Молодёжный вестник НФ БГТУ. 2025. Том 05. № 02 (18)
https://rio-nb-bstu.science

ОТРАСЛЕВАЯ СТРУКТУРА ЭКОНОМИКИ, ЭКОНОМИКА И ОРГАНИЗАЦИЯ ПРЕДПРИЯТИЯ

doi: 10.51639/2713-0576_2025_5_2_09
Научная статья
УДК 338: 001.895
ГРНТИ 06.54.31
ВАК 5.2.3

Способы парсинга и обоснование целесообразности их применения к отдельной социальной сети

Олеся Николаевна Панамарева [1]*, Владислав Ринатович Хусаинов [2], Никита Васильевич Зайцев [3]

[1, 2] *Военный инновационный технополис «ЭРА», Анапа, Россия,* *era_otd1@mil.ru
[3] *Войсковая часть 55060, Москва, Россия,* 55060-406@mil.ru

Аннотация

Объемы информации в цифровой форме растут экспоненциально, при этом существенно сокращается время на принятие управленческих решений, что сопровождается увеличением рисков, связанных с качеством и достаточностью данных, получаемых из различных источников. Особую роль в обеспечении безопасности, достижении устойчивости сложных организационно-технических систем, функционирующих как в гражданской, так и в военной сферах, в их развитии на сбалансированной инновационной основе, играет информация, содержащаяся и полученная в том числе и из социальных сетей. Контент, представленный в социальных сетях, может оказывать, наряду с положительным, и негативное влияние. Сегодня наблюдается большой интерес к его использованию со стороны агентов, хозяйствующих в различных отраслях экономики, а также в силовых ведомствах и в сфере ответственности оборонно-промышленного комплекса. В современных условиях обладание информацией, очищенной от излишнего шума и агрегированной, является залогом эффективности принимаемых решений, обеспечения устойчивости и безопасности. В данной научной работе раскрываются актуальные вопросы, связанные с автоматизированным сбором информации из такого рода источников. Обоснована актуальность парсинга новостных материалов из социальных сетей, представлены результаты анализа существующего инструментария парсинга. Сделан акцент на необходимости разработки отечественных решений в обозначенной области, что станет одной из важных составляющих фундамента обеспечения технологического и экономического суверенитета России. Выделены основные проблемы, возникающие при решении данной задачи, и способы их нивелирования. Приведены результаты оценки применимости средств автоматизированного сбора информации на примере социальной сети «Одноклассники».

Ключевые слова: информация, парсинг, автоматизация, инновации, технологический и экономический суверенитет, безопасность

Введение

Интенсивное развитие инфокоммуникационных технологий, технологий искусственного интеллекта, других ИТ-решений, исследованию которых уделяется достаточно большое внимание [1-4], нацелено на обеспечение формирования основы технологического и экономического национального суверенитета, безопасности и

устойчивости экономики России. При этом инновационная составляющая, сформированная и совершенствующаяся за счет и на базе отечественных разработок, является ключевым аспектом. Освещению особенностей создания инновационной экономики посвящен целый пласт работ отечественных ученых [5-8].

Сегодня мы наблюдаем разительный рост мирового объема данных (эксперты SaaSworthy отмечают его удвоение каждые два года [9]). Это неизбежно приводит к возрастающей потребности в автоматизации обработки информации и ее очистке от избыточного шума [10].

В современном обществе информация, а также инструменты ее обработки, передачи и хранения, приобрели беспрецедентно важное значение. В XXI веке она стала одним из ключевых ресурсов, сравнимым по важности с человеческими, финансовыми и материальными ресурсами, т.е. пятым фактором производства.

Мир вступил в эпоху цифровой экономики и сетевых сообществ, где сбор информации из интернета стал играть критически важную роль, особенно при обеспечении экономической, информационной и других видов безопасности и формировании условий устойчивого социально-экономического развития на динамичной инновационной основе.

По состоянию на 2024 г. в мире количество пользователей Интернета составило 5,52 млрд человек (это порядка 68 % всего населения планеты) [11], увеличившись в 1,4 раза с 2018 г. Среднее ежедневное время, проведенное людьми в Интернете, составило 6 часов и 36 минут. По состоянию на 2024 г. зарегистрировано 5,22 млрд пользователей социальных сетей [12] (рис. 1).

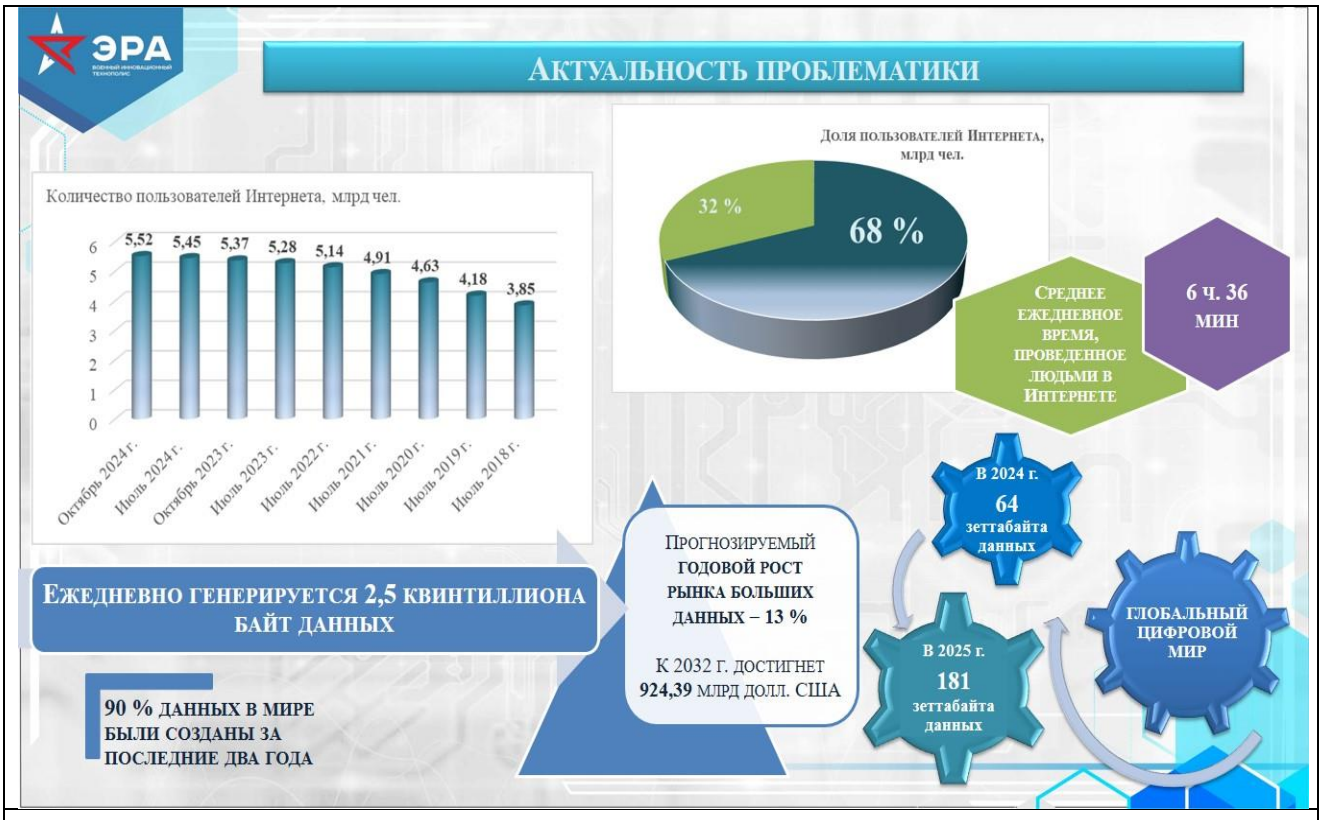

Рисунок 1 – Основные статистические данные в области цифрового развития общества в мире
Примечание: составлено Панамаревой О.Н. по данным источников [9, 11]

Таким образом, интернет, будучи практически неограниченным источником данных, представляет собой сложную среду, где обработка текстовых массивов вручную становится практически невозможной, что также нашло отражение в работе [13].

Это подчеркивает необходимость решения насущной научно-практической задачи – автоматизации процессов сбора и обработки информации из интернет-источников.

Обоснование актуальности разработки инструментов для автоматизированного сбора информации из интернет-источников

В мире наблюдается экспоненциальный рост объема цифровых данных, при этом порядка 70 % контента в мире генерируется пользователями [9], в т.ч.:

1) текстовые сообщения по электронной почте;
2) сообщения в социальных сетях;
3) блоги;
4) видео;
5) аудио;
6) журналы вызовов;
7) обзоры;
8) отзывы клиентов;
9) ответы на анкеты.

При этом примерно 90 % данных по всему миру не структурированы.

В октябре 2024 г. число активных авторов в социальных медиа (соцсетях, мессенджерах, форумах, блогах, маркетплейсах, геосервисах, отзовиках, UGC-площадкам и др.) в РФ составило 74,9 млн. Авторами было написано 1,815 млрд публичных сообщений (в т.ч. постов, репостов, комментариев). По сравнению с осенью (октябрем) 2023 г. наблюдался значительный прирост как активных авторов (порядка 16 %), так и объема создаваемого ими контента (примерно 17 %) [14].

В России можно выделить как наиболее популярные соцсети «Telegram», «ВКонтакте», «Одноклассники», «Instagram»*, «Facebook»* (*Facebook и *Instagram признаны экстремистскими организациями и запрещены на территории РФ), «Youtube», «X (Twitter)», «TikTok». Наибольший рост объема контента продемонстрировали «Одноклассники» (на 127 % до 154 млн сообщений) и «Дзен» (на 154,7 % до 18 млн), а Telegram (на 18,9 % до 1109,5 млн), Rutube (на 9,4 % до 5,2 млн).

Интерес аудитории к социальным сетям продолжает расти (количество пользователей в РФ порядка 119 млн, в КНР – 1070 млн). Они оказывают существенное влияние на настроение в обществе, затрагивая вопросы обеспечения национальной безопасности.

Именно для нивелирования гибридных угроз, связанных в т.ч. и с распространяемым контентом, а также для решения проблемных аспектов (снижения трудоемкости, ресурсозатратности поиска и обеспечения полноты информации) разрабатываются специальные методы автоматизированного сбора информации. Для нашей страны этот процесс новый и в первую очередь связан с обеспечением ее технологического и экономического суверенитета.

Актуальность решения обозначенной выше научно-практической задачи связана с рядом следующих аспектов:

1. Традиционный (т.е. в ручном режиме) поиск информации в интернете – это чрезвычайно медленный и трудоемкий процесс. Значительная часть времени тратится на поиск и последующее изучение найденных материалов.

2. Поисковые запросы по ключевым словам зачастую выдают огромные массивы данных, изучение которых может занять колоссально большое количество времени. При этом нет гарантии полноты охвата информации, что также критично в условиях ограниченных временных ресурсов.

3. Как для гражданской, так и военной сфер высока значимость информации, своевременность ее получения в достаточном объеме для анализа, оперативности разработки и эффективности принятия управленческих решений, для нивелирования гибридных угроз и др.

Для решения задачи эффективного поиска нужной информации в интернете разрабатываются методы, основанные на принципах поиска в базах данных и текстовых документах [15].

Парсинг (англ. parsing) – процесс автоматического извлечения и структурирования данных из открытых источников с помощью специальных программ. Он позволяет значительно ускорить и оптимизировать сбор информации, используя боты, скрипты и другие автоматизированные инструменты, обрабатывать большие объемы данных без применения ручного труда, нивелируя большую часть ошибок, связанных с человеческим фактором [16].

В настоящее время существует три основных способа автоматизированного сбора данных:

– сбор данных с помощью API (Application Programming Interface);

– сбор данных с помощью семантического разбора веб-страниц;

– сбор данных с помощью средств эмуляции поведения пользователя в браузере [17].

В рамках настоящей работы на базе общетеоретических методов и метода декомпозиции исследованы особенности обозначенных способов, результаты представлены ниже.

Описание способов автоматизированного сбора информации из интернет-источников

1. Сбор данных с помощью API.

API (Application Programming Interface, программный интерфейс) – это набор инструментов и протоколов, позволяющих различным программам и сервисам взаимодействовать друг с другом. Он предоставляет готовые блоки кода, функции и структуры данных, которые можно использовать для создания собственных приложений или интеграции с другими системами [18, 19].

API упрощает процесс получения информации из базы данных, позволяя отправлять HTTP-запросы к специальному серверу. Вместо изучения сложной структуры базы данных разработчики могут просто отправлять запросы и получать данные в стандартизированном формате. Правила и синтаксис запросов, а также формат возвращаемых данных определяются самим сервисом, предоставляющим API.

2. Сбор данных с помощью семантического разбора веб-страниц.

Этот метод включает последовательное изучение синтаксиса данных на вебсайтах.

Для анализа HTML используются:

– регулярные выражения (они предлагают высокую гибкость и настраиваемость, но требуют детальной разработки каждого шаблона; однако, использование только регулярных выражений может быть сложным, поскольку требует детальной разработки каждого шаблона, что не только усложняет задачу для программиста, но и увеличивает вычислительную нагрузку на систему);

– библиотеки BeautifulSoup и lxml (считаются оптимальными инструментами для эффективного анализа веб-страниц благодаря своей простоте использования; выбор в пользу какой-либо из них в основном зависит от личных предпочтений разработчика).

3. Сбор данных с помощью эмуляции поведения пользователя.

Кроме того автоматизированный сбор данных возможен с помощью средств эмуляции поведения пользователя в браузере. Одним из ключевых инструментов при этом является Selenium – проект с открытым исходным кодом (open source), включающий несколько

продуктов (в т.ч. Selenium WebDriver, Selenium RC, Selenium Server, Selenium Grid, Selenium IDE [17]). Некорректно обозначать любой из пяти продуктов семейства Selenium просто как «Selenium», однако это часто происходит (когда из контекста ясно, о каком продукте идет речь, или когда обсуждаются несколько или все продукты вместе). Среди них наиболее известен Selenium WebDriver, который представляет собой коллекцию библиотек для автоматизации действий в браузерах (т.е. целый набор драйверов для различных браузеров и клиентские библиотеки на разных языках программирования для взаимодействия с этими драйверами).

Эти библиотеки отправляют HTTP-запросы по протоколу JsonWireProtocol, в которых указываются инструкции для браузера в текущей сессии, такие как: навигация по URL, поиск элементов и взаимодействие с ними, а также парсинг контента страницы. Selenium WebDriver особенно эффективен для сайтов с динамическим контентом, так как запускает настоящий браузер, делая автоматизированную активность неотличимой от действий реального пользователя. Он запускает настоящий браузер для доступа к вебсайтам, что делает его активность неотличимой от действий реального пользователя. Когда страница загружается через Selenium WebDriver, браузер загружает все ресурсы и исполняет JavaScript, как если бы это делал пользователь [20].

Каждый из рассмотренных способов автоматизированного сбора информации имеет свои преимущества и недостатки (табл. 1), составленная с использованием результатов исследований, представленных в источниках [18-22]).

Исходя из результатов сравнения методов автоматизированного сбора информации, был сделан выбор наиболее оптимального способа, применительно к социальной сети «Одноклассники», т.е. способ с использованием API самой социальной сети на базе http-запросов, поскольку:

1. Не требуется взаимодействие с объектами, динамически генерируемыми JavaScript.
2. Социальная сеть «Одноклассники» сама имеет открытый API.
3. Этот метод обеспечивает простой и авторизованный доступ к информации.

Таким образом, парсинг с помощью http-запросов обеспечит простой и авторизованный доступ к информации, представленной на веб-страницах социальной сети «Одноклассники».

При этом следует понимать, что несмотря на все плюсы данного способа при парсинге придется столкнуться с рядом проблемных аспектов.

Проблемы автоматизированного сбора информации и способы их решения

Многие системы сбора информации прибегают к использованию взломанных или автоматически созданных аккаунтов для получения доступа к данным, защищенным от неавторизованного просмотра.

Для сбора общедоступной информации применяются веб-роботы. Эти программы посещают вебсайты и извлекают с них всю доступную информацию, включая метаданные, а также при необходимости скачивают статические и динамические ресурсы.

Разработка профессиональных веб-роботов – это сложный процесс, требующий учета множества факторов. Это обстоятельство затрудняет создание универсальных систем, поскольку решения, эффективные для одного ресурса, могут оказаться непригодными для другого.

Итак, основные проблемы, которые стоят перед разработчиками такого рода программных продуктов, можно классифицировать следующим образом:

1) проблема индивидуальной настройки: для каждого вебсайта требуется ручная настройка и отладка системы автоматизированного взаимодействия, особенно если ресурс имеет сложную структуру и верстку контента;

Таблица 1 – Преимущества и недостатки методов автоматизированного сбора информации

Методы автоматизированного сбора информации	Сбор информации с помощью *API*	Сбор информации с помощью семантического анализа страницы	Сбор информации с помощью средств автоматизации браузера
Преимущества	– простота использования и понимания – стандартное использование *HTTP* методов и кодов состояния – масштабируемость и гибкость – возможность использования веб-кэширования для оптимизации производительности	– автоматизация действий пользователя в браузере (от перехода по ссылке, до нажатия кнопок и ввода данных) – использование различных библиотек для обработки кода веб-страницы *(lxml-parser, html-parser*, собственные библиотеки) – доступ к страницам с закрытым *API* и возможность сбора информации с них	– полная имитация действий человека (от перехода по ссылке, до нажатия кнопок и ввода данных) – поддержка динамического контента, генерируемого *JavaScript* – естественность взаимодействия обеспечивается применением *Selenium WebDriver*
Недостатки	– ограниченность предоставляемых данных из соцсети (не все данные отдаются, которые видны пользователю) – отсутствие строгой спецификации, приводящее к разнообразию реализаций и сложностям в согласованности интерфейсов	– отсутствие поддержки *JavaScript* (нет возможности обрабатывать объекты, им сгенерированные) – отсутствие поддержки библиотеки *XPath* – язык запросов к элементам *XML*-документов – необходимость разработки алгоритмов обхода блокировок ботов, блокировок по *IP*-адресам	– высокие требования к ресурсам системы (к ее мощности) – низкая скорость работы (из-за ожидания полной загрузки страницы и выполнения *JavaScript* кода)

2) проблема обеспечения высокой производительности: сложные веб-роботы должны быть способны обрабатывать значительные объемы данных за ограниченное время, что особенно актуально при сборе часто обновляющейся информации (например, данных о рынке, аналитических отчетов и др.), а также при выполнении большого количества взаимосвязанных действий;

3) проблема необходимости адаптации: вебсайты подвержены постоянным изменениям в структуре и способах представления контента, в связи чем требуется регулярное обновление ИТ-решения для его адаптации к этим изменениям; кроме того, при одновременной работе с несколькими ресурсами эта задача становится весьма трудоемкой и зачастую требует вмешательства человека [23].

Для преодоления первой проблемы применяются методы адаптивного парсинга веб-страниц и извлечения слабоструктурированной информации. Однако, эти технологии имеют

свои ограничения и связаны с повышенными затратами на внедрение и настройку алгоритмов. Извлечение структурированных данных с веб-страниц сводится к решению следующих задач:

– определение и доступ к целевым страницам для извлечения информации (проблема навигации);

– идентификация областей, содержащих необходимые данные (проблема распознавания данных);

– выявление структуры обнаруженных данных (проблема поиска общей структуры данных);

– гарантия единообразия извлекаемых данных (проблема сопоставления атрибутов извлекаемых данных);

– консолидация данных из различных источников (проблема объединения данных) [24].

Как указано выше, вторая проблема заключается в недостаточной технической мощности. Процессы поиска, сбора и фильтрации данных предъявляют определенные требования к аппаратному обеспечению (т.е. объем памяти, механизмы отбора данных по заданным критериям, вычислительные ресурсы, способные обрабатывать большие объемы и потоки информации). Одним из решений данной проблемы является сужение области поиска, что, впрочем, может сказаться на конечном результате. Процесс обработки данных, как правило, состоит из трех этапов, на каждом из которых реализуется определенный набор задач. Состав этапов может меняться в зависимости от целей обработки:

1. Первый этап включает совокупность алгоритмов и процессов, необходимых для первичной подготовки данных к последующей обработке [15]:

– данные собираются из различных источников, которые могут быть представлены, например, перечнем веб-ресурсов;

– в зависимости от источников и способов получения данных формируются требования к их валидации;

– валидация информации необходима для исключения ошибок в исходных данных, которые в дальнейшем будут обрабатываться на следующих этапах;

– контроль может быть реализован путем простого сравнения данных.

После успешного сбора и проверки данные передаются на второй этап, где происходит основная обработка.

2. Второй этап включает набор алгоритмов и процессов, ориентированных на преобразование информации в удобный, стандартизированный формат.

Преобразование информации не подразумевает изменения ее смыслового содержания, а служит для форматирования данных с целью обеспечения удобства и эффективности их использования. В зависимости от поставленной задачи на данном этапе могут быть выделены наиболее значимые и релевантные данные. В ходе анализа могут быть представлены результаты преобразования и поиска запрашиваемой информации.

3. Третий этап – заключительный. На данном этапе, при необходимости, может осуществляться удаление избыточной информации, которая может мешать анализу данных, в зависимости от используемого алгоритма и поставленной задачи. Затем производится финальная проверка валидности обработанной информации, ее вывод и передача на хранение и дальнейшее использование.

В целом, три этапа процесса сбора и обработки данных требуют значительных вычислительных мощностей и предъявляют высокие требования к системе хранения данных.

Что касается третьей проблемы, то отметим, что на сегодняшний день не существует единого универсального решения, которое обеспечивало бы устойчивость системы извлечения данных к изменениям верстки веб-страниц, поэтому поиск эффективных подходов продолжается. Тем не менее, применение отдельных методов извлечения данных позволяет существенно снизить влияние данной проблемы на качество процесса извлечения.

Речь идет о ручных и полуавтоматических методах сбора информации с веб-страниц, которые, в отличие от полностью автоматизированного (интеллектуального) подхода к извлечению данных, предполагают определенную степень участия человека в рассматриваемом процессе.

Современные проблемы интеллектуального извлечения данных из веб-страниц имеют не только технический, но и правовой аспект. Последний связан с различиями в законодательстве об авторском праве и других правовых нормах на национальном и международном уровнях. Претензии в данной области являются вполне обоснованными, что обуславливает необходимость соблюдения этических норм при решении задач по сбору данных с вебсайтов [15].

Заключение

Стремительный рост объема данных в мире и возрастающая роль информации в современном обществе (ее влияние на национальную безопасность, устойчивость сложных организационно-технических систем разного уровня управления, их инновационное развитие и др.) обуславливают неэффективность ручного сбора данных. Такой традиционно применяемый метод поиска информации трудоемок, медленный и часто приводят к получению огромного объема нерелевантных данных, обработка, анализ которых достаточно затруднена.

В связи с этим автоматизированный сбор информации становится критически важным инструментом для хозяйствующих агентов в области бизнеса, науки, военного дела (российские концерны оборонной промышленности также ведут свои официальные аккаунты в социальных сетях; лидером в этой области является «Роскомос» [25]) и других сфер деятельности). Сервисы автоматизированного сбора информации позволяют эффективно собирать, обрабатывать и анализировать большие объемы данных из различных источников, включая социальные сети (в т.ч. из ставшей популярной сети «Одноклассники»). Развитие отечественных методов сбора информации открывает новые возможности для извлечения ценной информации, принятия обоснованных решений, нивелирования гибридных угроз и/или их последствий.

Однако, при автоматизированном сборе информации сталкиваются с рядом проблем, таких как необходимость адаптации к изменениям структуры сайтов, обеспечение высокой производительности программно-аппаратных комплексов и учет этических вопросов, связанных с использованием данных. Решение этих проблем требует разработки сложных алгоритмов и применения современных отечественных технологий, особенно в условиях беспрецедентного санкционного давления на Россию. Следовательно, разработка такого рода задач – важная научно-практическая проблема, решить которую предстоит в ближайшее время.

Конфликт интересов

Авторы статьи заявляют, что на момент подачи статьи в редакцию, у них нет возможного конфликта интересов с третьими лицами.

Список источников

1. Морозов А.В., Панамарев Г.Е., Гусеница Я.Н. Состояние и перспективы развития современной науки в области информационно-телекоммуникационных технологий в Военном инновационном технополисе «ЭРА» // Сб. статей II научно-технической конференции «Состояние и перспективы развития современной науки по направлению «ИТ-технологии». Т. 3. Высокопроизводительные вычислительные комплексы

и суперкомпьютерное моделирование в военно-научном сопровождении жизненного цикла вооружения, военной и специальной техники. – Анапа: ВИТ «ЭРА». – 2023. – С. 7-18.
2. Морозов А.В., Панамарев Г.Е. Вопросы защиты информации при применении технологий искусственного интеллекта: опыт Военного инновационного технополиса «ЭРА» // Вопросы защиты информации при применении технологий искусственного интеллекта на аппаратно-программных платформах российского и иностранного производства: сб. материалов круглого стола научно-деловой программы Международного военно-технического форума «АРМИЯ-2024», Кубинка, Московская область, 13 августа 2024 года. – Анапа: ФГАУ «Военный инновационный технополис «ЭРА». – 2024. – С. 4-12.
3. Пучков А.А., Панамарев Г.Е., Сень Г.А., Ивановский В.С. Искусственный интеллект в информационной безопасности // Состояние и перспективы развития современной науки по направлению «АСУ, информационно-телекоммуникационные системы»: сб. статей II Всероссийской научно-технической конференции, Анапа, 18 июня 2020 года / Военный инновационный технополис «ЭРА». Т. 3. – Анапа: ФГАУ «Военный инновационный технополис «ЭРА». – 2020. – С. 8-11.
4. Панамарева О.Н., Панамарев Г.Е., Шафеев А.А. Технологии искусственного интеллекта в АСУ предприятиями и комплексами // Наука в современном обществе: закономерности и тенденции развития: сб. статей международной научно-практической конференции: в 2 частях, Пермь, 25 февраля 2017 года. – Том Часть 1. – Пермь: ООО «Аэтерна». – 2017. – С. 96-102.
5. Ракова Н.Г., Балашова Е.С. Инновационная экономика как фактор повышения устойчивости (технологической безопасности) страны и благополучия населения // Счисляевские чтения: актуальные проблемы экономики и управления. – 2024. – № 12 (12). – С. 300-303.
6. Санжина О.П., Смирнов А.Ю. Принципы формирования механизма управления инновациями в современных условиях // Естественно-гуманитарные исследования. – 2024. – № 2(52). – С. 228-230.
7. Смирнов А.Ю. Развитие инновационной деятельности в России и факторы, ей препятствующие // Актуальные проблемы экономики и менеджмента. – 2023. – № 2 (38). – С. 50-57.
8. Кох Л.В., Кох Ю.В., Санжина О.П. Стратегическое управление цифровой трансформацией интеллектуальной экономики и промышленности в новой реальности: монография. – СПб. – 2024. – С. 315-343.
9. Big Data Statistics 2025: Growth and Market Data. By Naveen Kumar . November 13, 2024. – URL: https://www.demandsage.com/big-data-statistics/ (дата обращения: 12.03.2025).
10. Дубовик Т.С., Березовская Е.М. Автоматизация сбора данных с веб-ресурсов // Молодежная наука в XXI веке: традиции, инновации, векторы развития: материалы Международной научно-исследовательской конференции молодых ученых, аспирантов, студентов и старшеклассников: в 3 ч. Самара-Оренбург, 05 апреля 2017 г. Том Часть 1. – Самара-Оренбург: ООО «Аэтерна». – 2017. – С. 202-203. – URL: https://www.elibrary.ru/item.asp?id=30036788 (дата обращения: 17.02.2025).
11. How Many Use The Internet in 2025 (Statistics). Naveen Kumar / January 22, 2025. – URL: https://www.demandsage.com/internet-user-statistics/ (дата обращения: 12.03.2025).
12. 64 Social Media Statistics 2025 – Users & Growth. Naveen Kumar / December 26, 2024. . – URL: https://www.demandsage.com/social-media-marketing-statistics/ (дата обращения: 12.03.2025).
13. Закалин И.Ю. Автоматизация сбора информации в сети интернет // Вестник магистратуры. 2018. №5-4 (80). – URL: https://cyberleninka.ru/article/n/avtomatizatsiya-sbora-informatsii-v-seti-internet (дата обращения: 21.02.2025).

14. Социальные сети в России: цифры и тренды, осень 2024. – URL: https://brandanalytics.ru/blog/social-media-russia-autumn-2024 (дата обращения: 14.02.2025).
15. Костяшин Н.А., Колбина О.Н., Яготинцева Н.В. Применение автоматизированных средств сбора информации по сайтам // Информационные технологии и системы: управление, экономика, транспорт, право. – 2020. – № 3(39). – С. 11-17. – URL: https://www.elibrary.ru/item.asp?id=44383882 (дата обращения: 25.02.2025).
16. Меньшиков Я.С. Преимущества автоматического сбора данных в сети интернет над ручным сбором данных // Universum: технические науки: электрон. научн. журн. 2022. – URL: https://cyberleninka.ru/article/n/preimuschestva-avtomaticheskogo-sbora-dannyh-v-seti-internet-nad-ruchnym-sborom-dannyh/viewer (дата обращения: 19.01.2025).
17. Суханов А.А., Маратканов А.С. Анализ способов сбора социальных данных из сети Интернет // International scientific review. 2017. – URL: https://cyberleninka.ru/article/n/analiz-sposobov-sbora-sotsialnyh-dannyh-iz-seti-internet/viewer (дата обращения: 19.02.2025).
18. Что такое API и как он работает. – URL: https://skillbox.ru/media/code/chto_takoe_api/ (дата обращения: 12.03.2025).
19. Что такое API и что о нём нужно знать веб-разработчику. – URL: https://practicum.yandex.ru/blog/chto-takoe-api/ (дата обращения: 12.03.2025).
20. Веб-скрейпинг с нуля на Python: библиотека Beautiful Soup. – URL: https://nuancesprog.ru/p/14171/ (дата обращения: 12.03.2025).
21. Веб-скрейпинг с Python: Полное руководство. – URL: https://vc.ru/u/2726106-swift-stream/1012395-veb-skreiping-s-python-polnoe-rukovodstvo (дата обращения: 12.03.2025).
22. Москаленко А.А., Лапонина О.Р., Сухомлин В.А. Разработка приложения веб-скрапинга с возможностями обхода блокировок // Современные информационные технологии и ИТ-образование. 2019. – URL: https://cyberleninka.ru/article/n/razrabotka-prilozheniya-veb-skrapinga-s-vozmozhnostyami-obhoda-blokirovok (дата обращения: 19.02.2025).
23. Эшонкулов Х.И. Проблемы автоматизированного сбора информации // Вестник науки и образования. 2021. № 11-2 (114). – URL: https://cyberleninka.ru/article/n/problemy-avtomatizirovannogo-sbora-informatsii (дата обращения: 10.02.2025).
24. Коляда А.С., Гогунский В.Д. Извлечение информации из слабоструктурированных веб-страниц // ВЕЖПТ. 2014. № 9 (67). – URL: https://cyberleninka.ru/article/n/izvlechenie-informatsii-iz-slabostrukturirovannyh-veb-stranits (дата обращения: 12.02.2025).
25. Как ведут соцсети крупнейшие оборонные концерны. Исследование SMM-активностей в российском ОПК. – URL: https://www.cossa.ru/trends/324353/ (дата обращения: 12.02.2025).

Parsing methods and justification of the feasibility of their application to a particular social network

Olesya Nikolaevna Panamareva [1]*, Khusainov Vladislav Rinatovich [2] ,
Zaytcev Nikita Vasilyevich [3]

[1*, 2] Innovativ Technopolis «ERA», Anapa, Russia, *era_otd1@mil.ru
[3] Military Unit 55060, Moscow, Russia, 55060-406@mil.ru

Abstract

The volumes of information in digital form are growing exponentially. At the same time, the time for making management decisions is significantly reduced, which is accompanied by an increase in risks associated with the quality and sufficiency of data obtained from various sources. Information contained and received, including from social networks, plays a special role in ensuring security, achieving the sustainability of complex organizational and technical systems operating in both the civil and military spheres, in their development on a balanced innovative basis. The content

presented in social networks can have a negative impact along with a positive one. Today, there is a great interest in its use on the part of agents operating in various sectors of the economy, including those related to law enforcement agencies and the sphere of responsibility of the military-industrial complex. In modern conditions, the possession of information, cleared of unnecessary noise and aggregated, is the key to the effectiveness of decisions made, ensuring sustainability and security. This scientific work reveals current issues related to automated collection of such information. The relevance of parsing news materials from social networks is substantiated, the results of the analysis of the existing parsing tools are presented. The emphasis is placed on the need to develop and apply domestic solutions in this area, which will become one of the important components of the foundation for ensuring the technological and economic sovereignty of Russia. The main problems arising in solving the designated problem and ways to level them are highlighted. The results of assessing the applicability of automated information collection tools, such as innovations, are presented using the example of the social network «Odnoklassniki».

Keywords: information, collection, data collection, social networks, automation, innovation, economic and technological sovereignty, decision-making efficiency, sustainability, security