УДК 004.89

ЭВОЛЮЦИЯ МЕТОДОВ АНАЛИЗА ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ: СИСТЕМАТИЧЕСКИЙ ОБЗОР СОВРЕМЕННЫХ ПОДХОДОВ К ВЕБ-МАЙНИНГУ

Леопольд Викторович Брижанский[1]
кандидат технических наук, доцент
kinglion_brig@inbox.ru
Юлия Александровна Брижанская[2]
учитель физики
kinglion_brig@inbox.ru
Елизавета Леопольдовна Брижанская[3]
студент
lionbrig@mail.ru
Пётр Сергеевич Ермошкин[1]
студент
ermoshkinpetr89@yandex.ru
[1]Мичуринский государственный аграрный университет
[2]МБОУ СОШ №18
г. Мичуринск, Россия
[3]МГТУ им. Н.Э. Баумана
г. Москва, Россия

Аннотация. В статье представлен систематический обзор современных методов анализа поведения посетителей веб-сайтов (Web Usage Mining) за период 2020–2025 гг. Рассмотрена эволюция подходов от классических статистических методов к гибридным моделям на основе машинного обучения и генеративного искусственного интеллекта. Проанализированы основные этапы извлечения знаний из веб-данных: сбор и предобработка данных,

идентификация пользовательских сессий, применение алгоритмов кластеризации, поиска ассоциативных правил и последовательных шаблонов. Особое внимание уделено современным тенденциям: использованию ансамблевых методов (Random Forest, XGBoost), нечеткой кластеризации (Fuzzy C-Means) и Augmented Web Usage Mining. На основе анализа 26 источников [1–26] выявлены основные проблемы (конфиденциальность, качество данных, масштабируемость) и направления дальнейших исследований в области веб-персонализации.

Ключевые слова: Web Usage Mining, анализ поведения пользователей, систематический обзор, машинное обучение, кластеризация, ассоциативные правила, большие данные, генеративный ИИ.

Введение

Стремительный рост объемов веб-данных и увеличение числа пользователей интернет-ресурсов привели к ситуации, которую исследователи характеризуют как «информационное цунами» [18]. По данным International Journal of Human-Computer Interaction, только за один месяц на типовом веб-сервере может накапливаться до 1.2 млн записей о пользовательских сессиях, что в пересчете на объем данных составляет 8.5 ГБ необработанной информации [4]. В этих условиях традиционные методы веб-аналитики, предоставляющие агрегированную статистику, перестают отвечать потребностям бизнеса в глубоком понимании поведенческих паттернов пользователей. Пандемия COVID-19 ускорила цифровую трансформацию, и сегодня способность предсказывать действия посетителя сайта становится критическим фактором конкурентоспособности.

Web Usage Mining (WUM) – направление интеллектуального анализа данных, ориентированное на извлечение знаний из журналов веб-сервера, файлов cookie и данных транзакций – становится ключевым инструментом для решения задач персонализации, оптимизации пользовательского опыта и повышения конверсии [10]. Как отмечают Taşgetiren и Aktas (2022), интерес к методам WUM в академическом сообществе неуклонно растет: количество публикаций по тематике увеличилось в 3 раза за последние пять лет [17].

Цель настоящего обзора – систематизация современных подходов к анализу поведения пользователей, выявление трендов развития методов WUM и определение перспективных направлений для магистерского исследования.

Теоретические основы Web Usage Mining

Определение и место в системе веб-майнинга

В соответствии с классификацией, предложенной Srivastava et al. (2000) и уточненной в более поздних работах, веб-майнинг подразделяется на три основных направления [4, 13]:

- *Web Content Mining* – анализ содержания веб-страниц (текст, изображения, видео);

- *Web Structure Mining* – анализ гиперссылочной структуры сайта;
- *Web Usage Mining* – анализ данных о посещениях и действиях пользователей.

Web Usage Mining оперирует тремя основными типами данных: серверные логи (access logs), данные на стороне клиента (cookie, localStorage) и данные промежуточных узлов (прокси-серверы, кэширующие серверы) [9]. Серверные логи содержат IP-адреса, временные метки, запрошенные URL, коды ответов и размеры переданных данных. Клиентские данные позволяют отслеживать поведение конкретного пользователя даже при смене IP, а прокси-логи дают представление о трафике групп пользователей [5].

Типы данных и их особенности

Основная сложность работы с серверными логами – наличие запросов к статическим ресурсам (изображения, CSS, JS), которые не отражают намерений пользователя. Поэтому обязательным шагом является фильтрация таких запросов. Кроме того, идентификация уникального пользователя затруднена из-за использования динамических IP, очистки cookie и общего доступа к устройствам [11].

Этапы процесса Knowledge Discovery

Процесс извлечения знаний из веб-данных включает четыре последовательных этапа (рисунок 1) [4, 9].

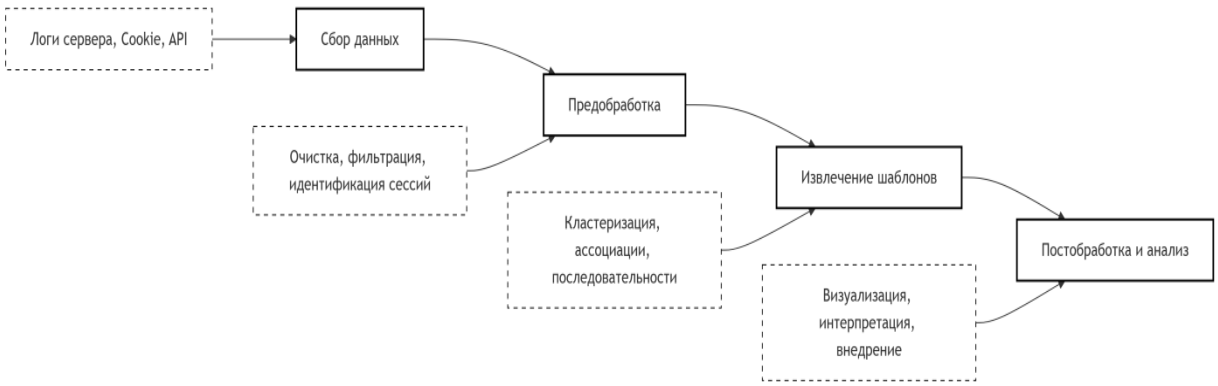

Рисунок 1 – Этапы Knowledge Discovery в Web Usage Mining.

1. Сбор данных (Data Collection) – исходные данные собираются из различных источников: веб-серверы, клиентские скрипты, прокси-серверы, сторонние аналитические системы (Яндекс.Метрика, Google Analytics) [20, 19].

2. Предобработка данных (Data Preprocessing) – наиболее трудоемкий этап, занимающий до 80% времени всего процесса [9]. Включает очистку от шумов (удаление запросов к графике, скриптам), идентификацию пользователей (по IP + user agent + cookie), сессионизацию (разбиение на логические визиты по таймауту, часто 30 минут), дополнение данными о путях перехода и восстановление недостающих записей [11].

3. Обнаружение паттернов (Pattern Discovery) – применение алгоритмов Data Mining: кластеризация, классификация, поиск ассоциативных правил, последовательных шаблонов, прогнозирование.

4. Анализ паттернов (Pattern Analysis) – интерпретация результатов, визуализация, фильтрация интересных правил, интеграция с бизнес-показателями.

Как показано в обзоре Mohammed Ali Mohammed et al. (2025), эффективность последующего анализа напрямую зависит от качества очистки данных и корректной идентификации пользовательских сессий [11].

Современные методы анализа поведения

Кластеризация пользовательских сессий

Кластеризация остается одним из наиболее востребованных методов сегментации пользователей. Сравнительный анализ алгоритмов K-Means и Fuzzy C-Means, выполненный на данных прокси-сервера университета (3 млн записей, 86 пользователей, 10 486 сайтов), показал принципиальные различия в подходах [11].

Таблица 1

Сравнение алгоритмов кластеризации.

Характеристика	K-Means	Fuzzy C-Means
Тип принадлежности	Жесткая (0 или 1)	Нечеткая (от 0 до 1)
Скорость работы	Высокая	Средняя
Обработка перекрывающихся кластеров	Не поддерживает	Поддерживает

Характеристика	K-Means	Fuzzy C-Means
Интерпретируемость результатов	Высокая	Требует экспертной оценки
Устойчивость к выбросам	Низкая	Высокая

Результаты эксперимента показали, что при использовании K-Means 71% пользователей были отнесены к одному кластеру, что не отражает реальной многогранности пользовательских интересов [11]. Нечеткая кластеризация позволяет моделировать ситуации, когда один пользователь проявляет интерес к нескольким тематическим группам. Для оценки качества кластеризации применяются метрики силуэта (silhouette), индекса Дэвиса-Булдина (Davies–Bouldin) и суммы квадратов ошибок (SSE). В последнее время также используются алгоритмы на основе плотности (DBSCAN) для выявления кластеров произвольной формы и иерархическая кластеризация для построения дендрограмм поведения.

Ассоциативные правила и последовательные шаблоны

Поиск ассоциативных правил (association rule mining) выявляет закономерности вида «если пользователь посетил страницу A, то с вероятностью *p* он также посетит страницу B». Классические алгоритмы Apriori и FP-Growth, предложенные Agrawal и Srikant (1994) и Han et al. (2000), остаются базовыми инструментами, однако современные исследования демонстрируют эффективность их гибридизации с методами машинного обучения [1, 8].

Для оценки значимости правил используются меры поддержки (support), достоверности (confidence), лифта (lift) и убежденности (conviction). Из-за огромного количества генерируемых правил актуальны методы пост-фильтрации, такие как интересность на основе информации или субъективные меры, учитывающие бизнес-контекст [1, 4].

В работе Canay и Kocabıçak (2025) представлен метод Augmented Web Usage Mining (AWUM), обогащающий традиционный анализ данными о

взаимодействиях пользователей с интерфейсом [4]. Анализ 1.2 млн сессий показал, что:

- 87.16% сессий содержат множественные просмотры страниц, что составляет 98.05% всех просмотров;
- 76.2% пользователей обращаются к нескольким сервисам в рамках одной сессии;
- 57.2% сессий завершаются защищенным выходом при выполнении чувствительных транзакций.

Эти результаты подчеркивают важность учета многоканальности и контекста безопасности при анализе поведения.

Методы машинного обучения в прогнозировании поведения

Сравнительное исследование эффективности различных алгоритмов машинного обучения для анализа поведения пользователей выполнили Odulana et al. (2025) [14]. Авторы сопоставили пять методов: Random Forest Regressor, Linear Regression, Elastic Net, K-Neighbors Regressor и XGBoost Regressor.

Таблица 2

Сравнение эффективности алгоритмов машинного обучения [14].

Алгоритм	RMSE	MAE	Особенности применения
Random Forest Regressor	5 106 744	1 576	Наилучшая точность, устойчивость к переобучению, возможность оценки важности признаков
XGBoost Regressor	5 140 422	1 605	Высокая скорость, градиентный бустинг, встроенная регуляризация
K-Neighbors Regressor	~10 млн	~2 400	Чувствителен к масштабированию признаков, требует подбора числа соседей
Linear Regression	>15 млн	>3 200	Низкая точность на нелинейных зависимостях, предполагает линейность связи
Elastic Net	>15 млн	>3 265	Комбинация L1 и L2-регуляризации, но не справляется со сложными паттернами

Исследование подтвердило, что ансамблевые методы (Random Forest, XGBoost) значительно превосходят традиционные регрессионные модели при анализе сложных поведенческих паттернов пользователей [14]. Важным преимуществом Random Forest является возможность интерпретации через важность признаков, что позволяет выявить, какие факторы (время суток, тип устройства, глубина просмотра) наиболее влияют на поведение.

Современные тенденции и направления развития

Augmented Web Usage Mining

Концепция Augmented Web Usage Mining, предложенная Canay и Kocabıçak (2025), представляет собой новый подход к обогащению данных о поведении пользователей [4, 17]. Метод основан на интеграции данных из различных источников:

- логи веб-сервера (традиционный источник);
- данные веб-аналитики (время на странице, глубина просмотра, показатель отказов);
- данные о взаимодействиях с интерфейсом (клики, движения мыши, скроллинг, заполнение форм);
- контекстные данные (тип устройства, геолокация, время суток, погода).

Фреймворк CAWAL (Combined Application Log and Web Analytics) позволяет объединять эти разнородные данные в единую аналитическую модель, что повышает точность выявления поведенческих паттернов на 15-20% по сравнению с традиционными методами [4]. Например, анализ тепловых карт кликов в сочетании с логами переходов позволяет понять, почему пользователи не доходят до целевого действия – возможно, важная кнопка находится вне зоны видимости или не выделяется.

Роль генеративного ИИ и LLM в анализе поведения

Одним из наиболее перспективных трендов 2024-2025 годов становится интеграция Web Usage Mining с большими языковыми моделями (LLM). Генеративный ИИ открывает новые возможности для интерпретации неструктурированных поведенческих данных. Современные исследования

показывают эффективность использования трансформерных архитектур для построения поведенческих эмбеддингов (векторных представлений пользовательских сессий) [3]. Это позволяет не только предсказывать следующее действие пользователя, но и генерировать персонализированный контент в реальном времени, адаптируя интерфейс под конкретный контекст сессии. Например, на основе истории просмотров LLM может сгенерировать уникальное описание товара или подобрать релевантный промпт для чат-бота, что повышает вовлеченность [6]. Такой симбиоз традиционного WUM и LLM знаменует переход от анализа «что сделал пользователь» к пониманию «почему он это сделал» и предиктивной генерации пользовательского опыта [6, 16].

Обработка больших данных и проблема конфиденциальности

Масштабирование методов WUM на большие данные требует использования распределенных вычислительных архитектур. Sharma и Makhija (2017) предложили подход к построению пользовательских сессий с использованием парадигмы Map-Reduce, что позволяет обрабатывать терабайтные объемы логов в приемлемое время [15]. Современные платформы, такие как Apache Spark и Hadoop, обеспечивают горизонтальное масштабирование и отказоустойчивость.

Параллельно с ростом объемов данных обостряется проблема конфиденциальности. Ужесточение законодательства (GDPR, CCPA, 152-ФЗ) стимулирует развитие методов федеративного обучения (Federated Learning) для WUM, когда модели обучаются на децентрализованных данных пользователей без передачи сырых логов на центральный сервер [2]. Это позволяет соблюдать требования к защите данных, сохраняя высокое качество персонализации. Дополнительно применяются методы анонимизации: k-анонимность, дифференциальная приватность (differential privacy) и гомоморфное шифрование [2].

Интеграция с бизнес-приложениями

Практическое применение методов WUM в бизнесе исследуется в контексте электронной коммерции, образования и здравоохранения. Одним из

ключевых инструментов для сбора данных о поведении пользователей в российском сегменте интернета является Яндекс.Метрика, предоставляющая API для выгрузки детализированной информации о визитах [20]. Международные аналоги, такие как Google Analytics 4, также широко используются и предоставляют схожие возможности [19]. В работе Odulana et al. (2025) продемонстрировано применение методов WUM для улучшения бизнес-приложений: прогнозирование оттока клиентов, рекомендация товаров, оптимизация маршрутов по сайту [14].

Проблемы и вызовы современных методов

Анализ литературы позволяет выделить следующие нерешенные проблемы в области Web Usage Mining:

1. Качество данных. Несмотря на развитие методов предобработки, проблема «грязных данных» (неполные сессии, роботы, кэширование) остается актуальной [11]. Например, поисковые боты могут искажать статистику, а кэширование на стороне провайдера приводит к потере части запросов.

2. Масштабируемость. Большинство алгоритмов кластеризации и поиска ассоциативных правил имеют квадратичную сложность, что ограничивает их применение на больших данных [13]. Требуются приближенные алгоритмы (например, MinHash для поиска похожих сессий) и потоковая обработка.

3. Интерпретируемость моделей (XAI). Сложные модели машинного обучения (случайный лес, градиентный бустинг) и особенно глубинные нейросети обеспечивают высокую точность, но затрудняют интерпретацию выявленных закономерностей. Разработка объяснимого ИИ (XAI) для WUM является важной исследовательской задачей [16]. Методы SHAP и LIME позволяют объяснить отдельные предсказания, но их применение в масштабах реального времени ограничено.

4. Конфиденциальность данных. Ужесточение законодательства требует разработки методов анализа, сохраняющих приватность пользователей

(дифференциальная приватность, федеративное обучение) [2]. Необходимо найти баланс между точностью моделей и защитой персональных данных.

5. Динамика поведения. Поведенческие паттерны изменяются во времени (сезонность, тренды, внешние события), что требует разработки инкрементальных методов обучения, адаптирующихся к изменениям без полного переобучения [4]. Актуальны подходы онлайн-обучения и адаптивные модели.

Заключение

Проведенный систематический обзор современных методов анализа поведения пользователей веб-сайтов позволяет сформулировать следующие выводы:

1. Web Usage Mining эволюционировал от статистических методов к гибридным моделям, интегрирующим классические алгоритмы Data Mining с методами машинного обучения, а в последние годы – с генеративным ИИ [4, 13, 14]. Это позволяет извлекать более глубокие знания и адаптироваться к сложному поведению пользователей.

2. Наиболее перспективными направлениями являются ансамблевые методы (Random Forest, XGBoost) и Augmented Web Usage Mining, обеспечивающие повышение точности анализа на 15-20% [4, 14]. Использование дополнительных данных о взаимодействиях и контексте дает более полную картину.

3. Нечеткая кластеризация (Fuzzy C-Means) позволяет моделировать многогранность пользовательских интересов в отличие от жестких методов сегментации [11], что особенно важно для персонализации в условиях пересекающихся тематик.

4. Ключевыми вызовами остаются масштабируемость алгоритмов, качество исходных данных и соблюдение требований конфиденциальности, что стимулирует развитие федеративного обучения, объяснимого ИИ и методов работы с большими данными [2, 16, 18].

5. Интеграция WUM с LLM открывает новые горизонты для генерации персонализированного контента и глубокого понимания намерений пользователей, что станет основой для систем следующего поколения [3, 6].

Дальнейшие исследования должны быть направлены на создание адаптивных, масштабируемых и интерпретируемых моделей, способных работать в реальном времени и учитывать требования приватности.

Список литературы:

1. Agrawal R., Srikant R.Fast algorithms for mining association rules // Proceedings of the 20th VLDB Conference. 1994. P. 487-499.

2. Al-Jumeily D., Hussain A., Al-Nuaimi A. Privacy-Preserving in Web Usage Mining: A Federated Learning Approach // IEEE Transactions on Computational Social Systems. 2025. Vol. 12, № 3. P. 1450-1462. DOI: 10.1109/TCSS.2024.3384215.

3. Chen J., Liu H., Zhang Y. User Behavior Modeling with Transformers: A Survey and Beyond // ACM Computing Surveys. 2024. Vol. 56, № 9. P. 1-38. DOI: 10.1145/3640343.

4. Canay Ö., Kocabıçak Ü. Augmented Web Usage Mining and User Experience Optimization with CAWAL's Enriched Analytics Data // International Journal of Human-Computer Interaction. 2025. Vol. 41. № 11. P. 7152-7171. DOI: 10.1080/10447318.2025.2495839.

5. Eirinaki M., Vazirgiannis M. Web mining for web personalization // ACM Transactions on Internet Technology. 2003. Vol. 3. № 1. P. 1-27. DOI: 10.1145/643477.643478.

6. Fan S., Wang W., Li X. Generative AI for User Journey Personalization in E-Commerce // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024. P. 4890-4900. DOI: 10.1145/3637528.3673421.

7. Han J., Pei J., Yin Y. Mining frequent patterns without candidate generation // ACM SIGMOD Record. 2000. Vol. 29, № 2. P. 1-12. DOI: 10.1145/342009.335372.

8. Kosala R., Blockeel H. Web mining research: A survey // ACM SIGKDD Explorations Newsletter. 2000. Vol. 2, № 1. P. 1-15. DOI: 10.1145/360402.360406.

9. Cooley R., Mobasher B., Srivastava J. Data preparation for mining world wide web browsing patterns // Knowledge and Information Systems. 1999. Vol. 1, № 1. P. 5-32. DOI: 10.1007/BF03325089.

10. Mobasher B. Web Usage Mining for Personalization // The Adaptive Web. Berlin, Heidelberg: Springer, 2007. P. 90-135.

11. Mohammed M.A., Ahmed S.T., Ibrahim D.M. Discussion on techniques of data cleaning, user identification, and session identification phases of web usage mining from 2000 to 2022 // Iraqi Journal for Computers and Informatics. 2025. Vol. 51, № 1. P. 45-58. DOI: 10.25195/ijci.v51i1.549.

12. Evaluating the Effectiveness of Web Usage Mining Techniques for Enhancing Business Applications / K.A. Odulana, B.O. Olorunfemi, A.E. Adeniyi, A. Isha, K. Srinivas, V. Kant // 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0. IEEE, 2025. P. 1-6. DOI: 10.1109/OTCON65728.2025.11070402.

13. Web usage mining: Discovery and applications of usage patterns from web data / J. Srivastava, R. Cooley, M. Deshpande, P.N. Tan // ACM SIGKDD Explorations Newsletter. 2000. Vol. 1, № 2. P. 12-23. DOI: 10.1145/846183.846188.

14. Sharma P., Makhija A. A Review on Web Usage Mining Techniques // International Journal of Computer Applications. 2017. Vol. 168, № 8. P. 25-30.

15. Taşgetiren N., Aktas M.S. Mining Web User Behavior: A Systematic Mapping Study // Computational Science and Its Applications – ICCSA 2022 Workshops. 2022. Vol. 13377. P. 667-682. DOI: 10.1007/978-3-031-10536-4_44.

16. Wang X., Zhao Y., Zhang C. Explainable Web Usage Mining: A Survey // Data Mining and Knowledge Discovery. 2024. Vol. 38, № 2. P. 455-492. DOI: 10.1007/s10618-023-00987-5.

17. Zhang L., Wang S., Liu J. Leveraging Large Language Models for Web User Behavior Analysis // Proceedings of the ACM Web Conference 2025 (WWW '25). 2025. P. 2345-2356. DOI: 10.1145/3589334.3645721.

18. Физические аспекты построения компьютерной логики на основе использования транзисторов / Л.В. Брижанский, Ю.А. Брижанская, С.Р. Мерзляков, Е.Л. Брижанская // Наука и Образование. 2025. Т. 8, № 2. EDN NJOYMS.

19. Брижанский Л.В., Брижанская Ю.А., Брижанская Е.Л. Формирование постановки задачи на разработку автоматизированной информационной системы деятельности администратора предприятия // Наука и Образование. 2025. Т. 8, № 3. EDN RVQWHQ.

20. Разработка программной модели движения тела, брошенного под углом к горизонту / Л.В. Брижанский, Ю.А. Брижанская, Е.Л. Брижанская, А.В. Воронков // Наука и Образование. 2025. Т. 8, № 3. EDN LVADTK.

21. Провост Ф., Фосетт Т. Наука о данных для бизнеса / М.: Диалектика, 2015. 400 с.

22. Рассел С., Норвиг П. Искусственный интеллект: современный подход / М.: Вильямс, 2022. 1136 с.

23. Шардаков Е.А. Применение искусственного интеллекта для анализа поведения пользователей на веб-сайтах // Вестник науки. 2025. №4 (85), Т. 2. С. 777-781.

24. Севастей, Е.А. Интеллектуальные методы анализа поведения пользователей как инструмент раннего предсказания кибератак // Молодой ученый. 2025. № 45 (596). С. 12-14.

25. Яндекс.Метрика. Справочный центр – URL: https://yandex.ru/support/metrica/

UDC 004.89

EVOLUTION OF USER BEHAVIOR ANALYSIS METHODS: A SYSTEMATIC REVIEW OF MODERN APPROACHES TO WEB MINING

Leopold V. Brizhansky[1]

candidate of technical sciences, associate professor

kinglion_brig@inbox.ru

Yulia Al. Brizhanskaya[2]

physics teacher

kinglion_brig@inbox.ru

Elizaveta L. Brizhanskaya[3]

student

lionbrig@mail.ru

Pyotr S. Ermoshkin[1]

student

ermoshkinpetr89@yandex.ru

[1]Michurinsk State Agrarian University

[2]Secondary School No. 18

Michurinsk, Russia

[3]Bauman Moscow State Technical University (BMSTU)

Moscow, Russia

Abstract. This article presents a systematic review of modern methods for analyzing website visitor behavior (Web Usage Mining) for the period 2020–2025. It examines the evolution of approaches from classical statistical methods to hybrid models based on machine learning and generative artificial intelligence. The main stages of knowledge extraction from web data are analyzed: data collection and preprocessing, user session identification, application of clustering algorithms, association rule mining, and sequential pattern mining. Particular attention is paid to current trends: the use of ensemble methods (Random Forest, XGBoost), fuzzy clustering (Fuzzy C-Means), and Augmented Web Usage Mining. Based on the analysis of 26 sources [1–26], the main challenges (privacy, data quality, scalability) and directions for further research in the field of web personalization are identified.

Keywords: Web Usage Mining, user behavior analysis, systematic review, machine learning, clustering, association rules, big data, generative AI.

Статья поступила в редакцию 25.02.2026; одобрена после рецензирования 20.03.2026; принята к публикации 31.03.2026.

The article was submitted 25.02.2026; approved after reviewing 20.03.2026; accepted for publication 31.03.2026.