УДК 004.42

А. Р. Аргинбаев[1], П. Ю. Бугаков[1]*

Разработка программного обеспечения для определения оригинальности текстовых электронных работ обучающихся

[1] Сибирский государственный университет геосистем и технологий, г. Новосибирск, Российская Федерация
* e-mail: arthur.arginbaev@gmail.com

Аннотация. Проблема плагиата в научной и образовательной среде существует уже долгое время. Стремительный рост и развитие цифровых технологий только усугубил данную проблему. Одним из наиболее популярных решений, используемых в высших учебных заведениях, является «Антиплагиат.ВУЗ», разработанный специально для использования в образовательных учреждениях для облегчения труда преподавателей. Использование даннного сервиса породило ряд проблем: существенные финансовые расходы; ограниченный перечень видов работ, подвергаемых проверке на наличие плагиата; ограниченное количество попыток проверки одной работы. В результате значительное количество видов самостоятельных работ обучающихся не подвергается проверке на плагиат и зачастую списывается учащимися друг у друга. В связи с этим было разработано программное обеспечение, определяющее оригинальность текстовых электронных работ без ограничений на количество попыток проверки, используя для этого внутреннюю гибко конфигурируемую коллекцию письменных работ учебного заведения для пресечения плагиата среди обучающихся. Данное программное обеспечение уже используется в учебном процессе на кафедре прикладной информатики и информационных систем СГУГиТ.

Ключевые слова: плагиат, учебный процесс, самостоятельные работы, антиплагиат, поиск заимствований, оценка оригинальности

A. R. Arginbaev[1], P. Yu. Bugakov[1]*

Development of Software for Determining the Originality of Students' Text Electronic Works

[1] Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation
* e-mail: arthur.arginbaev@gmail.com

Abstract. The problem of plagiarism in the scientific and educational environment has existed for a long time. The rapid growth and development of digital technologies has exacerbated this problem. One of the most popular solutions used in higher education institutions is "Antiplagiat.VUZ", developed specifically for use in educational institutions to make the work of teachers easier. The use of this service gave rise to a number of problems: significant financial costs; a limited list of types of work subject to checking for plagiarism; limited number of attempts to check one work. As a result, a significant number of types of independent work by students are not checked for plagiarism and are often copied by students from each other. In this regard, software was developed that determines the originality of text electronic works without restrictions on the number of verification attempts, using the internal flexibly configurable collection of written works of the educational institution to prevent plagiarism among students. This software is already used in the educational process at the Department of Applied Informatics and Information Systems of SSUGT.

Keywords: plagiarism, educational process, independent work, anti-plagiarism, search for borrowings, assessment of originality

Введение

Плагиат – это присвоение плодов чужого творчества: опубликование чужих произведений под своим именем без указания источника или использование без преобразующих творческих изменений, внесенных заимствователем [1–3]. Хотя данное явление может быть обусловлено множеством причин и вызвано различными факторами, очевидно, что оно пагубно влияет в первую очередь на автора. Проблема плагиата в научной и образовательной среде существует многие века [4–6]. К сожалению, стремительный рост и развитие цифровых технологий только усугубил данную проблему. В целом признано, что онлайн-плагиат действительно высок из-за легкой доступности информации [7–10]. Более того, широкое распространение получили сервисы, специализирующиеся в написании текстовых работ на заказ.

Для противодействия обострившейся проблеме было разработано множество сервисов: antiplagiat.ru, text.rucont.ru, text.ru, etxt.ru. Антиплагиат используется для борьбы с нарушениями авторских прав и защиты интеллектуальной собственности, установления подлинности и оригинальности текста, лингвистической экспертизы и установления авторства текста [11, 12].

Одним из наиболее популярных решений, используемых в высших учебных заведениях, является «Антиплагиат.ВУЗ», разработанный специально для использования в образовательных учреждениях. Программа «Антиплагиат» – это бесспорно хороший помощник, но проблема в том, что она используется в трудах, выносимых на обсуждение: курсовых работ, дипломных проектов, научных статей [13–15].

Согласно Федеральному закону «Об образовании в Российской Федерации» от 29.12.2012 N 273, а также Федеральным государственным образовательным стандартам, в рамках реализации учебной дисциплины до 50% от обязательной учебной нагрузки приходится на самостоятельную работу. Из этого можно сделать вывод, что большая часть выполняемых в процессе обучения работ никак не контролируется на наличие плагиата и это приводит к повсеместному списыванию.

В связи со сложившейся ситуацией, целью работы является разработка программного обеспечения для определения оригинальности текстовых электронных работ обучающихся, используя для этого внутреннюю, гибко конфигурируемую, коллекцию работ учебного заведения. Для достижения поставленной цели необходимо выполнить следующие задачи: составить список требований к разрабатываемому ПО; разработать алгоритм определения оригинальности; выполнить проектирование и практически реализовать программу; выполнить тестирование и апробацию.

Требования к разрабатываемому ПО: клиент-серверная архитектура; коллекция источников для поиска создается конечным пользователем в рамках учетной записи; систематизация и гибкая конфигурация работ, используемых в процессе поиска заимствований; выбор анализируемых символов, режима производительности, режима анализа, вида отчета, требуемого уровня оригинальности; сохранение отчетов в формате pdf-докуметов.

Требования к алгоритму поиска заимствований: определение точных и частичных совпадений; нормализация русскоязычного текста; обнаружение замаскированных заимствований; определение подделки информации о документе; формирование анализируемых значений в зависимости от анализируемых символов.

Для хранения текстовых электронных работ обучающихся была спроектирована модель отношения сущностей. Она позволяет хранить основную информацию об исходном документе, его оригинальный и нормализованный текст, а также ключевые параметры выполненной работы и информацию об ее авторе. Словарь нормализации составлен из данных OpenCorpora. В их индексе содержится информация о пяти миллионах русскоязычных слов. Из данных индекса был составлен словарь нормализации, содержащий сто семьдясят нормальных и три миллиона производных форм слов.

Поскольку разрабатываемое программное обеспечение предполагает поиск заимствований среди небольшого количества документов был разработан уникальный метод поиска заимствований, состоящий из трех этапов. Входными параметрами алгоритма является проверяемый и сравниваемый документы, список анализируемых символов (кириллица, латиница, цифры, прочие), а также выбранный режим анализа (полный или быстрый). Каждый документ имеет уникальный идентификатор.

На первом этапе поиска совпадений сравнивается идентификатор проверяемого и сравниваемого документов. В случае их совпадения поиск совпадений завершается.

На втором этапе выполняется сравнение хэшей для поиска точных совпадений. Процесс формирования хэшей документа является последовательным. Наличие анализируемых символов в конкретном слове определяет будет ли оно анализируемым, а также его анализируемое значение. Если предложение содержит анализируемые слова, то его хэш формируется из упорядоченной последовательности их хэшей (рис. 1). Хэш абзацев и преложений формируются по тому же принципу.

Сначала сравниваются хэши анализируемых документов. В случае их совпадения, в список источников точных совпадений проверяемого документа добавляется идентификатор сравниваемого и на этом поиск совпадений завершается. Затем выполняется пересечение множеств анализируемых хэшей абзацев.

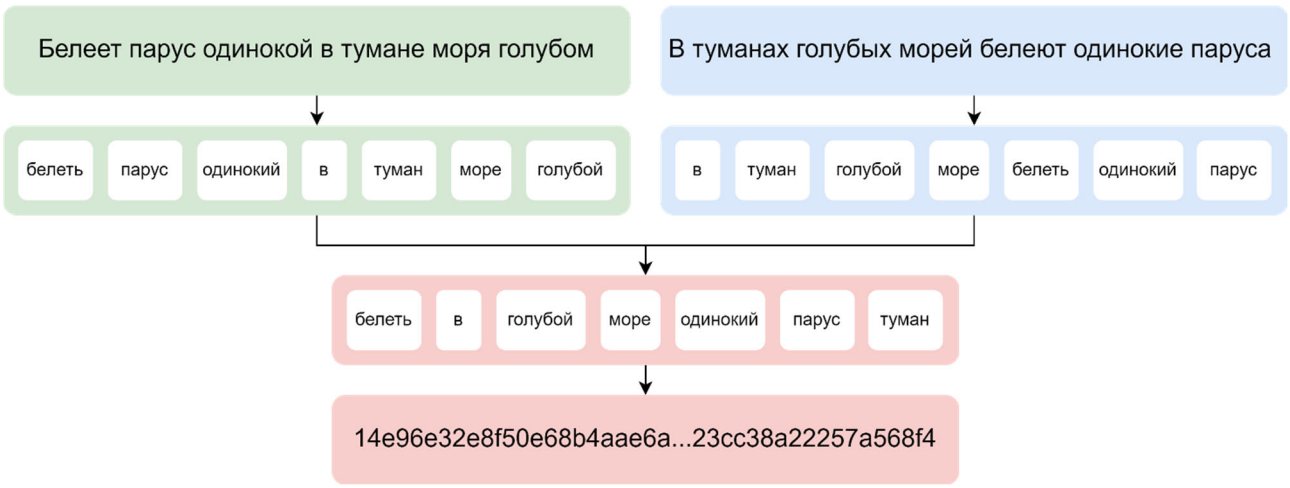

Рис. 1. Формирование одинакового хэша у двух разных предложений

Если множества пересекаются, в список источников точных совпадений пересекающихся абзацев проверяемого документа добавляется идентификатор сравниваемого. После этого выполняется пересечение множеств анализируемых предложений из не пересекшихся абзацев проверяемого документа с множеством анализируемых предложений сравниваемого. Если множества пересекаются, в список источников точных совпадений пересекающихся предложений проверяемого документа добавляется идентификатор сравниваемого. Если выбран быстрый режим анализа на этом поиск совпадений завершается.

На последнем этапе выполняется поиск частичных совпадений для каждого предложения проверяемого документа, не совпавшего на прошлом этапе. Для каждого проверяемого предложения выбираются анализируемые предложения сравниваемого документа, содержащие слова с теми же анализируемыми значениями. Если такие предложения найдены, из них выбирается единственное, с наибольшим количеством пересекающихся элементов (рис. 2).

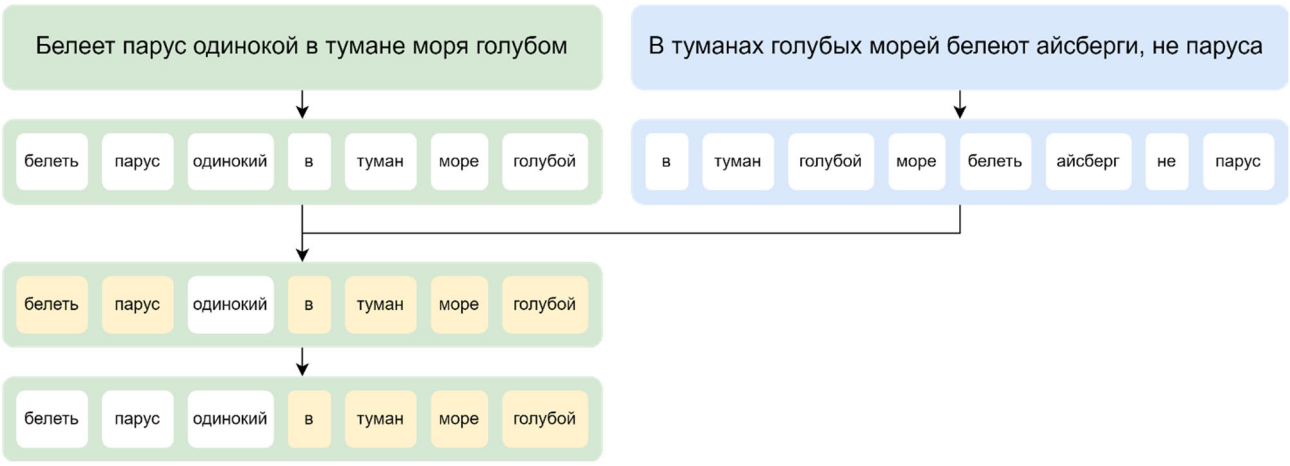

Рис. 2. Формирование пересекающихся элементов

Пересекающимися элементами считаются слова сравниваемых предложений, с совпадающими анализируемыми значениями, образующими последовательность из трех и более элементов в проверяемом предложении. Затем в список источников частичных совпадений пересекающихся слов проверяемого документа, образующих последовательность из трех и более элементов, добавляется идентификатор выбранного сравниваемого документа.

Для реализации приложения использовались следующие средства: язык программирования C#, платформа пользовательского интерфейса Windows Presentation Foundation, СУБД PostgreSQL, библиотека Entity Framework Core, библиотека Haukcode.WkHtmlToPdfDotNet.

Спроектированная модель отношения сущностей была реализована в PostgreSQL. Для ускорения выполнения запросов к базе данных были созданы btree индексы. Для упрощения процесса удаления работ и повышения целостности данных был наложен ряд ограничений, в том числе ограничивающие и каскадные удаления [16–18]. Ограничивающие предотвращают удаление связанной строки, а каскадные указывают, что при удалении связанных строк зависимые от них будут так же автоматически удалены. Для доступа к базе данных используется EF Core [19].

Поскольку в качестве платформы пользовательского интерфейса был выбран Windows Presentation Foundation (WPF), поддерживающий привязку данных, оптимальным архитектурным паттерном приложения является Model-View-ViewModel (рис. 3) [20].

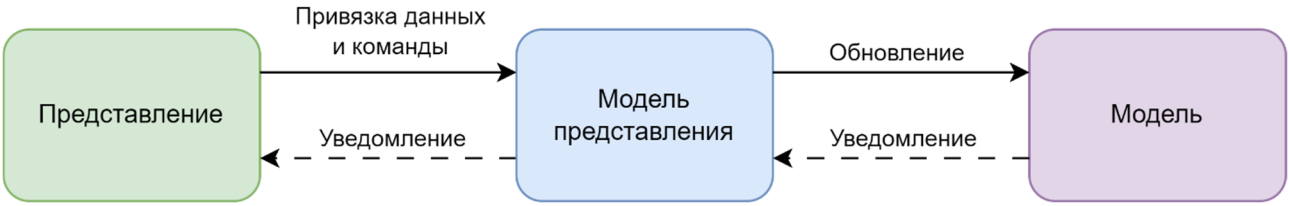

Рис. 3. Взаимодействие основных элементов паттерна

Результаты

Был разработан прототип программного обеспечения, позволяющий определить оригинальность текстовых электронных работ обучающихся в рамках группы или потока (рис. 4).

Для тестирования производительности выполнялся полный перекрестный анализ 60 курсовых работ (более 100 страниц в каждой) в трех режимах производительности на трех разных аппаратных конфигурациях (табл. 1).

От выбранного режима зависит количество потоков (8, 4 или 2), используемых в процессе анализа. Существенного влияния одновременного выполнения анализа с 10 учетных записей на итоговую скорость не выявлено. Потребление оперативной памяти варьировалось от 1 до 1,3 Гб в зависимости от конфигурации и режима анализа. Результаты тестирования свидетельствуют о прямой за-

висимости между производительностью процессора и временем выполнения. (табл. 2).

Результаты анализа каждой проверяемой работы сохраняются в виде отчетов, содержащих информацию о документе, параметры и результат анализа. Список источников, а также текст работы (с отображением заимствований) добавляются в случае выбора подробного вида отчета в параметрах анализа.

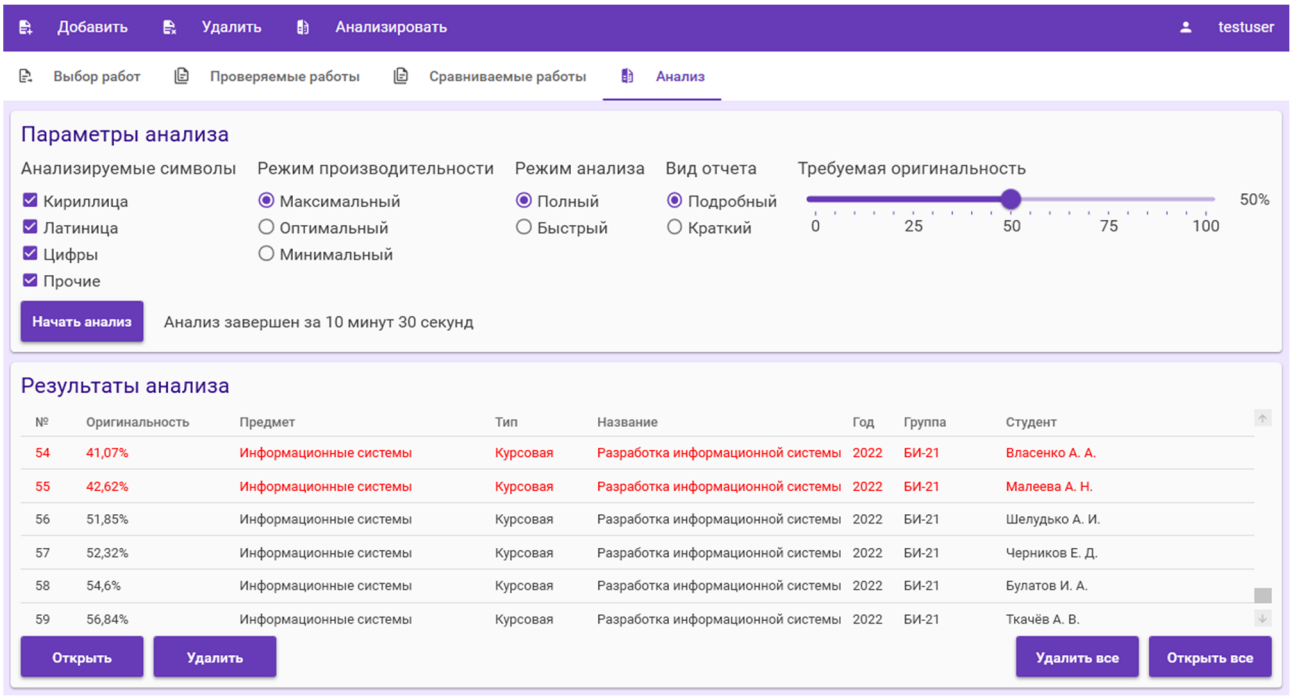

Рис. 4. Интерфейс прототипа

Таблица 1

Аппаратные конфигурации

№	ЦП	Кол-во ядер/потоков	Базовая частота ЦП, ГГц	Тип ОЗУ	Объем ОЗУ, Гб	Тип ПЗУ
1	Ryzen 7 2700X	8/16	3,7	DDR4	16	SSD
2	Ryzen 5 5500U	6/12	2,1	DDR4	16	SSD
3	Core i5 7400	4/4	3,0	DDR3	16	HDD

Таблица 2

Результаты тестирования

Режим	Конфигурация №1		Конфигурация №2		Конфигурация №3	
	Время	Исп. проц., %	Время	Исп. проц., %	Время	Исп. проц., %
Максимальный	00:09:03	26	00:12:22	53	00:12:27	80
Оптимальный	00:10:38	22	00:13:58	38	00:13:02	67
Минимальный	00:14:45	12	00:17:15	26	00:15:30	48

Обсуждение

Разработанный прототип позволяет пользователю самостоятельно формировать коллекцию работ для сравнения и гибко конфигурировать ее для каждой проверки. Выбор анализируемых символов позволяет изменять долю анализируемого текста, режим производительности – выбирать оптимальное соотношение между производительностью и потребляемыми ресурсами, режим анализа – выполнять поиск точных и частичных, либо только точных совпадений, вид отчета – степень детализации результатов анализа, а требуемая оригинальность – минимально допустимую долю авторского текста.

Сравнив результаты анализа одной и той же работы, полученные в системе «Антиплагиат.ВУЗ» и в результате выполнения перекрестного анализа в рамках потока, можно прийти к выводу, что далеко не всегда обучающиеся списывают из сторонних источников, не менее распространен и плагиат внутри коллектива (рис. 5).

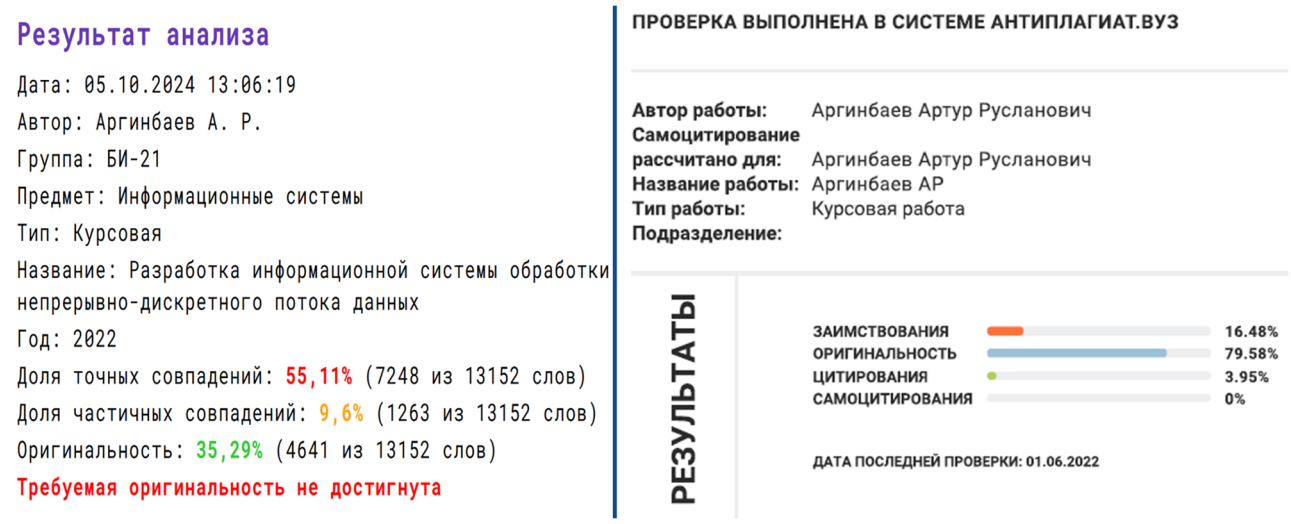

Рис. 5. Сравнение результатов

Заключение

Был разработан и протестирован прототип программного обеспечения, позволяющий определить оригинальность текстовых электронных работ обучающихся в рамках группы или потока. Получено свидетельство о государственной регистрации программы для ЭВМ [21]. Использование данной разработки в учебном процессе позволит значительно увеличить количество проверяемых на оригинальность работ, пресечь часть плагиата, пораждаемую списыванием, с помощью перекрестной проверки в рамках группы или потока. Она может быть проведена как среди работ текущего года, так и среди работ предыдущих лет.

Данное программное обеспечение планируется использовать в учебном процессе на всех кафедрах СГУГиТ для проверки оригинальности отчетов, лабораторных и курсовых работ.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Васильева, В. А. Плагиат глазами студентов: мошенничество или норма / В. А. Васильева, А. А. Шабаева // Социально-гуманитарные знания. – 2023. – № 3. – С. 20-29.
2. Силаев, П. В. Грамотная текстообразующая деятельность студентов на принципах кооперативного научного общения (избавление от плагиата) / П. В. Силаев // Родной язык в лингвокультурологическом аспекте. – Смоленск : Смоленский государственный университет, 2023. – С. 283-289.
3. Стаценко, Ю. Ю. Проблема плагиата. Как помочь студентам избежать плагиат при выполнении самостоятельной работы / Ю. Ю. Стаценко // Вестник Филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе. – 2021. – № 4(20). – С. 131-138.
4. Яркова, И. С. Соотношение норм права и морали при разрешении споров, связанных с академическим плагиатом / И. С. Яркова // Социальные нормы и практики. – 2021. – № 1(1). – С. 50-56.
5. Кацко, С. Ю. Проверка ВКР: корректные заимствования, плагиат и оригинальность текста / С. Ю. Кацко, И. П. Кокорина // Актуальные вопросы образования. – 2021. – № 1. – С. 142-145.
6. Кондрашова, Е. В. Влияние дистанционного обучения на плагиат и списывание / Е. В. Кондрашова // Современные наукоемкие технологии. – 2021. – № 3. – С. 156-161.
7. Витко, В. С. О содержании понятия "самоплагиат" / В. С. Витко // Вестник Томского государственного университета. – 2021. – № 467. – С. 235-243.
8. Бажанов, В. А. Феномен плагиата и его восприятие в академической среде / В. А. Бажанов, О. А. Козина // Вестник Томского государственного университета. Философия. Социология. Политология. – 2019. – № 48. – С. 225-235.
9. Плещенко, В. И. О плагиате в научных публикациях и выпускных работах / В. И. Плещенко // Высшее образование в России. – 2018. – Т. 27, № 8-9. – С. 62-70.
10. Левин, В. И. Плагиат, его сущность и борьба с ним / В. И. Левин // Высшее образование в России. – 2018. – Т. 27, № 1. – С. 143-150.
11. Алдохина, Е. Д. Системы и сервисы антиплагиата: общие аспекты / Е. Д. Алдохина // Развитие современной науки и технологий в условиях трансформационных процессов : Сборник материалов XIII Международной научно-практической конференции, Москва, 28 июля 2023 года. – Санкт-Петербург: Печатный цех, 2023. – С. 546-550.
12. Павлов, Е. М. Обзор возможностей и технологий внедрения систем защиты от плагиата / Е. М. Павлов, А. В. Рыжов, С. А. Петров // Инженерный вестник Дона. – 2023. – № 12(108). – С. 33-42.

13. Стрелкова, И. Б. Система «Антиплагиат»: проблема академической честности в условиях цифровизации образования / И. Б. Стрелкова // Университет - территория опережающего развития : Сборник научных статей Международной научно-практической конференции, посвящённый 80-летию ГрГУ им. Янки Купалы, Гродно, 19–20 февраля 2020 года / Редколлегия: Ю.Я. Романовский (гл. ред.) [и др.]. – Гродно: Гродненский государственный университет имени Янки Купалы, 2020. – С. 335-337.

14. Шарапова, Е. В. Сравнительный анализ сервисов проверки оригинальности текстов / Е. В. Шарапова // Машиностроение и безопасность жизнедеятельности. – 2019. – № 1(39). – С. 48-51.

15. Канатникова, Е. А. Интернет-технологии в образовании и проблема плагиата / Е. А. Канатникова // Russian Agricultural Science Review. – 2015. – Т. 6, № 6-3. – С. 155-160.

16. Моргунов, Е. П. PostgreSQL. Основы языка SQL : учебное пособие / Е. П. Моргунов ; под редакцией. Е. В. Рогова, П. В. Лузанова. – Санкт-Петербург : БХВ-Петербург, 2018. – 336 с. – ISBN 978-5-9775-4022-3. – Текст : непосредственный.

17. Рогов, Е. В. PostgreSQL 16 изнутри. – Москва : ДМК Пресс, 2024. – 664 с. – ISBN 978-5-93700-305-8. – URL: https://edu.postgrespro.ru/postgresql_internals-16.pdf (дата обращения: 05.02.2024). – Текст : электронный.

18. Лесовский, А. В. Мониторинг PostgreSQL / А. В. Лесовский. – Москва : Бумба, 2024. – 247 с. – ISBN 978-5-907754-42-3. – Текст : непосредственный.

19. Смит, Дж. П. Entity Framework Core в действии / пер. с англ. Д. А. Беликова. – Москва : ДМК Пресс, 2022. – 690 с. – ISBN 978-5-93700-114-6. – Текст : непосредственный.

20. Рыбанов, А. А. Паттерны проектирования на C# : учебное пособие / А. А. Рыбанов. – Волжский : ВПИ (филиал) ВолгГТУ, 2023. – 94 с. – ISBN 978-5-9948-4548-6. – URL: http://lib.volpi.ru:57772/csp/lib/PDF/739364773.pdf – Текст : электронный.

21. Свидетельство о государственной регистрации программы для ЭВМ 2024660592 Российская Федерация. Defori / А. Р. Аргинбаев, П. Ю. Бугаков; заявитель и правообладатель Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет геосистем и технологий». – № 2024619874; заявл. 08.05.2024; опубл 08.05.2024.