Особенности использования цифрового следа обучающихся в системах искусственного интеллекта в образовании

М.С. Дьяченко

ФГУ ФНЦ НИИСИ РАН, Москва, Россия, mdyachenko@niisi.ru

Аннотация. В статье рассмотрены особенности сбора и использования данных цифрового следа обучающихся на этапах внедрения новых учебных технологий на основе искусственного интеллекта. Исследуемые особенности характерны для разработки решений, требующих большого объёма предварительно накопленных данных цифрового следа обучающихся. Рассмотрены вопросы ускорения накопления данных, обезличивания данных цифрового следа обучающихся, вопрос обучения модели по федеративной схеме без извлечения данных, рассмотрены схемы исследований без достаточного объема данных, а также подходы к адаптации внедренных решений при переносе их между учебными заведениями или предметными областями.

Ключевые слова: цифровой след обучающихся, интеллектуальные системы обучения, машинное обучение, обезличивание цифрового следа обучающихся, адаптация решений ИИ

1. Введение

Необходимость технологической независимости Российской Федерации становится в наше время стимулом для дальнейшего повышения уровня подготовки выпускников школ и вузов. В системе образования РФ активно готовятся к внедрению в учебный процесс перспективных методов искусственного интеллекта. Одним из признаков такой подготовки можно считать появление стандартов для образовательных продуктов с алгоритмами искусственного интеллекта. Стандарты в том числе охватывают требования по работе с данными, а также требования к контрольным выборкам для испытания систем искусственного интеллекта в образовании. Такое внимание к данным не случайно, поскольку для создания решений на основе современных методов искусственного интеллекта необходимы большие наборы данные об обучаемых, собираемых учебной системой.

Решения на основе ИИ в образовании используются для реализации адаптивного обучения, элементов предсказательной учебной аналитики, систем поддержки принятия решения и автоматической генерации заданий. Проводившиеся в РФ эксперименты по внедрению технологий ИИ в образовании основывались на реализации методов, которые могут быть разработаны без больших объёмов предварительно накопленных данных об обучаемых, например, как в исследовании Кречетова И.А. с использованием генетических алгоритмов и кривых забывания [1], или адаптивных технологий без использования предварительно накопленных данных, как в исследованиях Шамсутдиновой Т.М. [2] и Вайнштейн Ю.В. [3]. Однако с появлением перспективных технологий, например, таких как глубокая трассировка знаний [4], её развития для учёта модели забывания [5] или содержания заданий [6], становится актуальным накапливать и использовать большие объёмы данных об обучении.

Требования к сбору, хранению, обработке, передаче и защите данных для систем ИИ в образовании приводятся в ГОСТ Р 59897-2021. Однако в рассмотренных исследованиях отсутствует достаточно информации об особенностях внедрении в РФ технологий обучения на основе ИИ, использующих для разработки большие объёмы накопленных данных. В связи с этим становится актуальной задача исследования особенностей использования цифрового следа обучающихся в процессе внедрения технологий ИИ в образовании, использующих большие объёмы предварительно накопленных данных.

2. Методология исследования

В исследовании используется описанный в работе автора [7] процесс внедрения новых учебных технологий, разрабатываемых с использованием больших объёмов данных об обучении. Укрупненно процесс состоит из последовательных этапов: накопления данных об обучении, краткосрочной оценки эффективности, долгосрочной оценки эффективности и масштабирования учебной технологии, использующей ИИ. На каждом из этих этапов рассматриваются особенности, вызванные использованием большого объёма данных. Выполнен анализ проведенных различными исследователями результа-

тов внедрения рассматриваемых учебных технологий и использования накопленных данных об обучении.

3. Результаты исследования

3.1. Сбор данных цифрового следа обучающихся

Рассмотрим некоторые особенности сбора данных цифрового следа обучаемых (далее ЦСО), происходящего на этапах подготовки к разработке и оценке эффективности внедряемой учебной технологии, использующей ИИ.

Данные для разработки решений на основе ИИ поступают из внутренних и внешних источников данных. Примеры внешний источников – это социальные сети [8] или библиотечная информационная система. Внешние источники данных должны предоставлять интерфейс для получения данных по результатам использования обучаемым внешней системы, например, это данные поисковых запросов в библиотечной информационной системе или данные диалога обсуждения задания с одногруппниками в социальной сети. При работе с внешними источниками данных необходимо учитывать время гарантированного хранения данных во внешней системе, возможность сопоставления идентификатора обучаемого во внутреннем и внешнем источниках данных, возможность упорядочивания данных внутренних и внешних источников в порядке возникновения событий. Например, если время гарантированного хранения данных во внешней системе меньше необходимого времени накопления данных, то данные из внешней системы следует копировать в хранилище, обеспечивающее необходимое время хранения данных ЦСО. Данные из каждого источника должны обладать метаданными с описанием содержания, технических деталей хранения, процесса обработки и доступа к данным.

В учебных системах, использующих ИИ, регистрации подлежат также результаты работы ИИ решений и действия преподавателя в системе. К этим данным также должны прилагаться метаданные. Для анализа эффективности внедряемого ИИ решения также необходимо собирать результаты работы ИИ алгоритмов (например, прогноз результатов обучения, результаты трассировки знаний – все, что влияет на поведение автоматизированной системы обучения) и регистрировать действия преподавателя, например, такие как ручная корректировка действия, предложенного автоматизированной системой с использованием алгоритма ИИ. Эти данные необходимы для анализа эффективности внедряемого ИИ решения.

Следующей особенностью сбора данных ЦСО для разработки и оценки эффективности является значительная длительность подготовительного этапа, обусловленная ограниченной пропускной способностью экспериментальных курсов. Одной из характеристик данных является их изменчивость – промежуток времени, в рамках которого данные остаются актуальными для целей исследования, например, изменение структуры учебного курса или учебных материалов может привести к потере актуальности данных результатов проверки знаний. С учётом изменчивости данных можно определить необходимую минимальную пропускную способность экспериментальных курсов для сбора требуемого объёма данных в пределах периода их актуальности [7]. Ускорить сбор данных можно за счёт подключения к эксперименту большего количества учебных заведений или запуска параллельного онлайн курса с менторами, который не привязан к расписанию учебного заведения, но при этом повторяет структуру курса учебного заведения.

Следующей особенностью сбора данных ЦСО является необходимость продолжать собирать данные после завершения курса для автоматизации контрольного среза остаточных знаний [9]. Для реализации контрольного среза необходимо собрать данные экспериментального курса и данные одного из следующих за ним курсом, который использует знания, полученные на экспериментальном курсе. В результате становится возможным оценить не только краткосрочный эффект от внедряемой учебной технологии, но и долгосрочный эффект.

К особенностям решений на основе ИИ, использующих большие объёмы накопленных данных, можно отнести необходимость сбора данных после завершения внедрения для обеспечения непрерывной оценки качества решения и доработки решения с целью улучшения его характеристик с использованием вновь полученных данных.

3.2. Использование накопленных данных цифрового следа обучающихся

Рассмотрим некоторые особенности использования накопленных данных ЦСО, которые в первую очередь нужны для разработки (обучения) решения на основе ИИ, а также для анализа данных и оценки эффективности применения разработанного решения.

Данные ЦСО для исследований и разработки можно использовать только в обезличенном виде. Обезличивание данных позволяет убрать из них компоненты, позволяющие однозначно идентифицировать принадлежность блока дан-

ных к конкретному обучаемому. Согласно исследованию Грацского технического университета [10] учебные данные обладают большим количеством перекрестных связей, что затрудняет их обезличивание и создает высокие риски для повторной идентификации. Например, имея доступ к данным учебной группы с временными метками событий можно однозначно определить группу, данные которой зарегистрированы, а анализируя сами данные (например, полученные оценки) сопоставить результаты с данными из других источников, таких как социальные сети и открытые источники информации. В результате появляется возможность однозначно сопоставить обезличенные данные с конкретными обучаемыми. В связи с этим обработка даже обезличенных данных должна выполняться в автоматическом режиме без участия человека. Например, обобщение данных для анализа можно выполнять без извлечения данных, также как и обучение решений на основе ИИ может выполняться по федеративной схеме [11] с использованием данных непосредственно в месте их хранения без необходимости их извлекать или перемещать.

Состав и формат сохраняемых в учебных системах данных ЦСО определяется на этапе разработки, например, это база данных в СДО Moodle [12] или журналы OpenEdu [13], а также возможности по регистрации данных ограничены доступными интерфейсами взаимодействия с внешними системами (библиотечные информационные системы, системы видеоконференций, диалоговые ассистенты преподавателя, социальные сети), которыми обучаемый пользуется в процессе обучения. Хотя учебные системы поддерживают унифицированные форматы регистрации результатов обучения, такие как спецификация Experience API (xAPI, https://xapi.com/), в настоящий момент в системе образования РФ нет единого общепризнанного формата хранения детальных данных ЦСО. Инициатива «Университета 20.35» по созданию стандарта ЦСО (https://standard.2035.university/), вероятно, направлена на решение специфичных задач, поскольку предлагаемый формат ЦСО содержит отсылки к специализированным коммерческим решениям, таким как Trello, Zoom, Discord, Slack. Внедрение общепринятого стандарта хранения данных ЦСО и реализация его в распространённых в РФ системах обучения позволит расширить возможности для исследований за счет охвата большего объёма данных детального ЦСО, а также даст возможность в перспективе создать большие открытые наборы данных для исследований и разработки в области применения ИИ в обучении.

3.3. Проблема недостаточного для разработки объёма данных цифрового следа обучающихся

Процесс накопления больших объёмов данных эффективно решён только в МООК масштаба таких систем, как Coursera (https://www.coursera.org/) и Stepik (https://stepik.org/), или в коммерческих системах обучения, таких как Пларио (https://plario.ru/) и KNewton (https://www.knewton.com/), учебные курсы которых одновременно используются в нескольких учебных заведениях.

Накопление необходимых данных в рамках одного учебного заведения занимает много времени [7], поэтому разработчики вначале запускают систему с автоматизированным учебным курсом, который реализует элементы ИИ, не требующие для разработки использования больших объёмов предварительно накопленных данных, после чего начинается предварительный этап сбора данных, в ходе которого недостающие данные ЦСО накапливаются, опять же за продолжительное время. При этом существуют организационные особенности, ограничивающие возможности по накоплению данных ЦСО в учебном заведении: начальные курсы в учебных заведениях как правило имеют большую пропускную способность поскольку являются частью общей базовой подготовки, через которую проходят все обучаемые, тогда как специализированные курсы при завершении обучения имеют существенно меньший охват.

По причине сложности накопления больших объёмов данных в сжатые сроки исследователи используют следующий подход при внедрении новых технологий [14]: берутся открытые для исследований наборы данных, которые нужны для получения предварительных результатов, после чего выполняют запуск эксперимента в учебных заведениях для накопления реальных данных и оценки эффективности внедряемой технологии. В случае отсутствия доступных для исследования наборов открытых данных формируются синтетические данные с требуемыми характеристиками, которые используются для получения предварительных результатов, после чего также начинается экспериментальная проверка с использованием реальных данных в учебном заведении.

Особенностью этапа масштабирования использования внедряемой технологии является также необходимость повторного сбора данных для разработки (или доработки) и оценки эффективности применения решения с элементами ИИ. Если решение с элементами ИИ зависит от структуры входных данных, например, как решения с использованием трассировки знаний,

зависящее от структуры предметной области, то изменение учебной программы или корректировка учебных материалов приведёт к необходимости доработки решения или к необходимости повторной разработки решения, начиная с этапа накопления данных. Для упрощения распространения решений на основе ИИ разрабатываются методы адаптации решения к новым условиям применения – адаптация предметной области. Например, для решений на основе глубокой трассировки знаний разработан алгоритм для переноса реализации из одного учебного заведения в другое с учётом изменения структуры и содержания данных [14], примечательно, что для переноса требуется меньше данных, чем для повторной разработки решения на данных нового учебного заведения.

4. Заключение

Наличие больших объёмов накопленных данных ЦСО является предусловием для обеспечения возможности исследования и внедрения в системы автоматизации учебного процесса перспективных методов искусственного интеллекта.

Накопление данных ЦСО в рамках одного учебного заведения – длительный процесс из-за ограниченной пропускной способности экспериментальных курсов. Для ускорения накопления данных ЦСО в эксперименте должны участвовать несколько учебных заведений, что возможно при создании гетерогенной экспериментальной среды или при использовании централизованной системы обучения, внедрённой в учебные процессы нескольких учебных заведений. Перспективным направлением является создание открытых больших наборов данных для исследований и разработки новых решений в области применения ИИ в образовании.

Использование накопленных данных ЦСО возможно только в обезличенном виде, при этом обезличивание данных ЦСО учебного заведения является сложной процедурой из-за большого количества перекрестных связей. По этой причине для учебных данных высок риск повторной идентификации (восстановление привязки к личности по обезличенным данным). В связи с этим необходимо исключать возможность выгрузки даже обезличенных данных в процессе исследований и разработки, выполняя процессы анализа данных и разработки в автоматическом режиме без участия человека. Например, за счет использования методов федеративного обучения решений с использованием ИИ, при котором распределенно хранимые данные используются непосредственно в месте их хранения без необходимости извлекать или перемещать данные на время разработки.

Внедренные решения с элементами ИИ не обязательно являются универсальными и могут зависеть от структуры входных данных. В результате этого при переносе разработанного решения в другое учебное заведение или при изменении материалов учебного курса необходимо дорабатывать решение за счет использования актуальных данных ЦСО. Для исключения необходимости повторять весь процесс накопления данных ЦСО и следующей за ним повторной разработки, исследуются методы по смене предметной области, применения которых позволит, собрав небольшой объём актуальных данных, адаптировать созданное решение для использования в новой предметной области.

Работа выполнена в ФГУ ФНЦ НИИСИ РАН в рамках фундаментальных исследований по теме FNEF-2022-0010.

Specificities of Using the Digital Footprint of Students in Artificial Intelligence Systems in Education

M.S. Dyachenko

Abstract. The article discusses the features of collecting and using digital footprint data of students at the stages of introducing new educational technologies based on artificial intelligence. The studied specificities are common for the development of solutions that require a large amount of pre-accumulated data of the digital footprint of students. The issues of accelerating data accumulation, depersonalization of the digital footprint of students, the issue of using a federated learning without data extraction, research and development schemes without a sufficient amount of data, as well as approaches to adapting implemented solutions when transferring them between educational institutions or subject areas are considered.

Keywords: digital footprint of students, intelligent learning systems, machine learning, depersonalization of the digital footprint of students, AI solutions adaptation

Литература

1. И. А. Кречетов, В. В. Романенко Реализация методов адаптивного обучения // Вопросы образования. 2020. № 2. С. 252–277.

2. Т. М. Шамсутдинова. Формирование индивидуальной образовательной траектории в адаптивных системах управления обучением // Открытое образование. 2021. Т. 25. № 6. С. 36–44.

3. В. А. Шершнева, Ю. В. Вайнштейн, Т. О. Кочеткова. Адаптивная система обучения в электронной среде // Программные системы: теория и приложения. 2018. Т. 9. № 4 (39). С. 159–177.

4. C. Piech и др. Deep Knowledge Tracing., NIPS, 2015.

5. W. Zhao и др. A novel framework for deep knowledge tracing via gating-controlled forgetting and learning mechanisms // Information Processing & Management. 2023. Т. 60. № 1. С. 103-114.

6. Q. Liu и др. EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction // IEEE Transactions on Knowledge and Data Engineering. 2021. Т. 33. № 1. С. 100–115.

7. М. С. Дьяченко, А. Г. Леонов. Архитектура учебной системы с индивидуализацией обучения на основе накопленных данных результатов автоматизированной проверки заданий // Успехи кибернетики. 2023. Т. 4. № 1(13). С. 39–48.

8. E. Pozdeeva и др. Assessment of Online Environment and Digital Footprint Functions in Higher Education Analytics // Education Sciences. 2021. Т. 11. № 6. С. 256.

9. С. М. Григорьев и др. Анализ сущности и содержания контроля успеваемости обучающихся // Национальная ассоциация ученых. 2015. № 4– 3 (9). С. 102–104.

10. M. Khalil, M. Ebner De-Identification in Learning Analytics // Learning Analytics. 2016. Т. 3. № 1.

11. C. Fachola и др. Federated Learning for Data Analytics in Education // Data. 2023. Т. 8. № 2. С. 43.

12. Т. А. Кустицкая и др. Цифровой след из LMS Moodle для прогнозирования результатов обучения // Информатизация образования и методика электронного обучения. 2022.

13. Н.Д. Барсуков и др. Анализ активности студентов на курсах онлайн-обучения на основе логов платформы «OpenEdu» // Труды Института системного программирования РАН. 2020. Т. 32. № 3. С. 91–100.

14. S. Cheng и др. AdaptKT: A Domain Adaptable Method for Knowledge Tracing // Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining WSDM ’22. New York, NY, USA: Association for Computing Machinery, 2022. С. 123–131.