УДК 681.3
DOI: 10.26102/2310-6018/2025.51.4.043

Разработка информационной системы для анализа наполнения веб-сайтов текстовыми документами

И.В. Поцебнева, К.А. Андреева✉, Н.А. Пуговкина, П.Ю. Гусев

Воронежский государственный технический университет, Воронеж, Российская Федерация

Резюме. Работа посвящена вопросам анализа наполнения веб-сайта организации текстовыми документами в целях осуществления поддержки принятия решений при управлении образовательными программами организации высшего образования. Наполнение веб-сайта организации текстовыми документами выступает одним из критериев эффективности работы сайта. Сами критерии эффективности, в свою очередь, определяются типом сайта и типом организации, которая это сайт создала и поддерживает. В работе рассматриваются веб-сайты организаций высшего образования и их особенности. Одной из рассматриваемых особенностей выступает необходимость наличия рабочих программ в виде текстовых документов. Помимо необходимости наличия, рабочие программы на веб-сайте организации являются информационным материалом для абитуриентов, что повышает ценность подобной информации. Анализ наличия и содержания рабочих программ позволит решить ряд управленческих задач, однако для этого требуется проектирование и разработка инструмента проверки наличия рабочих программ. Для решения задачи проверки наличия и анализа содержания рабочих программ осуществлены проектирование и разработка информационной системы. Этап проектирования информационной системы заключался в разработке контекстной и декомпозированной диаграмм IDEF0 и диаграммы действий. Контекстная диаграмма определила систему, входы, выходы, элементы управления и механизмы информационной системы. Декомпозированная диаграмма содержит следующие модули: парсинга, обработки документов, анализа учебных планов, интеграции данных, экспорта. Диаграмма действия включает следующие акторы: администратор, внешний сайт, база данных, система визуализации – и следующие прецеденты: парсинг сайта, обработка документов, анализ учебных планов, интеграция данных, экспорт данных, визуализация данных. Реализация информационной системы обеспечила возможность построения общего дашборда образовательной организации, дашборда факультета, дашборда кафедры. Результаты работы обеспечивают возможность принятия управленческих решений на основе информации о наличии рабочих программ на сайте образовательной организации.

Ключевые слова: информационная система, анализ веб-сайта, рабочие программы дисциплин, образовательные организации высшего образования, парсинг веб-сайта, обработка документов, диаграммы IDEF0, визуализация данных.

Для цитирования: Поцебнева И.В., Андреева К.А., Пуговкина Н.А., Гусев П.Ю. Разработка информационной системы для анализа наполнения веб-сайтов текстовыми документами. *Моделирование, оптимизация и информационные технологии.* 2025;13(4). URL: https://moitvivt.ru/ru/journal/pdf?id=2063 DOI: 10.26102/2310-6018/2025.51.4.043

Development of an information system for analyzing the content of websites with textual documents

I.V. Pocebneva, K.A. Andreeva✉, N.A. Pugovkina, P.Y. Gusev

Voronezh State Technical University, Voronezh, the Russian Federation

Abstract. This study focuses on analyzing the content of organizational websites with textual documents in order to support decision-making in the management of educational programs of a higher education

organization. The presence of textual documents on an organization's website is one of the key criteria for assessing website effectiveness. These effectiveness criteria, in turn, are determined by the type of website and the type of organization that created and maintains it. The paper examines websites of higher education institutions and their specific characteristics. One such characteristic is the necessity of having curricula (working programs) available in the form of textual documents. Besides being a mandatory requirement, these curricula serve as informational materials for prospective students, thereby increasing the value of such information. Analyzing the availability and content of curricula can help address various management tasks; however, this requires designing and developing a tool to verify the presence of curricula. To solve the problem of verifying the availability and analyzing the content of curricula, an information system was designed and developed. The design phase involved creating an IDEF0 context diagram, a decomposed IDEF0 diagram, and an action (use case) diagram. The context diagram defined the system, inputs, outputs, controls, and mechanisms of the information system. The decomposed diagram includes the following modules: web parsing, document processing, curriculum analysis, data integration, and data export. The action diagram identifies the following actors: administrator, external website, database, visualization system, and includes the following use cases: website parsing, document processing, curriculum analysis, data integration, data export, and data visualization. The implementation of the information system enabled the creation of comprehensive dashboards for educational organizations, faculty-level dashboards, and department-level dashboards. The results of the system's operation support managerial decision-making based on information about the availability of curricula on educational institution websites.

Keywords: information system, website analysis, work programs of disciplines, higher education institutions, website parsing, document processing, IDEF0 diagrams, data visualization.

For citation: Pocebneva I.V., Andreeva K.A., Pugovkina N.A., Gusev P.Y. Development of an information system for analyzing the content of websites with textual documents. *Modeling, Optimization and Information Technologies.* 2025;13(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=2063 DOI: 10.26102/2310-6018/2025.51.4.043

Введение

Исследование вопросов эффективности наполнения интернет-ресурсов началось и активно развивается параллельно с возникновением интернета и первых веб-сайтов [1]. Во время развития интернета сформировалось множество типов веб-сайтов, что связано, в первую очередь, со значительным их количеством. Типологию веб-сайтов можно провести как по объему: одностраничный сайт (landing), информационный сайт, веб-портал, форум, социальная сеть и т. д., так и по целевому назначению [2]. Целевое назначение веб-сайта формируется организацией или лицом, которые создали веб-сайт.

Сайт организации является источником информации для потенциальных клиентов и партнеров, который доступен в любое время. В связи с этим, вопросы оценки эффективности наполнения веб-сайта при решении различных управленческих задач приобретают ключевую роль [3]. В зависимости от типа организации наполнение сайта и, как следствие, его эффективность будут различаться. Создание универсальных метрик оценки эффективности сайта при решении управленческих задач существуют, однако не всегда отвечают поставленным критериям [4, 5].

В рамках данной работы рассмотрены веб-сайты образовательных организаций высшего образования. Сложность наполнения веб-сайта образовательной организации высшего образования заключается в необходимости одновременно учитывать требования учредителя и являться информативным информационным сайтом для абитуриентов. Анализ наполнения веб-сайтов образовательных организаций высшего образования позволил выделить задачу наполнения веб-сайта, которая отвечает требованиям учредителя и раскрывает особенности образовательных программ для абитуриентов. Такой задачей является размещение рабочих программ дисциплин на веб-сайте образовательной организации. Анализ наличия рабочих программ заключается в оценке содержания и проверке соответствия указанному названию дисциплины.

Оценка наличия документов в виде рабочих программ на веб-сайте и их анализ позволяют осуществить поддержку принятия решений об открытии или закрытии образовательных программ, что ключевым образом влияет на оценку эффективности работы образовательных подразделений. Независимой количественной метрикой оценки наполнения веб-сайта образовательным подразделением выступает количество уникальных посетителей раздела сайта, относящегося к подразделению, за период в 1 месяц. Применение указанной метрики объясняется интересом абитуриентов к образовательным подразделениям, которые наиболее полно представляют информацию по своим образовательным программам.

В настоящее время существуют подходы и инструменты для автоматического сбора информации с веб-сайтов [6], проверки текстовых документов [7], анализа текстовых данных [8]. Однако, следует отметить, что указанные подходы и инструменты не объединены едиными подходами к применению.

Таким образом, целью настоящей работы является формирование подходов для поддержки принятия решений при управлении образовательными программами, отличающихся автоматическим сбором, анализом и использованием текстовой информации, и позволяющих улучшить целевые метрики оценки эффективности наполнения веб-сайта организации.

Для достижения поставленной цели осуществлены проектирование и разработка информационной системы, которая обеспечивает постоянный контроль наличия документов и их проверку по эталонной базе документов. В рамках данной работы поставлены и решены следующие задачи:

– формирование схемы функционирования информационной системы анализа наполнения веб-сайта организации высшего образования;

– разработка контекстной и декомпозированной диаграммы информационной системы;

– формирование результатов представления работы информационной системы;

– проведение практической апробации разработанной информационной системы.

Материалы и методы

На Рисунке 1 представлена структурная схема информационной системы для поддержки принятия решений при управлении образовательными программами. Представленная схема учитывает основные этапы формирования данных, получаемых путем анализа текстовой информации с веб-сайта образовательной организации.

Рисунок 1 – Структурная схема информационной системы для поддержки принятия решений
Figure 1 – Structural diagram of the information system for decision support

В рамках проектирования информационной системы разработаны диаграммы в нотации IDEF0 [9, 10] и диаграмма действий [11]. Для представления работы информационной системы разработаны два типа диаграмм IDEF0: контекстная и декомпозированная. Контекстная диаграмма служит для отображения работы системы в целом, определяя ее границы и взаимодействие с внешней средой. Контекстная диаграмма представлена на Рисунке 2.

Рисунок 2 – Контекстная диаграмма IDEF0
Figure 2 – IDEF0 context diagram

Контекстная диаграмма IDEF0 содержит следующие компоненты:

1) систему, которая включает в себя ключевые процессы: парсинг сайта, обработку документов и учебных планов, сравнение и объединение данных, а также экспорт и визуализацию результатов;

2) входы: ссылка на сайт образовательного учреждения, файлы в форматах PDF и PLX/XML, содержащие учебные планы и нормативные документы, а также параметры базы данных для подключения, сохранения и выгрузки данных;

3) выходы: файлы CSV со структурированной информацией, данные, загруженные в базу данных, итоговые отчеты и графики, визуализирующие результаты проверки документов и их соответствие учебным планам;

4) управления: законы и регламенты, регулирующие образовательный процесс и определяющие требования к документам;

5) механизмы: модули программного обеспечения, регулирующие полный цикл работы системы.

Декомпозированная диаграмма IDEF0, представленная на Рисунке 3, детализирует процессы контекстной диаграммы, разделяя их на подпроцессы.

Одной из основных подзадач выделяется парсинг сайта. Модуль парсинга веб-сайта выполняет автоматизированный сбор данных с сайта вуза. На основе заданного URL система последовательно переходит по страницам, извлекает соответствующую информацию, а затем сохраняет структурированные данные в CSV-файл.

Модуль обработки документов анализирует результаты парсинга, скачивает PDF-документы по полученным ссылкам, извлекает текстовое содержимое и отыскивает дисциплины и практики. Результаты обработки сохраняются в отдельный CSV-файл для дальнейшего анализа.

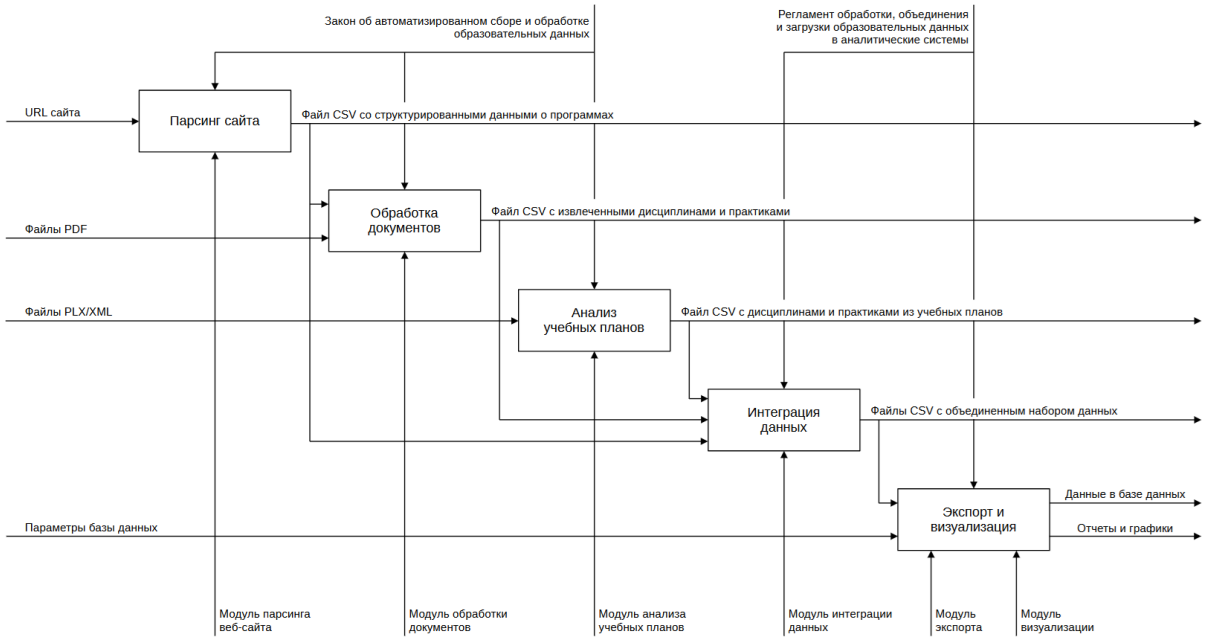

Рисунок 3 – Декомпозированная диаграмма IDEF0
Figure 3 – Decomposed IDEF0 diagram

Модуль анализа учебных планов обрабатывает файлы PLX/XML, содержащие эталонный перечень дисциплин и программ, предусмотренных учебными планами. После извлечения данных система формирует CSV-файл с полным списком документов дисциплин и практик, которые должны присутствовать на сайте.

Модуль интеграции данных сравнивает данные, полученные из PDF-файлов (т.е. фактическое наличие документов на сайте), с информацией из учебных планов (PLX/XML). В результате подобной работы формируются две итоговые таблицы: CSV-файл с перечнем отсутствующих и присутствующих документов и CSV-файл со сводной информацией о содержимом документов на страницах специальностей.

Модуль экспорта обеспечивает передачу итоговых данных в базу данных вуза с использованием заданных параметров подключения. Полученная информация передается в систему визуализации, где автоматически генерируются отчеты, графики и диаграммы для наглядного представления результатов проверки с помощью модуля визуализации.

На основе разработанных диаграмм IDEF0 продемонстрирована структура информационной системы: контекстная диаграмма определила границы и взаимодействие с внешней средой, а декомпозированная детализировала ключевые процессы. Благодаря этому обеспечена наглядность взаимодействия между компонентами системы, что упрощает ее дальнейшую реализацию и внедрение.

Для отображения процессов автоматизированного сбора, обработки и анализа данных о рабочих программах разработана диаграмма вариантов использования, представленная на Рисунке 4.

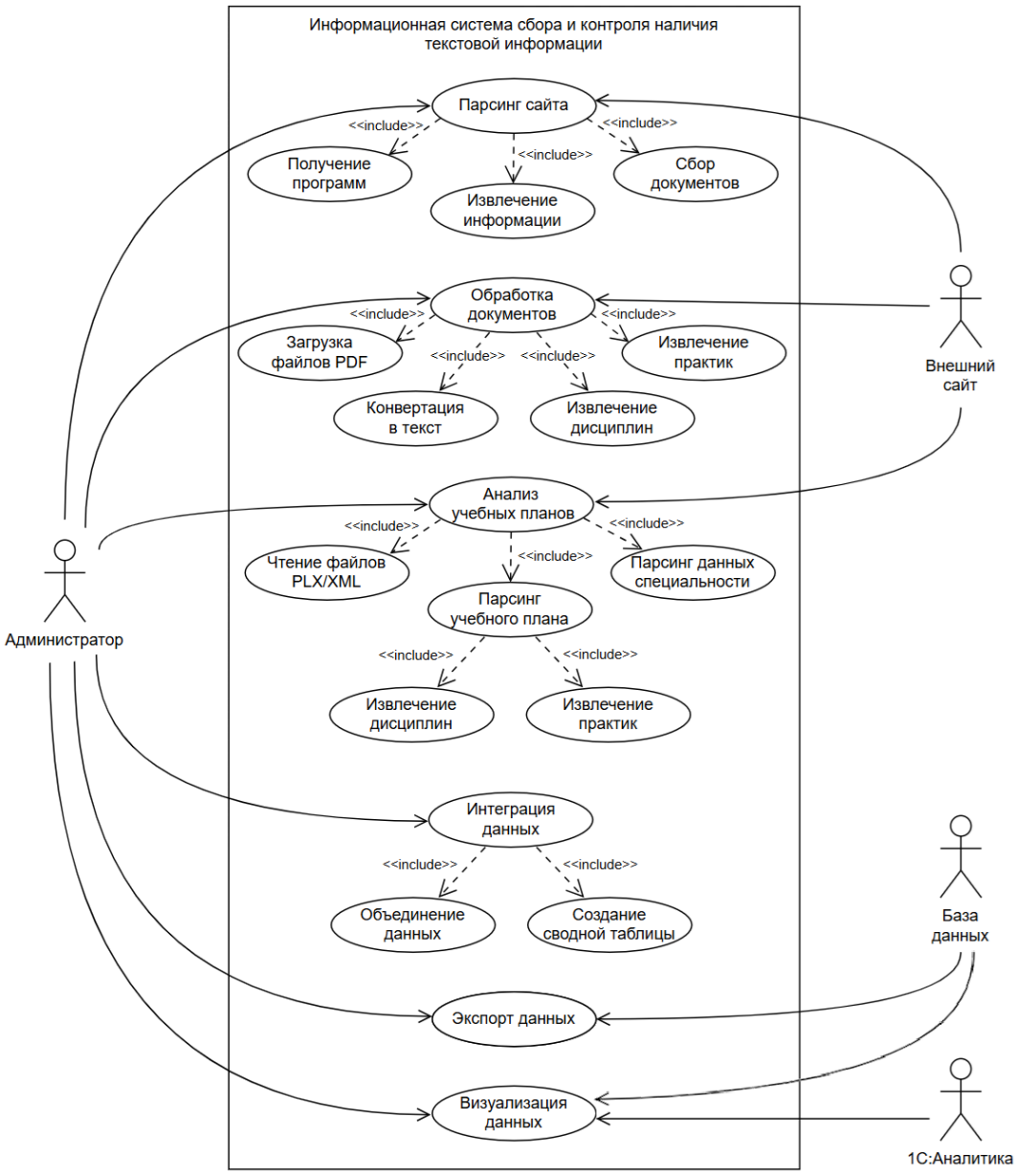

Рисунок 4 – Диаграмма вариантов использования
Figure 4 – Use case diagram

Из разработанной диаграммы вариантов использования видно, что в системе выделены следующие акторы:

1) администратор – отвечает за настройку, запуск и мониторинг работы системы. Администратор взаимодействует со всеми модулями, включая парсинг сайта, обработку документов, анализ учебных планов, интеграцию данных и визуализацию результатов;

2) внешний сайт вуза – источник данных, содержащий информацию об образовательных программах, факультетах, кафедрах, руководителях, а также документы дисциплин и практик и файлы учебных планов. Сайт предоставляет данные для парсинга и обработки собираемых документов;

3) база данных вуза – хранилище структурированной информации, куда загружаются результаты работы системы. База данных используется для хранения и последующего анализа данных;

4) система визуализации информации – внешняя система, предназначенная для визуализации данных. Она получает результаты анализа и формирует соответствующие отчеты, графики и диаграммы.

Диаграмма содержит следующие основные прецеденты:

– парсинг сайта – выполняется администратором и использует данные внешнего сайта. Данный прецедент получает список образовательных программ, извлекает информацию о программах, а также собирает ссылки на документы (рабочие программы, аннотации, программы практик);

– обработка документов – выполняется администратором, использует данные внешнего сайта. Такой вариант использования позволяет загружать PDF-файлы с сайта, конвертировать их в текстовый формат и извлекать из них дисциплины и практики;

– анализ учебных планов – выполняется администратором, использует заранее загруженные данные с внешнего сайта. Такой прецедент производит чтение учебных планов PLX/XML, отыскивает из файлов названия специальности и профиля и производит парсинг самого учебного плана (извлечение дисциплин и практик);

– интеграция данных – выполняется администратором. Благодаря такому варианту использования производится объединение данных, полученных с образовательного сайта и из учебных планов, а также создается сводная таблица с результатами проверки документов;

– экспорт данных – выполняется администратором, взаимодействует с базой данных вуза. Данный прецедент, в отличие от ранее рассмотренных, не включает в себя другие подчиненные прецеденты. Он занимается загрузкой структурированных данных в базу данных;

– визуализация данных – просматривается администратором, выполняется системой инструментами средств визуализации, используя данные из базы данных. Данный прецедент также не имеет включений. Его функционал предполагает автоматическую генерацию отчетов, графиков и диаграмм на основе полученных данных.

Благодаря спроектированной диаграмме вариантов использования удалось наглядно продемонстрировать взаимодействие между акторами системы и ее ключевыми процессами. Диаграмма подтвердила функциональную полноту решения, обеспечив четкое представление о работе системы и взаимодействии ее компонентов. Это значительно упрощает последующие этапы реализации и тестирования рассматриваемого решения.

В качестве инструментов практической реализации спроектированной системы выступили Python в качестве языка программирования; 1С: Аналитика в качестве платформы аналитики; Debian в качестве операционной системы сервера.

Результаты

Представленная на Рисунке 5 схема иллюстрирует целостный цикл принятия решений, основанный на автоматизированном анализе контента веб-сайта вуза.

Каждый сценарий транслируется в конкретное управленческое решение: от закрытия/расширения программы до материального поощрения кафедры или плана мероприятий по улучшению сайта. Все решения проходят через цикл обратной связи – мониторинг через 1 месяц. Повторный анализ определяет, улучшились ли метрики: если да – практики закрепляются и KPI обновляются; если нет – проводится корректировка решений и углубленный анализ факторов.

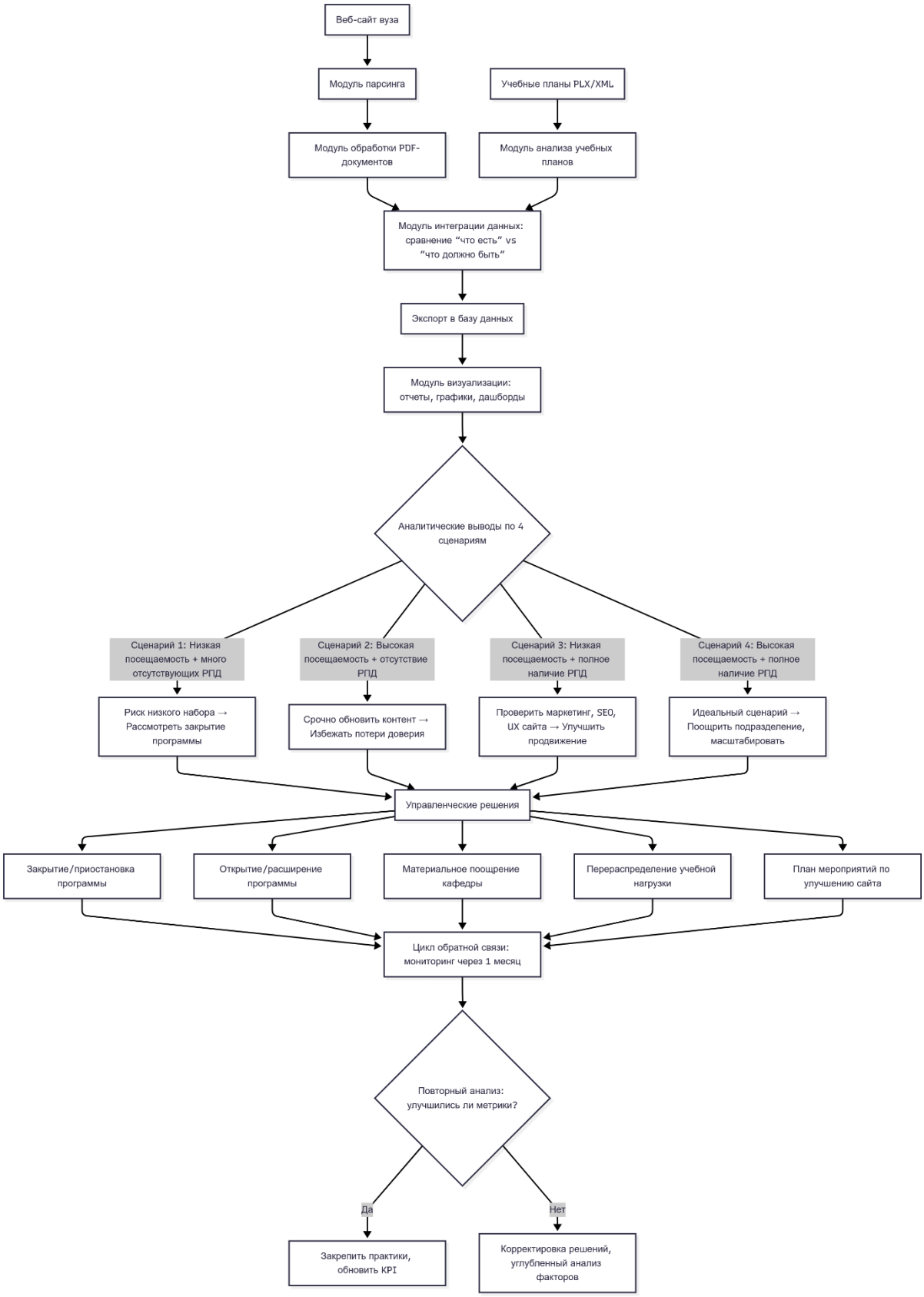

Рисунок 5 – Целостный цикл принятия решений
Figure 5 – Holistic decision-making cycle

Практическая апробация системы показала положительную динамику роста посещаемости целевых разделов сайта, что подтверждает эффективность предложенного подхода. Однако следует отметить, что посещаемость является комплексной метрикой, зависящей от множества внешних и внутренних факторов, требующих дальнейшего исследования для выявления причинно-следственных связей.

Заключение

В результате работы спроектирована и реализована информационная система анализа наполнения веб-сайта организации. Результаты работы информационной системы позволяют в автоматическом режиме собирать данные по наличию рабочих программ на сайте образовательной организации, что обеспечивает возможность применения данной информации в принятии решений при управлении образовательными программами в организации высшего образования. Управленческие решения на основе собранной информации могут затрагивать вопросы открытия и закрытия образовательных программ, распределения материального поощрения, изменения учебной нагрузки и т. д. Дальнейшее развитие информационной системы предполагает расширение собираемой информации и углубленный анализ содержания рабочих программ.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Brügger N. Website History and the Website as an Object of Study. *New Media & Society.* 2009;11(1-2):115–132. https://doi.org/10.1177/1461444808099574
2. Cebi S. Determining Importance Degrees of Website Design Parameters Based on Interactions and Types of Websites. *Decision Support Systems*. 2013;54(2):1030–1043. https://doi.org/10.1016/j.dss.2012.10.036
3. Вейс Л.Д., Живоглядов В.П. Информационная система поддержки принятия управленческих решений на основе ГИС и WEB-технологий. *Информатика и системы управления.* 2001;(2):50–57.
4. Ударцева О.М., Рыхторова А.Е. Использование инструментов веб-аналитики в оценке эффективности способов продвижения библиотечных ресурсов. *Библиосфера.* 2018;(2):93–99. https://doi.org/10.20913/1815-3186-2018-2-93-99
Udartseva O.M., Rykhtorova A.E. Using Web Analytics Tools to Assess the Effectiveness of Means for Promoting Library Resources. *Bibliosphere.* 2018;(2):93–99. (In Russ.). https://doi.org/10.20913/1815-3186-2018-2-93-99
5. Butkiewicz M., Madhyastha H.V., Sekar V. Understanding Website Complexity: Measurements, Metrics, and Implications. In: *IMC '11: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, 02–04 November 2011, Berlin, Germany.* New York: Association for Computing Machinery; 2011. P. 313–328. https://doi.org/10.1145/2068816.2068846
6. Khder M. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & Its Applications.* 2021;13(3):145–168.
7. Tang L., Laban Ph., Durrett G. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In: *EMNLP 2024: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 12–16 November 2024, Miami, FL, USA.* Association for Computational Linguistics; 2024. P. 8818–8847.
8. Apu K.U. AI-Driven Data Analytics and Automation: A Systematic Literature Review of Industry Applications. *Strategic Data Management and Innovation.* 2025;2(01):21–40.

9. Li X., Pu W., Zhao X. Agent Action Diagram: Toward a Model for Emergency Management System. *Simulation Modelling Practice and Theory.* 2019;94:66–99. https://doi.org/10.1016/j.simpat.2019.02.004
10. Van Rossum G., Drake F.L. *An Introduction to Python.* Bristol: Network Theory Ltd.; 2003. 164 p.
11. Mateos-Garcia J., Steinmueller W.E. The Institutions of Open Source Software: Examining the Debian Community. *Information Economics and Policy.* 2008;20(4):333–344. https://doi.org/10.1016/j.infoecopol.2008.06.001

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Поцебнева Ирина Валерьевна, кандидат технических наук, доцент кафедры систем управления и информационных технологий в строительстве, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: ipozebneva@cchgeu.ru
ORCID: 0000-0002-4659-0726

Irina V. Pocebneva, Candidate of Engineering Sciences, Docent at the Department of Control Systems and Information Technologies in Construction, Voronezh State Technical University, Voronezh, the Russian Federation.

Андреева Кристина Алексеевна, аспирант, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: kandreeva@cchgeu.ru
ORCID: 0009-0003-9318-8152

Kristina A. Andreeva, Postgraduate, Voronezh State Technical University, Voronezh, the Russian Federation.

Пуговкина Наталия Александровна, магистрант кафедры искусственного интеллекта и цифровых технологий, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: pugovka01112002@gmail.com
ORCID: 0009-0009-4132-0194

Natalia A. Pugovkina, Master's Degree student at the Department of Artificial Intelligence and Digital Technologies, Voronezh State Technical University, Voronezh, the Russian Federation.

Гусев Павел Юрьевич, доктор технических наук, доцент, заведующий кафедрой искусственного интеллекта и цифровых технологий, Воронежский государственный технический университет, Воронеж, Российская Федерация.
e-mail: gusevpvl@gmail.com
ORCID: 0000-0002-3752-0152

Pavel Y. Gusev, Doctor of Engineering Sciences, Docent, Head of the Department of Artificial Intelligence and Digital Technologies, Voronezh State Technical University, Voronezh, the Russian Federation.

Статья поступила в редакцию 02.09.2025; одобрена после рецензирования 05.11.2025; принята к публикации 12.11.2025.

The article was submitted 02.09.2025; approved after reviewing 05.11.2025; accepted for publication 12.11.2025.