DOI 10.37882/2223-2966.2025.02.21

АНАЛИЗ ЦИФРОВОГО СЛЕДА ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ ВКОНТАКТЕ С ИСПОЛЬЗОВАНИЕМ ВОЗМОЖНОСТЕЙ VK API

ANALYSIS OF THE DIGITAL FOOTPRINT OF VKONTAKTE USERS USING THE VK API

I. Murashkin
M. Aliev

Summary. The study focuses on investigating the digital footprints of VKontakte users by leveraging VK API tools. The research aims to develop a systematic approach to data analysis for identifying user preferences and audience segmentation. Key stages of working with open data, such as data collection, preprocessing, and analysis, are explored in detail. The integration of clustering techniques, thematic analysis, and natural language processing (NLP) significantly reduced manual intervention and streamlined the research process. Findings highlight the practical potential of digital footprint analysis in constructing detailed user profiles applicable in marketing, HR analytics, and social research. The study emphasizes the critical importance of adhering to ethical and legal standards during data analysis.

Keywords: digital footprint, VK API, data processing, clustering, personalization, social platforms.

Мурашкин Илья Николаевич
инженер по обеспечению качества (QA) полного стека (VK), магистрант, Адыгейского государственного университета (АГУ), Майкоп
iluxa9494@gmail.com
Алиев Марат Вячеславович
кандидат физико-математических наук, Адыгейский государственный университет (АГУ), Майкоп
alievmarat@mail.ru

Аннотация. Данная статья посвящена исследованию особенностей цифрового следа пользователей социальной сети ВКонтакте с использованием функционала VK API. Цель работы заключается в создании методологического подхода к анализу данных, направленного на выявление предпочтений пользователей и сегментацию аудитории. В материале рассматриваются ключевые этапы работы с открытыми данными, включая их сбор, обработку и дальнейший анализ. Применение методов кластеризации, тематического анализа и алгоритмов обработки естественного языка (NLP) позволило минимизировать участие человека и значительно повысить эффективность исследования. Результаты демонстрируют возможность использования цифрового следа для формирования точных пользовательских профилей, что находит применение в маркетинговых стратегиях, HR-аналитике и социальных исследованиях. В работе также акцентируется внимание на важности соблюдения законодательных и этических норм при анализе данных.

Ключевые слова: цифровой след, VK API, обработка данных, кластеризация, персонализация, социальные платформы.

Введение

Социальные сети, включая ВКонтакте, стали значимым источником данных о предпочтениях и активности пользователей. Инструменты VK API позволяют собирать информацию о профилях, подписках и активности, но работа с данными ограничена требованиями конфиденциальности, лимитами API и качеством информации. Это требует разработки эффективных методов анализа данных. Цифровой след пользователей выступает важным инструментом для сегментации аудитории и прогнозирования её поведения. Методы кластеризации (например, K-Means) и обработка естественного языка (NLP) в сочетании с API позволяют глубже анализировать текстовые данные и выявлять предпочтения. Цель исследования — изучение цифрового следа пользователей ВКонтакте и определение их предпочтений на основе анализа данных, извлечённых через VK API. Примеры применения анализа данных включают повышение эффективности рекламных кампаний через персонализацию, адаптацию образовательных программ и определение актуальных тем для культурных мероприятий. Для достижения цели изучены возможности VK API, разработаны методы анализа данных, автоматизированы процессы сбора информации с учётом этических стандартов. Научная новизна работы заключается в предложении подхода, который интегрирует методы обработки естественного языка, кластеризацию и API для автоматизации анализа данных социальных сетей. Практическая значимость исследования связана с повышением эффективности маркетинговых кампаний, выявлением интересов аудитории для HR-аналитики и определением трендов в социальной аналитике. Работа охватывает теоретические основы VK API, используемую методологию, результаты, практическую значимость и перспективы дальнейших исследований.

Обзор теоретических основ и возможностей VK API

Цифровой след — это данные, оставленные пользователями в интернете, включая профили, подписки,

лайки и комментарии. Анализ этих данных позволяет сегментировать аудиторию, прогнозировать поведение и разрабатывать персонализированные стратегии. Исследования подтверждают значимость цифрового следа как инструмента изучения предпочтений и социального взаимодействия. Персонализация контента является одной из ключевых областей применения цифрового следа. Chen и Xu (2023) предложили методы извлечения пользовательских интересов на основе машинного обучения и тематического анализа, демонстрирующие высокую точность сегментации. Garcia и Schweitzer (2022) показали, что социальные взаимодействия и цифровой след могут предсказывать вовлечённость аудитории. Brown и Green (2023) подчеркнули важность соблюдения принципов конфиденциальности при использовании обезличенных данных, что актуально для работы с платформами, такими как VK API. Williams и Brown (2021) разработали техники автоматизированного анализа больших данных, минимизирующие участие человека. Garcia и Schweitzer (2022) также предложили методы анализа социальных графов, позволяющие выявлять ключевые фигуры и строить сетевые модели, что полезно в маркетинге и HR-аналитике. ВКонтакте предоставляет доступ к данным через VK API, который позволяет исследователям строить поведенческие профили. Методы API, такие как users.get и wall.get, дают доступ к информации о пользователях, публикациях и комментариях. Эти данные используются для анализа предпочтений и выделения ключевых тем. Несмотря на ограничения, включая доступ только к публичным данным, лимиты запросов и необходимость адаптации методов к изменениям API, VK API остаётся мощным инструментом для исследования цифрового следа. Комбинация методов анализа, таких как кластеризация и NLP, позволяет создавать точные профили пользователей. Применение зарубежных подходов, интегрированных с VK API, открывает новые возможности для изучения поведения аудитории. Таким образом, анализ цифрового следа через VK API остаётся перспективным направлением, обеспечивая высокую точность данных при соблюдении этических и технических стандартов.

Методология исследования

Для достижения целей исследования применялся структурированный подход, включающий этапы сбора, предобработки, анализа и интерпретации данных. VK API использовался для получения открытой информации о пользователях, включая профили, подписки и публикации. Реализация на языке программирования Java обеспечивала высокую производительность и автоматизацию. Критерии выборки включали возраст пользователей до 25 лет, интересы в технической, гуманитарной и экономической сферах, а также активность в сети. Пример запроса к API для извлечения данных:

```
String url = «https://api.vk.com/method/users.get» +
 «?user_ids=12345» +
 «&fields=sex,bdate,city,interests» +
 «&access_token=YOUR_ACCESS_TOKEN» +
 «&v=5.131»;
HttpClient client = HttpClient.newHttpClient();
HttpRequest request = HttpRequest.newBuilder()
 .uri(URI.create(url))
 .build();
HttpResponse<String> response =
      client.send(request,      HttpResponse.BodyHandlers.
ofString());
System.out.println(«Ответ: « + response.body());
```

Данные проходили предобработку, включавшую очистку (удаление дубликатов и неполных записей), анонимизацию (замена идентификаторов обезличенными метками) и унификацию (преобразование текстовых значений в числовые). Например, для преобразования даты рождения в возраст применялись встроенные библиотеки Java. Анализ данных включал тематический анализ, кластеризацию и статистическую обработку. Методы NLP, такие как TF-IDF, использовались для выделения ключевых слов и тем публикаций. Алгоритм K-Means применялся для группировки пользователей по интересам, а оптимальное количество кластеров определялось эмпирически. Статистический анализ охватывал распределение пользователей по возрасту, полу и интересам. Техническая реализация включала модули сбора, предобработки и анализа данных. Модуль сбора взаимодействовал с VK API, обеспечивая автоматизацию. Модуль предобработки занимался очисткой и анонимизацией, а модуль анализа объединял тематическое моделирование, кластеризацию и статистическую обработку. Данные сохранялись в PostgreSQL для обеспечения консистентности. Архитектура системы предусматривала строгую последовательность этапов: сбор, предобработка, анализ и визуализация данных. Интерпретация результатов включала визуализацию с помощью диаграмм и графиков, демонстрировавших распределение интересов и активности пользователей. Полученные данные использовались для формулирования выводов о поведенческих моделях и предпочтениях. Система, реализованная на Java, обеспечивает адаптацию для различных исследовательских задач, высокую точность анализа и воспроизводимость, что делает предложенную методологию применимой в аналогичных исследованиях.

Результаты исследования

В рамках исследования была сформирована выборка из 500 пользователей социальной сети ВКонтакте. Возрастной состав показал, что 87 % участников младше 25 лет. Интересы распределились следующим образом: технические темы, такие как программирование, кос-

мос и технологии, привлекли 45 %; гуманитарные темы, включая литературу, искусство и историю, заинтересовали 32 %; экономические темы, такие как финансы и маркетинг, оказались важны для 23 %. В плане активности 68 % пользователей подписаны на тематические группы, а 52 % публиковали записи. Эти данные подтверждают разнообразие интересов и активности аудитории, предоставляя основу для дальнейшего анализа.

Результаты представлены на рисунке 1.

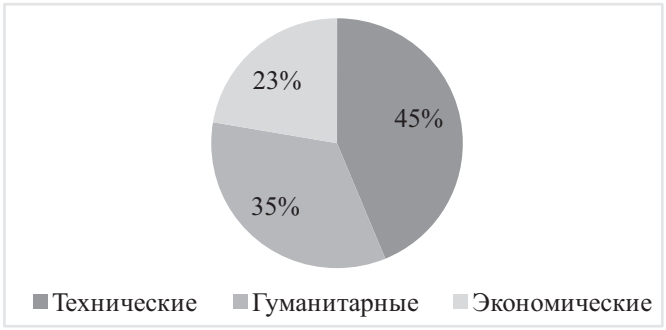

Рис. 1. Диаграмма распределения интересов пользователей

В рамках исследования был проведён процесс кластеризации пользователей на основе их интересов с использованием алгоритма K-Means, что позволило выделить три основных кластера. Первый, технический кластер, включал 45 % пользователей, интересующихся программированием, технологиями и космосом. Второй, гуманитарный кластер, составили 32 % участников, чьи интересы связаны с литературой, искусством и историей. Третий, экономический кластер, объединил 23 % пользователей, увлечённых финансами, маркетингом и предпринимательством. Для выполнения кластеризации ключевыми признаками стали подписки на группы и использование характерных ключевых слов. Кластеризация пользователей с использованием алгоритма K-Means выделила три основных кластера. Технический кластер включал 45 % пользователей, интересующихся программированием и космосом. Гуманитарный кластер составили 32 % участников, чьи интересы связаны с литературой и искусством. Экономический кластер объединил 23 % пользователей, увлечённых финансами и предпринимательством. Для кластеризации использовались подписки и характерные ключевые слова. Тематический анализ текстов публикаций выявил основные темы в каждом кластере. В техническом доминировали темы программирования, робототехники и космоса. В гуманитарном ключевыми были литература, музыка и искусство. В экономическом — финансы, инвестиции и бизнес. Это позволило глубже понять предпочтения каждой группы и определить направления их активности.

Пример программы для извлечения ключевых слов из текста:

```
public static String[] extractKeywords(String text) {
   Set<String> stopWords = Set.of(«и», «это», «на», «в», «с», «а», «о», «по», «за», «но», «что», «у», «мы», «вы», «он», «она», «они», «я», «ты»);
    return Arrays.stream(text.toLowerCase().replaceAll(«[^a-яа-z0-9 ]», «»).split(«\\s+»))
            .filter(word -> !stopWords.contains(word) && word.length() > 2)
                .collect(Collectors.groupingBy(word -> word, Collectors.counting()))
            .entrySet().stream()
                .sorted((e1, e2) -> e2.getValue().compareTo(e1.getValue()))
            .map(Map.Entry::getKey)
            .limit(5)
            .toArray(String[]::new);
}
```

Статистический анализ показал, что 67 % выборки составляют пользователи в возрасте 18–22 лет. Популярные группы включают «Программирование для всех», «Космические технологии», «Литература и искусство» и «Финансовая грамотность». Эти данные отражают интересы и активность аудитории, подчёркивая её разнообразие.

Теоретическая и практическая значимость

Исследование расширяет научное понимание цифрового следа как инструмента изучения личностных характеристик. Основной вклад заключается в разработке универсальной методологии, объединяющей статистический анализ, тематическое моделирование и анализ социальной активности. Этот подход может быть адаптирован для изучения данных на других платформах, открывая новые возможности для анализа взаимосвязей в пользовательской активности. Практическая значимость анализа данных ВКонтакте заключается в возможности формировать целевые аудитории для рекламных кампаний с учётом возраста, интересов и географии пользователей. Это позволяет персонализировать сообщения, например, предлагать образовательные курсы для программистов, и оптимизировать расходы на рекламу за счёт узкой сегментации и повышения конверсии. Цифровой след также помогает выявлять тренды, оценивать социальные предпочтения и планировать образовательные, культурные или спортивные мероприятия на основе подписок и активности пользователей. В HR-аналитике это позволяет оценивать интересы кандидатов, создавать портреты сотрудников и разрабатывать программы обучения и профессионального роста. Анализ VK API автоматизирует коммуникации, включая отправку персонализированных сообщений. Например, можно использовать следующий код для приглашения пользователя на вебинар:

```
String url = «https://api.vk.com/method/messages.send» +
 «?user_id=12345» +
 «&message=Здравствуйте! Приглашаем Вас на наш вебинар.» +
 «&access_token=YOUR_ACCESS_TOKEN» +
 «&v=5.131»;
HttpClient client = HttpClient.newHttpClient();
HttpRequest request = HttpRequest.newBuilder()
 .uri(URI.create(url))
 .build();
HttpResponse<String> response = client.send(request, HttpResponse.BodyHandlers.ofString());
System.out.println(«Результат отправки: « + response.body());
```

Этические аспекты использования данных играют ключевую роль. Данные должны быть обезличены, а их публикация исключать возможность идентификации пользователей. Взаимодействие с пользователями, включая отправку сообщений, возможно только с их согласия, что гарантирует прозрачность и соблюдение стандартов конфиденциальности.

Технические и этические ограничения исследования

Анализ данных через VK API имеет ряд технических ограничений, которые влияют на процесс и результаты исследования. Доступ предоставляется только к публичному контенту, что исключает возможность работы с закрытыми профилями и личными сообщениями. Это делает исследование зависимым от доступности данных, которые могут быть неполными или устаревшими. Например, информация о возрасте, интересах или активности пользователей не всегда заполняется корректно, что требует дополнительных этапов очистки данных. Лимит запросов, установленный VK API, ограничивает объём данных, обрабатываемых за определённый период, а его превышение может привести к временной блокировке доступа. Для минимизации этих рисков применялись оптимизированные алгоритмы, включая параллельное выполнение запросов и исключение избыточных операций. Методы анализа также имеют ограничения. Алгоритм K-Means, используемый для кластеризации, чувствителен к выбросам и требует предварительного задания числа кластеров, что делает результаты анализа зависимыми от субъективного выбора. Тематический анализ с использованием TF-IDF может не учитывать семантические связи между словами, что снижает точность выделения ключевых тем. Кроме того, обновления VK API могут изменить параметры или доступные методы, требуя постоянной адаптации инструментов. Ограничения вычислительных ресурсов также могут усложнять работу с большими выборками и применять ресурсоёмкие методы анализа, такие как глубокое обучение. Этические ограничения включают соблюдение политики конфиденциальности платформы и использование только публичной информации. Все данные обезличивались: идентификаторы заменялись уникальными метками, что исключало возможность идентификации пользователей. Результаты представлялись в агрегированном виде, соответствуя принципам анонимности. Пример соблюдения технических и этических норм — отправка сообщений пользователям через официальные методы VK API с их согласия:

```
String url = «https://api.vk.com/method/messages.send» +
 «?user_id=12345» +
 «&message=Здравствуйте! Приглашаем Вас на мероприятие.» +
 «&group_id=67890» +
 «&access_token=YOUR_ACCESS_TOKEN» +
 «&v=5.131»;
HttpClient client = HttpClient.newHttpClient();
HttpRequest request = HttpRequest.newBuilder()
 .uri(URI.create(url))
 .build();
HttpResponse<String> response = client.send(request, HttpResponse.BodyHandlers.ofString());
System.out.println(«Результат отправки: « + response.body());
```

Таким образом, несмотря на описанные ограничения, предложенные подходы минимизируют их влияние, обеспечивая надёжность и воспроизводимость результатов. Эти аспекты важно учитывать при интерпретации данных и планировании дальнейших исследований.

Заключение

Исследование подтвердило возможность анализа цифрового следа пользователей ВКонтакте с использованием VK API. Открытые данные успешно применяются для выявления предпочтений и формирования целевой аудитории. VK API предоставляет доступ к публичным данным, таким как профили, подписки и публикации. Ограничения преодолеваются соблюдением политики API и оптимизацией обработки данных. Разработанная методология включает сбор, предобработку и анализ данных с использованием Java. Кластеризация и тематический анализ позволили выделить ключевые группы пользователей и их интересы. Практическая значимость заключается в применении результатов в маркетинге, HR и социальной аналитике. Автоматизация делает подход эффективным и масштабируемым. Исследование соответствовало этическим нормам и законодательству РФ благодаря использованию обезличенных данных. Дальнейшие исследования могут включать анализ текстов с использованием глубокого обучения, изучение мультимедийных данных и разработку моделей для прогнозирования поведения. Интеграция с другими платформами позволит проводить сравнительный анализ

цифрового следа и выявлять общие закономерности поведения. Цифровой след ВКонтакте является ценным источником данных для прикладных задач. Разработанная методология может быть адаптирована для других платформ, открывая новые возможности для анализа и дальнейших исследований.

ЛИТЕРАТУРА

1. Гайдаш О.В. Феномен цифрового следа в современном обществе // Вестник магистратуры. — 2020. — № 3. — С. 44–49. — URL: https://cyberleninka.ru/article/n/fenomen-tsifrovogo-sleda-v-sovremennom-obschestve (дата обращения: 18.10.2024).
2. ВКонтакте API — Документация для разработчиков [Электронный ресурс]. — URL: https://dev.vk.com/ru/reference (дата обращения: 18.10.2024).
3. Панферова Е.В., Матюшин Р.А. Сравнительная оценка методов кластеризации в работе с большими данными // Вестник Пермского университета. Математика. Механика. Информатика. — 2024. — Вып. 2 (65). — С. 61–67. — URL: https://cyberleninka.ru/article/n/sravnitelnaya-otsenka-metodov-klasterizatsii-v-rabote-s-bolshimi-dannymi (дата обращения: 18.10.2024).
4. Федеральный закон РФ от 27.07.2006 № 152-ФЗ «О персональных данных» // КонсультантПлюс [Электронный ресурс]. — URL: http://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 18.10.2024).
5. Гасанов И. З., Ликсаков М. В. Эффективная работа с данными сообществ на примере API ВКонтакте // Инновации и инвестиции. — 2023. — № 3. — С. 45–50.
6. Низомутдинов Б.А., Видясова Л.А. Применение автоматизированного сбора информации из сообществ социальных сетей для выявления активных пользователей // International Journal of Open Information Technologies. — 2021. — Т. 9. — № 4. — С. 15–20.
7. Реут А.В. Анализ социальных сетей с помощью Big Data: выявление трендов // Материалы конференции «Информационные технологии и системы». — Казань, 2023. — С. 120–125.
8. Литвинова Е.С., Гаврилюк М.О., Неробелова М.В. Анализ возможностей использования цифрового следа обучающегося для разработки рекомендательных систем // Образовательные технологии и общество. — 2022. — Т. 25. — № 2. — С. 45–60.
9. Witten I.H., Frank E., Hall M.A. Data Mining: Practical Machine Learning Tools and Techniques. — 4th ed. — San Francisco: Morgan Kaufmann Publishers, 2017. — 654 p.
10. Johnson M., Smith L. Social Network Analysis: Methods and Applications // Journal of Social Structure. — 2022. — Vol. 23. — No. 2. — pp. 112–130.
11. Williams R., Brown T. Big Data Techniques for Analyzing Social Media // International Journal of Data Science. — 2021. — Vol. 5. — No. 3. — pp. 98–110.
12. Chen H., Xu X. Mining User Interests from Social Media Profiles // IEEE Transactions on Knowledge, and Data Engineering. — 2023. — Vol. 35. — No. 4. — pp. 789–802.
13. Garcia D., Schweitzer F. Modeling User Behavior in Online Social Networks // Social Network Analysis and Mining. — 2022. — Vol. 12. — No. 1. — pp. 1–15.
14. Brown A., Green N. Ethical Considerations in Social Media Data Analysis // Ethics, and Information Technology. — 2023. — Vol. 25. — No. 1. — pp. 23–35.
15. Stanford Natural Language Processing Group. Stanford NLP Tools for Java [Электронный ресурс]. — URL: https://stanfordnlp.github.io/ (дата обращения: 18.10.2024).

Журнал «Современная наука: актуальные проблемы теории и практики»