Молодёжный вестник НФ БГТУ. 2024. Том 04. № 02 (14)
https://rio-nb-bstu.science

doi: 10.51639/2713-0576_2024_4_2_72
УДК 004.6, 519.252
ГРНТИ 83.77.00
ВАК 1.2.3

Использование предобработки данных для эффективной сегментации абитуриентов на основе цифрового следа

[*] Бабичева Н. Б., Кирчева А. С., Мамедов И. В.

ФГБОУ ВО «Сибирский государственный индустриальный университет», 654007, Россия, г. Новокузнецк, ул. Кирова 42

email: [*] babicheva_nb@mail.ru, alinakircheva@mail.ru, mamedowilkin15@gmail.com

Аннотация

Цифровой след абитуриентов представляет собой ценную информацию, которая может быть использована для решения задачи сегментации образовательных услуг в соответствии с их разнообразными потребностями. Для достижения задачи необходима предварительная обработка данных, этапы которой приведены в статье. Использование парсера позволяет автоматизированно собирать информацию об интересах пользователя. Результаты работы представлены в виде графиков, отображающих распределение возрастов и соотношение открытых и закрытых профилей людей в диапазоне 15-19 лет. Дальнейшее исследование направлено на анализ активности и интересов участников групп с целью более точной сегментации и разработки эффективных стратегий привлечения абитуриентов в образовательные учреждения.

Ключевые слова: цифровой след, сегментация, парсер, обработка, абитуриенты.

Теория и методы исследования

В настоящее время цифровая технология становится все более важной и неотъемлемой частью нашей повседневной жизни. Вместе с ростом числа пользователей интернета и развитием онлайн-сервисов, люди оставляют за собой цифровой след, который представляет собой информацию о действиях и активности пользователя в цифровой среде. Он включает в себя данные о посещенных веб-сайтах, онлайн-взаимодействиях и о вкладе пользователя в цифровом мире [1].

Одной из интересных возможностей, которые открывает перед абитуриентами цифровой след, является его использование для решения задачи сегментирования. Сегментация абитуриентов и образовательных услуг представляет собой особый вид анализа, начинающийся с определения ключевых факторов сегментации. Выбор факторов зависит от конкретных проблем, которые учебное заведение стремится решить, а также от потребителей образовательных услуг.

Молодёжный вестник НФ БГТУ. 2024. Том 04. № 02 (14)
https://rio-nb-bstu.science

Однако сегментация абитуриентов и их потребностей является наиболее сложной задачей, требующей тщательного и взвешенного подхода. Это обусловлено вариативностью их требований, приоритетов и проблем, которые могут повлиять на выбор образовательной услуги, а также недостаточным уровнем информированности, организованности и целенаправленности личностей в процессе принятия решений.

Потребности абитуриентов можно распределить, то есть они поддаются структуризации и сегментации, поскольку потребители регулярно осуществляют выбор в соответствии с их стратегиями и планами действий. Важно понимать, что сегментация абитуриентов позволяет определить образовательную траекторию, которая включает в себя образовательные цели и интересы, академические успехи, географическое и социокультурное окружение, а также особые потребности и обстоятельства. Понимание этих факторов помогает образовательным учреждениям адаптировать свои программы и услуги к различным потребностям абитуриентов.

Для успешного сегментирования необходима предварительная обработка данных. Этапы предобработки представлены на рисунке 1.

Чаще всего требуется получить данные в годной к употреблению форме. Было подсчитано, что специалисты по работе с данными тратят на их подготовку 80% времени [2]. Сложно сформулировать проблему таким образом, чтобы к ней можно было применить методы машинного обучения и получить имеющие практическую ценность и измеримые результаты. Очень важно правильно подобрать признаки для качественных результатов.

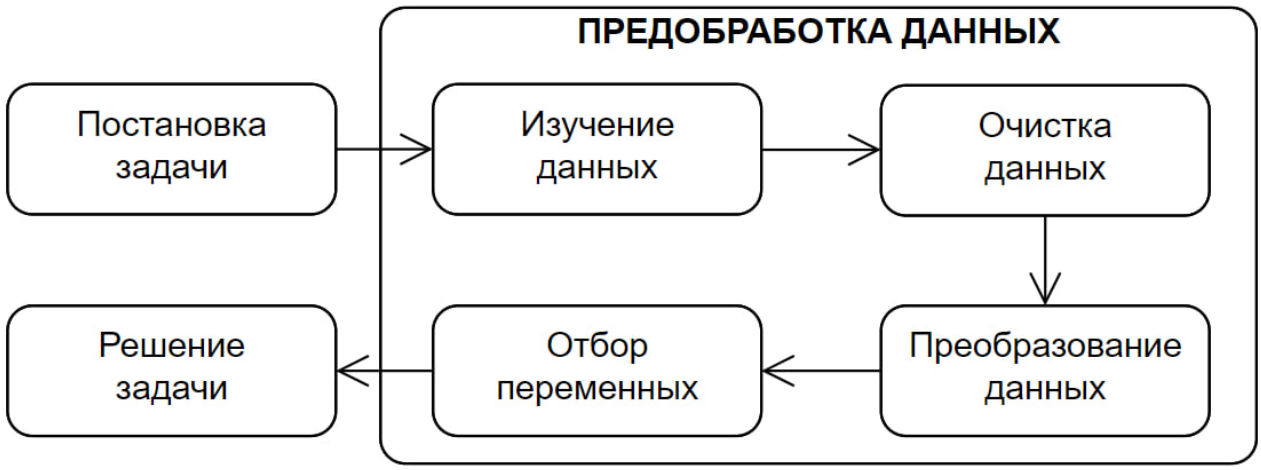

Рис. 1. Место предобработки в типовом процессе решения задач на данных

Этап изучения данных представляет собой важный этап в предобработке данных, поскольку от качества и полноты данных зависит точность и достоверность выводов и результатов анализа. Систематическое и глубокое изучение данных позволит исследователям получить более полное представление о рассматриваемом объекте или процессе и сделать выводы на основе имеющейся информации.

Исследование данных является важным этапом в научных исследованиях и аналитических процессах. Перед тем как проводить любые процедуры на данных, необходимо ответить на ряд ключевых вопросов. Прежде всего, необходимо понять, что именно описывают данные, и к какой конкретной предметной области они относятся. Это позволяет определить экспертов в данной области и основные параметры, характеризующие объект или процесс, описываемый данными. Также важно оценить стабильность объекта или процесса во времени и сформулировать цель исследования с учетом мнения специалистов предметной области.

Анализ полноты данных играет ключевую роль, позволяя оценить, насколько данные хорошо описывают изучаемый объект или процесс, а также учитывать влияние внешних факторов. Для описания данных важно знать их размер, состав и типы переменных, статистические характеристики и количество пропусков.

Очистка данных направлена на обеспечение полноты, истинности, корректности и непротиворечивости данных. Однако не всегда можно убрать все некачественные данные, так как таких данных может быть много, что может повлиять на результаты предобработки данных. В этом процессе применяются различные процедуры.

Одной из ключевых задач предобработки следует отнести заполнение пропусков. Методы обработки пропусков включают в себя удаление строк или столбцов с пропусками, заполнение пропусков логически обоснованными значениями, использование самого частого значения или расчетных значений, восстановление значений с использованием различных методов, таких как регрессия и классификация, а также другие методы.

Бывает такое, что в процессе получения данных встречаются дубликаты записей и выбросы, которые могут повлиять на результаты. Причины появления таких аномалий могут быть разными: возникновение технических сбоев, ошибок ввода или неправильный подход к решению задачи.

Выбор подхода к обработке данных зависит от того, что именно означают значения переменных, как возникают некорректные данные и их тип, поэтому важно различать между собой нулевые значения, пустые ячейки и неправильные данные. Важно учитывать различия в методологии обработки ошибок для целевых переменных.

Разнообразие данных также может оказать влияние на качество результатов. Преобразование данных в предобработке данных – это процесс изменения или преобразования исходных данных для подготовки их к анализу или использованию в моделях машинного обучения.

Чаще всего категориальные признаки поддаются преобразованию данных. Категориальные данные относятся к данным, которые не представлены в числовой форме, и могут включать в себя как бинарные признаки (имеющие два уникальных значения), так и признаки с более чем двумя уникальными значениями. Для обработки таких признаков необходимо провести процесс кодирования категориальных признаков, который заключается в преобразовании категориальных данных в числовое представление в соответствии с определенными заранее установленными правилами или методами.

Также к процессу преобразованию данных относится инженерия признаков – процесс преобразования и модификации необработанных данных с целью создания новых признаков, которые более точно отражают суть задачи или являются более информативными для моделей машинного обучения.

Процесс отбора данных заключается в обеспечении максимальной эффективности модели на предварительно подготовленном наборе данных. Его целью является определение набора переменных, которые наиболее важны для получения оптимальных результатов предсказания.

Методы отбора переменных могут различаться в зависимости от алгоритма, используемого для построения модели. Они могут включать в себя анализ важности признаков, методы отбора на основе статистических тестов, алгоритмы рекурсивного и прямого отбора признаков и другие подходы. Поэтому выбор подходящего метода отбора переменных зависит от конкретной задачи, характеристик данных и используемой модели.

Чтобы данные прошли процесс предобработки, необходимо собрать данные исходя из постановки задачи с помощью нужных алгоритмов и средств. Парсер, способный извлекать данные из социальных сетей, играет ключевую роль в сегментации абитуриентов. Социальные сети становятся основным источником цифрового следа, оставленного абитуриентами в Интернете. Взаимодействия в социальных сетях, такие как посты, подписанные группы, комментарии, фотографии, лайки и другие активности, могут содержать ценную информацию о предпочтениях, интересах, увлечениях и даже личных характеристиках абитуриентов.

Использование парсера для анализа данных из социальных сетей позволяет унифицировать и автоматизировать процесс сегментации абитуриентов. Парсер может извлекать информацию о пользовательских профилях, такую как место жительства, интересы, образование, группы и другие факторы, которые могут быть полезны для понимания профиля потенциальных абитуриентов. Это позволяет учебным заведениям создавать более точные и персонализированные стратегии привлечения абитуриентов, а также более эффективно взаимодействовать с ними в процессе набора на обучение.

В результате этого была сформирована цель проекта: разработать парсер, который будет занимать поиском будущих абитуриентов среди школьных групп Новокузнецка в социальной сети ВКонтакте. Данный парсер собирает всех пользователей в единый файл, который в будущем пройдет процесс предобработки данных.

Исходя из целей работы были определены задачи, которые заключались в следующем:

- необходимо выявить сообщества школ Новокузнецка;
- создать способ получения данных в социальной сети Вконтакте для дальнейшей обработки полученных результатов.

Для достижения цели были использованы методы социальной сети ВКонтакте [3], функция расчета возраста на языке программирования Python и инструменты MicrosoftExcel.

На рисунке 2 представлена диаграмма состояний парсера.

При запуске парсера, его первым шагом является установление связи с сервером социальной сети Вконтакте. Данный шаг критически важен, поскольку от него зависит возможность получения данных. Если сервер ВКонтакте недоступен, парсер выдаст соответствующее сообщение об ошибке, после чего процесс набора данных завершится.

Если сервер доступен, парсер приступает к обработке данных. Этот этап начинается с активации методов VK API, которые позволяют извлечь необходимую информацию о группе и её участниках. После получения идентификатора группы парсер осуществляет запрос на сервер ВКонтакте для получения информации о самом сообществе.

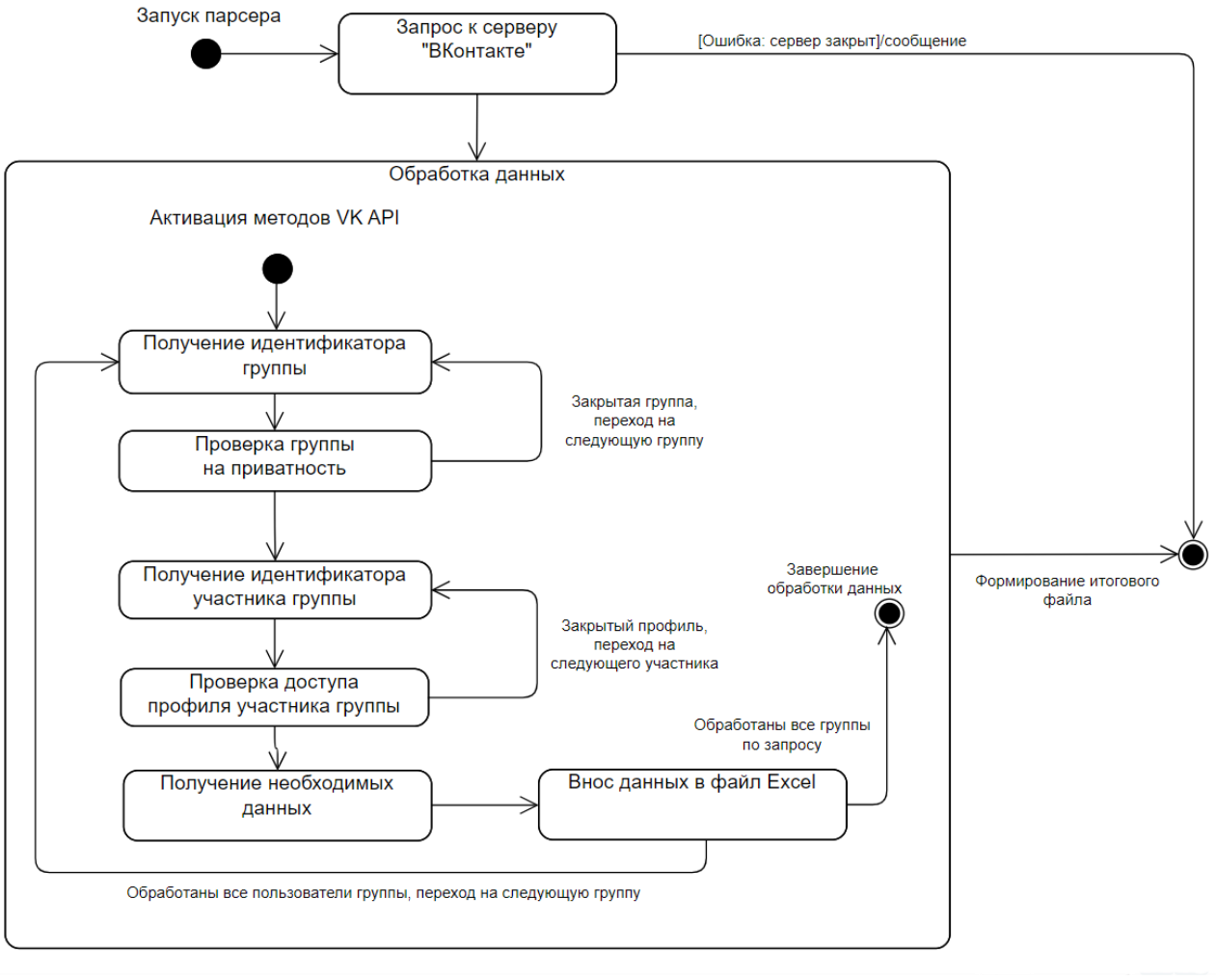

Рис. 2. Диаграмма состояний парсера

Важным шагом является проверка приватности группы. Если группа закрытая и её данные недоступны для парсера, он автоматически переходит к следующей группе в списке. Однако, если группа открытая, парсер приступает к извлечению данных участников данной группы.
Для каждого участника группы парсер получает идентификатор и проверяет доступность его профиля. Парсер не способен получить полную информацию, если профиль пользователя закрытый. Однако, если профиль пользователя открытый, парсер извлекает необходимые данные пользователя, которые зашифрованы в процессе получения информации с целью избежания утечки, во время обработки и статус приватности профиля. Если некоторые данные неизвестны, то программа выдаст «Не указано».
Собранные данные о каждом пользователе группы заносятся в специально созданный Excel файл. Каждая группа имеет свой собственный лист в этом файле. После того, как парсер обработал все указанные группы и всех их участников, он завершает свою работу.
В заключение программа формирует итоговый файл, содержащий собранные данные. Этот файл предоставляет информацию об участниках групп, что может быть полезно для анализа.

Молодёжный вестник НФ БГТУ. 2024. Том 04. № 02 (14)
https://rio-nb-bstu.science

Полученные результаты и их обсуждение

Исследование включало анализ 33 школьных групп в социальной сети «ВКонтакте» с использованием парсера, который просматривал 32 открытые группы. В результате было парсер обнаружил 23987 пользователей, где 21771 уникальных. Диаграмма распределения возрастов представлена на рисунке 3. Среди пользователей старше 25 лет можно сделать вывод, что это, вероятно, преподаватели или студенты.

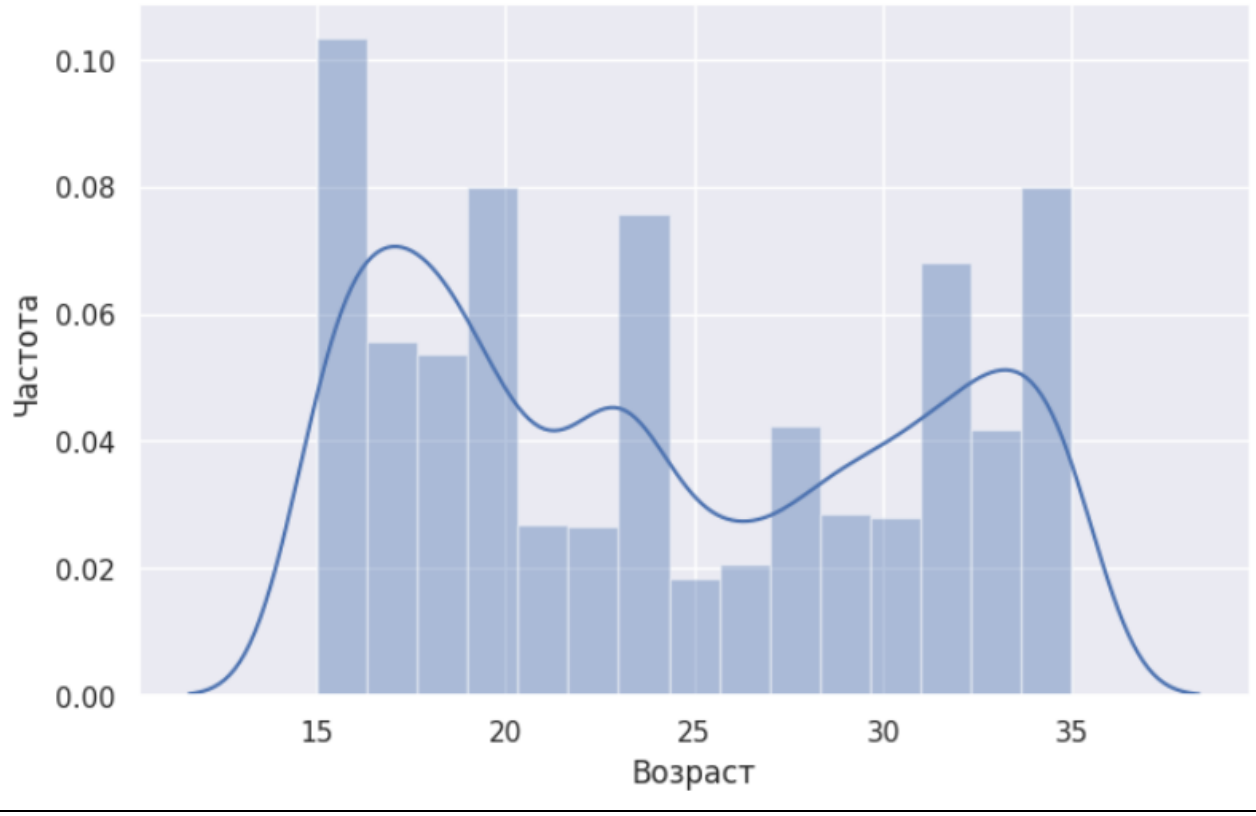

Рис. 3. Диаграмма распределения возрастов

Также было выделено 2380 человек в возрасте от 15 до 19 лет, с дальнейшим разбиением на подгруппы по возрасту. На рисунке 4 показано соотношение процентов пользователей с открытым и закрытым профилем. С увеличением возраста наблюдается рост процента пользователей с закрытым профилем.

В настоящее время проводится дальнейшее исследование с целью выделения дополнительных групп школьников для будущего анализа и сегментации. Это включает в себя анализ активности и взаимодействия в социальной сети «ВКонтакте» среди различных групп пользователей. Кроме того, осуществляется выделение характерных особенностей и интересов участников этих групп для более точной сегментации. Данный процесс позволит более глубоко понять социальное взаимодействие школьников в сети и поможет разработать стратегии обучения и взаимодействия с этой аудиторией.

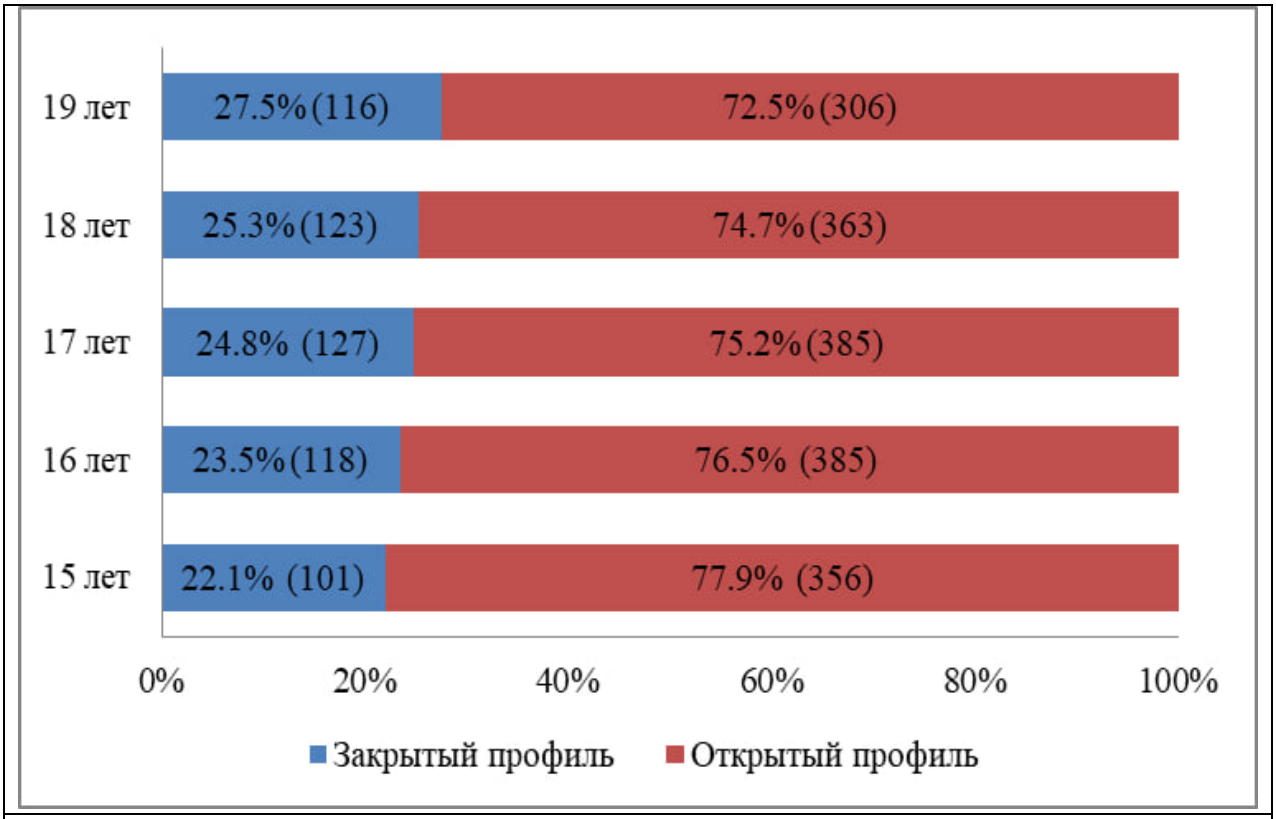

Рис. 4. Соотношение открытых и закрытых профилей людей в диапазоне 15-19 лет

Конфликт интересов

Авторы статьи заявляют, что у них нет конфликта интересов по материалам данной статьи с третьими лицами на момент подачи статьи в редакцию журнала, и им ничего не известно о возможных конфликтах интересов в настоящем со стороны третьих лиц.

Список литературы

1. Бабичева, Н. Б. Применение цифрового следа в построении непрерывной образовательной траектории / Н. Б. Бабичева, А. С. Кирчева, И. В. Мамедов // Системы автоматизации (в образовании, науке и производстве) AS'2023 : труды Всероссийской научно-практической конференции (с международным участием), Новокузнецк, 12–14 декабря 2023 года. – Новокузнецк: Сибирский государственный индустриальный университет, 2023. – С. 248-254.
2. Начальное руководство по данным для обучения ИИ [Электронный ресурс] - URL: https://ru.shaip.com/blog/the-only-guide-on-ai-training-data-you-will-need-in/ (дата обращения 13.03.2024)
3. Описание методов API | VK для разработчиков [Электронный ресурс] - URL: https://dev.vk.com/ru/method (дата обращения 16.03.2024)

Молодёжный вестник НФ БГТУ. 2024. Том 04. № 02 (14)
https://rio-nb-bstu.science

UTILIZING DATA PREPROCESSING FOR EFFECTIVE PROSPECTIVE STUDENT SEGMENTATION BASED ON DIGITAL FOOTPRINT

Babicheva N. B., Kircheva A. S., Mamedov I. V.

Siberian State Industrial University
654007, Russia, Novokuznetsk, Kirova st., 42

email: * babicheva_nb@mail.ru, alinakircheva@mail.ru, mamedowilkin15@gmail.com

The digital footprint of prospective students represents valuable information that can be used for segmenting educational services according to their diverse needs. To achieve this task, data preprocessing is necessary, the stages of which are outlined in the article. The use of a parser allows for the automated collection of information about user interests. The results of the study are presented in the form of graphs showing the distribution of ages and the ratio of open and closed profiles of individuals aged 15-19. Further research is aimed at analyzing the activity and interests of group participants to enable more precise segmentation and the development of effective strategies for attracting prospective students to educational institutions.

Keywords: digital footprint, segmentation, parser, processing, prospective students.

References

1. Babicheva, N. B., Kircheva, A. S., & Mamedov, I. V. (2023). Application of digital footprint in designing continuous educational trajectory.In Proceedings of the All-Russian Scientific and Practical Conference on Automation Systems (in Education, Science, and Industry) AS'2023 (with international participation), December 12–14, 2023 (pp. 248-254). Novokuznetsk: Siberian State Industrial University.
2. Initial Guide to Data for AI Training - [Electronic resource]. URL: https://ru.shaip.com/blog/the-only-guide-on-ai-training-data-you-will-need-in/ (13.03.2024)
3. Description of VK API Methods for Developers - [Electronic resource]. URL: https://dev.vk.com/ru/method (16.03.2024)