MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 29, NO. 4, 2022
journal homepage: www.mais-journal.ru

THEORY OF DATA

Analysis of Students' Academic Performance using LMS Event Logs

N. D. Shaimov[1], I. A. Lomazova[1], A. A. Mitsyuk[1], I. Y. Samonenko[1]

DOI: 10.18255/1818-1015-2022-4-286-314

[1]HSE University, 20 Myasnitskaya str., Moscow 101000, Russia.

MSC2020: 68U35
Research article
Full text in Russian

Received June 5, 2022
After revision August 23, 2022
Accepted August 26, 2022

Modern educational process involves the use of electronic educational environments. These are special information systems that are both a means for storing educational materials and a tool for conducting tests, collecting homework, keeping a grade book, and working together. Such environments produce a large amount of data containing the recorded behavior of students and teachers within the educational process. This paper proposes an approach that allows one to analyze such data and discover typical student trajectories that lead to successful or unsuccessful learning outcomes. It is shown how process mining can be used to build models of the educational process based on the available data. We also show how you can evaluate the extent to which the synthesized model reflects the actual behavior of the system recorded in event logs. The paper contains not only a description of the proposed approach, but also a case study with its application to a real data set for an undergraduate educational program. It is clearly shown how, using our approach, it is possible to find out what factors lead to the formation of successful and unsuccessful student trajectories. The bottlenecks of the educational process were identified, as well as errors in the data, indicating the incorrect operation of the system. As a result of the analysis, points of special attention for administrators of the educational program were identified, as well as some signal events, the appearance of which in a student's individual trajectory can be an alarm. The application of the approach involves the use of free open source software, which further facilitates its deployment in a variety of educational organizations.

Keywords: process analysis; process mining; learning management systems; event logs

INFORMATION ABOUT THE AUTHORS

Nikita D. Shaimov	orcid.org/0000-0003-3843-5379. E-mail: nshaimov@hse.ru Postgraduate student.
Irina A. Lomazova correspondence author	orcid.org/0000-0002-9420-3751. E-mail: ilomazova@hse.ru Professor, Doctor of Sciences in Theoretical Foundations of Computer Science.
Alexey A. Mitsyuk	orcid.org/0000-0003-2352-3384. E-mail: amitsyuk@hse.ru Associate Professor, PhD in Computer Science.
Ilya Yu. Samonenko	orcid.org/0000-0002-3063-4640. E-mail: isamonenko@hse.ru Associate Professor, PhD in Sociology.

Funding: This work is supported by the Basic Research Program at the National Research University Higher School of Economics.

For citation: N. D. Shaimov, I. A. Lomazova, A. A. Mitsyuk, and I. Y. Samonenko, "Analysis of Students' Academic Performance using LMS Event Logs ", *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 286-314, 2022.

© Shaimov N. D., Lomazova I. A., Mitsyuk A. A., Samonenko I. Y., 2022
This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 29, № 4, 2022
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Анализ академической успеваемости студентов с использованием журналов событий электронной образовательной среды

Н. Д. Шаимов[1], И. А. Ломазова[1], А. А. Мицюк[1], И. Ю. Самоненко[1]

DOI: 10.18255/1818-1015-2022-4-286-314

[1]Национальный исследовательский университет «Высшая школа экономики», ул. Мясницкая, д. 20, г. Москва, 101000 Россия.

УДК 004.04
Научная статья
Полный текст на русском языке

Получена 5 июня 2022 г.
После доработки 23 августа 2022 г.
Принята к публикации 26 августа 2022 г.

Современный образовательный процесс предполагает использование электронных образовательных сред. Это специальные информационные системы, которые являются как средством для хранения учебных материалов, так и инструментом для проведения проверочных работ, сбора домашних заданий, ведения журнала оценок, совместной работы. Такие среды производят большое количество данных о поведении учащихся и преподавателей в рамках учебного процесса. В данной работе предлагается подход, позволяющий анализировать такие данные, извлекать из них типичные траектории учащихся, которые ведут к успешным или неудачным результатам обучения. Показано, как для построения моделей образовательного процесса на основе имеющихся данных могут быть использованы алгоритмы process mining. Также показано, как можно оценить, насколько синтезированная модель отражает реальное поведение системы, записанное в журналах событий. Работа содержит не только описание предлагаемого подхода, но и пример его применения к реальному набору данных для образовательной программы бакалавриата. Наглядно показано, как с использованием нашего подхода можно выяснить, какие факторы приводят к формированию успешных и неудачных траекторий студентов. Выявлены узкие места образовательного процесса, а также ошибки в данных, свидетельствующие о некорректной работе системы. В результате анализа выявлены точки особого внимания для администраторов образовательной программы, а также определены некоторые сигнальные события, появление которых в индивидуальной траектории студента может быть тревожным сигналом. Применение подхода предполагает использование только свободных программных инструментов с открытым исходным кодом, что дополнительно облегчает его внедрение в самых разных образовательных организациях.

Ключевые слова: моделирование процессов; извлечение и анализ моделей процессов; электронная образовательная среда; журналы событий

ИНФОРМАЦИЯ ОБ АВТОРАХ

Никита Денисович Шаимов	orcid.org/0000-0003-3843-5379. E-mail: nshaimov@hse.ru аспирант.
Ирина Александровна Ломазова автор для корреспонденции	orcid.org/0000-0002-9420-3751. E-mail: ilomazova@hse.ru профессор, доктор физ.-мат. наук.
Алексей Александрович Мицюк	orcid.org/0000-0003-2352-3384. E-mail: amitsyuk@hse.ru доцент, канд. комп. наук.
Илья Юрьевич Самоненко	orcid.org/0000-0002-3063-4640. E-mail: isamonenko@hse.ru доцент, канд. соц. наук.

Финансирование: Работа выполнена при поддержке Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики».

Для цитирования: N. D. Shaimov, I. A. Lomazova, A. A. Mitsyuk, and I. Y. Samonenko, “Analysis of Students’ Academic Performance using LMS Event Logs ”, *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 286-314, 2022.

© Шаимов Н. Д., Ломазова И. А., Мицюк А. А., Самоненко И. Ю., 2022
Эта статья открытого доступа под лицензией CC BY license (https://creativecommons.org/licenses/by/4.0/).

Введение

В настоящее время активно происходит цифровизация традиционных процессов во всех областях хозяйственной деятельности: от банковского дела, промышленности и продаж до юриспруденции и систем голосования. Современное образование всех ступеней также всё больше строится на базе специализированных информационных систем. Особенно ускорился процесс внедрения цифровых технологий в образовании после начала пандемии COVID-19, когда все школы и вузы существенную часть учебного процесса вынуждены были перенести в дистанционный формат.

Сегодня образовательный процесс в высших учебных заведениях, как правило, построен с использованием специализированных информационных систем, которые также называют *системами управления обучением* или *цифровыми образовательными средами* [1]. Первоначально такие системы разрабатывались специально для организации дистанционного образования, но сегодня они используются и для поддержки обучения в традиционном формате. Студенты не только получают учебные материалы в электронной системе, но выполняют в ней остальные виды образовательной деятельности: делают индивидуальные и командные задания, проходят тестирования, контрольные работы и экзамены. Современная зачётная книжка студента также перешла в электронный вид, как и экзаменационные ведомости.

Переход образовательного процесса в электронный вид несёт с собой как риски, так и возможности. Обсуждение рисков такого перехода находится за рамками данной работы. Вместо этого обратим внимание на то, что одна из неотъемлемых особенностей современного образовательного процесса — фиксация всех действий его участников. Так как все действия как студента, так и преподавателя осуществляются через информационную образовательную среду и могут записываться. Можно сохранить факт входа в систему, обращения к тому или иному материалу, прохождение теста, просмотр видеолекции и так далее. Событийные данные сочетаются с другими, представляющими журналы оценок, программы курсов, тексты переписки студентов с преподавателями и т.д. В результате формируются огромные объёмы данных, которые обычно складируются на защищённых серверах образовательных организаций. Такие массивы данных могут быть использованы с пользой для администраторов образовательного процесса, преподавателей и студентов, что и демонстрируется в данной работе.

На основе анализа данных образовательного процесса можно, например, выявлять студентов или преподавателей, которым требуется помощь. Анализ такого рода данных позволяет совершенствовать структуру образовательной программы в целом, выявляет слабые связки дисциплин, неэффективные организационные решения и так далее. Дополнительные возможности возникают в том случае, если образовательный процесс устроен вариативным образом, когда студенты могут самостоятельно строить свою траекторию обучения. Задача исследователей в области информационных систем — предложить инструменты, которые бы помогли участникам образовательного процесса, включая администраторов, действительно эффективно использовать имеющиеся массивы данных. При этом, так как речь идёт о социальном процессе, в который вовлечены многие действующие лица, важно не только найти что можно улучшить в процессе, но ещё и не навредить никому из его участников.

В данной работе рассматривается, как данные об академической успеваемости студентов университета, получаемые из системы управления обучением, могут использоваться для выявления проблемных мест и ошибок при построении образовательной программы, которые приводят к неуспеху и неудачам студентов. Для анализа данных используется подход интеллектуального анализа процессов или, как его называют на английском языке, process mining [2]. Методы, объединяемые этим названием, которые будут рассмотрены далее в разделе 1, позволяют не просто выявлять зависимости в наборах данных, но разработаны с целью выявления динамики и причинно-следственных связей между событиями, происходящими в исследуемой системе. Это особенно полезно для

анализа образовательного процесса, в рамках которого неудачное выстраивание траектории обучения или неудачное прохождение какой-то дисциплины могут привести к провалу студента не сразу, а по прошествии существенного временного промежутка.

1. Анализ процессов

Приведём теперь некоторые базовые сведения из области интеллектуального анализа процессов, которые познакомят с ней неподготовленного читателя. А затем введём основные понятия и определения, которые потребуются далее.

1.1. Общее описание

Интеллектуальный анализ процессов (process mining) активно развивается с начала XXI века и включает в себя составляющие элементы [2], которые показаны на Рис. 1.

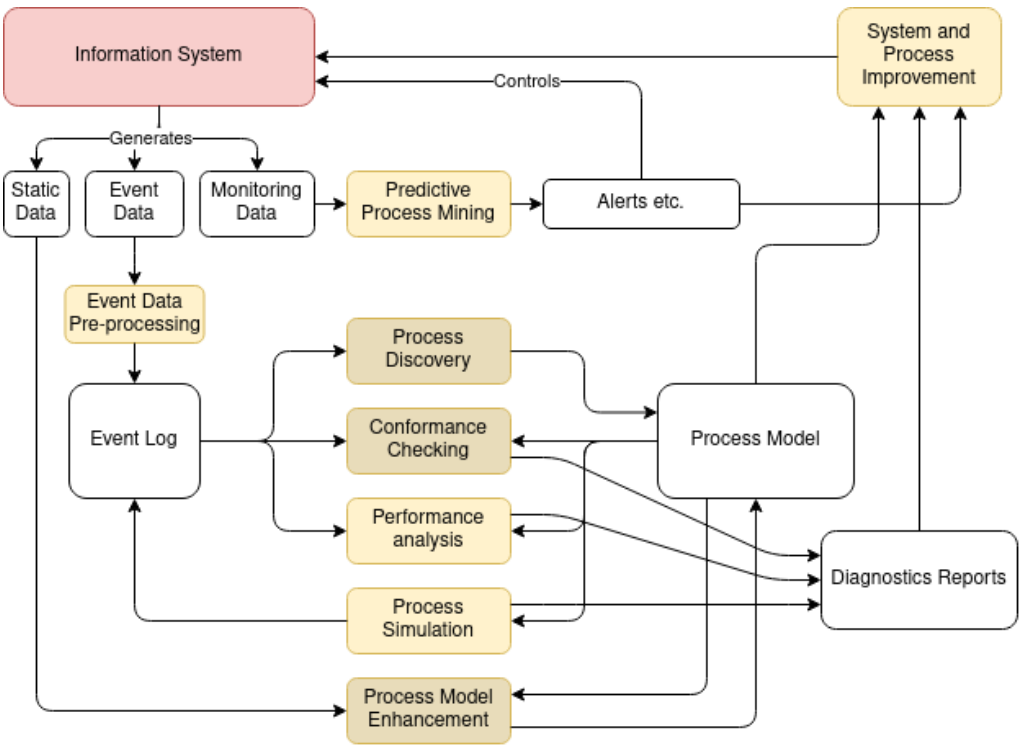

Fig. 1. Process mining

Рис. 1. Интеллектуальный анализ процессов

К ключевым задачам интеллектуального анализа процессов относят:

- автоматический синтез модели процесса по журналу событий (process discovery);
- оценку соответствия модели процесса и реального поведения процесса журнала событий информационной системы (conformance checking) [3];
- обогащение модели процесса дополнительными данными (process model enhancement).

Методы интеллектуального анализа и оптимизации процессов применяются в страховании [4], в программной инженерии [5], в электронной торговле [6], в промышленности [7—9], в медицине [10—12] и многих других областях человеческой деятельности [13].

Кроме вышеназванных основных задач, к ключевым следует отнести задачу предварительной обработки данных для получения журналов событий, с которыми могут работать алгоритмы анализа процессов. Большинство информационных систем всё ещё не настроены для генерации готовых журналов событий требуемого формата. В частности, именно так обстоит дело с данными,

получаемыми из электронных образовательных сред. В данной работе подготовке данных уделено особое внимание в разделе 4.2.

По мере внедрения методов анализа процессов в коммерческую деятельность многих компаний в дополнение к основным классическим составляющим дисциплины всё более активно развиваются новые направления, связанные с анализом производительности систем и предсказательными алгоритмами, которые позволяют анализировать данные о работе системы, что называется, «на лету» с выдачей предупреждений и предсказаний [14—16]. Не менее значимым является «А что, если ... ?» анализ, базирующийся на использовании методов симуляции моделей процессов [17—19].

Получаемые модели процессов, а также диагностические данные должны использоваться для настройки и усовершенствования работы информационной системы, которая поддерживает тот или иной процесс. Например, это может быть образовательный процесс и система управления образованием. Заметим, что усовершенствованию могут подвергаться как технические инструменты, так и сам поддерживаемый процесс, которые в действительности в процессно-ориентированных системах неразрывны [20].

Отметим, что большинство методов интеллектуального анализа процессов направлены не на анализ конкретных кейсов, а дают обобщённую картину того, как действуют и как взаимодействуют участники процесса. Это позволяет анализировать процесс в целом и соблюдать требования приватности данных.

1.2. Основные понятия

Анализ процесса начинается с журнала событий, который содержит записи о поведении информационной системы, поддерживающей этот процесс. Журналы событий разных систем могут выглядеть по-разному. В последние годы ассоциацией IEEE был разработан и стандартизирован общий формат журналов событий XES [21], который поддерживается всеми основными инструментами анализа процессов. В данной работе мы используем этот стандарт, впрочем, не все его возможности.

Процесс выполняется в соответствии с некоторой унифицированной процедурой. Выполнение этой процедуры для некоторого конкретного случая называется экземпляром процесса или кейсом. Каждый экземпляр имеет идентификатор. В случае образовательного процесса, экземпляр процесса обычно соответствует студенту.

Процесс определяется как множество событий. Каждое событие представляет собой запись (кортеж), включающий идентификатор экземпляра (case ID), отметку о времени события (timestamp), действие (activity) и набор дополнительных полей с необходимой информацией о событии. В случае образовательных процессов типом события может быть «выбор курса», «сдача экзамена», «переход на индивидуальный план» и прочее. Выделение действий и дополнительных полей обычно зависит от особенностей процесса и его аспектов, которые мы хотим анализировать. Например, дополнительные поля могут содержать информацию об исполнителе действия, используемых ресурсах, и пр.

Последовательность событий, относящихся к одному конкретному экземпляру процесса и упорядоченных в соответствии с временными отметками, называется трассой. При построении модели процесса важно разделение событий на трассы и их упорядочение в каждой трассе. Поэтому поля идентификатора экземпляра, временной отметки и поля с дополнительной информацией на этом этапе могут быть отброшены. Тогда трасса есть конечная последовательность действий. В свою очередь, упрощенный лог представляет собой мультимножество трасс, поскольку возможно, что трассы нескольких экземпляров процессов совпадают. После построения модели процесса опущенная информация может быть вновь привязана к элементам модели и использована для анализа процесса.

Таблица 1 содержит пример простого журнала событий L_{ex}. Трассы этого журнала определяются экземплярами 1, 2 и 3. Мультимножество трасс журнала L_{ex} есть $[\langle a, b, c, d\rangle^1, \langle a, c, b, d\rangle^1, \langle a, b, c, a, c, b, d\rangle^1]$, где верхний индекс 1 означает, что соответствующая трасса входит в мультимножество с кратностью 1.

Table 1. Simple event log **Таблица 1.** Простейший журнал событий

Case	Activity	Time	Actor
1	a	2022-06-29T10:00+03:00	Ivan
1	b	2022-06-29T10:02+03:00	Maria
2	a	2022-06-29T10:20+03:00	Ivan
1	c	2022-06-29T10:21+03:00	Elena
2	c	2022-06-29T10:25+03:00	Elena
1	d	2022-06-29T10:30+03:00	Fyodor
3	a	2022-06-29T10:31+03:00	Ivan
2	b	2022-06-29T10:36+03:00	Maria
2	d	2022-06-29T10:40+03:00	Fyodor
3	b	2022-06-29T10:42+03:00	Maria
3	c	2022-06-29T10:45+03:00	Elena
3	a	2022-06-29T10:46+03:00	Ivan
3	c	2022-06-29T10:48+03:00	Elena
3	b	2022-06-29T10:51+03:00	Fyodor
3	d	2022-06-29T10:55+03:00	Fyodor

В области process mining для моделирования процессов применяются различные формализмы и нотации: сети Петри, BPMN-модели, системы переходов и др. В этой работе мы используем графы частотного следования (dependency-frequency graphs).

Граф частотного следования — это ориентированный граф, в котором вершины представлены действиями и помечены их весом в журнале событий, а дуги обозначают причинно-частотную зависимость между соответствующими действиями. Дуги также помечены. Пометки на дугах в таком графе описывают силу причинно-частотной зависимости, которая вычисляется с помощью некоторого эвристического алгоритма.

На Рис. 2 показан пример графа частотного следования, построенного на основе журнала событий L_{ex} с помощью эвристического алгоритма [22] с настройками по умолчанию. Вершина a в этой модели помечена числом 4, так как действие a встречается в журнале событий 4 раза. Другие вершины помечены аналогичным образом.

Заметим, что пометки на дугах могут отображать силу причинно-следственной зависимости в разном виде. Это может быть либо частотность соответствующего отношения, либо мера относительной надёжности связи в диапазоне от 0 до 1, которая показывает степень уверенности относительно наличия причинно-следственной связи между соответствующими действиями (определение метрики см. в работе [22]) . Например, модель на Рис. 2 показывает, что действия a и b могут быть связаны причинно-следственным отношением (a является причиной b), но степень уверенности в этом не высока (0.667).

Чем больше примеров следования одного действия за другим имеется в журнале событий, тем больше будет степень уверенности в наличии отношения причинно-следственной зависимости между ними. Например, добавим в журнал событий дополнительные трассы и получим журнал L'_{ex}, содержащий такое поведение: $[\langle a, b, c, d\rangle^7, \langle a, c, b, d\rangle^1, \langle a, b, c, a, c, b, d\rangle^1]$. Тогда эвристический алгоритм со стандартными настройками выдаст модель, показанную на Рис. 3. В новой модели

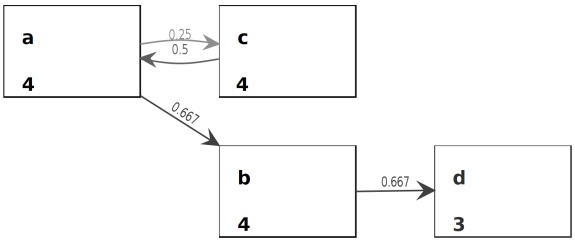

Fig. 2. Dependency-frequency graph for L_{ex} event log

Рис. 2. Граф частотного следования для журнала событий L_{ex}

степень уверенности в наличии причинно-следственной связи между действиями a и b повысилась до 0.889.

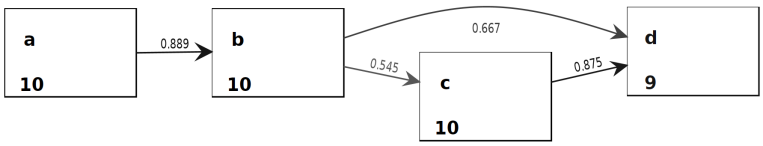

Fig. 3. Dependency-frequency graph for L'_{ex} event log

Рис. 3. Граф частотного следования для журнала событий L'_{ex}

Модель процесса не всегда точно отражает реально исполняемый процесс. Алгоритмы синтеза абстрагируются от некоторых деталей процесса с целью улучшения читаемости модели, выделения ключевых аспектов процесса и т.д. Важно уметь проверить, что модель соответствует журналу событий и в этом смысле верна.

Для оценки того, насколько хорошо данная модель процесса отражает реальное поведение системы, представленное в журнале событий, используются специальные метрики [3]. Наиболее важными являются метрики *соответствия* (fitness) между моделью и журналом событий и *точности* (precision) модели по отношению к журналу событий.

Метрика соответствия показывает, насколько хорошо модель позволяет воспроизвести трассы журнала событий. Степень соответствия измеряется числом в интервале от 0 до 1. В том случае когда модель позволяет исполнить в точности все трассы, записанные в журнале событий, т.е. степень соответствия равна 1, говорят об идеальном соответствии между моделью и журналом событий.

Как правило, модель может допускать поведение, не представленное в журнале событий, поскольку она нужна, в частности, для обобщения информации, содержащейся в журнале событий. Однако слишком общая модель может допускать много «лишнего» поведения. Для оценки такого «лишнего» поведения, допускаемого моделью, используется метрика точности. Точность модели также измеряется числом в интервале от 0 до 1. При точности 1 модель допускает только поведение, представленное в журнале событий. При точности 0 модель допускает любое поведение.

С практической точки зрения хорошими считаются модели, для которых их соответствие журналу событий близко или даже равно 1, но точность не слишком высока.

Отметим, что граф частотного следования отображает зависимости между событиями, но не всегда явно представляет допускаемые трассы. Например, графовая модель на Рис. 2 не имеет пути, соответствующего трассе $\langle a, b, c, d \rangle$ из журнала событий, на основе которого она построена. Чтобы было удобнее «накладывать» трассы на модель для вычисления метрик соответствия и точности, причинно-частотные модели переводят в эквивалентные сети Петри. Для сетей Петри разработаны различные алгоритмы вычисления этих метрик [23—26]. Далее в разделе 5.1 мы используем такой двух-этапный способ вычисления метрик соответствия и точности для графов частотного следования относительно журналов событий.

1.3. Программные инструменты

В области интеллектуального анализа процессов имеются как свободные, так и коммерческие программные инструменты, реализующие упомянутые выше алгоритмы анализа и синтеза моделей процессов.

Графические изображения моделей процессов, показанные на Рис. 2 и Рис. 3, получены с помощью реализации эвристического алгоритма синтеза в среде ProM Tools Framework [27]. Это свободный фреймворк для реализации алгоритмов анализа процессов, функционирующий на базе JVM, который появился раньше всех среди существующих на данный момент инструментов. Система ProM[1] разрабатывается открытым для участия исследовательским сообществом и доступна каждому. Именно ProM содержит реализации наибольшего количества (более 600) разных алгоритмов синтеза, анализа и усовершенствования моделей процессов.

Другой популярный свободный инструмент для анализа процессов — библиотека PM4Py[2]. Это относительно новая библиотека включает реализации базовых алгоритмов синтеза моделей процессов (среди прочих, например, и эвристического алгоритма) и их анализа на языке программирования Python [28]. Данная библиотека также разрабатывается исследовательским сообществом и довольно активно развивается. Что особенно важно, сообществом прилагаются существенные усилия для повышения эффективности работы реализаций алгоритмов, как с использованием чисто технологических подходов, например, путём исполнения задач на GPU [29], за счёт эффективного использования СУБД [30], так и на основе применения более эффективных алгоритмов [23, 31].

Существует также большое количество коммерческих инструментов, предназначенных для анализа и синтеза моделей процессов. В данной работе они не используются, а потому не рассматриваются.

2. Анализ процессов в образовании: обзор литературы

Обсудим некоторые работы, опубликованные в области анализа образовательных процессов.

Первым делом стоит упомянуть большой обзор, опубликованный в 2018 году [32]. В этом обзоре составлен каталог различных методов, которые могут применяться в задачах, характерных для образовательных процессов, а также рассматривается большое число применений. Всё это сделано на базе изучения существенного количества публикаций. Из доклада, в частности, следует, что одними из самых популярных алгоритмов синтеза в данной области являются нечёткий (Fuzzy miner) и эвристический (Heuristics miner) алгоритмы, а самая популярная методика проверки соответствия модельного и наблюдаемого поведения — проигрывание журнала событий на модели (token-based replay).

Исходные данные, по которым синтезируются модели, обычно получаются из разнообразных электронных образовательных систем [33—35].

Существенное внимание в области уделяется анализу образовательных траекторий с целью выявления и последующего устранения неудачных вариантов организации образовательного процесса, которые ведут к нежелательным результатам [33, 35—37]. Симптомами такой нежелательной организации могут быть, например, общая низкая академическая успеваемость или большое число отчисляющихся по разным причинам студентов. Так как область исследования образовательных процессов достаточно молода, многие работы последних лет посвящены сравнению эффективности различных алгоритмов автоматического синтеза моделей в применении к задачам анализа образовательных процессов [34, 35]. Чаще всего рассматриваются нечёткий или эвристический алгоритмы синтеза, но иногда исследователи включают в рассмотрение и индуктивный алгоритм. Некоторое исследователи [38] предлагают новые способы интерпретации классических моделей для

[1] http://promtools.org
[2] https://pm4py.fit.fraunhofer.de

применения в задачах анализа и моделирования образования или предлагают специализированные модели процессов [39].

В [38] предложена довольно любопытная модель. Набор академических задолженностей студента сравнивается с «рюкзаком», наполненным камнями, который, очевидно, затрудняет движение. По мере прохождения студентом тех или других контрольных мероприятий по одной из дисциплин рюкзак может наполняться новыми задолженностями или опустошаться. События изменения его наполнения фиксируются в журнале. Таким образом, становится возможным синтезировать разные траектории (представляются графами непосредственного следования), демонстрирующие типичные жизненные циклы рюкзаков студентов в ходе их обучения. Данная работа оказала существенное влияние на наше исследование, которое развивает идею построения траекторий на основе информации об академической успеваемости.

В некоторых работах рассматривается не организация процесса в целом, но поведение студентов в рамках конкретной составляющей образовательного процесса. Например, это может быть анализ выполнения студентами тестов в образовательной системе [40], или анализ поведения студентов в совместной работе над командным заданием [41], или выявление типичных шаблонов предоставления обратной связи в процессе обучения [42]. Авторы показывают, как на основе моделей, синтезированных с использованием нечёткого алгоритма синтеза, могут быть выявлены типичные варианты прохождения тестов студентами, а также отклоняющиеся от обычного порядка варианты поведения, которые потенциально могут приводить и приводят к провалу теста.

Характерные фрагменты поведения студентов в образовательной системе могут рассматриваться как шаблоны [43]. Поиск нежелательных (или, наоборот, желательных) шаблонов в типичных траекториях студентов позволяет получить более полное понимание того, как устроен образовательный процесс, а также выявить недостатки организации образовательной программы [44].

В некоторых случаях делаются даже попытки формирования индивидуальных рекомендаций по выбору наиболее актуальных для студента образовательных ресурсов (лекций, задач для самопроверки и т.д.) в рамках дисциплин, построенных по принципам самоуправляемого обучения [45]. При этом синтезированные модели процессов используются вместе с результатами анализа статистических данных о пользователях и позволяют выявлять характерные стратегии студентов и строить рекомендации на их базе. Авторы на практических данных демонстрируют, что построение рекомендаций с учётом в том числе и моделей процессов, делает предложения богаче и потенциально уменьшает риск неудачи с тестом и отчисления студентов. Когда речь идёт об обучении сложным навыкам, например, на курсах по моделированию, автоматически синтезированные модели могут повысить и качество преподавательских советов студенту, обогатить обратную связь, которая даётся по результатам выполнения заданий [46].

В работе [34] сделан вывод, что индуктивный алгоритм синтеза хорошо подходит для синтеза моделей образовательных процессов и дает более содержательные результаты по сравнению с Альфа-алгоритмом, эвристическим алгоритмом и алгоритмом синтеза эволюционных деревьев. Важным для нас выводом авторов этой работы является заключение о том, что при анализе образовательных процессов совершенно невозможно обойтись без предварительной обработки сырых данных. Впрочем, эта мысль является общей для process mining [2].

3. Постановка задачи

Основная цель нашей работы — сформулировать и описать подход для выявления ключевых характеристик образовательной программы, влияющих на успешное завершение студентом учебного года, на основании журналов событий, содержащих данные с результатами отдельных экзаменов.

Как конкретный пример мы анализируем данные одной из образовательных программ бакалавриата НИУ «Высшая школа экономики» за 2020/2021 год. Учебный год разделен на 4 модуля. Каждая дисциплина завершается экзаменом в одном из модулей. Оценка за экзамен ставится по

десятибалльной шкале, при этом оценка ниже 4 баллов считается неудовлетворительной. В случае получения неудовлетворительной оценки у студента возникает академическая задолженность по данной дисциплине и он имеет право на две пересдачи. Если у студента возникли три академические задолженности или если он не устранил академическую задолженность по итогу двух пересдач, то студент отчисляется с образовательной программы или ему предлагается индивидуальный учебный план для повторного изучения дисциплины. Существуют также и другие причины, по которым студент может не завершить учебный год успешно. Среди таких причин, например, уход в академический отпуск, перевод на другую образовательную программу и другие.

В нашей работе для каждого студента экзамен (или переэкзаменовка) может быть представлен двумя различными типами событий. Или студент «сдал экзамен» (тип события pass), т.е. получил оценку 4 балла или выше, или студент «не сдал экзамен» (тип события fail), т.е. получил оценку ниже 4 баллов. В случае неявки студента определяется была ли эта неявка по уважительной причине или нет.

Журнал событий, сформированный на основе данных из информационной системы университета, содержит следующую информацию:

1. Идентификатор (ID) студента;
2. Тип события;
3. Дата события.

Событиями являются, например, успешная или нет сдача конкретного экзамена (или переэкзаменовка), запись о пропуске экзамена, запись о неуспешном завершении учебного года по той или иной причине (больше деталей приводится в разделе 4).

Построенная на основании полученного журнала событий модель может быть использована для ответа на следующие вопросы:

1. Как выглядят образовательные траектории успешных студентов, и как устроены отклонения от данных траекторий?
2. Какие последовательности событий в наибольшей степени могут привести к тому, что студент не сможет успешно завершить курс? В результате будут выявлены наиболее «проблемные» для студентов дисциплины.
3. Существуют ли некорректные последовательности событий? Их наличие означает возможные ошибки в информационной системе университета или организации учебного процесса, которые необходимо устранить.

Предлагаемая модель может быть инструментом для руководителей образовательных программ и университетской администрации для анализа образовательных программ в целях их дальнейшего совершенствования и развития. Отдельно отметим, что большинство известных авторам университетских информационных систем в той или иной форме содержат информацию, используемую в данном исследовании, что придаёт нашему подходу универсальность.

4. Анализ академической успеваемости методами process mining: подход к решению задачи

Рассмотрим теперь в подробностях подход, который мы предлагаем для анализа данных об академической успеваемости и поиска ответов на вопросы, сформулированные в разделе 3. На Рис. 4 представлена общая схема подхода, предлагаемого в данной работе.

Сырые данные из разных источников предварительно обрабатываются, что даёт исходный набор данных. Описание используемых данных и метода их предварительной обработки содержится в разделе 4.1. Затем с помощью специально разработанных программных инструментов из сырого набора данных конструируется журнал событий (детали приводятся в разделе 4.2). Модели процессов синтезируются на основе журнала событий и, при необходимости, обогащаются дополнительной информацией, содержащейся в исходном наборе данных, но не принятой во внимание в ходе

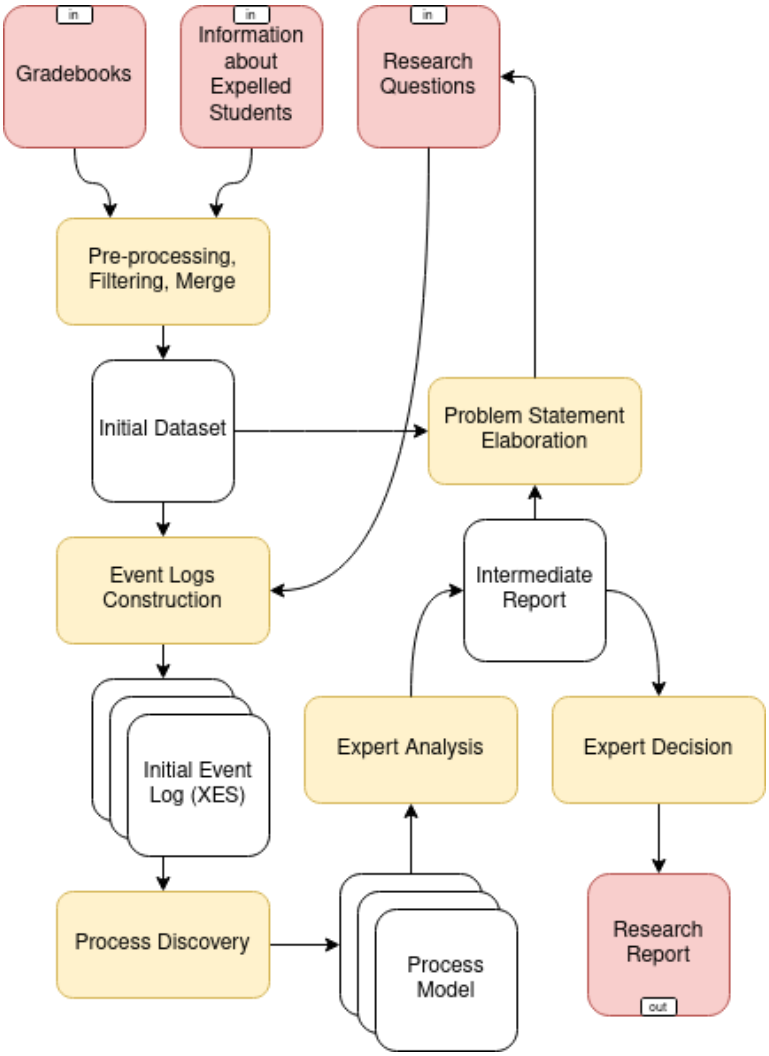

Fig. 4. General scheme for analysis of academic data

Рис. 4. Общая схема анализа данных об академической успеваемости студентов

конструирования первичного журнала событий (см. раздел 4.3). Наконец, получившиеся модели используются для формулировки ответов на поставленные в разделе 3 вопросы. Заметим тот важный факт, что сама формулировка задачи может изменяться, уточняться и детализироваться в ходе применения указанного подхода. Заметим также, что используемый подход в целом согласуется с описанным в литературе подходом PM^2 [47] для проведения проектов по анализу процессов.

4.1. Исходные данные

Как уже было отмечено, данные для данного исследования получены из внутренней электронной образовательной среды НИУ ВШЭ. Вся личная информация была анонимизирована, и студентам был присвоен уникальный идентификатор с помощью хэш-функции Python.

Для объединения и предобработки таблиц со студенческими оценками и экзаменационными ведомостями применена библиотека Pandas[3]. Вся несущественная для нашего исследования информация (например, информация о факультете или типе дисциплины) при этом была исключена из рассмотрения.

[3] https://pandas.pydata.org/

В результате получена таблица с таким набором *исходных данных*: [уникальный идентификатор студента]; [оценка по пятибалльной шкале]; [оценка по десятибалльной шкале]; [отметка присутствия на экзамене]; [дата экзамена]; [номер экзаменационной ведомости]; [тип записи (экзамен, переэкзаменовка, пересдача с комиссией)]; [учебный год]; [модуль (1–4)]; [название дисциплины]; [подразделение (кафедра), поддерживающее дисциплину]; [имя заполнившего ведомость профессора].

В том случае, если поле оценки по какой-то причине не заполнено, оно заполняется «0». В исходных данных даты заполнения ведомостей и проведения экзаменов были в разных форматах. Все они приведены в единый формат. Некоторые записи не содержали важной информации. Такие записи полностью удалены из набора данных. Обработанная таблица использована для конструирования журнала событий. Данный процесс описан далее в разделе 4.2.

4.2. Подготовка журнала событий

Рассмотрим теперь подробно процесс конструирования журнала событий из подготовленных данных образовательной информационной системы.

Для применения методов process mining нам необходимо преобразовать полученный массив данных из образовательной системы в более компактный журнал событий с унифицированным форматом данных. Для этого прежде всего определим формат и структуру журнала событий. Как уже было отмечено в разделе 1, каждая запись о событии представляет собой кортеж. Для определения события в составе записи обязательно должны быть представлены поля с идентификатором экземпляра процесса (case ID), отметкой о времени события (timestamp), а также указан тип действия (activity). При необходимости в записи также может содержаться набор дополнительных полей с другой информацией о событии. В качестве идентификатора экземпляра процесса мы пользуемся уникальным идентификатором (ID) студента. Отметкой о времени события послужит дата экзамена или пересдачи. Поле, фиксирующее тип действия, может быть заполнено разными способами, выбор каждого из которых определяется вопросами, для ответа на которые производится анализ процесса. Данное поле должно содержать всю информацию о действии, вызвавшем событие, которая может понадобится в ходе анализа процесса. В то же время необходимо избегать чрезмерного усложнения.

В данной работе поле тип действия для журнала событий формировалось экспериментальным путем. Отправной точкой стал вариант журнала событий, представленный в работе [38], где в качестве действий рассматриваются текущие задолженности студента. Однако при проведении первых экспериментов было выявлено множество недостатков такого формата журнала событий в контексте нашей задачи. В отличии от образовательной системы, рассматриваемой в [38], правила анализируемой в данной работе образовательной среды предполагают наличие ограничений на максимальное число задолженностей и сроки их устранения. Что ещё более важно, студенты в нашем случае имеют возможность пересдавать экзамены. Также в течение одного учебного года может быть несколько экзаменов по учебному курсу. Дальнейшие эксперименты привели к выбору другого варианта заполнения поля типа действия. Фокус сместился с задолженностей непосредственно на академическую траекторию студентов. Отслеживая экзамены и пересдачи в течение учебного года, мы получили модель, отражающую полную картину событий в ходе года. Полученная из полученного лога модель схожа с вариантами, используемыми в работах [40, 45], и позволяет выявлять шаблоны и взаимосвязи действий в траекториях студентов.

В конечном итоге мы определили общую структуру для поля типа действия так:

- номер учебного модуля (1-4);
- название учебного курса;
- тип события (экзамен, пересдача, пересдача с комиссией);
- результат действия (pass — успех, fail — неудача);

• (необязательный атрибут) причина отсутствия (уважительная, неуважительная).

Приведём несколько примеров заполнения поля тип действия у записей с использованием такой структуры:

• «4 module Algebra exam fail nonvalid miss»;
• «2 module Algebra retake pass»;
• «4 module Programing exam pass».

В ходе дальнейших экспериментов была выявлена необходимость отслеживать итоговое состояние студента, чтобы формулировать выводы об успешности траектории. Студенты, которые были отчислены по тем или иным причинам, взяли академический отпуск или перевелись на другую специальность, будут иметь соответствующее событие с указанным действием. В том случае, если студент успешно завершил учебный год и продолжит обучение, в конце трассы будет добавлено событие с действием «pass». Благодаря наличию таких событий можно также отследить студентов, которые были восстановлены в данном учебном году. Эти студенты будут иметь событие с отчислением в начале трассы.

Наконец, из собранных данных на основе приведенной выше структуры полей и столбцов был сконструирован журнал событий. Для этого нами был разработан алгоритм, использующий средства библиотеки Pandas. Алгоритм добавляет в журнал события для каждого студента в указанном формате преобразовывая данные из таблицы с данными. Если студент присутствует в таблице отчисленных, то будет добавлено событие с соответствующим типом действия и указанной временной меткой события. В противном случае в трассу, соответствующую студенту, добавляется событие с действием «pass» с временной отметкой конца учебного года.

Так как мы не собираемся строить модель для всех студентов сразу, наиболее рациональным решением будет разбить журнал на части. В нашем случае данные содержат информацию за один учебный год. Мы можем разделить студентов по курсам и получить 4 журнала, каждый из которых соответствует конкретному курсу. При наличии наборов данных за больший промежуток времени появляются возможности иного разделения студентов. Например, можно сгруппировать траектории студентов по году начала обучения, что позволяет проследить за группами студентов на протяжении нескольких курсов.

В журналах событий, которые мы получаем в данной работе, присутствует большое число различных событий: успешные или неудачные экзамены, переэкзаменовка, переэкзаменовка с комиссией, уход в академический отпуск по собственному желанию или по медицинским основаниям, отчисление в связи с академической задолженностью и т. п. Такая подробная детализация позволяет более точно моделировать траекторию студента в учебном процессе, но в то же время делает модель сложнее для восприятия. Поэтому одновременно мы также анализируем два упрощённых журнала событий.

В итоге мы получаем и анализируем далее три варианта построения журнала событий.

Базовый вариант конструирования предполагает учёт всех имеющихся событий без объединения различных событий. Такой подход особенно полезен для выявления различного рода аномалий в данных. Если в учетной системе образовательной организации произошла ошибка (например, была ошибочно указана дата экзамена так, что студент как будто бы сдавал экзамены уже после своего отчисления), модель процесса покажет соответствующий переход от отчисления к сдаче экзамена, что невозможно в реальности. Все подобные аномалии могут быть выявлены, проанализированы и, при необходимости, устранены.

Первый вариант модификации журнала событий предполагает объединение всех событий, связанных с конкретной учебной дисциплиной, в две группы: дисциплина успешно закрыта студентом (pass) и дисциплина не закрыта (fail). К первому виду событий относятся успешная сдача экзамена по дисциплине с первого раза, а также успешные пересдачи. Ко второму виду относятся все

события, связанные как с провалом экзамена по дисциплине, так и с провалом переэкзаменовок. В отличии от базового подхода в данном случае возможен повтор события с провалом дисциплины. Это означает появление циклов в модели. События, описывающие завершение обучения на данном курсе, также относятся к двум группам: успешное окончание (pass) и его отсутствие (non-pass). К первой группе относятся варианты успешного завершения учебного года: перевод на следующий курс или успешное завершение обучения. Ко второй группе относятся все прочие события: уход в академический отпуск, отчисление по собственному желанию, отчисление в связи с недобросовестным освоением образовательной программы и т. п. В рамках данного подхода группируются редкие события. Благодаря этому выделяются общие тенденции, а траектории студентов представляются в более общем виде без лишней детализации.

Второй вариант модификации журнала событий предполагает группировку всех событий, связанных с конкретной дисциплиной, такого вида: или дисциплина сдана с первого раза (pass), или с дисциплиной возникли проблемы проблемы (fail). Отличие данного подхода от предыдущего заключается в том, что любая пересдача (даже успешная) или пропуск экзамена (даже по уважительной причине) рассматриваются как проблемные ситуации. Заметим, что такие ситуации являются проблемными не только для студента, но и для учебной администрации, для которой усложняется организация образовательного процесса. Такой вариант журнала событий позволяет выделить наиболее проблемные области в образовательных траекториях.

Второй и третий варианты модификации исходного журнала событий представляют собой только некоторые возможные способы модификации. Для получения наиболее полной картины процесса рекомендуется использовать наиболее полный журнал событий, так как при модификации часть важных деталей будет утеряна. Модификация журнала уместна в тех случаях, когда необходимо рассмотреть определенный аспект образовательного процесса, а также когда детали не несут существенной ценности или когда алгоритм построения модели не способен качественно обработать зашумленные данные.

4.3. Построение моделей образовательного процесса

Рассмотрим теперь, какие методы используются в данной работе для построения моделей образовательного процесса.

Мы воспользуемся эвристическим алгоритмом синтеза (Heuristics miner) [48] в той его версии, что реализована в библиотеке PM4Py. Данный алгоритм позволяет получить модели в виде графа частотного следования, а также в виде сети Петри, что позволяет произвести расчеты соответствия (fitness) и точности (precision) модели по отношению к заданному журналу событий. Более того, это можно сделать разными методами: путём проигрывания журнала событий на модели (token-based replay) [23, 24] или с использованием выравниваний (aligment-based) [25, 26]. Мы используем в работе оба этих метода для повышения достоверности результатов.

Алгоритм синтеза имеет несколько входных параметров, которые влияют на вид получаемой модели. Среди прочего данные параметры позволяют задать пороговые значения для учёта событий и связей между ними в зависимости от их частоты. Это позволяет отсекать менее значимые события и связи в модели, что сделано для уменьшения влияния шума в данных на получаемую модель. Кроме того, восприятие и интерпретация модели, построенной без фильтрации редких вариантов поведения, крайне затруднительны. Пример подобной модели приведен на Рис. 5.

В нашей работе мы рассматриваем процесс без параллелизма. В связи с этим максимальное значение устанавливается для параметра эвристического алгоритма, отвечающего за уровень отсечения параллельных событий. Это обеспечивает полное исключение возможных параллельных событий в генерируемой модели, что является артефактом синтеза, а не реальным феноменом анализируемого процесса.

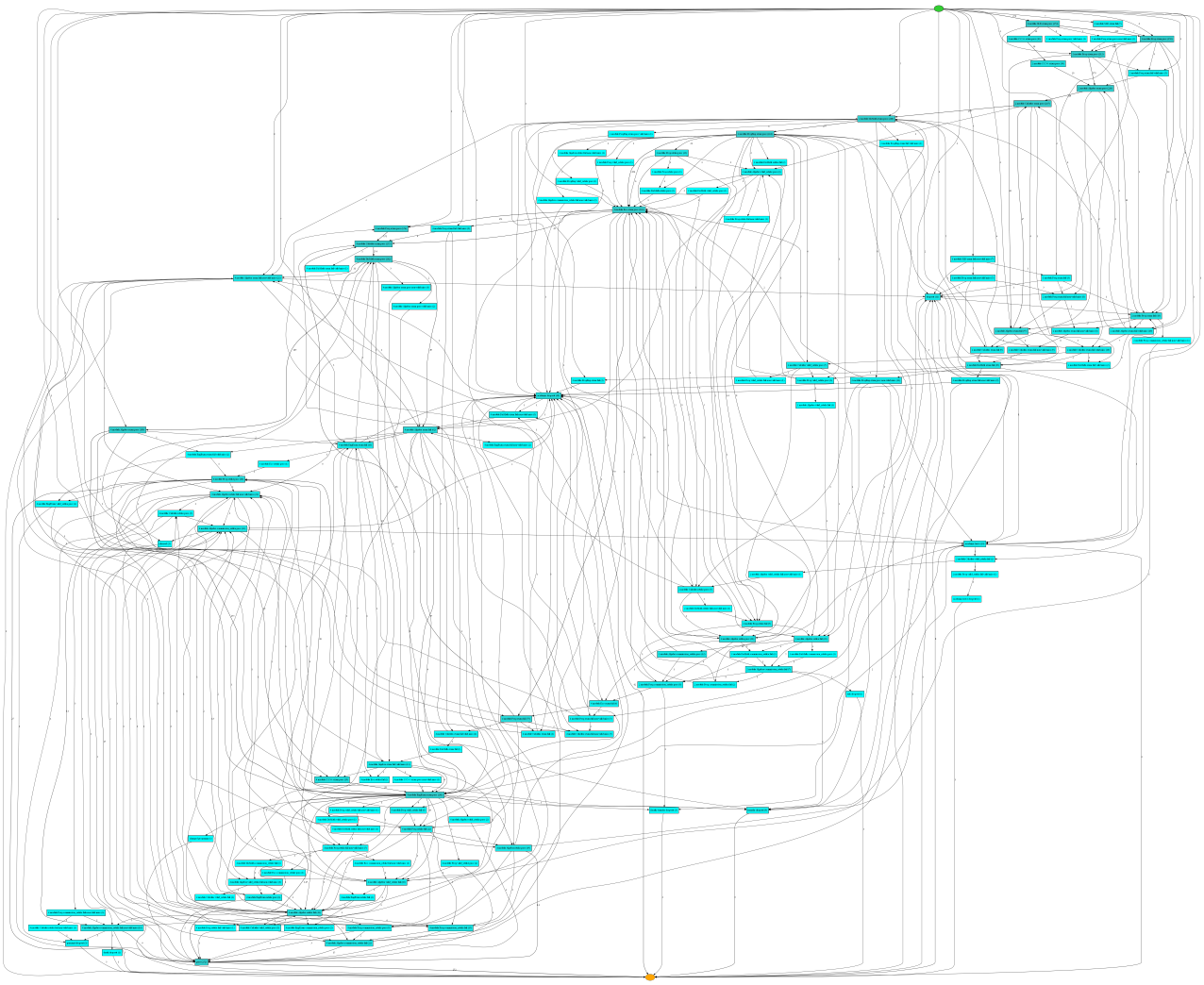

Fig. 5. Process model example without specifying threshold

Рис. 5. Пример модели процесса без использования пороговых значений

Возможность изменять входные параметры алгоритма позволяет выделить наиболее важные аспекты в модели. Для того чтобы изучить конкретные аспекты процесса более детально, можно прибегнуть также к фильтрации журнала событий. При помощи фильтров можно оставить в журнале событий только определенные трассы, соответствующие заданному критерию. Таким образом можно, например, получить модель, которая содержит только трассы студентов, которые провалили определенный экзамен.

Для поиска отклонений в учебных траекториях студентов и характерных закономерностей мы используем следующий подход к построению моделей.

На *первом этапе* анализируются модели, синтезированные из журнала событий. При этом применяются только различные параметры алгоритма синтеза. Обозначим подобные модели как *полные*. Путем задания пороговых значений для алгоритма можно получить достаточно компактную модель, выделяющую наиболее важные части всех траекторий.

Например, по журналу событий для первого курса на подобной модели можно выделить наиболее проблемные учебные курсы. На фрагменте модели, показанном на Рис. 6, можно выделить провалы экзаменов по программированию и алгебре во 2 модуле. Связь между данными событиями провалов демонстрирует, что в половине подобных случаев присутствуют оба провала. Далее

для всех найденных подобных аномалий можно задействовать фильтрацию журнала событий и построить модели для анализа только траекторий, содержащих конкретную аномалию. При необходимости можно вновь установить пороговые значения алгоритма для получения более читаемых моделей.

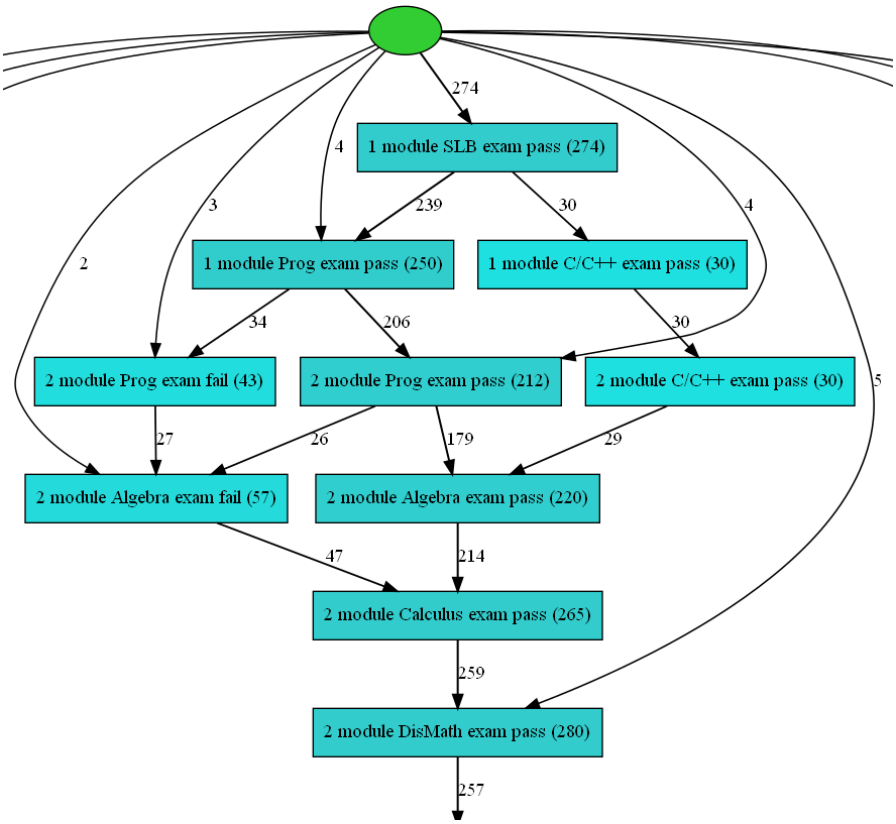

Fig. 6. Example fragment of process model for the 1-year students

Рис. 6. Пример фрагмента модели процесса для студентов 1 курса

Второй этап предполагает исключение (при помощи фильтров) из журнала событий «идеальных» траекторий студентов. Это траектории без отклонений, где все экзамены сданы с первого раза без пересдач. Обозначим данные модели как *модели отклонений*. Ввиду особенностей эвристического алгоритма, расчет относительной важности событий и переходов ведётся с учетом их частоты. Исключение траекторий части студентов из журнала событий позволяет выделить некоторые отклонения, которые в первом варианте не имели достаточного веса в модели, а потому были не видны.

Для уже упомянутого журнала событий для студентов первого курса полученная модель, содержащая только траектории с отклонениями, показала проблемы с некоторыми пересдачами экзаменов, которые отсутствовали в модели на первом этапе.

Как уже было сказано ранее, найденные аномалии можно дополнительно изучить после фильтрации журнала событий.

Третий этап развивает идею, использованную на втором этапе. В журнале событий, который содержит только траектории студентов с отклонениями, можно дополнительно исключить траектории с уникальными вариантами поведения. Таким образом, мы получим журнал событий, который содержит только повторяющиеся траектории отклонений. Такие модели обозначим как *модели повторных отклонений*. Модель, построенная по данному журналу событий, позволит рассмотреть

отклонения, которые встречались более одного раза и выделить области пересечений соответствующих траекторий.

На примере подобной модели для первого курса удалось обнаружить траекторию студентов, которые пропустили все экзамены во втором модуле и были отчислены за академическую неуспеваемость. Более того, в этой траектории один из экзаменов во втором модуле не был указан как пропущенный. При дополнительном рассмотрении выяснилось, что это ошибка в исходных данных, а действительные пропуски для данного курса не были отмечены в данных. Таким образом, различные подходы к анализу могут также помогать обнаруживать ошибки в данных. Как и на прошлых этапах, аномалии можно дополнительно изучить при помощи фильтрации журнала событий.

Используя данный подход из трех этапов можно получить полную картину траекторий студентов и изучить все возможные отклонения и аномалии. Результаты применения подхода к конкретным данным реальной образовательной программы рассмотрим в следующем разделе.

5. Анализ моделей образовательного процесса

Анализ будем проводить в соответствии с подходом, который описан в разделе 4.3. Для различения двух видов модификации журнала событий, описанных в разделе 4.2, обозначим первый вариант модификации журнала событий как вариант А, а второй вариант модификации — как вариант Б.

Данный раздел организован следующим образом. Первым делом в подразделе 5.1 приводится оценка синтезированных моделей с использованием классических для process mining метрик. Затем в подразделе 5.2 собраны различные наблюдения, сформулированные в ходе анализа синтезированных моделей образовательного процесса. Наконец, подраздел 5.3 содержит обобщённые результаты анализа и выводы.

5.1. Оценка синтезированных моделей процесса

Классические для process mining метрики соответствия и точности, упоминавшиеся ранее в разделе 1.2, позволяют выразить численно то, насколько модель процесса отражает наблюдаемую реальность, представленную журналом событий.

Результаты расчёта значений этих метрик для синтезированных моделей процессов приводятся в Таблице 2. При расчёте использованы два метода из библиотеки PM4Py — алгоритм, основанный на проигрывании журнала событий на модели (соответствующие результаты расчёта помечены в таблице с помощью латинской буквы «T»), а также алгоритм, использующий выравнивания (соответствующие результаты помечены буквой «A»).

В большинстве случаев оба алгоритма выдают схожий результат. Особенно это заметно для расчета точности модели. В наших экспериментах значение соответствия, посчитанное с использованием алгоритма, проигрывающего журнал событий на модели, в среднем выше, чем при использовании алгоритма, основанного на использовании выравниваний.

Для лога без модификаций характерно снижение значений метрик при переходе от полной модели к модели отклонений и увеличение при переходе от полной модели к модели повторных отклонений. Данное соотношение прослеживается на всех трех журналах событий. Пониженные значения метрик у модели отклонений можно объяснить большим числом уникальных трасс с отклонениями, которые тяжело выразить на одной общей модели без потерь точности и соответствия. В полной модели данные трассы составляют меньшую долю и оказывают меньше влияния при расчете метрик. В случае с моделями повторных отклонений причина повышенных значений может быть в малом числе кейсов у данных моделей.

Модифицированные логи показывают несколько иную картину. С точки зрения соответствия соотношения значений между полной моделью, моделью отклонений и моделью повторных

Table 2. Calculated fitness and precision values for discovered educational process models

Таблица 2. Рассчитанные значения соответствия и точности для синтезированных моделей образовательного процесса

Модель	Fitness T	Fitness A	Precision T	Precision A
Первый курс				
Полная модель	0.963	0.892	0.641	0.639
Модель отклонений	0.945	0.835	0.621	0.619
Модель повторных отклонений	0.979	0.959	0.746	0.746
Лог А. Полная модель	0.939	0.811	0.777	0.768
Лог А. Модель отклонений	0.888	0.635	0.786	0.780
Лог А. Модель повторных отклонений	0.966	0.902	0.620	0.611
Лог Б. Полная модель	0.941	0.851	0.794	0.794
Лог Б. Модель отклонений	0.897	0.715	0.767	0.767
Лог Б. Модель повторных отклонений	0.966	0.900	0.642	0.642
Второй курс				
Полная модель	0.989	0.967	0.617	0.616
Модель отклонений	0.982	0.940	0.579	0.577
Модель повторных отклонений	1.0	1.0	0.849	0.849
Лог А. Полная модель	0.946	0.855	0.872	0.867
Лог А. Модель отклонений	0.864	0.607	0.803	0.790
Лог А. Модель повторных отклонений	0.933	0.885	0.731	0.728
Лог Б. Полная модель	0.944	0.847	0.857	0.857
Лог Б. Модель отклонений	0.886	0.643	0.799	0.799
Лог Б. Модель повторных отклонений	0.977	0.935	0.692	0.692
Третий курс				
Полная модель	0.998	0.997	0.843	0.843
Модель отклонений	0.993	0.989	0.769	0.769
Модель повторных отклонений	1.0	1.0	0.930	0.930
Лог А. Полная модель	0.987	0.971	0.822	0.822
Лог А. Модель отклонений	0.935	0.846	0.646	0.646
Лог А. Модель повторных отклонений	1.0	1.0	0.902	0.902
Лог Б. Полная модель	0.992	0.974	0.896	0.896
Лог Б. Модель отклонений	0.961	0.879	0.704	0.704
Лог Б. Модель повторных отклонений	1.0	1.0	0.854	0.854

отклонений в большинстве случаев идентичны соотношению у полного журнала событий. Однако точность у модели повторных отклонений чаще всего оказывается ниже, чем у полной модели и модели отклонений, что связано с наличием циклов на модели в отличии от полного журнала событий, где циклы отсутствуют. Единственное исключение — это модель третьего курса. Для этой модели значения ведут себя также, как и у полного журнала событий. Данное исключение можно объяснить крайне низким числом экземпляров процесса (студентов) и отклонений для модели повторных отклонений у третьего курса.

При сравнении значений для одного и того же типа моделей в случае полного журнала событий и его модификаций можно заметить, что значения соответствия для модифицированных журналов ниже, чем у полной версии журнала событий. Снижение значения соответствия можно объяснить всё тем же наличием циклов в моделях, синтезированных на основе модификаций журнала событий. Циклы значительно усложняют как построение качественной модели, так и сам расчёт метрик. Однако значение точности в моделях, синтезированных на основе модифицированных журналов

событий, в целом выше, чем у моделей на основе полного журнала событий, что объясняется меньшим разнообразием событий.

Рассмотрим также значения для моделей, полученных после изменения параметров алгоритма построения моделей. Единственный параметр алгоритма, который будет изменяться в ходе исследования, — это порог значительности для отношения взаимной зависимости между событиями (dependency threshold). Данный параметр устанавливает минимально необходимый уровень значительности для включения в модель того или иного отношения между событиями [48]. Уровень значительности определяет используемую эвристику алгоритма. Данный параметр позволяет уменьшить объём модели без потери наиболее важных событий. Рассмотрим влияние разных значений данного параметра на значение соответствия и точности модели по отношению к журналу событий на основе данных, приводимых в Таблице 3.

Table 3. Dependence of fitness and precision from algorithm threshold

Таблица 3. Зависимость значений метрик соответствия и точности от порогового значения алгоритма

Модель	Fitness T	Fitness A	Precision T	Precision A
Полная модель второго курса	0.989	0.967	0.617	0.616
Пороговое значение 0.2	0.989	0.965	0.606	0.606
Пороговое значение 0.4	0.985	0.958	0.607	0.606
Пороговое значение 0.6	0.968	0.915	0.377	0.379
Пороговое значение 0.8	0.936	0.861	0.312	0.320
Полная модель третьего курса	0.998	0.997	0.843	0.843
Пороговое значение 0.2	0.996	0.990	0.845	0.845
Пороговое значение 0.4	0.996	0.990	0.845	0.845
Пороговое значение 0.6	0.978	0.946	0.624	0.632
Пороговое значение 0.8	0.944	0.916	0.527	0.547

Исходя из полученных данных можно заключить, что данный параметр оказывает заметное влияние на значение соответствия модели только при сравнении крайних значений порогового значения. При высоком пороговом значении заметно некоторое снижение значения соответствия относительно модели, для которой пороговое значение равно нулю. С точки зрения точности модели параметр значительно уменьшается при пороговом значении выше 0.5.

Полностью обойтись без использования данного параметра возможно лишь в редких случаях, тогда как чаще модели реального образовательного процесса получаются крайне объемными. Использование данного параметра алгоритма позволяет значительно упростить внешний вид модели и, соответственно, подчеркнуть ключевые особенности процесса в ходе анализа.

5.2. Наблюдения об образовательном процессе, сформулированные в ходе анализа моделей процессов

Приведём теперь некоторые наблюдения, которые были сделаны в ходе анализа большого количества моделей в виде графов частотного следования, синтезированных на базе журналов событий реального образовательного процесса.

Первый курс. Полная модель.

С использованием параметра эвристического алгоритма можно значительно сократить объем полной модели журнала событий и получить модель, отображающую наиболее важную часть анализируемого процесса. Анализ полной модели даёт возможность выделить значимые события. Например, в ходе анализа данной модели можно выделить проблемы с экзаменами по алгебре и программированию во 2 и 4 модулях. При этом для экзамена по программированию в 1 модуле

практически не наблюдается провалов. Также алгоритм выделяет на модели провалы по пересдаче экзамена по программированию в 4 модуле. После определения точек интереса на полной модели можно сделать определенные гипотезы для дальнейшего анализа и продолжить более подробное изучение выявленных точек при помощи фильтрации журнала событий. Далее при помощи фильтров получим урезанные журналы событий, оставляя только студентов с конкретным событием в траектории, и построим на их основе модели.

Рассмотрим модель с траекториями студентов, у которых был провал экзамена по программированию во 2 модуле. Это позволит нам обнаружить взаимосвязь данного события с другими событиями процесса. Половина таких студентов также имеет провал экзамена по алгебре во 2 модуле. А четверть студентов имеют провалы пересдач по алгебре или программированию за 2 модуль. Однако провал пересдач по двум этим предметам одновременно встречается редко. Лишь четверть студентов успешно сдаёт экзамен по алгебре в 4 модуле. Такое же количество проваливает программирование в 4 модуле и пересдачу алгебры с комиссией за 4 модуль. Всего в итоге только 25% студентов успешно завершает учебный год. Оставшиеся отчисляются, переводятся или уходят в академический отпуск.

Изучая траектории с провалом экзамена по алгебре во 2 модуле мы видим схожую картину. Половина таких студентов провалило экзамен по программированию во 2 модуле. Треть студентов имеют проблемы с пересдачей алгебры за 2 модуль и провал экзамена по программированию в 4 модуле. Только шестая часть студентов успешно сдают экзамен по алгебре в 4 модуле. В итоге к успешному завершению года приходит только треть из таких студентов.

По траекториям с провалом экзамена по программированию в 4 модуле можно заметить, что большинство студентов успешно сдали экзамен по программированию и по алгебре во 2 модуле, а провалы обычно наблюдались в паре: сразу и по программированию, и по алгебре. Однако экзамен по алгебре в 4 модуле успешно сдали только 40% студентов. При этом две трети студентов успешно завершили учебный год.

Можно обнаружить, что половина студентов, имеющих провал экзамена по алгебре в 4 модуле, также имеет провал экзамена по программированию в 4 модуле. Кроме того, лишь половина студентов успешно сдали экзамены по алгебре и программированию во 2 модуле с первой попытки. В итоге 60% из данных студентов успешно завершили учебный год.

Последняя выявленная точка интереса — это провалы пересдачи экзамена по программированию за 4 модуль. В траекториях, которые включают данное событие, у половины студентов наблюдаются провалы во 2 модуле сразу и экзамена по алгебре, и экзамена по программированию. Подавляющее большинство имели проблемы с экзаменом по алгебре в 4 модуле. Половина студентов также не справились с пересдачей экзамена по программированию за 4 модуль с комиссией, а среди траекторий часто наблюдаются провалы пересдач по иным учебным курсам. Только половина таких студентов смогли успешно завершить учебный год.

Важно также отметить, что для первого курса предусмотрены два варианта учебного курса по программированию. Часть студентов вместо обычного курса программирования проходит углублённый курс программирования на C/C++. Все траектории студентов, выбравших данный курс, приходят в точку с успешным завершением учебного года.

Первый курс. Модель отклонений.

После исключения из журнала событий траекторий студентов без каких-либо отклонений на новой модели в зависимости от используемого алгоритма могут появится новые детали, которые не проявлялись на фоне более часто происходящих успешных событий. Данная модель полезна для анализа журналов событий с траекториями большого числа студентов и отклонений в процессе обучения.

Сравнивая обнаруженные отклонения и аномалии на модели можно заметить, что многие из них совпадают с полной моделью траекторий. Однако есть некоторые отличия. Так, например, можно заметить повышенное число проблемных пересдач по алгебре за 2 и 4 модуль в добавок к уже обнаруженным ранее проблемам с пересдачами по программированию за 4 модуль. Кроме того, во втором модуле наблюдаются провалы с дискретной математикой, которые к тому же имеют связь с пропусками других экзаменов.

Рассматривать уже рассмотренные отклонения при помощи фильтрации журнала событий не имеет смысла, так как получаемые модели не будут отличаться друг от друга. Поэтому обратим внимание на студентов, проваливших пересдачу по алгебре.

Студенты, провалившие пересдачу по алгебре во 2 модуле, в половине случаев имели провал экзамена по программированию за 2 модуль, а четверть из них провалила вторую пересдачу с комиссией за 2 модуль алгебры. К тому же эти студенты в половине случаев имели провал экзаменов по алгебре и программированию в 4 модуле. Только половина таких студентов успешно завершили учебный год.

Студенты, провалившие пересдачи по алгебре в 4 модуле, показывают схожую картину. Четверть студентов имели провал экзаменов по алгебре и программированию за 2 модуль, а половина от этого количества имели провал по двум этим курсам одновременно. Половина студентов имели провал экзамена по программированию в 4 модуле. Интересно, что только половина студентов с провалом пересдачи по алгебре в 4 модуле провалили экзамен по данному предмету. Вторая половина от проваливших пересдачу пропустили первый экзамен по уважительной или неуважительной причине. Половина студентов провалили и вторую пересдачу с комиссией за 4 модуль алгебры. Однако в итоге 75% данных студентов смогли успешно завершить учебный год.

Траектории, следующие через провал экзамена по дискретной математике во 2 модуле, в большинстве своём содержат множество провалов или пропусков. Ни один из следующих по данным траекториям студентов не закончил год успешно. Практически все траектории обрываются отчислением во втором модуле.

Первый курс. Модель повторных отклонений.

Эта модель даёт возможность рассмотреть наиболее важные отклонения, так как здесь присутствуют только те отклонения от «идеальной» траектории, которые повторялись многократно. Воспроизводимость отклонения обычно является показателем распространенности проблемы или наличия у проблемы какой-то фундаментальной причины. Рассматривая траектории в виде единой модели можно наглядно увидеть точки пересечения повторных отклонений и выявить наиболее проблемные места образовательного процесса.

Для первого курса на подобной модели видна обособленная траектория с пропусками экзаменов, которая оканчивается отчислением за академическую неуспеваемость после 2 модуля. Более детальное изучение данной траектории выявило ошибку в исходных данных, где пропуск экзамена по дискретной математике во 2 модуле был указан как провал. В остальных случаях повторяющиеся отклонения чаще всего связаны с провалом экзамена по программированию в 4 модуле, реже — с провалом экзаменов по алгебре во 2 и 4 модулях. Во всех случаях, за исключением траектории с пропусками, все студенты в представленной модели успешно завершили учебный год.

Первый курс. Варианты журнала событий А и Б.

Как было указано ранее, любой вариант модификации журнала событий будет скрывать часть информации по сравнению с полным вариантом журнала событий. По результатам анализа моделей, синтезированных на основе данных журналов событий, не было выявлено ничего нового по сравнению с наблюдениями, сформулированными в ходе анализа моделей на основе полного журнала событий.

Однако данные журналы событий можно использовать несколько иначе. С использованием метрики *rework* из библиотеки PM4Py можно посчитать количество повторений событий в журнале. С учетом особенностей построения данных журналов событий, в таком случае мы получим список экзаменов, имеющих наибольшее число провалов. Обнаруженные события можно рассмотреть детальнее на моделях, синтезированных по полному журналу событий, с фильтром, который оставляет только траектории студентов с определенной проблемой. Таким образом можно ускорить процесс анализа образовательного процесса.

В результате применения метрики rework для данных журналов событий можно сделать вывод, что наибольшим числом повторений обладают провалы по алгебре и программированию во 2 и 4 модуле. Среди этих событий при рассмотрении по варианту журнала событий A особенно часто повторяются события, связанные с курсом алгебры в 4 модуле. Данная дисциплина в условном рейтинге повтора событий оказывается выше прочих минимум в 2 раза. Это является показателем необычно высокого числа неудачных пересдач.

Второй курс. Полная модель.

При анализе полной модели для второго курса, полученной при помощи эвристического алгоритма с установленным параметром отсечки, можно выделить несколько ключевых точек интереса. В 1 модуле наблюдаются многочисленные провалы экзамена курса «Архитектура вычислительных систем», затем во 2 модуле также наблюдаются провалы экзаменов по курсам «Конструирование программного обеспечения», «Алгоритмы и структуры данных» и «Теория вероятностей и математическая статистика». Наконец, в 4 модуле присутствуют провалы экзамена по курсу «Конструирование программного обеспечения». Важно отметить, что частота для всех указанных событий практически равная. При изучении модели также можно заметить несколько событий, связанных с провалами пересдач экзаменов по курсам «Алгоритмы и структуры данных» и «Теория вероятностей и математическая статистика» за 2 модуль. Более того, событие с провалом пересдачи по курсу «Теория вероятностей и математическая статистика» имеет повторы, что не предусмотрено образовательной системой и явно указывает на ошибку в исходных данных.

Воспользуемся фильтрацией журнала событий и рассмотрим более подробно провал экзамена по курсу «Архитектура вычислительных систем» в 1 модуле. Только половина таких студентов смогли успешно сдать с первого раза экзамен по данному курсу во 2 модуле. Дальнейшие траектории довольно разнообразны. Среди всех событий преобладают пропуски и провалы экзаменов. В итоге лишь 14% от общего числа таких студентов успешно завершили учебный год.

Рассмотрим также отдельно траектории студентов с провалом экзамена по курсу «Конструирование программного обеспечения» во 2 модуле. Среди подобных траекторий наблюдаются провалы экзаменов, упомянутых выше. Эти провалы затрагивают от четверти до половины студентов. Среди таких провалов больше всего выделяется провал экзамена по курсу «Алгоритмы и структуры данных» во 2 модуле. Также можно отметить провал курса «Теория вероятностей и математическая статистика» во 2 модуле. 30% таких студентов провалили экзамен и еще 30% пропустили экзамен в основном по уважительным причинам. 40% студентов с указанным провалом успешно завершили учебный год.

Траектории студентов с провалом экзамена по курсу «Алгоритмы и структуры данных» во 2 модуле также содержат провалы уже упомянутых выше курсов, затрагивающих от четверти до половины студентов. На данной модели выделяется большое число провалов пересдачи курса «Алгоритмы и структуры данных» за 2 модуль. Она была неудачна для около 60% от числа студентов, проваливших экзамен. А 40% провалили и вторую пересдачу с комиссией. Только половина студентов успешно завершили учебный год.

Модель траекторий с провалом экзамена по курсу «Теория вероятностей и математическая статистика» во 2 модуле показывает картину, идентичную двум, которые обсуждались ранее. Однако

для данного курса на полной модели можно также заметить большое число провалов данного экзамена из-за пропуска по уважительной причине. Среди траекторий тех, кто пропустил экзамен по уважительной причине, наблюдается малое число отклонений, за исключением того, что половина таких студентов провалили пересдачу после пропуска. Также около половины не закончили учебный год. Половина от этого числа перевелась в другое учебное учреждение, а другая половина была отчислена за академическую неуспеваемость.

Провалы экзамена по курсу «Конструирование программного обеспечения» в 4 модуле связаны с той же самой группой провалов всё тех же курсов, которые были явно видны на полной модели. Однако в отличии от прошлых случаев, частота отклонений ниже. Вторым отличием является присутствие в модели группы различных пересдач с комиссией дисциплин 2 курса. Их частота довольно низка, однако представлены все экзамены за 2 модуль. Также 60% из студентов имеющих указанный провал экзамена провалили и его пересдачу, а 28% от общего числа студентов также провалили и вторую пересдачу с комиссией. Несмотря на это 74% студентов смогли успешно завершить учебный год.

Рассмотрим также и обнаруженные проблемы с пересдачами. Среди студентов с провалом пересдачи экзамена по курсу «Алгоритмы и структуры данных» во 2 модуле заметно большое число провалов и второй пересдачи с комиссией данного экзамена, около 65% от числа таких студентов. Еще для 35% траекторий студентов можно выделить связь с провалами экзаменов по курсу «Конструирование программного обеспечения» во 2 или 4 модуле. Только половина указанных студентов успешно завершили учебный год.

По траекториям с провалом пересдачи экзамена по курсу «Теория вероятностей и математическая статистика» за 2 модуль видно, что большинство студентов с данным провалом имели пропуск по уважительной причине. Как было указано ранее, у данного события имеются повторы, а на основе анализа траекторий видна возможная причина ошибки в данных. Вполне вероятно, что данные повторы компенсируют отсутствующее событие с дополнительной пересдачей ввиду пропуска по уважительной причине. У четверти студентов имеется провал экзамена по курсу «Конструирование программного обеспечения» в 4 модуле. В итоге 40% студентов успешно завершили учебный год.

Второй курс. Модель отклонений.

Модель, построенная на основе журнала событий после исключения траекторий без отклонений, не показывает ничего нового относительно предыдущей модели, так как для второго курса наблюдается меньшее число студентов и меньшее число отклонений от «идеальной» траектории относительно первого курса.

Второй курс. Модель повторных отклонений.

Модель повторных отклонений в сравнении с первым курсом имеет меньшее число студентов, но, как и ранее, большинство успешно завершают учебный год.

На модели можно заметить обособленную траекторию, ведущую к отчислению за академическую неуспеваемость, содержащую провалы трех экзаменов во 2 модуле по курсам: «Алгоритмы и структуры данных», «Теория вероятностей и математическая статистика», «Конструирование программного обеспечения».

В основном все повторяющиеся траектории отклонений содержат пропуски экзаменов по уважительной или неуважительной причине. Среди пропусков можно отметить сдачу экзамена по курсу «Алгоритмы и структуры данных» во 2 модуле при пропуске по неуважительной причине. Такое возможно ввиду особенностей образовательного процесса. Студенты получают общую удовлетворительную оценку по дисциплине без успешной сдачи экзамена благодаря баллам, накопленным в ходе семестра.

Второй курс. Варианты журнала событий А и Б.

В сравнении с первым курсом повторений стало меньше, что говорит о большей частоте успешных пересдач. Среди повторений можно наблюдать группу экзаменов по курсам «Алгоритмы и структуры данных» и «Теория вероятностей и математическая статистика» за 2 модуль, а также «Конструирование программного обеспечения» за 2 и 4 модуль.

Третий курс. Полная модель.

В журнале событий, содержащем события образовательного процесса третьего года обучения, количество студентов в данных заметно меньше. Однако даже несмотря на это невооруженным глазом видно практически полное отсутствие отклонений на модели. Все присутствующие отклонения — это, в основном, единичные случаи. Единственный относительно часто встречающийся в траекториях вариант отклонения связан с курсом «Математические методы анализа данных» в 4 модуле. Также модель содержит сдачу экзамена по курсу «Проектирование архитектуры программных систем» в 4 модуле при наличии пропуска по уважительной или неуважительной причине.

Третий курс. Модель отклонений.

Модель отклонений в данном случае позволила взглянуть на все имеющиеся отклонения, так как все они имеют низкую частоту и теряются на фоне основной траектории. В то же время частота данных событий слишком мала для того, чтобы можно было сделать содержательные выводы.

Третий курс. Модель повторных отклонений.

Модель повторных отклонений еще раз подчеркнула то, что было выявлено при анализе полной модели и позволила сделать дополнительные выводы относительно упомянутых курсов. Так, провал экзамена по курсу «Математические методы анализа данных» в 4 модуле среди повторяющихся траекторий представлен только в виде пропуска по неуважительным причинам. При этом траектории студентов, содержащиеся в модели повторных отклонений, приводят к успешному, хоть и непростому, завершению учебного года.

Третий курс. Варианты журнала событий А и Б.

Повторения событий в основном представляют собой единичные случаи, что показывает практически полное отсутствие неуспешных пересдач.

Четвертый курс.

Количество траекторий студентов в журнале событий, как и число отклонений этих траекторий, в сравнении с третьим курсом уменьшилось настолько, что уже не позволяет провести какой-либо существенный анализ.

5.3. Результаты анализа образовательного процесса

Подведем итоги проведённого и сделаем обобщающие выводы из полученных наблюдений.

Анализ синтезированных моделей образовательного процесса с применением предложенного подхода позволяет в краткие сроки получить большое количество информации о зависимостях между конкретными дисциплинами, выявить проблемы и аномалии в учебных траекториях студентов. В комбинации с другими методами анализа данных можно легко обнаружить причины тех или иных нежелательных событий. Результаты анализа также позволяют выявить события в студенческих траекториях, которые сигнализируют о возможных дальнейших отклонениях, которые могут привести к негативным исходам.

Для первого курса удалось установить, что наибольшее число проблем связано с учебными курсами по алгебре и программированию. Чаще всего проблемы возникают с экзаменом по алгебре за 4 модуль. Дальнейший анализ показал, что между курсами по алгебре и программированию наблюдается корреляция, а их провал часто взаимосвязан. Провалы по алгебре и программированию во 2 модуле более опасны, чем в 4 модуле, так как чаще приводят к безуспешному завершению

учебного года. Важно отметить, что результат экзамена по программированию в 1 модуле не оказывает влияния на последующие провалы экзаменов по программированию во 2 и 4 модулях, что странно. Студенты, которые выбрали альтернативный курс программирования на C/C++, могут иметь отклонения в своих учебных траекториях. Однако даже не смотря на наличие отклонений в траекториях, все подобные студенты успешно завершают учебный год. Провал экзамена курса по дискретной математике в 2 модуле наблюдается только в траекториях с большим числом отклонений. То есть у среднего студента проблем с курсом не возникает. Среди студентов первого курса наблюдается группа неактивных студентов, пропускающих экзамены по неуважительной причине, что приводит к их отчислению за академическую неуспеваемость в конце 2 модуля. К неудачному завершению учебного года, в основном, приводят уникальные траектории отклонений. Траектории отклонений, для которых зарегистрировано несколько случаев, приводят рано или поздно к успешному завершению учебного года. Таким образом, можно дать рекомендацию администраторам академической программы с особой внимательностью относиться ко всем необычным проблемам, которые возникают у студентов. Обычные проблемы существенно реже приводят к отчислению, чем уникальные.

Для второго курса удалось установить, что наибольшее число проблем у студентов вызывают учебные курсы «Алгоритмы и структуры данных», «Теория вероятностей и математическая статистика» и «Конструирование программного обеспечения». Провалы данных курсов часто встречаются в студенческих траекториях вместе, что показывает взаимосвязь между ними. Также это говорит о том, что студенты, попавшие на отклоняющуюся траекторию, вероятнее будут иметь более одного отклонения на этой учебной траектории. Среди провалов экзаменов выделяется провал экзамена по курсу «Архитектура вычислительных систем» в 1 модуле. Студенты, провалившие данный экзамен, часто имеют множество других отклонений в своих учебных траекториях. Кроме того, эти студенты редко успешно заканчивают учебный год. Можно порекомендовать администраторам академической программы сразу же брать студентов, которые споткнулись на данном курсе, под индивидуальную «опеку», чтобы помочь справиться и не отчислиться. Наконец, удалось обнаружить ошибки в исходных данных, допущенных на этапе сбора информации.

Для третьего курса наблюдается низкое число отклонений. Студенты на данном этапе обучения «оступаются» гораздо реже, а все отклонения наблюдаются ближе к концу учебного года. К тому же провалы экзаменов редко получают продолжение в виде провалов пересдачи и большинство студентов с отклонениями в траекториях всё же пересдают экзамены и успешно завершают учебный год.

Заключение

В данной работе предложен общий подход, пригодный для анализа данных, получаемых из электронных образовательных сред, которые лежат в основе современного образования. Данный подход позволяет извлечь пользу из большого количества данных о движении студентов по их индивидуальным образовательным траекториям, которые записываются подобными системами. Данная работа не только описывает новый подход, основанный на использовании методов process mining, но и содержит пример применения этого подхода к реальному набору данных учебного процесса на одной из программ бакалавриата, который получен из информационной образовательной среды НИУ «Высшая школа экономики».

Синтезированные модели процессов наглядно представляют образовательные траектории успешных студентов, а также то, как устроены отклонения от данных траекторий. Показано, что в комбинации с другими методами анализа данных можно легко обнаружить причины тех или иных событий и траекторий. В результате анализа выявлены точки особого внимания для администраторов образовательной программы, а также определены некоторые сигнальные события, появление которых в индивидуальной траектории студента может быть тревожным сигналом. Наконец,

показано, что наш подход позволяет выявлять и некорректные последовательности событий, наличие которых в учебных траекториях свидетельствует об ошибках в данных или о некорректной работе образовательной среды.

Конечно, не на все открытые вопросы даны ответы в данной работе, а потому отметим также несколько направлений дальнейших исследований.

Увеличение разнообразия рассматриваемых типов действий в журнале событий позволяет строить большое количество разнообразных моделей процессов и отвечать на самые разные вопросы. Например, можно рассматривать журналы событий, отдельными действиями в которых будут получение отличной, хорошей, удовлетворительной оценок. Такие модели можно использовать для выявления типичных траекторий не просто успевающих студентов, но отличников.

Интерес представляет также анализ моделей прохождения отдельных дисциплин. Для этого необходимо использовать журналы событий, содержащие данные не только об итоговых экзаменах, но и о промежуточных контрольных и домашних работах. Интересным является также проведение сравнительного анализа журналов событий для различных образовательных программ.

Анализ данных позволяет выявлять цепочки событий в учебных траекториях, которые с большей вероятностью приведут к положительному исходу. Это означает, что есть возможность построения прогноза для образовательной траектории студента. Модуль такого прогнозирования может быть основой для рекомендательной системы, осуществляющей помощь в выборе индивидуальной учебной траектории из вариативных дисциплин учебного плана. Представляется, что разработка и внедрение подобных систем в образовательный процесс может помочь снизить число отчислений, а значит, сделать жизнь студентов, преподавателей, администраторов образования немного проще.

References

[1] R. Jaakonmäki, J. vom Brocke, S. Dietze, H. Drachsler, A. Fortenbacher, R. Helbig, M. D. Kickmeier-Rust, I. Marenzi, A. Suarez, and H. Yun, *Learning Analytics Cookbook - How to Support Learning Processes Through Data Analytics and Visualization*, ser. Springer Briefs in Business Process Management. Springer, 2020.

[2] W. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016, ISBN: 978-3-662-49850-7. DOI: 10.1007/978-3-662-49851-4. [Online]. Available: https://doi.org/10.1007/978-3--49851-4.

[3] J. Carmona, B. van Dongen, A. Solti, and M. Weidlich, *Conformance Checking - Relating Processes and Models*. Springer, 2018, ISBN: 978-3-319-99413-0. DOI: 10.1007/978-3-319-99414-7. [Online]. Available: https://doi.org/10.1007/978-3-319-99414-7.

[4] S. Suriadi, M. T. Wynn, C. Ouyang, A. H. M. ter Hofstede, and N. J. van Dijk, «Understanding Process Behaviours in a Large Insurance Company in Australia: A Case Study», in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 7908, Springer, 2013, pp. 449–464.

[5] M. Mittal and A. Sureka, «Process mining software repositories from student projects in an undergraduate software engineering course», in *ICSE Companion*, ACM, 2014, pp. 344–353.

[6] A. Mitsyuk, A. Kalenkova, S. Shershakov, and W. van der Aalst, «Using process mining for the analysis of an e-trade system: A case study», *Biznes-informatika*, no. 3 (29), pp. 15–27, 2014.

[7] S.-k. Lee, B. Kim, M. Huh, S. Cho, S. Park, and D. Lee, «Mining transportation logs for understanding the after-assembly block manufacturing process in the shipbuilding industry», *Expert Syst. Appl.*, vol. 40, no. 1, pp. 83–95, 2013.

[8] Á. Valencia-Parra, B. Ramos-Gutiérrez, A. J. Varela-Vaca, M. T. G. López, and A. G. Bernal, «Enabling process mining in aircraft manufactures: extracting event logs and discovering processes from complex data», in *BPM (Industry Forum)*, ser. CEUR Workshop Proceedings, vol. 2428, CEUR-WS.org, 2019, pp. 166–177.

[9] K. Smit and J. Mens, «Process Mining in The Rail Industry: A Qualitative Analysis of Success Factors and Remaining Challenges», in *Bled eConference*, University of Maribor Press / Association for Information Systems, 2019, p. 25.

[10] J. Munoz-Gama, N. Martin, C. Fernández-Llatas, O. A. Johnson, M. Sepúlveda, E. Helm, V. Galvez-Yanjari, E. Rojas, A. Martinez-Millana, D. Aloini, I. A. Amantea, R. Andrews, M. Arias, I. Beerepoot, E. Benevento, A. Burattin, D. Capurro, J. Carmona, M. Comuzzi, B. Dalmas, R. de la Fuente, C. D. Francescomarino, C. D. Ciccio, R. Gatta, C. Ghidini, F. Gonzalez-Lopez, G. Ibáñez-Sánchez, H. B. Klasky, A. P. Kurniati, X. Lu, F. Mannhardt, R. Mans, M. Marcos, R. M. de Carvalho, M. Pegoraro, S. K. Poon, L. Pufahl, H. A. Reijers, S. Remy, S. Rinderle-Ma, L. Sacchi, F. Seoane, M. Song, A. Stefanini, E. Sulis, A. H. M. ter Hofstede, P. J. Toussaint, V. Traver, Z. Valero-Ramon, I. van de Weerd, W. van der Aalst, R. J. B. Vanwersch, M. Weske, M. T. Wynn, and F. Zerbato, «Process mining for healthcare: Characteristics and challenges», *J. Biomed. Informatics*, vol. 127, p. 103 994, 2022.

[11] A. Guzzo, A. Rullo, and E. Vocaturo, «Process mining applications in the healthcare domain: A comprehensive review», *WIREs Data Mining Knowl. Discov.*, vol. 12, no. 2, 2022.

[12] M. R. Dallagassa, C. dos Santos Garcia, E. E. Scalabrin, S. O. Ioshii, and D. R. Carvalho, «Opportunities and challenges for applying process mining in healthcare: a systematic mapping study», *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 1, pp. 165–182, 2022.

[13] C. dos Santos Garcia, A. Meincheim, E. R. F. Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos, and E. E. Scalabrin, «Process mining techniques and applications - A systematic mapping study», *Expert Syst. Appl.*, vol. 133, pp. 260–295, 2019.

[14] M. Dumas and F. M. Maggi, «Enabling Process Innovation via Deviance Mining and Predictive Monitoring», in *BPM - Driving Innovation in a Digital World*, J. vom Brocke and T. Schmiedel, Eds., Springer, 2015, pp. 145–154.

[15] I. Teinemaa, M. Dumas, F. M. Maggi, and C. D. Francescomarino, «Predictive Business Process Monitoring with Structured and Unstructured Data», in *Business Process Management - 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, September 18-22, 2016. Proceedings*, M. L. Rosa, P. Loos, and O. Pastor, Eds., ser. Lecture Notes in Computer Science, vol. 9850, Springer, 2016, pp. 401–417.

[16] I. Teinemaa, N. Tax, M. de Leoni, M. Dumas, and F. M. Maggi, «Alarm-Based Prescriptive Process Monitoring», in *Business Process Management Forum - BPM Forum 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings*, M. Weske, M. Montali, I. Weber, and J. vom Brocke, Eds., ser. Lecture Notes in Business Information Processing, vol. 329, Springer, 2018, pp. 91–107.

[17] W. van der Aalst, «Business Process Simulation Survival Guide», in *Handbook on Business Process Management 1, Introduction, Methods, and Information Systems, 2nd Ed*, ser. International Handbooks on Information Systems, J. vom Brocke and M. Rosemann, Eds., Springer, 2015, pp. 337–370.

[18] A. A. Mitsyuk, I. S. Shugurov, A. A. Kalenkova, and W. van der Aalst, «Generating event logs for high-level process models», *Simul. Model. Pract. Theory*, vol. 74, pp. 1–16, 2017. DOI: 10.1016/j.simpat.2017.01.003. [Online]. Available: https://doi.org/10.1016/j.simpat.2017.01.003.

[19] W. van der Aalst, «Process mining and simulation: a match made in heaven!», in *Proceedings of the 50th Computer Simulation Conference, SummerSim 2018, Bordeaux, France, July 09-12, 2018*, ACM, 2018, 4:1–4:12. [Online]. Available: http://dl.acm.org/citation.cfm?id=3275386.

[20] W. van der Aalst, «Process-Aware Information Systems: Design, Enactment, and Analysis», in *Wiley Encyclopedia of Computer Science and Engineering*, B. W. Wah, Ed., John Wiley & Sons, Inc., 2008.

[21] G. Acampora, A. Vitiello, B. N. D. Stefano, W. van der Aalst, C. W. Günther, and E. Verbeek, «IEEE 1849: The XES Standard: The Second IEEE Standard Sponsored by IEEE Computational Intelligence Society [Society Briefs]», *IEEE Comput. Intell. Mag.*, vol. 12, no. 2, pp. 4–8, 2017.

[22] A. J. M. M. Weijters and W. van der Aalst, «Rediscovering workflow models from event-based data using little thumb», *Integr. Comput. Aided Eng.*, vol. 10, no. 2, pp. 151–162, 2003. DOI: 10.3233/ica-2003-10205. [Online]. Available: https://doi.org/10.3233/ica-2003-10205.

[23] A. Berti and W. van der Aalst, «A Novel Token-Based Replay Technique to Speed Up Conformance Checking and Process Enhancement», *Trans. Petri Nets Other Model. Concurr.*, vol. 15, pp. 1–26, 2021.

[24] J. Munoz-Gama and J. Carmona, «A Fresh Look at Precision in Process Conformance», vol. 6336, Sep. 2010, pp. 211–226, ISBN: 978-3-642-15617-5. DOI: 10.1007/978-3-642-15618-2_16.

[25] J. Buijs, B. Dongen, and W. Aalst, «Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity», *International Journal of Cooperative Information Systems*, vol. 23, p. 1 440 001, Mar. 2014. DOI: 10.1142/S0218843014400012.

[26] A. Adriansyah, J. Munoz-Gama, J. Carmona, B. Dongen, and W. Aalst, «Measuring precision of modeled behavior», *Information Systems and e-Business Management*, vol. 13, Jan. 2014. DOI: 10.1007/s10257-014-0234-7.

[27] E. Verbeek, J. C. A. M. Buijs, B. F. van Dongen, and W. van der Aalst, «ProM 6: The Process Mining Toolkit», in *Proceedings of the Business Process Management 2010 Demonstration Track, Hoboken, NJ, USA, September 14-16, 2010*, M. L. Rosa, Ed., ser. CEUR Workshop Proceedings, vol. 615, CEUR-WS.org, 2010. [Online]. Available: http://ceur-ws.org/Vol-615/paper13.pdf.

[28] A. Berti, S. J. van Zelst, and W. van der Aalst, «Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science», *CoRR*, vol. abs/1905.06169, 2019. arXiv: 1905.06169. [Online]. Available: http://arxiv.org/abs/1905.06169.

[29] A. Berti, M. P. Nghia, and W. van der Aalst, «PM4Py-GPU: A High-Performance General-Purpose Library for Process Mining», in *Research Challenges in Information Science - 16th International Conference, RCIS 2022, Barcelona, Spain, May 17-20, 2022, Proceedings*, R. S. S. Guizzardi, J. Ralyté, and X. Franch, Eds., ser. Lecture Notes in Business Information Processing, vol. 446, Springer, 2022, pp. 727–734.

[30] S. A. Shershakov, «VTMine for Visio: A Graphical Tool for Modeling in Process Mining», *Autom. Control. Comput. Sci.*, vol. 55, no. 7, pp. 847–865, 2021. DOI: 10.3103/S0146411621070282. [Online]. Available: https://doi.org/10.3103/S0146411621070282.

[31] I. S. Shugurov and A. A. Mitsyuk, «Applying MapReduce to conformance checking», *Proceedings of ISP RAS*, vol. 28, no. 3, pp. 103–122, 2016. [Online]. Available: https://ispranproceedings.elpub.ru/jour/issue/download/9/17#page=104.

[32] A. Bogarín, R. Cerezo, and C. Romero, «A survey on educational process mining», *WIREs Data Mining Knowl. Discov.*, vol. 8, no. 1, 2018. DOI: 10.1002/widm.1230. [Online]. Available: https://doi.org/10.1002/widm.1230.

[33] J. C. Vidal, B. Vázquez-Barreiros, M. Lama, and M. Mucientes, «Recompiling learning processes from event logs», *Knowl. Based Syst.*, vol. 100, pp. 160–174, 2016. DOI: 10.1016/j.knosys.2016.03.003. [Online]. Available: https://doi.org/10.1016/j.knosys.2016.03.003.

[34] A. Bogarín, R. Cerezo, and C. Romero, «Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs).», *Psicothema*, vol. 30 3, pp. 322–329, 2018.

[35] H. Al-Qaheri and M. Panda, «An Education Process Mining Framework: Unveiling Meaningful Information for Understanding Students' Learning Behavior and Improving Teaching Quality», *Inf.*, vol. 13, no. 1, p. 29, 2022. DOI: 10.3390/info13010029. [Online]. Available: https://doi.org/10.3390/info13010029.

[36] E. M. Real, E. P. Pimentel, L. V. de Oliveira, J. C. Braga, and I. Stiubiener, «Educational Process Mining for Verifying Student Learning Paths in an Introductory Programming Course», in *IEEE Frontiers in Education Conference, FIE 2020, Uppsala, Sweden, October 21-24, 2020*, IEEE, 2020, pp. 1–9. DOI: 10.1109/FIE44824.2020.9274125. [Online]. Available: https://doi.org/10.1109/FIE44824.2020.9274125.

[37] J. P. Salazar-Fernandez, M. Sepúlveda, J. Munoz-Gama, and M. Nussbaum, «Curricular Analytics to Characterize Educational Trajectories in High-Failure Rate Courses That Lead to Late Dropout», *Applied Sciences*, vol. 11, no. 4, 2021, ISSN: 2076-3417. DOI: 10.3390/app11041436. [Online]. Available: https://www.mdpi.com/2076-3417/11/4/1436.

[38] J. P. Salazar-Fernandez, J. Munoz-Gama, J. Maldonado-Mahauad, D. Bustamante, and M. Sepúlveda, «Backpack Process Model (BPPM): A Process Mining Approach for Curricular Analytics», *Applied Sciences*, vol. 11, no. 9, 2021, ISSN: 2076-3417.

[39] I. A. Lomazova, A. A. Mitsyuk, and A. M. Sharipova, *Modeling MOOC learnflow with Petri net extensions*, 2021. DOI: 10.48550/ARXIV.2111.04419. [Online]. Available: https://arxiv.org/abs/2111.04419.

[40] L. Juhanák, J. Zounek, and L. Rohlíková, «Using process mining to analyze students' quiz-taking behavior patterns in a learning management system», *Comput. Hum. Behav.*, vol. 92, pp. 496–506, 2019. DOI: 10.1016/j.chb.2017.12.015. [Online]. Available: https://doi.org/10.1016/j.chb.2017.12.015.

[41] V. Southavilay, K. Yacef, and R. A. Calvo, «Process Mining to Support Students' Collaborative Writing», in *EDM*, www.educationaldatamining.org, 2010, pp. 257–266.

[42] G. Deeva and J. D. Weerdt, «Understanding Automated Feedback in Learning Processes by Mining Local Patterns», in *Business Process Management Workshops*, ser. Lecture Notes in Business Information Processing, vol. 342, Springer, 2018, pp. 56–68.

[43] D. Codish, E. Rabin, and G. Ravid, «User behavior pattern detection in unstructured processes - a learning management system case study», *Interact. Learn. Environ.*, vol. 27, no. 5-6, pp. 699–725, 2019.

[44] J. Maldonado-Mahauad, M. Pérez-Sanagustín, R. F. Kizilcec, N. Morales, and J. Munoz-Gama, «Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses», *Comput. Hum. Behav.*, vol. 80, pp. 179–196, 2018.

[45] W. Hachicha, L. Ghorbel, R. Champagnat, C. A. Zayani, and I. Amous, «Using Process Mining for Learning Resource Recommendation: A Moodle Case Study», in *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES-2021, Virtual Event / Szczecin, Poland, 8-10 September 2021*, J. Watróbski, W. Salabun, C. Toro, C. Zanni-Merk, R. J. Howlett, and L. C. Jain, Eds., ser. Procedia Computer Science, vol. 192, Elsevier, 2021, pp. 853–862. DOI: 10.1016/j.procs.2021.08.088. [Online]. Available: https://doi.org/10.1016/j.procs.2021.08.088.

[46] G. Sedrakyan, J. D. Weerdt, and M. Snoeck, «Process-mining enabled feedback: "Tell me what I did wrong" vs. "tell me how to do it right"», *Comput. Hum. Behav.*, vol. 57, pp. 352–376, 2016.

[47] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. van der Aalst, «PM^2: Process Mining Project Methodology», in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 9097, Springer, 2015, pp. 297–313.

[48] A. Weijters, W. Aalst, and A. Medeiros, *Process Mining with the Heuristics Miner-algorithm*. Jan. 2006, vol. 166, pp. 1–34.