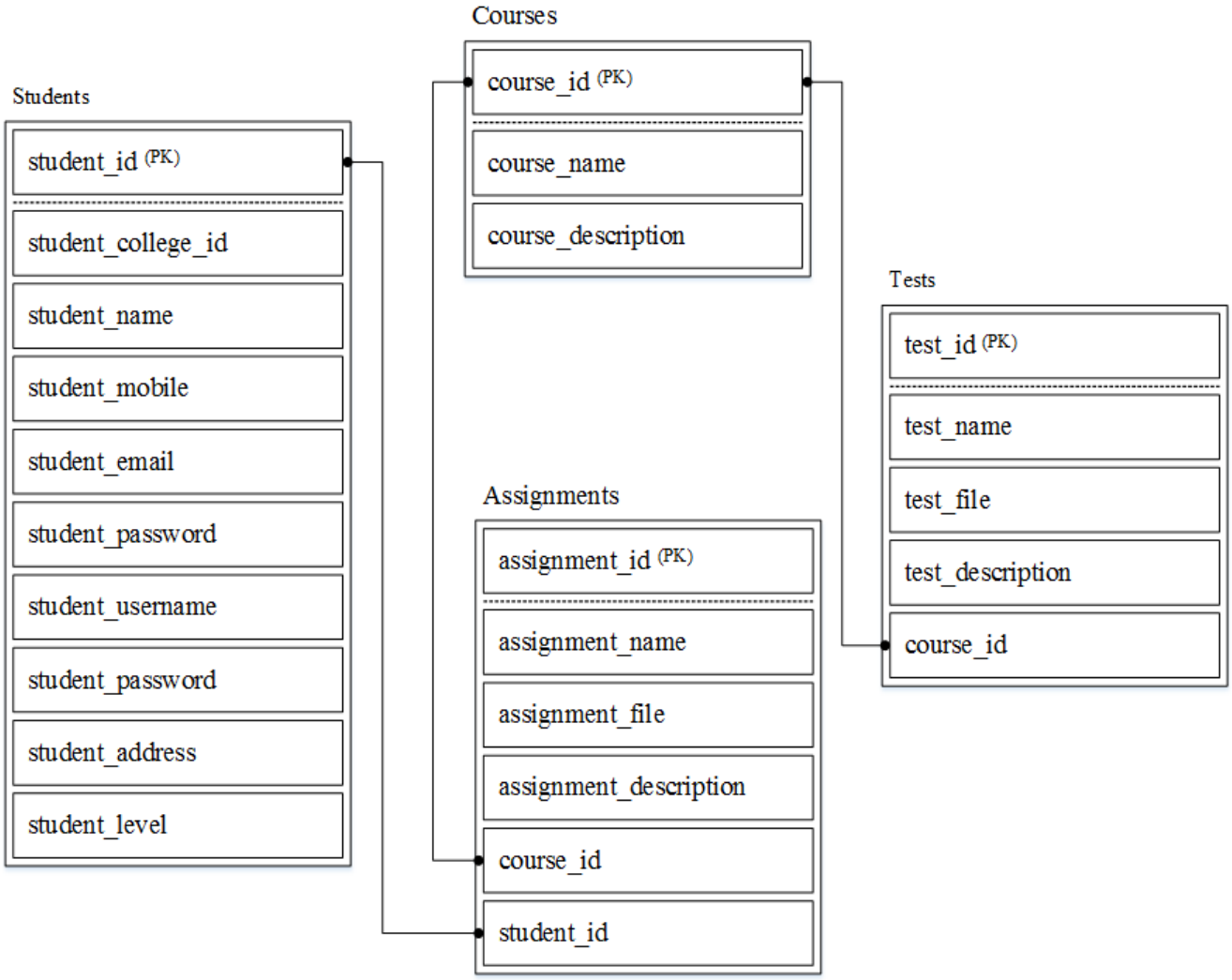

Рисунок 8 – Логическая модель данных системы электронного образования

Сущности платформы электронного обучения и их атрибуты:

1. Сущность студента имеет атрибуты: student_id, student_college_id, student_name, student_mobile, student_email, student_password, student_username, student_password, student_address, student_level.
2. Сущность курса имеет атрибуты: course_id, course_name, course_description.

3. Тесты для оценки знаний имеют атрибуты: test_id, test_name, test_file, test_description, course_id.
4. Отзыв о студенте имеет атрибуты: assignment_id, assignment_name, assignment_file, assignment_description, course_id, student_id.

Опишем базу данных платформы электронного обучения:

1. Детали курсов хранятся в таблицах курсов, соответствующих всем таблицам.
2. Данные о студентах хранятся в таблицах студентов, соответствующих всем таблицам.
3. Данные о специальностях хранятся в таблицах специальностей, соответствующих всем таблицам.
4. Каждый объект (курсы, студенты, задания, тесты, специальности) содержит первичный ключ.
5. Сущность тестов, задание имеет отношение к курсу, студентам по внешнему ключу.
6. Между курсами, студентами, заданиями, тестами и специальностями доступны отношения один к одному и один ко многим.
7. Все сущности курсов, студентов, заданий, тестов, специальностей нормализуются и не дублируются.
8. Все таблицы платформы электронного обучения имеют индексирование для быстрого выполнения запросов.

Рассмотрев структуру данных в системе электронного образования, перейдем к выбору и сравнению наиболее популярных методов интеллектуального анализа данных для оценки поведения студентов в системах электронного образования.

Глава 3 АЛГОРИТМЫ АНАЛИЗА BIG DATA

3.1 Интеллектуальный анализ данных для BIG DATA

Big data - это популярный термин, описывающий набор данных, объем которых растет с экспоненциальной скоростью, эта информация может быть как структурированной, так и полуструктурированной или неструктурированной. В образовании это дает шанс получить более точные рекомендации по курсам большого объема.

Новые методы анализа необходимы для обработки данных, так как объем данных растет беспрецедентно быстро, а также их сложность. Модели прогнозирующей аналитики или прогнозирования в среде больших данных позволяют учреждениям принимать правильные инвестиционные решения для более высокого институционального воздействия. Алгоритмы машинного обучения применяются для построения картографической функции для прогнозного анализа.

Big data все чаще применяется в онлайн образовании, в основном с целью извлечения образовательных данных и аналитики обучения. Данные по образованию в значительной степени расширились, что дало беспрецедентный шанс для исследований в соответствующих учреждениях. Многие образовательные учреждения имеют хранилища данных и используют аналитические инструменты для составления персонализированных планов обучения, а также для анализа потенциальных студентов.

Учебная аналитика больше зависит от семантических сетей, интеллектуальных учебных программ и системных вмешательств. Интеллектуальный анализ данных в большей степени основан на результатах образовательного программного обеспечения, моделирования учащихся и прогностических учебных программ.

Разработка больших данных может реализовать оптимизацию обучения. Преподаватели могут получить реальную информацию о каждом студенте. Содержание, метод и процесс обучения могут быть настроены в соответствии с полученными характеристиками об учащихся. Восприятия потребностей

учащихся в обучении, управления процессом обучения учащихся, диагностики результатов обучения учащихся и т. д.

Анализ больших данных поможет определить стили обучения и данные о поведении большого количества студентов посредством отслеживания и сбора больших данных о них. Так что учащиеся, вероятно, получат наиболее подходящие для их уровня материалы. Это не только повысит эффективность работы преподавателей и эффективность обучения студентов, но также сможет по-настоящему обучать студентов в соответствии с их способностями и развивать индивидуальные и инновационные таланты, которые отвечают потребностям информационной эпохи.

Исследователи обсуждают возможности применения больших данных в образовании в следующих аспектах [6]:

- расширение возможностей для учащихся – большой выбор программ и индивидуализация учебных предложений;
- своевременные и надлежащие корректировки для удовлетворения различных потребностей в обучении учащихся с разными стилями обучения и связанными с этим проблемными областями;
- методы, позволяющие сократить издержки в процессе, требующем много времени, за счет избыточности информации в системе глубоких знаний.

Интеллектуальный анализ данных в системах электронного обучения является критически важной задачей для извлечения, анализа и управления большими объемами данных. Прогнозирование академического поведения студентов позволяет определить, насколько хорошо отдельные лица и команды будут выполнять различные учебные задания. Обработка больших данных включает в себя важные ресурсы данных и подходящие аналитические инструменты для создания эффективных подходов, основанных на данных. Эти подходы могут обеспечить раннюю обратную связь, информирование преподавателей и менеджеров, что может помочь улучшить показатели студентов, результаты формальной оценки, также могут быть использованы для

выявления подверженных риску студентов, которые могут не сдать итоговые экзамены.

Процедуру прогнозирования можно упростить следующим образом:

1. Определение студентов, в том числе их уровень знаний, эрудиция, личностное соответствие, способность к обучению, начальные знания и другие аспекты.
2. Сбор и извлечение данных о студентах на первом шаге в наборе данных и выполнение предварительной обработки, уделяя особое внимание времени и усилиям, которые влияют на результаты обучения.
3. Интерпретация привычек обучения, чтобы усовершенствовать кривые прогнозирования.

Прогнозирование успеваемости может создать равные шансы для образования и однозначно повысить интерес учащихся.

Студентам обычно предлагают широкий выбор курсов и модулей. У них есть большой выбор как для формального обучения, так и для неформального обучения на платформе электронного обучения [7]. Сделать правильный выбор - может стать для них реальной дилеммой. Система рекомендаций курса, полученная из анализа больших данных, может помочь.

Руководствуясь глубокими исследованиями и анализом образовательных данных, преподаватели могут количественно оценить процесс обучения и статус обучения студента. Они могут обратить внимание на корреляцию и определить причинно-следственные связи. Можно выявлять актуальные проблемы в системе образования, быстрее и точнее находить факторы влияния и стратегии вмешательства, открывать действительно новые правила и значительно расширять возможности обучения.

Кроме того, большие данные в образовании могут также анализировать данные, сгенерированные учащимися в процессе обучения, прогнозировать их режимы обучения и способности к обучению, выявлять потенциальные проблемы в учебе и систематически улучшать модель обучения [6].

Разнообразные взаимодействия и обратная связь в зависимости от курсов имеют важное значение для повышения эффективности образования, что приводит к важности аналитики обучения. Начальные знания и привычки к учебе, индивидуальные планы обучения служат способом максимизации результатов обучения с учетом моделей мышления разных учащихся.

Большие данные позволяют исследователям в сфере образования судить о целесообразности, а также о преимуществах и недостатках с совершенно новой точки зрения. Это демонстрирует, что каждое состояние обучения можно заметить, используя методы, более совершенные, чем традиционные, и что каждый студент может сосредоточиться на своем собственном пути к новым знаниям.

Описание системы аналитики обучения может состоять из семи этапов:

- понимание потребностей в обучении;
- понимание образовательных данных;
- подготовка и предварительная обработка данных;
- выбор и планирование моделей;
- разработка функций и построение моделей;
- оценка, рефрейминг и оптимизация;
- мониторинг и анализ.

Кеннет изучал влияние учебного поведения и рефлексивного обучения на онлайн-бизнес-курсах. Приор проанализировал влияние онлайн-обучения по трем аспекта: отношение к обучению, информационная грамотность и самоэффективность, Бачер изучил различные уровни начальных знаний и отношения этих результатов показывают, что более высокий уровень начальных знаний может привести к тому, что учащиеся будут более успешны.

На средний балл или процентную долю успешных учащихся могут влиять различные факторы, такие как методика обучения и внимание учителей к некоторым конкретным учащимся. Общим является то, что учителя в основном сосредотачиваются на тех учениках, которые принимают участие в занятиях и показывают удовлетворительные результаты. Кроме того, среди студентов

существуют некоторые скрытые паттерны. Студенты могут быть разделены на разные группы в зависимости от их успеваемости. Один и тот же метод обучения может быть неэффективен для разных групп учащихся.

Меры сходства и кластеризации являются важными задачами для поиска аналогичных групп в образовательных больших данных. Подобные образцы данных в различных областях могут быть полезны для исследователей и учащихся, чтобы легко получить знания из различных областей.

Сложность анализа больших данных заключается в специфике их сбора, курирования, разделения, хранения, передачи, визуализации и сохранении конфиденциальности информации. Под анализом больших данных часто понимается применение прогнозной аналитики или других передовых методов с целью извлечения из множества данных определенной полезной информации. Точность при анализе больших данных помогает принимать более рациональные решения. В свою очередь, принятие наилучших решений позволяет увеличить производственную эффективность, сократить расходы и снизить риски.

Обычные реляционные базы данных подходят для достаточно быстрых и однотипных запросов, а на сложных и гибко построенных запросах нагрузка превышает разумные пределы и использование СУБД становится неэффективным.

Некоторые методы и техники анализа, применимые к Big Data:

- Data Mining: обучение ассоциативным правилам, классификации (методы категоризации новых данных на основе принципов, ранее примененных к уже наличествующим данным), кластерный анализ, регрессионный анализ;
- смешение и интеграция данных - набор техник, позволяющих интегрировать разнородные данные из разнообразных источников для возможности глубинного анализа;
- машинное обучение, включая обучение с учителем и без учителя, а также использование моделей, построенных на базе

статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей;

- искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы;
- пространственный анализ - использование топологической, геометрической и географической информации в данных;
- статистический анализ: A/B-тестирование и анализ временных рядов;
- визуализация аналитических данных - представление информации в виде рисунков, графиков, схем и диаграмм с использованием интерактивных возможностей и анимации как для результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

Для обработки данных о студентах был выбран интеллектуальный анализ данных, рассмотрим его подробнее.

3.2 Методики экспертной оценки поведения студента в системах электронного образования

Существуют различные методики оценки поведения студента в системах электронного образования. Никерина Е.А. [8] в своей работе рассматривает несколько из них, например, методики по концепции:

- Стреляу Я;
- Кузнецовой;
- Киркпатрика.

Методика оценки поведения студента по концепции Стреляу Я. предлагает на основе 10 разных критериев оценивать поведение студента. Каждый критерий может быть оценен от 1 до 5 баллов. В итоге студент может получить максимум пятьдесят баллов, а минимум – десять баллов. При этом уровень однотипного поведения тем ниже, чем ниже количество баллов

получит испытуемый. Это сделано для облегчения восприятия количественных результатов.

После оценки всех десяти критериев, полученные результаты суммируются и сохраняются. Затем полученный результат сопоставляют с усредненными результатами прошедших тестов. Чем ближе оценка к усредненному значению, тем с большей уверенностью можно определить, что у студента уровень поведения попадает в те же рамки, что и в прошлые разы. Согласно концепции Стреляу Я. процент итоговой оценки от усредненной оценки необходимо определить в одну из областей поведения:

1. 0-40, 161-200 единиц – можно с уверенностью сказать, что это не тот же самый студент.
2. 41-50, 151-160 единиц – скорее всего, это не тот же самый студент.
3. 51-70, 131-150 единиц – можно предположить, что это тот же самый студент.
4. 71-80, 121-130 единиц – скорее всего это тот же самый студент.
5. 81-120 единиц - можно с уверенностью сказать, что это тот же самый студент.

Методика экспертной оценки поведения студента по концепции Кузнецовой А.М. предполагает с помощью 6 различных свойств оценивать поведение учащегося. Каждое свойство оценивается от 1 до 4 баллов. В итоге студент может получить максимум двадцать четыре балла, а минимум – ноль баллов. При этом уровень однотипного поведения тем ниже, чем ниже количество баллов получит испытуемый. Это сделано для облегчения восприятия количественных результатов. Итак, число 24 свидетельствует о максимальном значении однотипного поведения, 0 – о минимальном значении [2].

При оценке следует исходить из конкретных, наблюдаемых форм и способов поведения. Один балл назначается в том случае, если свойство совсем не соответствует для данного студента. Максимальные четыре балла

назначаются в том случае, если свойство полностью соответствует для конкретного студента. Например, два балла – это средняя оценка, означает умеренную интенсивность данного свойства [8].

После оценки всех шести свойств, необходимо суммировать полученные результаты. Чем ближе оценка к максимальным двадцати четырем баллам, тем с большей уверенностью можно определить, что у студента уровень поведения попадает в те же рамки, что и в прошлые разы. Согласно концепции Кузнецовой А.М. процент итоговой оценки от усредненной оценки необходимо определить в одну из областей поведения:

1. 0-50, 151-200 единиц – можно с уверенностью сказать, что это не тот же самый студент.
2. 51-65, 136-150 единиц – скорее всего, это не тот же самый студент.
3. 66-80, 121-135 единиц – скорее всего это тот же самый студент.
4. 81-120 единиц – можно с уверенностью сказать, что это тот же самый студент.

Методика экспертной оценки поведения студента по концепции Киркпатрика предполагает с помощью теста, состоящее из 5 вопросов определить, как изменилось поведение участников в результате обучения, насколько полученные знания и навыки применяются на рабочем месте[10].

Каждый вопрос состоит из определённого количества предлагаемых ответов, каждый из ответов оценивается в процентах. В итоге студент может получить максимум сто процентов, а минимум – 0 процентов. При этом уровень однотипного поведения тем ниже, чем ниже количество процентов получит испытуемый. Это сделано для облегчения восприятия количественных результатов. Итак, число 100 свидетельствует о максимальном значении однотипного поведения, 0 – о минимальном значении [12].

Необходимо ответить на каждый из пяти вопросов. Киркпатрик предлагает воспользоваться инструментом – обзор (отслеживание) поведения для оценки однотипного поведения студента. В тоже время Киркпатрик

указывает на то, что отсутствие изменений в поведении участников не означает, что курс был завершен ровно также, а также сильное отклонение поведения не означает, что курс прошел совсем другой человек. Возможны ситуации, когда реакция на курс была позитивной или негативной, но не были созданы необходимые условия и их поведение дальнейшем не изменилось. Киркпатрик отмечает, что в этих случаях необходимо проверить наличие следующих условий:

- желание участников изменить поведение;
- понимание участниками, что и как делать;
- создание соответствующего социально-психологического климата;
- поощрение участников за изменение поведения.

Говоря о социально-психологическом климате, Киркпатрик имеет в виду прежде всего непосредственных руководителей участников обучения и рекомендует с целью создания позитивного климата вовлекать руководителей в разработку учебных программ.

После ответа на все пять вопросов, необходимо суммировать полученные проценты, а затем поделить на количество вопросов, а данной методике – пять вопросов. Чем ближе оценка к максимальным ста процентам, тем с большей уверенностью можно определить, что у студента уровень поведения попадает в те же рамки, что и в прошлые разы при прохождении им курсов. Согласно концепции Киркпатрика процент итоговой оценки от усредненной оценки необходимо определить в одну из областей поведения:

1. 0-25, 176-200 единиц – можно с уверенностью сказать, что это не тот же самый студент.
2. 26-50, 151-175 единиц – скорее всего, это не тот же самый студент.
3. 51-75, 126-150 единиц – скорее всего это тот же самый студент.
4. 76-125 единиц – можно с уверенностью сказать, что это тот же самый студент.

Более подробно данные методики и их реализации рассматриваются в работе Никерина Е.А. [8].

3.3 Методы интеллектуального анализа данных в системе электронного обучения

В АОД в основном используются методы интеллектуального анализа данных. Часто используемыми методами являются классификация, кластеризация, поиск связующих правил, поиск последовательных шаблонов, и интеллектуальный анализ текстов. Эти методы делятся на три основные группы:

- прогнозирование;
- обнаружение структуры;
- выявление взаимосвязей.

Рассмотрим их более подробно.

3.2.1 Прогнозирование

Прогнозирование используется для разработки модели, которая будет предсказывать значение искомой величины по значениям, которыми обладают независимые переменные. Если выходная переменная принимает непрерывное значение, то такая модель называется регрессией - поиском зависимости между входными и выходной данными.

Регрессия - это зависимость матожидания случайной величины от одной или нескольких других случайных величин, то есть $E\ y\ x\ = f(x)$. Регрессионный анализ - это поиск функции f, описывающей эту зависимость. Регрессию обычно представляют, как сумму – неслучайной и случайной величины.

$$y = f\ \ x\ \ + v, \text{ где} \qquad (1)$$

- $f\ x$ – регрессионная функция зависимости;
- v — аддитивная случайная величина, которая имеет нулевое математическое ожидание.

Гипотеза порождения данных - это предположение о характере распределения функции регрессивной зависимость. Считают, что величина v имеет гауссово распределение с нулевым средним и дисперсией δ_v^2.

Постановка задачи для нахождения регрессионной модели нескольких свободных переменных. Задана выборка — множество $x_1, x_2, \dots, x_N\ x \in R^M$ значений свободных переменных и множество $y_1, y_2, \dots, y_N\ y \in R$ соответствующих им значений зависимой переменной. Множества обозначаются символом D, множество исходных данных - (x, y) . Задана регрессионная модель — параметрическое семейство функций $f(w, x)$, которая зависит от параметров $w \in R$ и свободных переменных x. Требуется найти наиболее вероятные параметры w:

$$w = argmax\ p\ y\ x, w, f\ = p(D|w, f) \quad (2)$$

Функция вероятности p зависит от гипотезы порождения данных, она задается Байесовским выводом.

О задаче классификации говорят, если выходная переменная имеет конечный дискретный набор значений, опираясь на значения входных переменных, выходная будет определена к одному или другому классу.

Используя прогнозирование можно предположить, что часть данных размечена, для них заданы значения входных переменных и подходящие выходные переменные. "Научившись" на таком наборе данных, с помощью алгоритмов регрессии или классификации, можно предугадывать значение выходной переменной для других данных. Например, опираясь на данные по обращению к учебным материалам и итоговой оценке, которую студент получил на экзамене у реального преподавателя, можно спрогнозировать полученную оценку студенту со сходной учебной активностью.

3.2.2 Обнаружение структуры

Алгоритмы обнаружения структуры применяется для обнаружения структуры в данных, без использования априорных представлений о ней. Самая известная группа подобных алгоритмов - алгоритмы кластеризации. Кластеризация - это логическое продолжение идеи классификации. Уникальность кластеризации состоит в том, что классы объектов в начальных условиях не заданы, тогда как при классификации, прежде чем перейти к созданию модели, требуется разметить некоторое подмножество данных. Результатом кластеризации является разбиение множества объектов на группы близких в каком-либо смысле объектов.

Постановка задачи кластеризации выглядит следующим образом: пусть X - множество объектов, Y - множество меток кластеров. Функция расстояния между объектами определена, как $p(x, x')$. Конечная обучающая выборка представляет собой $X^m = \{x_1, x_2, \ldots, x_m\} \subset X$. Требуется разделить выборку на непересекающиеся подмножества, кластеры, таким образом, чтобы каждый их них состоял из объектов, близких по заданной метрике p, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

3.2.3 Выявление взаимосвязей

Целью выявления взаимоотношений является установление взаимосвязи между переменными в большом объеме исходных данных со значительным числом переменных. Например, можно попробовать определить, какие переменные связаны сильнее с интересующей или в какой из пар связь сильнее, чем в других. Выявление взаимоотношений в АОД часто используется в виде поиска связующих правил или поиска последовательных шаблонов. В поиске связующих правил цель – это нахождение правила вида "if-then", которое показывает, что если (if) одни переменные принимают какое-то множество значений, то (then) другая переменная будет иметь определенное значение. Например, если в покупке имеется набор товаров А, то можно сделать вывод,

что в этой же покупке должен иметься товар B. Поиск последовательных шаблонов является развитием поиска связующих правил и предназначен для выявления взаимоотношений между проявлениями последовательных во времени событий.

Данные методы применяются для оценки поведения студентов в системе электронного обучения, с их помощью можно понять, честно ли студент проходит итоговые тесты, спрогнозировать, какое количество студентов окончит обучение и с какими оценками. Рассмотрим существующие программные разработки на эту тему.

3.4 Алгоритмы интеллектуального анализа данных в системе электронного обучения

Интеллектуальный анализ данных, также известный как обнаружение данных или обнаружение знаний, представляет собой процесс анализа данных с разных точек зрения и обобщения их в полезную информацию. Эта информация используется предприятиями для увеличения доходов и снижения операционных расходов. Программное обеспечение, используемое в интеллектуальном анализе данных, относится к числу инструментов, используемых при анализе данных.

При анализе данных, как правило, нет возможности рассмотреть всю интересующую нас совокупность объектов. Изучение очень больших объемов данных является дорогостоящим процессом, требующим больших временных затрат, а также неизбежно приводит к ошибкам, связанным с человеческим фактором. Вполне достаточно рассмотреть некоторую часть всей совокупности, то есть выборку, и получить интересующую нас информацию на ее основании. Однако размер выборки должен зависеть от разнообразия объектов, представленных в генеральной совокупности. В выборке должны быть представлены различные комбинации и элементы генеральной совокупности, т.е. всей совокупности изучаемых объектов. Часть генеральной совокупности, отобранная на основе свойств и характеристик – это выборка (sample).

Интеллектуальный анализ данных отвечает за поиск ассоциаций, рекомендаций и интеллекта, чтобы обеспечить индивидуальный и мощный механизм обучения для студентов. Например, выбор подходящего контента на основе интересов студентов является большой проблемой. Эту проблему можно решить, сгруппировав все содержимое, просто применив подход кластеризации для фильтрации содержимого в соответствии с индивидуальным профилем студента. Кроме того, ключевые компоненты вывода в таких системах электронного обучения основаны на методах интеллектуального анализа данных, которые анализируют профиль пользователя и предлагают какие-то действия с применением искусственного интеллекта.

В EDM использовались различные методы кластеризации, такие как сдвиг среднего значения, K-средние, медоид (K-medoids), основанная на плотности пространственная кластеризация для приложений с шумами (англ. Density-based spatial clustering of applications with noise, DBSCAN) в [21,48] и иерархическая кластеризация в [49–52]. однако эти подходы также не являются надежными для идентификации значительных кластеров в неоднозначных и зашумленных наборах данных [22]. Рассмотрим данные методы подробнее.

3.3.1 Сдвиг среднего значения

Сдвиг среднего значения — это непараметрическая техника анализа пространства признаков для определения местоположения максимума плотности вероятности, так называемый алгоритм поиска моды. Это подход на основе скользящих окон; пытается обнаружить уплотненные области.

Целью этого подхода является обнаружение центра каждого кластера на основе метода центроида. В методе сдвига среднего значения кандидаты обновляются для центральных точек (среднего значения точек) в скользящем окне. На этапе постобработки окна-кандидаты фильтруются для удаления дубликатов и образуют окончательный набор центральных точек и их совпадающих кластеров.

Сдвиг среднего значения, имеющая радиус «r» (как ядро), начинается с круглого скользящего окна, центрированного в случайно выбранной точке C. Метод сдвига среднего значения сдвигает «r» в область более высокой плотности (на каждом шаге) до сходимости. Плотность каждого раздвижного окна пропорциональна его размеру (точкам внутри него). При смещении плотность точек постепенно перемещается к областям с более высокой плотностью точек. Сдвиг скользящего окна продолжается до тех пор, пока сдвиг не сможет вместить больше точек внутри ядра (больше не увеличивая плотность). В случае, когда несколько раздвижных окон перекрываются, точки данных группируются в соответствии со скользящим окном, в котором они находятся, и скользящим окном, содержащим наиболее хорошо поддерживаемые точки.

В системах электронного обучения, работая с big data, вычисление сложных функций занимает большое количество времени. Так же, имея невысокую плотность данных, результат работы метода не всегда точен.

Алгоритм реализован в следующих программных средствах:

- ELKI – Java-приложение для интеллектуального анализа данных со многими алгоритмами кластеризации;
- ImageJ – создан для того, чтобы фильтровать изображения на основе фильтра сдвига среднего значения;
- OpenCV – включает в себя реализацию сдвига среднего значения про помощи метода cvMeanShift;
- Orfeo – представляет собой инструментальный набор, разработанный на языке програмирования C++;
- Scikit-learn – разработан на языке программирования Python и применяет шаровое дерево для результативного просмотра соседних точек.

Метод сдвиг среднего значения имеет четыре недостатка:

- сложный подбор выбора размера окна;

- неправильный размер окна может привести к слиянию мод или образованию дополнительных «теневых» мод;
- сложные вычисления;
- он достоверен на высокой плотности данных (с идеальным градиентом, чтобы найти центр кластера).

3.3.2 K-means

K-means [20] - это современный алгоритм кластеризации на основе разделов. В K- means входные данные делятся на k различных групп, где k - входной параметр, используемый для указания количества выходных кластеров. K-means итеративно улучшает начальные разделы, пока оптимальные кластеры не будут найдены. Математически мы можем выразить K- means используя следующее выражение:

$$argmin_S \sum_{i=1}^{n} \sum_{x \in S_i} \| x - \mu_i \|^2 \text{, где} \quad (3)$$

μ_i - среднее значение точек данных в S. S - начальное разбиение набора данных $\{x_1, x_2, x_3, \dots, x_n\}$.

K-means - лучший выбор для обнаружения интересующего сигнала из образовательных наборов данных, если значительное число кластеров уже определено. Тем не менее, поиск соответствующих групп с использованием K-means может быть невозможен без предварительного анализа и вычисления существующего числа кластеров или при наличии шумов, или сложных данных.

В образовательных данных, выбор количества кластеров и начальную установку центроидов K-means сложно угадать.

Недостатки алгоритма:

- не гарантируется достижение глобальных минимумов (максимумов), а только локальных;
- результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен;

- число кластеров нужно знать заранее.

Таким образом, требуются более сложные и пограничные методы кластеризации для сравнения образовательных данных и получения внутренней информации.

3.3.3 K-medoids

K-medoids [53] используется для разделения набора данных на кластеры, подобно k-means. Целью обоих методов (K-средних и K-медоидов) является минимизация суммы расстояний между точками данных кластера и центральной точкой данных того же кластера. В отличие от k-средних, k-medoids выбирает точки данных посредством центров (медоидов) и обработок по общему правилу Манхэттена, чтобы выразить расстояние между точками данных. Он группирует набор данных из «n» точек в «k» группы или кластеры. K-medoids уменьшает количество попарных вариаций вместо суммирования квадратов евклидовых расстояний. Вот почему он считается более устойчивым (к шуму и выбросам), чем k-средних. Тем не менее, данные метод также нуждается в предварительной обработке данных и нахождении числа кластеров.

Метод k-medoids эффективен для обратного преобразования данных. Также подходит для кластеризации определенных данных, где среднего не существует.

3.3.4 DBSCAN

DBSCAN [21] - это основанная на плотности пространственная кластеризация для приложений с шумами, которая начинается через случайную начальную точку. Соседние точки данных находятся на расстоянии - ε. Если в окрестности существует достаточное количество точек данных, начинается процедура кластеризации, и текущая точка данных считается первой. Другие данные будут помечены как шум. Позже эти зашумленные точки данных могут стать частью кластера. В обеих ситуациях точки данных отмечаются как

посещенные. Для текущего кластера процедура создания всех точек данных в ε -окрестности повторяется, чтобы добавить новые точки. Этот процесс повторяется до тех пор, пока все точки данных в текущем кластере не будут распознаны и помечены как «посещенная». Тот же процесс повторяется для всех кластеров.

Алгоритм DBSCAN реализован и применяется в следующих программных средствах:

- Apache Commons Math включает в себя реализацию на Java алгоритма, который работает за квадратичное время;
- ELKI - предоставляет реализацию DBSCAN, GDBSCAN и других видов данного алгоритма. В этой реализации используются произвольные функции расстояния и произвольные типы данных, добиться ускорения можно оптимизацией на низком уровне, а так же применяя специальные методы на малых наборах данных;
- PostGIS содержит в себе ST_ClusterDBSCAN — двумерную реализацию DBSCAN, которая использует R-дерево в качестве индекса, реализованы такие геометрические типы, как «точка», «отрезок», «многоугольник» и т. д.;
- язык R содержит пакет fpc с DBSCAN, который поддерживает произвольные функции расстояния через матрицы расстояний;
- SPMF содержит реализацию алгоритма DBSCAN с поддержкой k-d дерева только для евклидового расстояния;
- Weka, в виде дополнительного пакета, содержит базовую реализацию DBSCAN, которая требует линейную память и работает за квадратичное время.

DBSCAN имеет следующие недостатки:

- DBSCAN неоднозначен — краевые точки, которые могут быть достигнуты из одного и более кластеров, могут относиться к любому из этих кластеров, что зависит от последовательности просмотра точек;

- DBSCAN плохо кластеризует наборы данных с большой разницей в плотности;
- DBSCAN не оповещает, когда кластеры переменной плотности установлены;
- при изменении плотности пороговое расстояние ε и точки для идентификации соседних точек данных будут отличаться от кластера к кластеру, для очень высокой размерности, пороговое расстояние ε становится сложной для оценки.

3.3.5 Иерархическая кластеризация

В этом пункте мы обсудим только восходящую иерархическую кластеризацию. Вначале он обрабатывает каждую точку данных как один кластер, а затем непрерывно объединяет пары кластеров, пока целые кластеры не будут объединены в отдельный кластер. Он также известен как агломеративное иерархическое кластерное дерево (AHCT) [54]. AHCT представляется в виде дерева. Корни дерева рассматриваются как уникальный кластер.

В начале каждая точка данных обрабатывается как один кластер. Это означает, что точки данных "k" обрабатываются как кластеры "k". На следующем шаге выбирается метрика расстояния для его измерения между двумя кластерами. Кроме того, два кластера объединяются в один итеративно для каждой пары. Объединенные кластеры выбираются с наименьшей средней связью; оба кластера имеют наименьшее расстояние и наиболее похожие точки данных. Этот шаг будет продолжаться до тех пор, пока корень дерева не будет явно задан в начале. Количество кластеров может быть выбрано путем распознавания заданного корневого номера, что помогает прекратить объединение кластеров.

Восходящий иерархический метод кластеризации не требует указания количества кластеров и имеет возможность выбрать лучший кластер из-за использования дерева.

3.3.6 CFSFDP и CFSFDP-HD

В 2014 году CFSFDP был предложен Алексом и Лайо [23]. Он обладает характеристиками для обнаружения значительных кластеров более интуитивным способом по сравнению с K-means. Предлагается новый эвристический подход, который расширяет процедуру кластеризации, в котором области высокой плотности идентифицируются как потенциальные кластеры, автоматически определяются исключения и организуются произвольные формы кластеров.

В K- means для получения значимых кластеров пользователи должны повторить процесс кластеризации несколько раз с различными параметрическими настройками. Тем не менее, уникальный подход, используемый в CFSFDP для адаптивного обнаружения кластеров и шума, станет важным инструментом кластеризации для анализа в сфере электронного образования. CFSFDP использует следующую методологию для обнаружения значительных кластеров.

Данные о каждом студенте представляют собой данные - data-point i, координатой которой является пара r, t, где r – рейтинг студента, вычисление которого происходит на основе отзывов преподавателей и среднего баллов студента по предмету за текущий семестр, а t – среднее количество времени, проведенное на платформе.

CFSFDP рассчитывает локальную плотность p_i и минимальное расстояние δ_i между каждым студентом i и точкой высокой плотности (предполагаемым центром кластера). Где p_i равно числу студентов, которые находятся на расстоянии меньшем, чем d_c, по отношению к точке i. d_c в свою очередь, это важный параметр, используемый для оценки p_i каждой точки i. Эффективность метода CFSFDP зависит от правильного подбора параметра d_c.

Локальная плотность может быть оценена с использованием следующего определения, где d_{ij} – расстояние от data-point i до data-point j (т.е. между студентами под номерами i, j):

$$p_i = \quad _i X(d_{ij} - d_c)^2 \text{ , где} \tag{4}$$

$$\text{X x } = \begin{matrix} 1, & \text{при } x < 0 \\ 0, & \text{при } x \geq 0 \end{matrix} \tag{5}$$

Расстояние δ_i может быть вычислено с использованием следующего определения:

$$\delta_i = \begin{matrix} min_{j:p_j>p_i} \ d_{ij} \ , & \text{если } \exists\, j, \text{такое что } p_j > p_i \\ max_{j:p_j>p_i} \ d_{ij} \ , & \text{если } \nexists\, j, \text{такое что } p_j > p_i \end{matrix} \tag{6}$$

Центры кластеров достигаются путем построения рассчитанных значений p_i и δ_i, что называется графом решений. В кластерном анализе ключевая задача - найти правильные кластерные центры в наборах данных [1]. Тем не менее, CFSFDP использует граф решений для выбора правильных центров кластеров с наименьшим количеством взаимодействия с человеком, что делает его более достойным для анализа больших данных или потоковых данных. CFSFDP имеет множество применений в образовании, а также во многих других областях, таких как биоинформатика [58], обработка изображений и анализ белка [23].

Поскольку CFSFDP обладает характеристиками, позволяющими обнаруживать скрытые данные из неоднозначных данных, он может применяться в существующих системах интеллектуального анализа образовательных данных и системах электронного обучения для создания более значимых кластеров, а в дальнейшем его можно использовать для кластеризации аналогичных документов, поиска плагиата в документы, а также анализировать профили студентов и находить аналогичные идеи в разных областях исследований. CFSFDP посредством диффузии (CFSFDP-HD) [21] был предложен в качестве варианта CFSFDP, где ограничения CFSFDP улучшены, и пользователи могут анализировать данные без какого-либо предварительного знания предметной области. В CFSFDP-HD для оценки

плотности базового набора данных использовался адаптивный метод, который представлен следующим образом:

$$f\ x;t\ = \frac{1}{n}\ \ \sum_{j=1}^{n}\ \ \sum_{k=-\infty}^{\infty} e^{-k^2\pi^2 t/2} \cos\ k\pi x\ \ \cos(k\pi x_j), \text{ где} \quad (7)$$

x представляет набор точек данных i, а начальная вероятность распределена на всем протяжении множества $\{x_1, x_2, x_3, \dots, x_n\}$. Преобразуем уравнение (7):

$$f\ x;t\ \approx\ \ \sum_{j=1}^{n} a_k e^{-k^2\pi^2 t/2} \cos\ k\pi x\ , \text{ где} \quad (8)$$

n - положительное большое целое число, а a_k:

$$a_k = \begin{cases} 1, & k = 0 \\ \dfrac{1}{n} \displaystyle\sum_{j=1}^{n} \cos(k\pi x_j), & k = 1,2,\dots,n-1 \end{cases} \quad (9)$$

Это процесс отвечает за сбор данных из баз данных, отфильтрованных по профилю студента с помощью методов интеллектуального анализа данных. Он также имеет возможность предотвратить дублирование информации, созданной ранее.

Этот алгоритм рекомендует оценку для преподавателя и предупреждает студента о возможных результатах следующим образом:

- в выборку для анализа попадают все студенты из базы данных, проходившие исследуемый курс:
  - студенты, завершившие курс успешно;
  - студенты завершившие курс неуспешно;
  - студенты, проходящие курс;
- алгоритм выполняет свою работу, распределяя студентов на кластеры на основе средних баллов за промежуточные задания курса, баллов, полученных в результате экспертной оценки,

описанных выше, и среднего времени $T_{\text{ср}}$, которое рассчитывается по формуле (10):

$$T_{\text{ср}} = \frac{T_{\text{к}}}{N_{\text{д}}}\text{, где} \quad (10)$$

- $T_{\text{к}}$ – общее время, проведенное студентом на материалах исследуемого курса;
- $N_{\text{д}}$ – количество дней с момента допуска к исследуемому курсу до даты проведения исследования;

- студенту, проходящему курс, и преподавателю предлагается оценка на основе оценок студентов, завершивших курс, в виде:
  - N% студентов, с похожими показателями среднего времени нахождения на материалах курса и средним баллом по промежуточным заданиям, по окончании курса получили M баллов.

В отличие от существующих систем электронного обучения, предлагается использовать методы кластеризации с помощью быстрого поиска и нахождения пиков плотности (CFSFDP) и CFSFDP с помощью диффузии тепла (CFSFDP-HD) для достижения надежных результатов.

Глава 4 КОМПЬЮТЕРНАЯ МОДЕЛЬ АЛГОРИТМА CFSFDP-HD ДЛЯ ОЦЕНКИ ПОВЕДЕНИЯ СТУДЕНТА В СИСТЕМЕ ЭЛЕКТРОННОГО ОБУЧЕНИЯ

4.1 Компьютерная модель алгоритма CFSFDP-HD

Основные этапы алгоритма CFSFDP-HD показаны на Рисунок 9. Представленный подход использует матрицу расстояний D набора данных в качестве входных данных.

Шаг 1: на первом этапе, предлагаемый подход оценивает плотность p_i через алгоритм CFSFDP-HD, используя формулу (6).

Шаг 2: предлагаемый подход вычисляет минимальное расстояние δ_i от более высоких ближайших точек, используя формулу (8).

Шаг 3: идентификация кластерных центров достигается с помощью графа решений. На графе решений изображены p_i и δ_i. Результатом этого шага являются кластерные центры.

Шаг 4: присвоение оставшихся точек идентифицированным кластерным центрам. Результатом этого шага являются организованные кластеры с шумами и перекрывающимися кластерами.

Шаг 5: на этом этапе представленный подход идентифицирует и исправляет ошибочно классифицированные точки, а также идентифицирует шум или выбросы организованных кластеров (шумы и перекрывающиеся кластеры). Результатом предлагаемого подхода являются организованные кластеры.

В результате работы алгоритма, пользователю представляются данные о студентах, распределенные по кластерам. Преимуществом будет являться то, что даже самые отдаленные точки не будут считаться шумами, а примкнут к созданным кластерам.

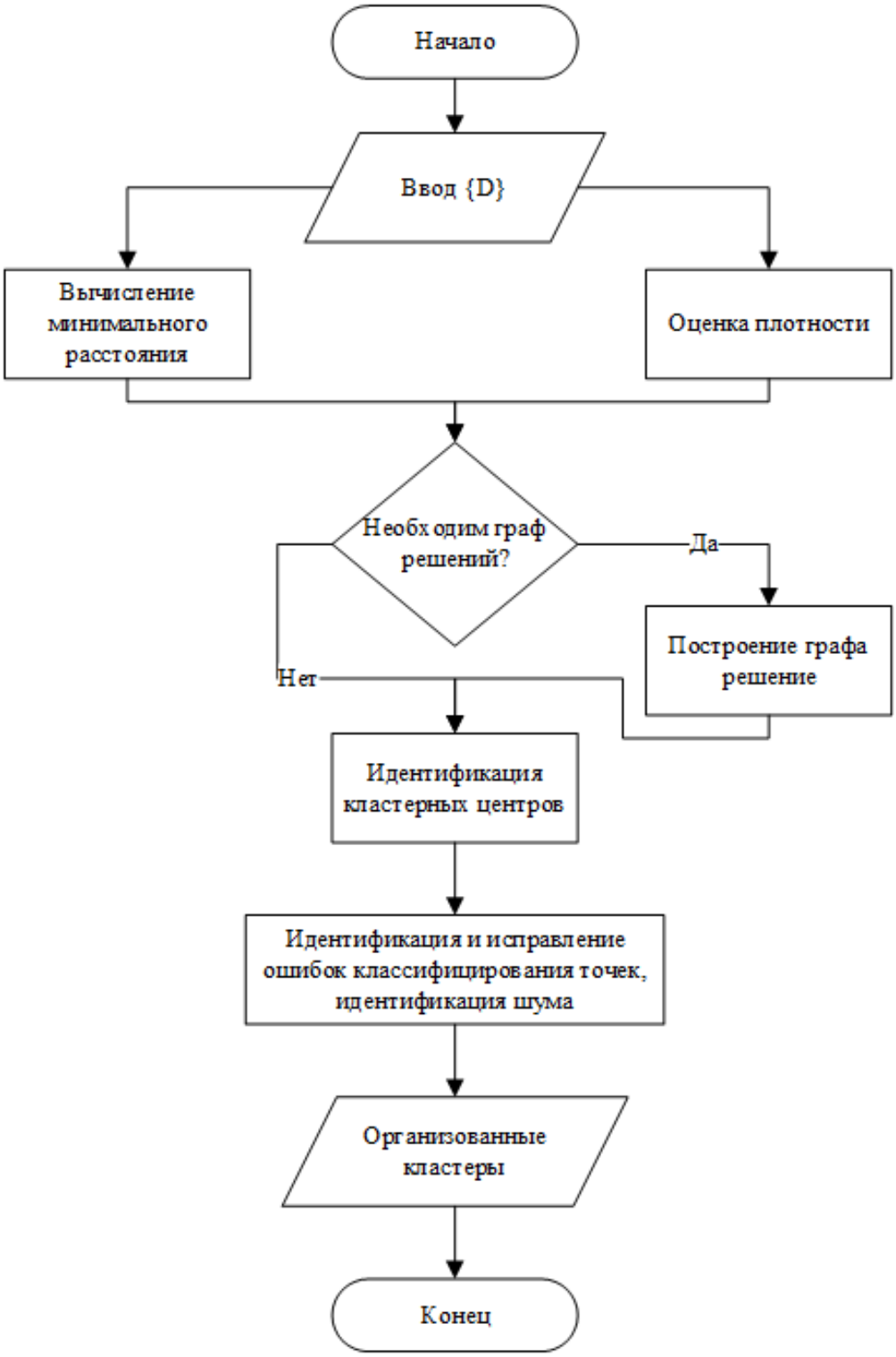

Рисунок 9 – Блок-схема алгоритма CFSFDP-HD

Рассмотрев блок-схему алгоритма CFSFDP-HD, перейдем к проведению анализа данных о студентах.

4.2 Анализ данных о студентах в системе электронного обучения с помощью алгоритмов K-means и CFSFDP-HD

4.2.1 Алгоритм CFSFDP-HD

В этом эксперименте набор данных состоит из 600 студентов разных годов обучения. Эксперимент моделируется с помощью подхода CFSFDP-HD для разделения учащихся на соответствующие группы и основан на полученных оценках учащихся и посещаемости занятий. Сегментация учащихся на основе прогресса необходима для разработки соответствующих методов обучения, направленных на устранение недостатков определенной группы. На Рисунок 10(a) эвристический подход на основе графа решений визуализируется для интуитивного выбора точного количества кластеров. Полные черные точки на Рисунок 10(a) рассматриваются как некластерные центральные точки, в то время как цветные точки рассматриваются как центральные точки ожидаемого кластера.

График принятия решений устанавливается после (1) оценки плотности p_i посредством термодиффузии и (2) расчета минимального расстояния δ_i от более высоких ближайших плотных точек. Идентификация кластерных центров достигается с помощью графа решений, показанного на Рисунок 10(a), где изображены p_i и δ_i.

При минимальной интерпретации эвристического подхода к выбору точного числа кластеров четыре отдельных кластера эффективно идентифицируются, как показано на Рисунок 10(a), где шумы рассматриваются как потенциальные центры кластеров и представлены разными цветами. После идентификации потенциальных кластерных центров оставшиеся точки данных присваиваются идентифицированным кластерным центрам. Обнаруженные кластеры показаны с помощью другой цветовой схемы на Рисунок 10 (b), где двумерное неклассическое многомерное масштабирование используется для визуализации набора данных. Рекомендуемый подход «CFSFDP-HD» носит адаптивный характер, поэтому нет необходимости явно устанавливать какой-либо параметр.

Вышеупомянутое разделение студентов на четыре значимые группы может сыграть важную роль в развитии навыков обучения, уделяя особое внимание определенной группе студентов. Самомотивированные и талантливые ученики отделяются от учеников с низким показателями. Основываясь на полученных разных категориях студентов, преподаватели могут адаптировать различные подходы к обучению для работы с соответствующей группой студентов. Следовательно, производительность студентов может быть улучшена путем применения различных методов для каждой группы студентов. Из вышеупомянутого тематического исследования, кластеризация может разделить данные об образовании на соответствующие группы, и эти группы могут быть использованы для дальнейшего анализа для улучшения общей системы образования.

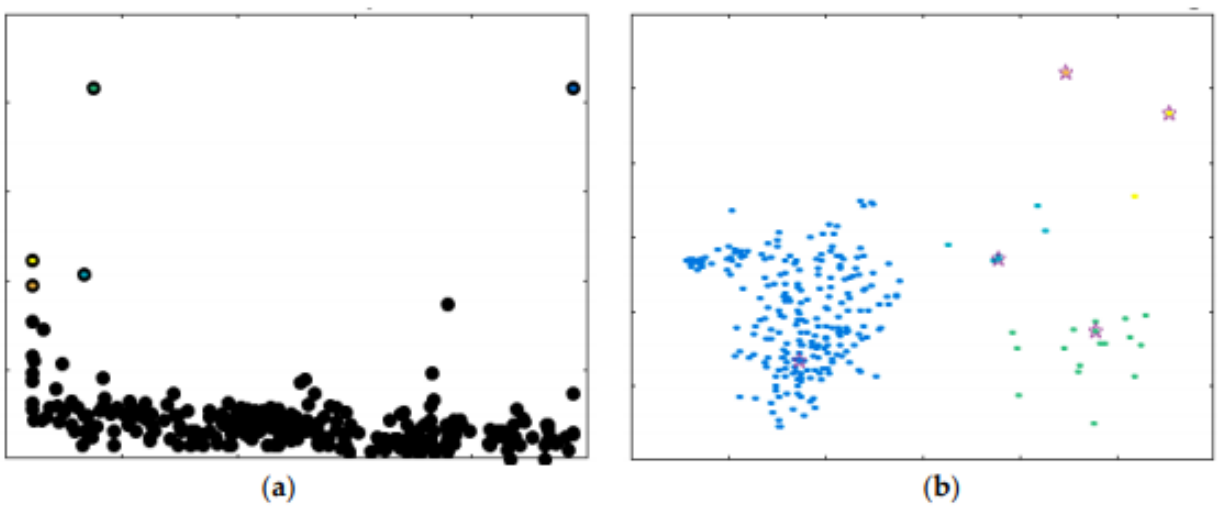

Рисунок 10 – Визуализация результатов эксперимента для метода CFSFDP-HD

(a) В графе решений показаны параметры p_i и δ_i. Идентификация кластерных центров достигается с помощью графа решений;

(b) Присвоение оставшихся точек идентифицированным кластерным центрам показано разными цветовыми схемами, где разные цвета представляют разные группы.

Как видим на Рисунок 10(b), центры кластеров расположились таким образом, что небольшое количество точек в верхнем правом углу не

игнорируются и не считаются за шумы, а организовываются в отдельные кластеры, что поможет спрогнозировать более точное количество баллов, которое может получить студент по окончании исследуемого курса.

4.2.2 Сравнение результатов анализа данных разными методами интеллектуального анализа

Существующие методы кластеризации, то есть сдвиг среднего значения, K-means, K-medoids, DBSCAN и иерархическая кластеризация, были протестированы с использованием одного набора данных из 600 учащихся и с учетом ограничений, описанных в таблице 3.

Таблица 1 - Список ограничений

Алгоритм	Ограничения
K-means	1. Количество кластеров k=4. 2. Количество итераций n=50.
K-medoids DBSCAN	1. Количество кластеров k=4. 2. Предопределенное количество итераций n=50. 3. ε= 0.5: определяет радиус окрестности вокруг точки данных X. 4. Минимальное количество точек (minPts = 10): минимальное количество соседей в пределах ε. 5. Не нужно указывать количество кластеров и итераций.
Иерархическая кластеризация	1. Количество кластеров k=4.
CFSFDP-HD	1. Не нужно явно указывать количество кластеров и количество итераций.

Путем сравнения CFSFDP-HD с K-means, K- medoids, DBSCAN и AHCT подход на основе графа решений CFSFDP-HD обеспечивает лучшее понимание для выбора потенциальных кластеров интуитивно. В обычной практике пользователи используют K-means и K-medoids более 1000 раз с различными настройками (т.е. числом кластеров и итераций), чтобы получить значимые

кластеры, однако подход на основе графа решений, используемый в CFSFDP-HD, обеспечивает эвристику чтобы получить точные решения за несколько итераций.

В то время как подход DBSCAN показывает некоторые зашумленные данные, также необходимо определить некоторые явные значения параметров, то есть ε и минимальные точки «minPts». ε определяет радиус окрестности вокруг точки данных, а минимальные точки данных представляют минимальное количество соседей в радиусе значения ε.

В подходе AHCT также необходимо явно указывать количество кластеров. Кроме того, четыре отдельные группы, показанные на Рисунок 10(b), могут быть легко изучены, визуализированы и сопоставлены с K-means, K-medoids, DBSCAN и AHCT на Рисунок 11(a), 10(b), 10(c), 10(d) соответственно. Представление точек данных на Рисунок 10 (b) и Рисунок 11 отличается из-за их отображаемой схемы. Учащиеся с хорошими оценками показаны слева на Рисунок 10(b), а учащиеся с хорошими оценками показаны в правом верхнем углу графиков на Рисунок 11.

Чтобы получить соответствующие кластеры с использованием обсуждаемых подходов, пользователи должны иметь предварительные знания существующих кластеров (количество кластеров) и в этом случае, не в состоянии обнаружить студентов с низкими и очень низкими показателями. Это ограничение делает неуместным обнаружение всех внутренних скрытых закономерностей в данных.

Для решения технических проблем рекомендуется использовать метод CFSFDP-HD для обнаружения существующих шаблонов, не имея технических знаний о базовых данных.

(a) - Результаты кластеризации K-means, черные звезды - центроиды скоплений, в то время как цветные точки - скопление данных;

(b) Показаны скопления K- medoids; черные звезды - это медоиды, в то время как разные цвета показывают разные скопления. Как K- means,

так и K- medoids имеют сходную природу и почти одинаковые результаты;

(c) Кластеры DBSCAN представлены разными цветами, в то время как зашумленные точки данных представлены черными кружками;

(d) Разные цвета используются для представления различных кластеров, идентифицированных агломеративным иерархическим деревом кластеров.

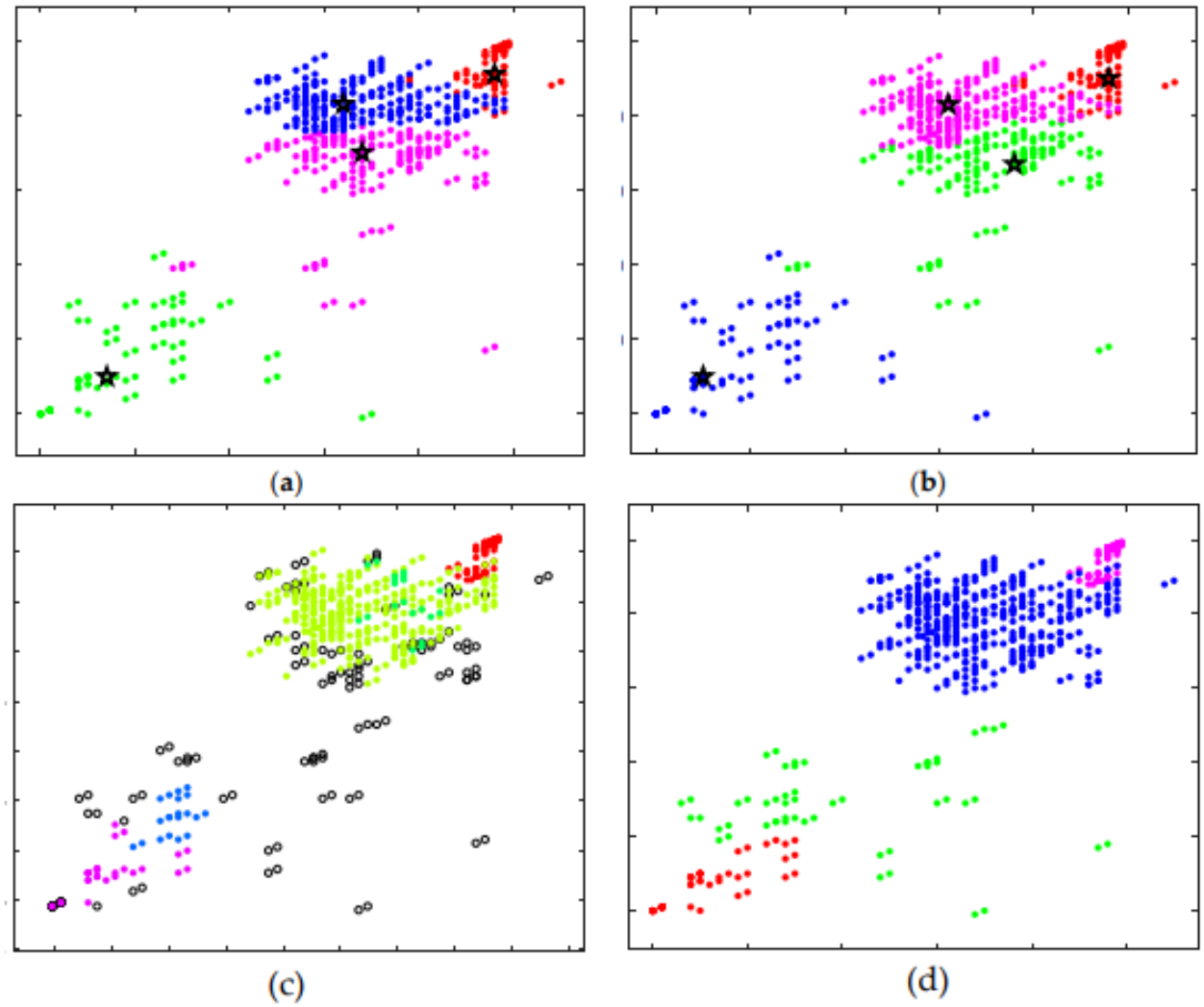

Рисунок 11 – Визуализация результатов кластеризации различными методами

Из экспериментов видно, что CFSFDP-HD является более адаптивным по своей природе, и его результаты более значительны по сравнению с некоторыми из существующих подходов.

4.2.3 Результаты исследования

Целью работы было применение методов и алгоритмов интеллектуального анализа данных для исследования поведения студента в системе электронного образования. В работе была представлена архитектура электронного обучения с использованием методов интеллектуального анализа данных. Также было рассмотрено потенциальное применение кластеризации в образовательных больших данных. Выше представлены результаты данного эксперимента, можно сказать, что цель работы достигнута.

Гипотеза исследования состояла в том, что можно применять методы и алгоритмы интеллектуального анализа для принятия решения о корректности поведения студента в системе электронного обучения. Для этого, в пункте 3.3.6 были обозначены методы оценки поведения студента. В пункте 4.1 реализована технология, которая обеспечивает достоверный анализ поведения студента при прохождении курса электронного обучения. В пункте 4.2 проведено исследование поведение студента при помощи методов и алгоритмов интеллектуального анализа данных. Исходя из изложенного выше, можно сказать, что гипотеза доказана.

Подходы интеллектуального анализа данных дают результаты в существующих системах электронного обучения, эффективно и результативно. Из литературы было отмечено, что традиционные системы электронного обучения в основном основаны на запросах, и на запросы работают без каких-либо интеллектуальных или эвристических данных. Точно так же K-means подходит для кластеризации образовательных данных, в которых известно количество кластеров, и сталкивается с недостатками при применении к кластерам неизвестного размера. Следовательно, более надежные подходы к интеллектуальному анализу данных (CFSFDP и CFSFDP-HD) включены в

предлагаемую систему электронного обучения для поиска кластеров в образовательных данных. Кроме того, было оценено, что методы интеллектуального анализа данных эффективны при анализе больших данных, что делает системы образования надежными и способными решать проблемы междисциплинарных исследований, эмоционального обучения и электронного обучения.

Для будущей работы подходы к интеллектуальному анализу данных могут быть еще улучшены, если они станут более чувствительны для генерирования знаний и оказания более объективной помощи студентам. Большие и реальные наборы данных могут быть смоделированы для анализа поведения предлагаемых подходов интеллектуального анализа данных. Обучение студентов может быть улучшено путем внедрения интеллектуальных игр. Общение студентов является важным аспектом обучения посредством группового обсуждения и обмена личным опытом.

ЗАКЛЮЧЕНИЕ

В ходе выполнения магистерской работы были сформулированы цели и задачи, определены объект и предмет, выдвинута гипотеза, обозначена актуальность темы исследования.

Выполнены задачи магистерской диссертации:

- проведен сравнительный анализ существующих способов и средств анализа поведения студента при прохождении курса электронного обучения;
- исследованы возможные критерии сравнения поведения студентов;
- исследованы методы и алгоритмы интеллектуального анализа данных;
- разработана математическая модель оценивания студента по данным его поведения;
- спроектирована и реализована математическая модель интеллектуального анализа поведения студента при прохождении курса электронного обучения;
- проверена эффективность разработанной модели.

Основываясь на анализе работ по теме диссертации, были описаны теоретико-методологические основы интеллектуального анализа данных поведения студентов в системе электронного обучения. Обозначены основные пункты анализа поведения студентов в системе электронного обучения. Рассмотрены методы интеллектуального анализа данных в системе электронного обучения. Выделены главные достоинства и недостатки существующих программных пакетов по оценке поведения студентов в системе электронного обучения. Описаны существующие программные пакеты, реализующие методы интеллектуального анализа данных в системе электронного обучения.

Был проведен анализ существующих разработок на тему «Методы и алгоритмы интеллектуального анализа данных для оценки поведения студента в системе электронного обучения», рассмотрены существующие системы

электронного обучения и их особенности. Исследованы модели анализа процесса онлайн-обучения, а также big data в системах электронного образования. Описана форма представления данных о студентах, алгоритмы выгрузки больших данных и представлена логическая модель данных в системах электронного образования.

Исследованы методы интеллектуального анализа данных:

- прогнозирование;
- обнаружение структуры;
- выявление взаимосвязей.

Исследованы алгоритмы интеллектуального анализа данных, обозначены их достоинства и недостатки:

- сдвиг среднего значения;
- K-means;
- K-medoids;
- DBSCAN;
- иерархическая кластеризация;
- CFSFDP и CFSFDP-HD.

Проведено математическое и компьютерное моделирование алгоритма CFSFDP-HD, представлен эксперимент и сравнение результатов анализа работы разных методов.

Основной научный результат магистерской диссертации заключаются в том, что применение методов и алгоритмов интеллектуального анализа данных для исследования поведения студента в системе электронного образования может быть улучшен за счет применения алгоритмов CFSFDP и CFSFDP-HD. Применение в системах электронного образования алгоритмов CFSFDP и CFSFDP-HD является эффективным и целесообразным.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

Научная и методическая литература

1. Балашова, Ю. В. Когнитивные и личностные особенности студентов очного и дистанционного обучения : диссертация ... кандидата психологических наук : 19.00.01 / Балашова Юлия Владимировна; [Место защиты: Моск. гос. гуманитар. ун-т им. М.А. Шолохова].- Москва, 2011.- 180 с.: ил.

2. Буваков К.В. Организация самостоятельной работы студентов по дисциплинам специализации с применением интернет-технологии в программной среде web c ourse t ools[Текст]. – ТПУ, 2010 С.-Петерб. нац. исслед. ун-т информац. технологий, механики и оптики]. - Санкт-Петербург, 2013. - 19 с.

3. Готская И. Б. Аналитическая записка «Выбор системы дистанционного обучения» / Готская И. Б., Жучков В. М., Кораблев А. В.// РГПУ им. А.И. Герцена.

4. Ильин, Е.П. Психология индивидуальных различий / Е.П. Ильин. — СПб.: Питер, 2011. — 701 с.: ил. — (Серия «Мастера психологии»).

5. Михалева, Г. В. Особенности дистанционного обучения в системе образования [Текст] / Г. В. Михалева, Т. В. Ромашова // Актуальные вопросы современной педагогики: материалы V междунар. науч. конф. (г. Уфа, май 2014 г.). — Уфа: Лето, 2014. — С. 39-41.

6. Муромцев, Д.И. Автоматизация оценки знаний студентов в системе электронного обучения ECOLE // Программные продукты и системы . 2015. №3 (111).

7. Муромцев, Д.И. Автоматизация оценки знаний студентов в системе электронного обучения ecole / Д.И. Муромцев, Ф.А. Козлов // Программные продукты и системы. - 2015. - №3 (111). - С.175-179.

8. Никерин, Е.А. Информационно-аналитическая система экспертной оценки поведения студента при прохождении курса дистанционной формы обучения - 2017. – С. 40-50.

9. Ниязова Г. Ж. Особенности использования lms moodle для дистанционного обучения [Текст] / Г. Ж. Ниязова, Г. А. Дуйсенова, Б. А. Иманбеков // Молодой ученый. — 2014. — №3. — С. 991-994.

10. Образование и XXI век: Информационные и коммуникационные технологии[Текст] / Под ред. И.М. Маркова — М.: Наука, 2010. — 191 с.

11. Охрименко, Е. И. Дидактические проблемы применения средств новых информационных технологий в системе дистанционного обучения / Е. И. Охрименко // Новые образовательные технологии в вузе : материалы XI международной научно-методической конференции. — Екатеринбург, 2014. – С. 2-4.

12. Савинов, А. Н. Методы, модели и алгоритмы распознавания клавиатурного почерка в ключевых системах : диссертация ... кандидата технических наук : 05.13.19 / Савинов Александр Николаевич;

13. Савченко, А.А. Особенности обучения финансовой математике по дистанционной форме обучения [Текст] / А. А. Савченко // Методика преподавания экономических дисциплин : научные статьи / Москва, 2013. — С. 92-96.

14. Смирнова, Н.А. Системы управления обучением в дистанционном образовании [Текст] / Н. А. Смирнова // Сборники конференций НИЦ Социосфера : научные статьи / Чехия, 2014. — С. 129-131.

15. Стреляу, Я. Роль темперамента в психическом развитии / Я. Стреляу. — М.: Прогресс, 2014. — С. 157-160.

16. Студеникина, Л.И. Педагогические условия эффективности использования элементов электронного обучения в вузовской профессиональной подготовке студентов: на материале математической подготовки: диссертация … кандидата педагогических наук: 13.00.08 / Студеникина Лариса Ивановна; [Место защиты: Моск. гос. гуманитар. ун-т им. М.А. Шолохова].- Москва, 2011.- 180 с.: ил.

17. Уддин, Актхер. Сравнительный анализ личностных и мотивационных особенностей студентов очного и дистанционного обучения :

на примере студентов-психологов : диссертация ... кандидата психологических наук : 19.00.07 / Уддин Мд. Актхер; [Место защиты: Моск. псих.-соц. ун-т] - Москва, 2014. - 234 с.

18. Новые образовательные технологии в вузе : материалы X международной научно-методической конференции. — Екатеринбург, 2013. - С. 2-4.

19. Ядов Г.Б. Информация и общество[Текст]// Вокруг света. - 2011. - №2.

Электронные ресурсы

20. Информационные технологии в дополнительном образовании [Электронный ресурс]. - Режим доступа: www.ict.edu.ru/ft/004444/sec3.pdf

21. Компьютерное тестирование в образовании. [Электронный ресурс]. - Режим доступа: http://www.slmini.narod.ru.

22. Обзор Мирового и российского рынка электронного обучения. [Электронный ресурс]. – Режим доступа: https://studfiles.net/preview/5764773/

23. Рузавин Г. И. Методология научного познания [Электронный ресурс]: учебное пособие для вузов / Г. И. Рузавин. - Москва : ЮНИТИ-ДАНА, 2012. - 287 с. - ISBN 978-5-238-00920.

24. Шипулина, Л. Г. Дистанционное обучение как форма самостоятельной работы студента [Электронный ресурс] / Л. Г. Шипулина //

25. Data Mining and Knowledge Discovery in the Real World [Электронный ресурс]. – Режим доступа: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/3695/0000/Data-mining-in-the-real-world/10.1117/12.339981.short?SSO=1

26. Data Mining [Электронный ресурс]. – Режим доступа: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/3787/0000/Data-mining

27. Data Mining от Oracle: настоящее и будущее [Электронный ресурс]. – Режим доступа: http://citforum.ru/database/oracle/data_mining_solutions/

28. From Data Mining to Knowledge Discovery in Databases [Электронный ресурс]. – Режим доступа: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

29. Online Behavior Analysis-Based Student Profile for Intelligent E-Learning [Электронный ресурс]. – Режим доступа: www.ejel.org/issue/download.html?idArticle=525

30. The Effectiveness of E-Learning: An Explorative and Integrative Review of the Definitions, Methodologies and Factors that Promote e-Learning Effectiveness [Электронный ресурс]. – Режим доступа: www.ejel.org/issue/download.html?idArticle=438

31. The E-Learning Setting Circle: First Steps Toward Theory Development in E-Learning Research [Электронный ресурс]. – Режим доступа: www.ejel.org/issue/download.html?idArticle=571

Литература на иностранном языке

32. Ahmed A. Saleh, Hazem M. El-Bakry, Taghreed T. Asfour and Nikos Mastorakis “Adaptive E-Learning Framework for Digital Design," Proc. of 9th WSEAS International Conference on Telecommunications and Informatics, Italy, May 29-31, 2010

33. Ahmed A. Saleh, Hazem M. El-Bakry, Taghreed T. Asfour and Nikos Mastorakis “Adaptive E-Learning Tools for Numbering Systems," Proc. of 9th WSEAS International Conference on Applications of Computer Engineering (ACE'10), Penang, Malaysia, March 23-25, 2010

34. Almeda, M. V. Clustering of Design Decisions in Classroom Visual Displays. In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge; ACM, 2014; pp 44–48.

35. Anaya, A. R. Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks. International Working Group on Educational Data Mining, 2009.

36. Antonenko, P. D. Using Cluster Analysis for Data Mining in Educational Technology Research. Educational Technology Research and Development 2012.

37. Baker, R. Data Mining for Education. International encyclopedia of education 2010.

38. Blanco, T. From the Islands of Knowledge to a Shared Understanding: Interdisciplinarity and Technology Literacy for Innovation in Smart Electronic Product Design. International Journal of Technology and Design Education. 2017.

39. Chang, W.-C. Learning Ability Clustering in Collaborative Learning. JSW 2010.

40. Chen, C.-M. Diagnosis of Students' Online Learning Portfolios. In Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007.

41. Cordeiro, M.Online Social Networks Event Detection: A Survey. In Solving Large Scale Learning Tasks. Challenges and Algorithms; Springer, 2016.

42. Daniel, B. B Ig D Ata and Analytics in Higher Education: Opportunities and Challenges. British journal of educational technology 2015.

43. Dede, C. Next Steps for" Big Data" in Education: Utilizing Data-Intensive Research. Educational Technology 2016.

44. Dutt, A. Systematic Review on Educational Data Mining. IEEE Access 2017.

45. Engström, S. Differences and Similarities between Female Students and Male Students That Succeed within Higher Technical Education: Profiles Emerge through the Use of Cluster Analysis. International Journal of Technology and Design Education 2018.

46. F. Matsebula, E. Mnkandla, "A big data architecture for learning analytics in higher education," IEEE africon: Science, Technology and Innovation for Africa, pp. 951-956, 2017

47. Gaeta, M. An Approach To Personalized E-Learning. Journal of Education, Informatics & Cybernetics 2013.

48. Hazem M. El-Bakry, Alaa M. Riad, Aziza S. Asem, Mohamed E. Ibrahim, Ahmed E. Hassan, Mahmod S. Kandel, and Nikos Mastorakis "Design and Implementation of Total Quality Assurance Management System for Universities, "

Proc. of Recent Advances in Business Administration, Cambridge, UK, February 20-22, 2010

49. Ivancevic, V. The Individual Stability of Student Spatial Deployment and Its Implications. In Computers in Education (SIIE), 2012 .

50. J. F. H. Barril, Q. Tan, "Integrating privacy in architecture design of student information system for big data analytics," IEEE Conference on Cloud Computing and Big Data Analysis, pp. 139-144, 2017

51. J. Liang, J. Yang, Y. Wu, C. Li, L. Zheng, "Big data application in education: dropout prediction in edx MOOCs," IEEE 2nd International Conference on Multimedia Big Data, pp. 440-443, 2016

52. J. Shu, et al, "Exploration on college education big data open service platform," 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA, pp. 161-165 2017

53. Kim, G.-H. Big-Data Applications in the Government Sector. Communications of the ACM 2014.

54. L. Cen, D. Ruta, J. Ng, "Big education: opportunities for big data analytic," International Conference on Digital Signal Processing, DS 2015

55. L. Lin. "A study on general education in local institutions of higher learning," 3rd International Conference on Management, Education Technology and Sports Science, METSS, 2016

56. M.S. Vyas, R. Gulwani, "Predictive analytics for e-learning system," International Conference on Inventive Systems and Control, ICISC 2017

57. Markowska-Kaczmar, U. Intelligent Techniques in Personalization of Learning in E-Learning Systems. In Computational Intelligence for Technology Enhanced Learning; Springer, 2010.

58. Mehmood, R. Clustering by Fast Search and Find of Density Peaks via Heat Diffusion. Neurocomputing 2016.

59. Prakash, B. R. Big Data in Educational Data Mining and Learning Analytics. Int. J. Innov. Res. Comput. Commun. Eng 2014.

60. Qian, G. Semisupervised Clustering by Iterative Partition and Regression with Neuroscience Applications. Computational intelligence and neuroscience 2016, 2016.

61. Saa, A. A. Educational Data Mining & Students' Performance Prediction. International Journal of Advanced Computer Science and Applications 2016.

62. Shah, G. H. An Empirical Evaluation of Density-Based Clustering Techniques. International Journal of Soft Computing and Engineering (IJSCE) ISSN 2012.

63. Siemens, G. Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE review 2011.

64. Tulasi, B. Significance of Big Data and Analytics in Higher Education. International Journal of Computer Applications 2013.

65. Wiwie, C. Comparing the Performance of Biomedical Clustering Methods. Nature methods 2015.

66. Yan-lin, L. L. Z. The Application of the Internet of Things in Education [J]. Modern Educational Technology 2010.

67. Ying, K. Clustering Students Based on Their Annotations of a Digital Text. In Technology for Education (T4E), 2012.