УДК 004.414.3

РАЗРАБОТКА МЕТОДОВ И ПРОГРАММНЫХ ИНСТРУМЕНТОВ ФОРМИРОВАНИЯ ЦИФРОВОГО ПОРТРЕТА УЧАЩИХСЯ

М. А. Солнцев[1] [0009-0002-4106-3035], М. М. Абрамский[2] [0000-0003-3063-8948]

[1, 2]*Институт информационных технологий и интеллектуальных систем, Казанский (Приволжский) федеральный университет.*

[1]mrt.solncev@gmail.com, [2]ma@it.kfu.ru

Аннотация

Рассмотрены вопросы возможности использования данных об обучающихся, представленных в электронном виде, для построения цифрового портрета. Предложен набор характеристик, необходимых для его построения, обозначена модель данных. Реализованы инструменты сбора данных об обучающихся из социальных сетей и других интернет-ресурсов. Предложены алгоритмы построения цифрового портрета. Проиллюстрировано применение алгоритмов машинного обучения для этих задач. Приведены примеры использования цифрового портрета в образовании.

Ключевые слова: социальные сети, сбор данных, портрет пользователя, образование

ВВЕДЕНИЕ

Согласно Постановлению Правительства Российской Федерации «О проведении эксперимента по внедрению цифровой образовательной среды» на территории отдельных субъектов РФ будут организованы мероприятия по внедрению ЦОС [1]. В рамках этого эксперимента планируется проверка применения возможности формирования «цифрового профиля обучающегося». Цифровой профиль станет обязательным, его будут регистрировать при первом обращении за образовательной услугой, например, при зачислении в детский сад или первый класс школы. В связи с этим становится актуальным вопрос формирования цифрового портрета обучающегося.

В настоящей работе представлены методы и программные инструменты

формирования цифрового портрета обучающихся. Для построения цифрового портрета могут быть использованы информация о социальной активности, а также цифровой след.

Решением схожих задач занимаются системы, созданные для поиска целевой аудитории в социальных сетях. В последнее время их количество стремительно растет, к наиболее популярным системам такого характера можно отнести «Pepper.ninja», «Segmento Target», «Target Hunter» и «Церебро Таргет» [2–5].

В настоящей работе рассмотрено использование текстовой, графической и медиа информации.

Статья построена следующим образом.

В разделе 1 выделены основные источники данных и характеристики, используемые для построения цифрового портрета, отмечены особенности их хранения.

В разделе 2 описаны способы извлечения и алгоритмы обработки данных, участвующих в построении.

Третий раздел посвящен методам и алгоритмам анализа данных, используемых для построения цифрового портрета, рассмотрены способы их применения.

1. ИСТОЧНИКИ ДАННЫХ ДЛЯ ЦИФРОВОГО ПОРТРЕТА

Задачи, связанные со сбором и анализом данных с целью последующего нахождения в них полезной информации, принято относить к типу *Data Mining* задач. Для решения такого класса задач в основном используются технологии *краулинга* и *скрейпинга*, а также методы *машинного обучения* [6–8].

В задачах формирования цифрового портрета личности большой популярностью пользуются алгоритмы классификации и кластерного анализа, а также ассоциативные правила.

В современном мире популярность социальных сетей растет с каждым годом всё динамичнее, во многом по этой причине их количество увеличивается. Вместе с этим люди оставляют всё больший цифровой след, который может точно описать их личность: характер, взгляды и интересы. Результаты статистического исследования ресурса statista.com, приведенные на рис. 1, показывают, что в настоящее время жители России активно используют порядка 15 социальных се-

тей каждый день [9]. Эти результаты свидетельствуют, что наибольшей популярностью пользуются такие социальные сети, как YouTube и ВКонтакте. Ниже рассмотрена социальная сеть ВКонтакте, так как в ней пользователи помимо подписок на различные тематические сообщества и каналы могут публиковать собственные медиа и текстовые публикации – это дает возможность применить множество методов анализа данных для получения полезной информации. Например, текстовая информация может быть использована для математической лингвистики с целью классификации публикуемых текстов.

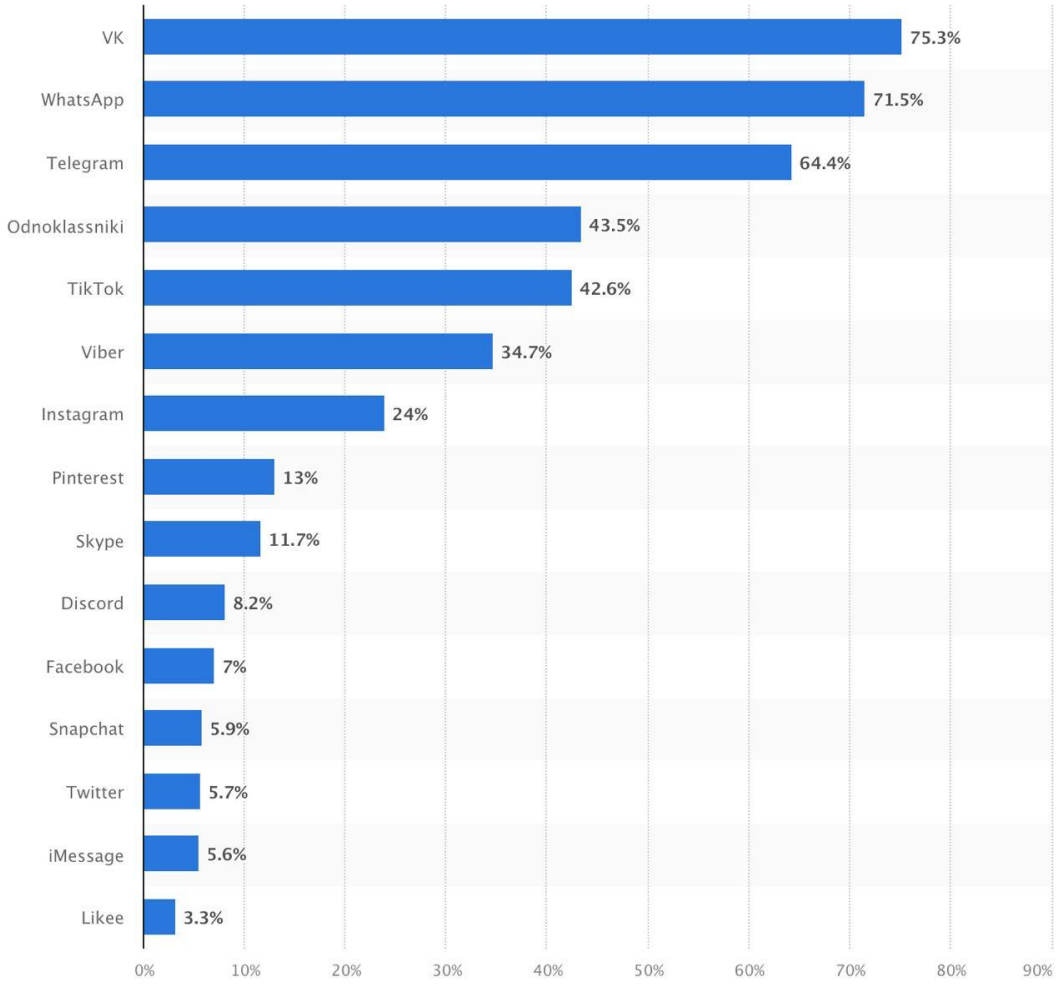

Рис. 1. Использование социальных сетей в России в 2022 году [9]

Помимо этого, при построении цифрового портрета необходимо учитывать информацию о получаемом образовании и курсах дополнительного образования. Успешность дополнительной образовательной и спортивной деятельности может быть подтверждена результатами выступлений на различных олимпиадах,

конкурсах и соревнованиях. Для этого могут быть использованы финальные протоколы участия, которые могут быть получены у организаторов такого рода мероприятий. Протоколы участия в большинстве случаев предоставляются в формате Excel и имеют вид, представленный на рис. 2.

	Фамилия уч	Имя	Отчество		Район об	Образовате	Педагог		пи	у	и	Статус
2	Гайнуллин	Әмир	Илнурович	2	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	8	32	40	Победитель
3	Имамиев	Камиль	Ильназович	2	Московский	МБОУ Татарска	Нагимулли	Оли	8	33	41	Победитель
4	Маулиханова	Ралина	Ранисовна	4	Тетюшский	МБОУ "Тетюшская СОШ №		Оли	6	29	35	Победитель
5	Гарипова	Исламия	Марселевна	1	Авиастроите	МБОУ "Гимназ	Закирова Л	Оли	7	16	23	Призер
6	Кашапова	Азалия	Айратовна	3	Кировский	МБОУ "СОШ"5	Шакирова	Оли	10	25	35	Призер
7	Низамова	Джамиля	Дамировна	1	Приволжски	МАОУ "Гимназ	Галаветдин	Оли	8	19	27	Призер
8	Рамазанова	Амина	Ренатовна	2	Кировский	МБОУ"Лицей 1	Мингалиев	Оли	7	23	30	Призер
9	Хамзина	Ранелия	Радиковна	2	Кировский	МБОУ "Полите	Мингалиев	Оли	7	30	37	Призер
10	Шагитова	Камиля	Ильфатовна	4	Лаишевский	МБОУ Пелевск	Шагитова Л	Оли	6	22	28	Призер
11	Шарифуллина	Диляра	Айратовна	4	Приволжски	МБОУ "Школа	Яковлева Р	Оли	10	20	30	Призер
12	Шафикова	Дина	Альбертовна	2	Авиастроите	МБОУ "Школа	Муктат Флё	Оли	4	29	33	Призер
13	Абдуллина	Самира	Наилевна	4	Зеленодоль	МБОУ "Гимназ	Зиганшина	Оли	9	14	23	Участник
14	Абдуллина	Алина	Рустамовна	4	Кировский р	МБОУ Политех	Гибадулли	Оли	4	12	16	Участник
15	Байбалаева	Самира	Баходуровна	4	Зеленодоль	МБОУ "Гимназ	Зиганшина	Оли	5	16	21	Участник
16	Билалова	Руфина	Рифатовна	3	Кировский р	МБОУ Политех	Гибадулли	Оли	4	28	32	Участник
17	Габдрахманова	Гульназ	Ильгамовна	4	Зеленодоль	МБОУ "Гимназ	Зиганшина	Оли	5	19	24	Участник
18	Галимов	Карим	Тауфийкович	1	Ново-Савин	Гимназия №155		Оли	7	11	18	Участник
19	Гарифуллина	Самира	Ильгамовна	4	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	2	13	15	Участник
20	Гилманова	Амелия	Марселевна	3	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	4	21	25	Участник
21	Гимадиев	Самир	Русланович	3	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	4	24	28	Участник
22	Ибрагимова	Газиза	Айтугановна	3	Кировский	МБОУ"Лицей 1	Мингалиев	Оли	8	7	15	Участник
23	Идиатов	Камил	Ринатович	2	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	3	14	17	Участник
24	Исмагилова	Зилә	Ленаровна	1	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	4	13	17	Участник
25	Касимова	Джамиля	Дамировна	2	Ново-Савин	МБОУ "Школа	Галимова Л	Оли	3	19	22	Участник
26	Латыпова	Диляра	Ильшатовна	4	Кировский р	МБОУ Политех	Гибадулли	Оли	3	21	24	Участник
27	Лотфуллина	Самира	Алмазовна	3	Кировский р	МБОУ Политех	Гибадулли	Оли	5	19	24	Участник
28	Любицкая	Сафия	Вадимовна	4	Советский	МБОУ "Гимназ	Насибулли	Оли	3	14	17	Участник
29	Масгутова	Алина	Фанилевна	4	Кировский р	МБОУ Политех	Гибадулли	Оли	6	15	21	Участник

Рис. 2. Пример финального протокола олимпиады

В Таблице 1 представлена модель данных, используемая при построении цифрового портрета.

Таблица 1. Модель данных для цифрового портрета

Хранимый объект	**Описание и источник**
ФИО	
Дата рождения	
Пол	
Текущее место обучения	

Посещаемые курсы дополнительного образования	
Результаты участия в различных конкурсах и олимпиадах	Информация из протоколов мероприятий и олимпиад
Образование	Указанная информация об образовании
Идентификаторы в социальных сетях	Идентификаторы в ВКонтакте
Информация в разделе «О себе»	Содержимое поля «О себе» из профилей в ВКонтакте
Интересы	Содержимое поля «Интересы» в ВКонтакте
Родной город	Родной город, указанный в ВКонтакте
Город проживания	Город, указанный в ВКонтакте
Знание языков	Содержимое поля «Языки» в ВКонтакте
Опыт работы	Указанные места работы и стаж работы
Любимые книги	Содержимое поля «Любимые книги»
Любимые фильмы	Содержимое поля «Любимые фильмы»
Список понравившихся публикаций	Список публикаций, которые пользователь пометил отметкой «Мне нравится» или разместил на своей странице
Список сообществ	Список сообществ, на которые подписан пользователь в ВКонтакте
Список медиа публикаций	Список медиа публикаций в ВКонтакте и их метаинформация (локация, хештеги, содержимое изображения и т. д.)

2. ИСПОЛЬЗУЕМЫЕ МЕТОДЫ СБОРА И ОБРАБОТКИ ДАННЫХ

Информация, выбранная для построения цифрового портрета, находится в разных источниках информации, имеет разные структуру и формат. Поэтому для

её получения необходимо поддерживать различные методы, используя:

- *API* внешнего ресурса;
- сканирование страниц с информацией;
- обработку документов в цифровом формате;
- ручной ввод и корректировку данных.

В случае, когда внешний источник информации предоставляет открытый *API*, информация может быть получена путем отправки соответствующего HTTP *REST* запроса. Ответ в таком случае будет получен в формате *JSON* и может быть использован без преобразований.

В случае, когда внешний источник информации не обладает открытым *API*, необходимо использовать специальные техники получения данным путем *краулинга* и *скрейпинга*. В таком случае необходимо обрабатывать статические HTML-страницы для получения интересующей информации в более понятном формате (например, JSON).

Еще одним рассматриваемым источником данных являются документы, представленные в цифровом или бумажном форматах. Бумажные документы для автоматической обработки необходимо предварительно оцифровать путем сканирования. Оцифрованные документы могут быть обработаны с использованием различных библиотек, независимо от формата документа (Excel, PDF и т. д.).

Для получения информации из профилей социальных сетей используется несколько методов. Для анализа извлеченной информации применяются методы машинного обучения: так, для анализа медиа контента из ВКонтакте применяется нейронная сеть. Схема преобразований данных изображена на рис. 3.

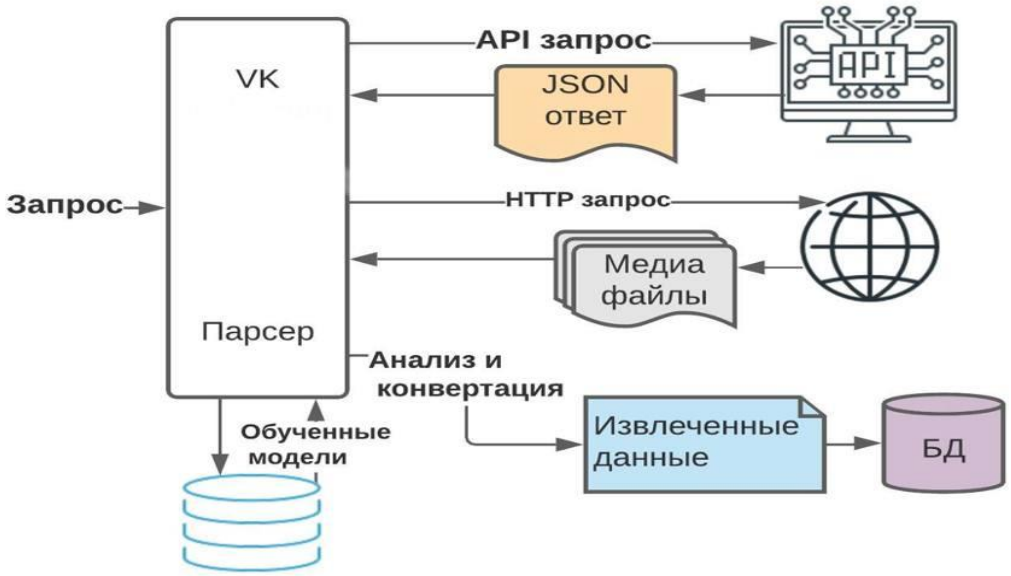

Рис. 3. Схема обработки данных из социальных сетей

Для обработки данных об учебной и творческой успешности используется информация о результатах участия в мероприятиях. Так как организаторы таких мероприятий, как правило, не имеют открытого API, необходимо применять скрейпинг их сайта. Такой метод позволяет извлечь из HTML-страниц необходимые URL-адреса файлов с финальными протоколами. После этого файлы, находящиеся в формате Excel, преобразуются в JSON. Соответствующая схема действий изображена рис. 4.

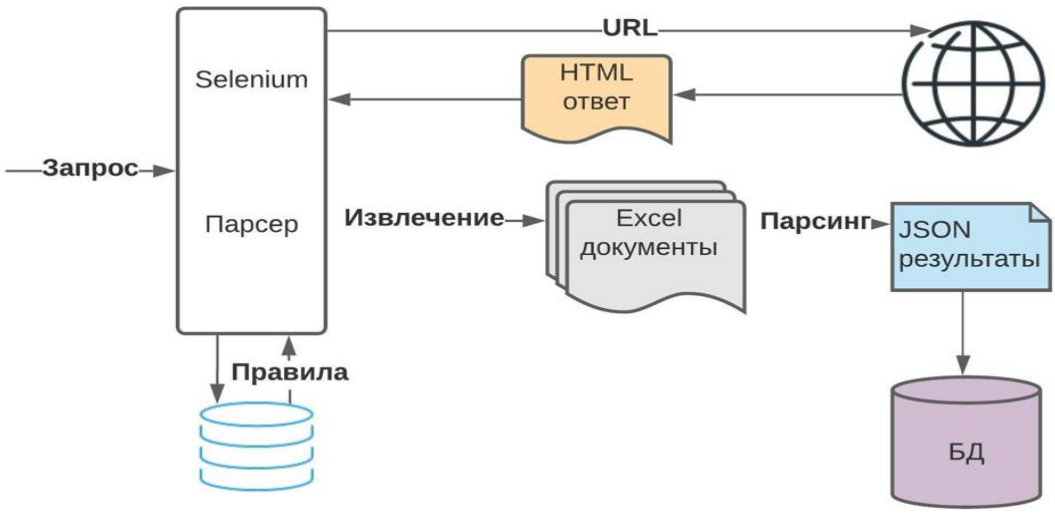

Рис. 4. Схема извлечения данных о достижениях

3. АНАЛИЗ ДАННЫХ ДЛЯ ЦИФРОВОГО ПОРТРЕТА

Анализ данных социальных сетей: данный алгоритм позволяет собирать и анализировать данные из профилей социальных сетей. Сервис состоит из трех модулей:

- анализа социальной активности – отвечает за получение и обработку данных из профилей «ВКонтакте» (рис. 5);
- анализа медиа контента – отвечает за анализ медиа информации из профилей (рис. 6);
- анализа тональности текстовых публикаций.

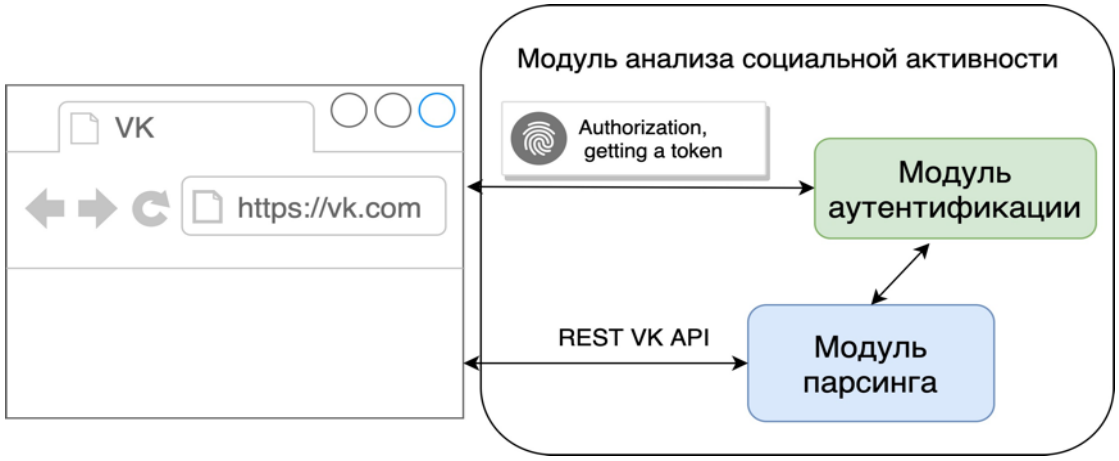

Рис. 5. Архитектура модуля анализа социальной активности

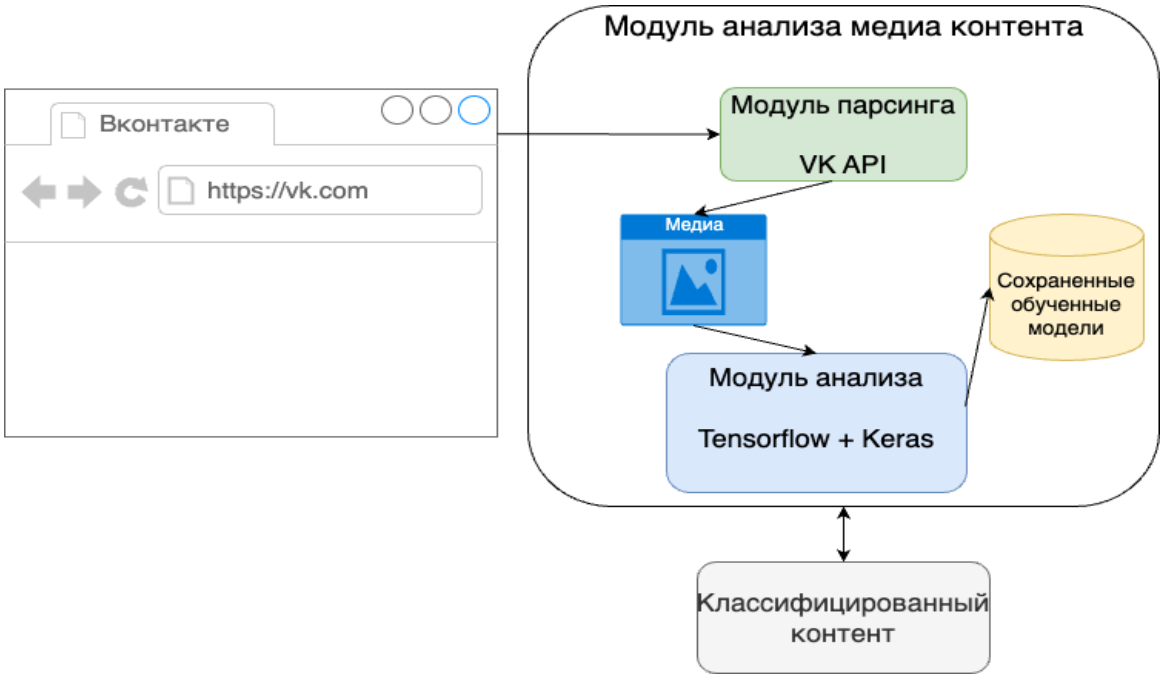

Рис. 6. Модуль анализа медиа контента

Сервис взаимодействует с внешними сервисами («ВКонтакте»), используя предоставляемый ими открытый API.

Для начала работы с VK API [10] необходимо зарегистрировать сервисное приложение в «ВКонтакте», от его лица совершаются все запросы для получения данных. Приложение имеет ряд настроек, которые в дальнейшем могут быть изменены. Зарегистрированному приложению присваиваются уникальный идентификатор, защищенный ключ и сервисный ключ доступа (рис. 7), которые используются в системе клиентом VK SDK [11].

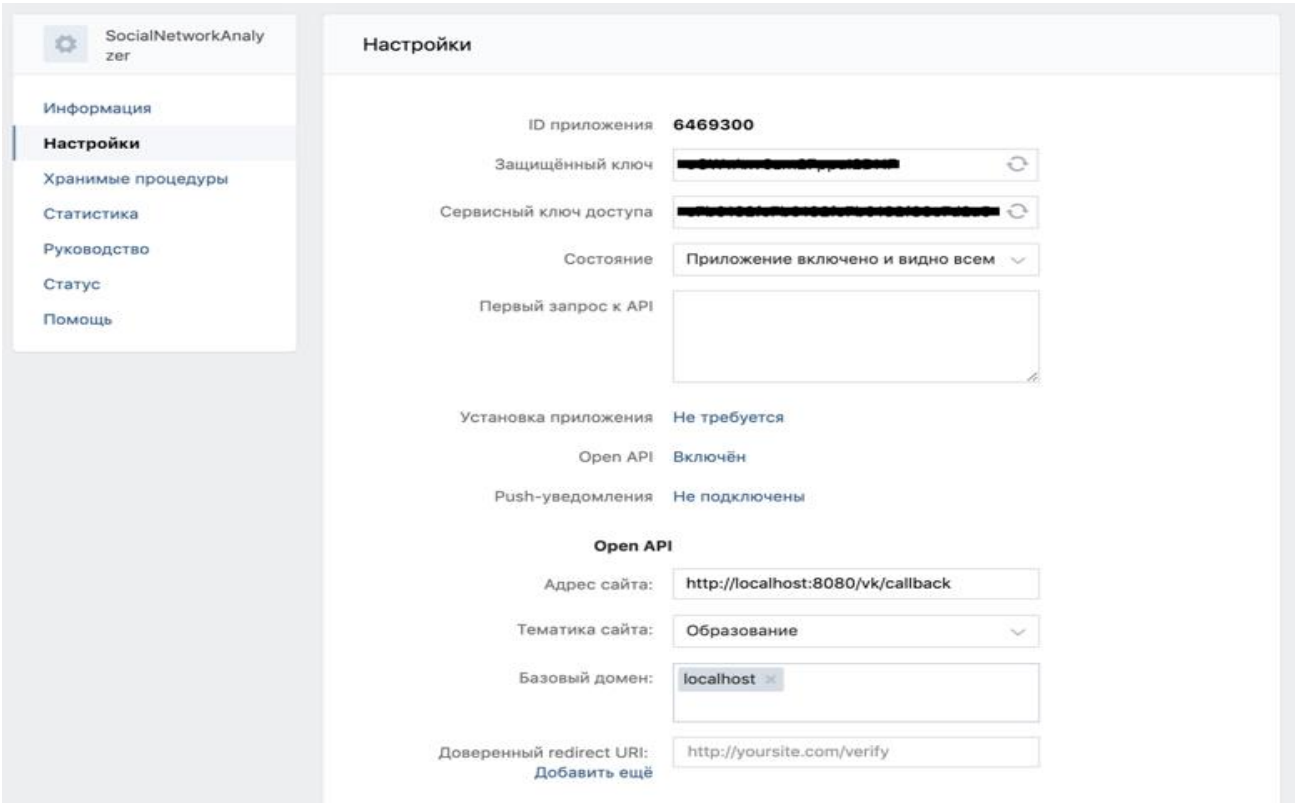

Рис. 7. Настройки параметров приложения ВКонтакте

Для начала выполнения запросов клиенту, выполняющему запросы к VK API, необходимо авторизоваться. Для этого используется OAuth-аутентификация (рис. 8).

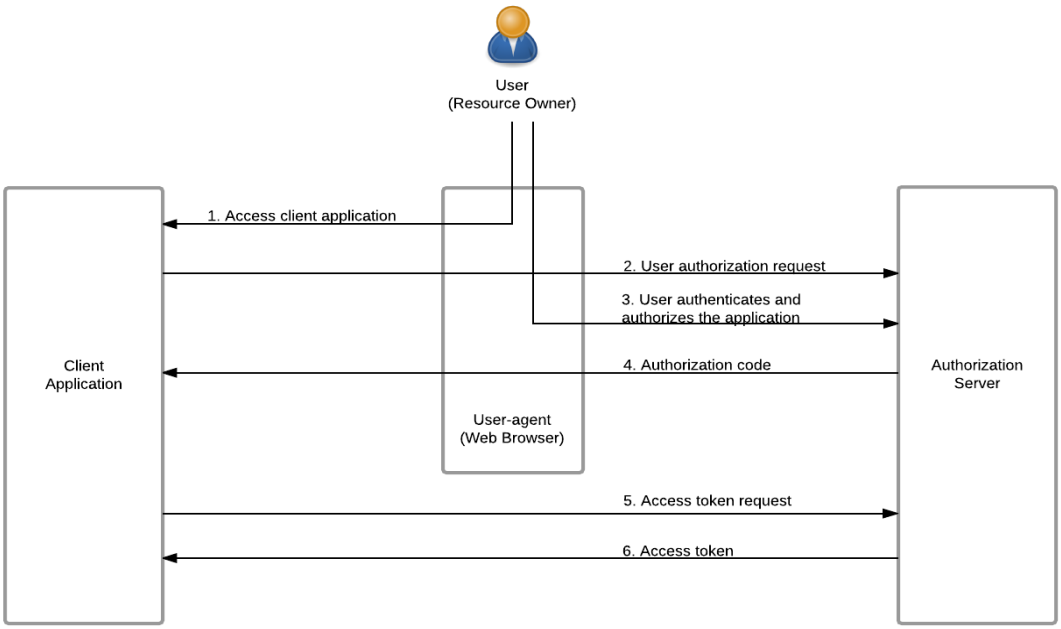

Рис. 8. Auth Code Flow для получения ключа доступа

Для более удобного процесса авторизации приложения используется ScribeJava [12] – клиент для работы с OAuth-авторизацией [13]. Для того чтобы клиент смог получить токен доступа, ему необходимо передать: идентификатор приложения; список разделов, к которым будет иметь доступ система при выполнении запросов; секретный ключ доступа и Callback URL. После получения токена система может выполнять запросы на получения данных. Пример ответа VK API изображен на рис. 9. VK API придерживается архитектуры REST [14], поэтому возвращает данные в формате JSON [15].

```
{
  "response": [{
    "first_name": "Lindsey",
    "id": 210700286,
    "last_name": "Stirling",
    "can_access_closed": true,
    "is_closed": false,
    "photo_50": "https://sun7-9.us...8,641,641&ava=1",
    "verified": 1,
    "city": {
      "id": 5331,
      "title": "Los Angeles"
    },
    "interests": "Family, Friends, Dancing, Music",
    "about": "http://www.lindse...com/LindseyStirling",
    "career": [],
    "university": 0,
    "university_name": "",
    "faculty": 0,
    "faculty_name": "",
    "graduation": 0
  }]
}
```


Рис. 9. Пример ответа VK API

Это позволяет проанализировать группы и сообщества группы пользователей ВКонтакте. Для этого собираются данные о подписках пользователей на различные страницы, определяются их название, тематика и описание. Для каждого пользователя подсчитывается количество групп, принадлежащих к заранее определенном наборам тематик. В данной системе были выделены две основные группы тематик: группы с развлекательным характером и группы, связанные с образованием, личностным и профессиональным ростом. К первой группе были отнесены сообщества с тематиками: 'Покупки', 'Туризм, путешествия', 'Развлечения', 'СМИ', 'Спорт', 'Юмор', 'Шоу, передача', 'Игры', 'Стиль, одежда, обувь', 'Музыка', 'Кино', 'Веб-сайт', 'Обмен музыкой', 'Интернет-СМИ', 'Городское сообщество', 'Искусство и развлечения', 'Молодёжное движение', 'Музыкальная группа'; ко второй – 'Искусство', 'IT', 'Наука', 'Образование', 'Саморазвитие', 'Техника', 'Экономика', 'Языки', 'Бизнес', 'Дизайн и графика', 'История', 'Финансы', 'Культура',

'Философия', 'Обучающие курсы', 'Литература', 'Творчество', 'Фотография', 'Культурный центр', 'Программное обеспечение', 'Образовательное учреждение', 'Программирование'. Данные списки в дальнейшем могут быть скорректированы, также можно добавить большее количество подгрупп тематик. После того, как выделены группы тематик, подсчитывается количество сообществ пользователя, принадлежащих к той или иной группе. На основе данных о процентном содержании сообществ всех подгрупп составляется характеристический вектор студента. Размерность вектора определяется количеством подгрупп сообществ, значения – долей сообществ соответствующей группы.

Для получения сведений о заинтересованности пользователей в определенных областях, нахождения тенденций и связей интересов пользователей используется кластеризация. Система применяет метод К-средних, который является одним из самых популярных и простых в реализации. Данный метод позволяет разделить множество объектов, имеющих определенные свойства, на количество кластеров, равное К [16]. Величина К является параметром и может задаваться вручную, но в системе определяется автоматически на основании мощности кластеризуемого множества. В реализации модуля анализа данных была использована реализация метода К-средних Scikit-Learn [17]. На вход подаются характеристические векторы пользователей, построенные на основании данных о подписках, на выходе – список принадлежностей объекта кластеризации (пользователя) к определенному кластеру (рис. 10).

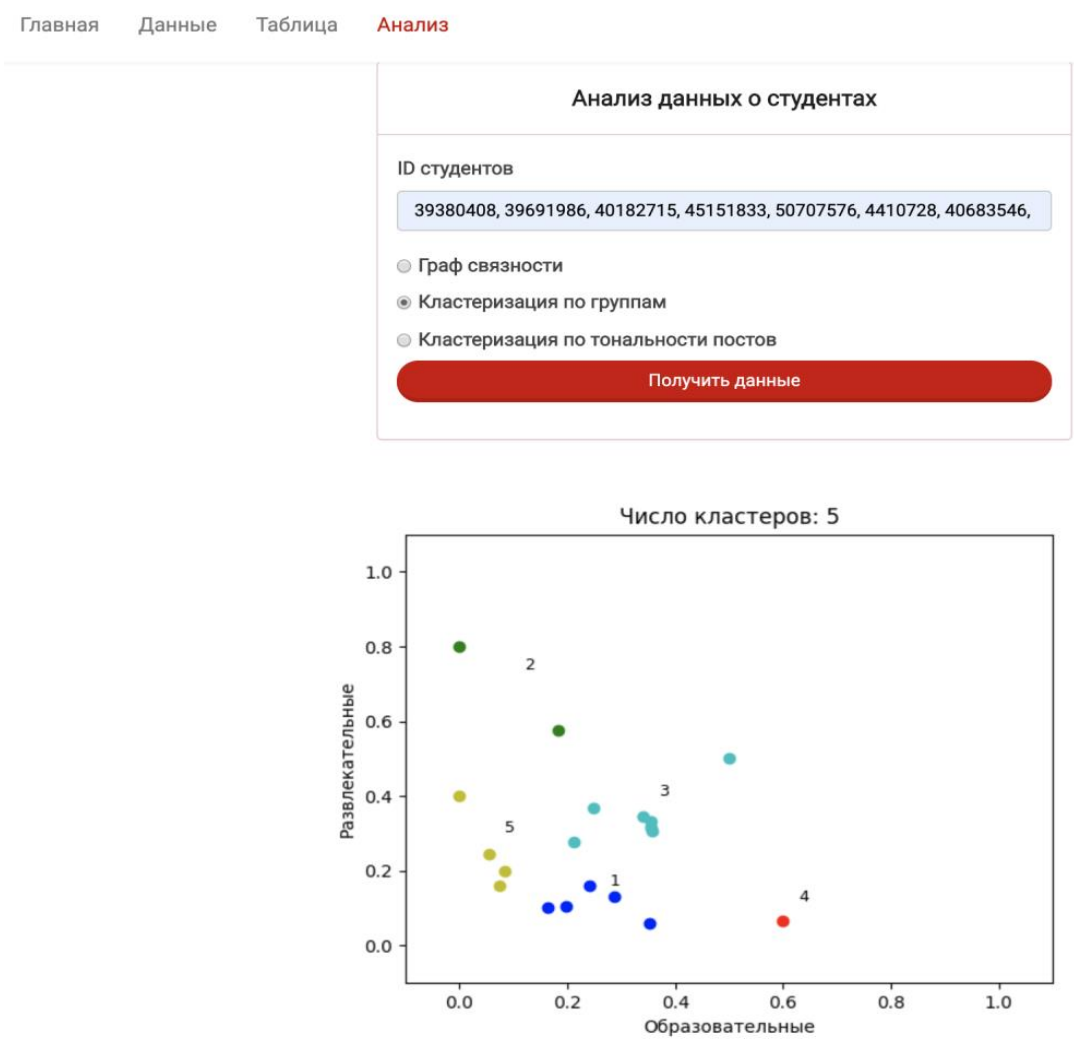

Рис. 10. Кластерный анализ по тематикам групп

Таким же способом получаются ссылки на медиаконтент пользователя. После того, как медиа файлы получены, они передаются в модуль анализа информации. Для классификации фото в нем используются библиотеки Numpy и Keras [18, 19]. В системе реализована многослойная сверточная сеть, обученная на датасете *cifar-100*, представленным Kaggle [20]. Этот датасет содержит 100 классов, модуль анализа определяет принадлежность медиа контента к одному из этих классов.

Для анализа текстовых публикаций используется модель SocialNetworkModel, которая поставляется в библиотеке с открытым исходным кодом Dostoevsky, предназначенной для анализа русскоязычных текстов [21]. Данная модель обучается на наборе текстов, оставленных в социальной сети ВКонтакте. Система использует эту модель для определения тональности текстов

публикаций группы пользователей. На вход подается список идентификаторов пользователей, затем из базы данных достаются тексты публикаций, выложенные этими пользователями. Тексты проходят обработку – они токенизируются, из полученного набора токенов удаляются стоп-слова, затем из оставшихся токенов выделяются леммы. По набору лемм обученная модель определяет тональность текста. Для каждого пользователя подсчитывается процентное соотношение текстов, имеющих позитивную, негативную и нейтральную тональности. Полученные сведения могут быть прочитаны в текстовом формате, а также могут быть отображены на графике. Осями графиков служат две выбранные тональности, каждая точка на графике – пользователь, координаты которой определяются количеством текстов соответствующей тональности.

Информация о тональности публикаций применяется для дальнейшей кластеризации группы пользователей. Для кластеризации пользователей по тональности публикаций также используется метод К-средних (рис. 11).

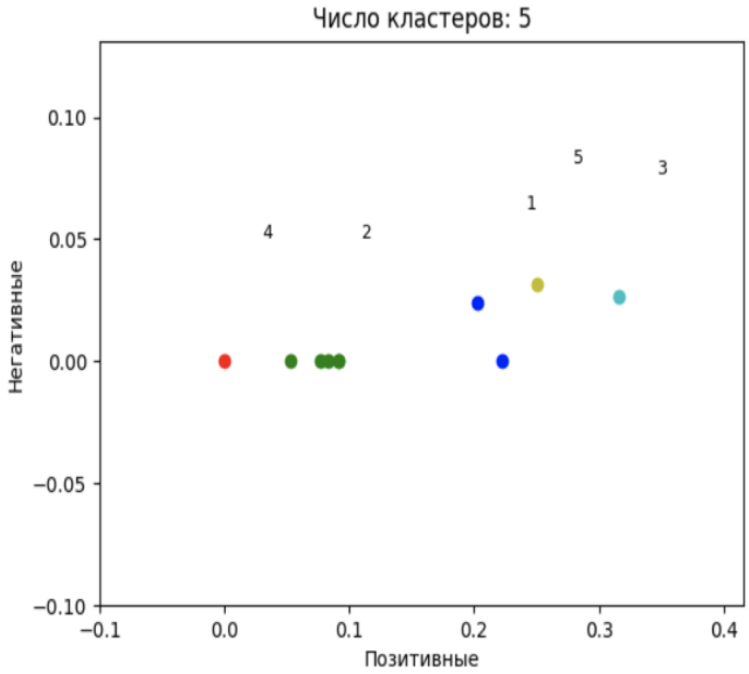

Рис. 11. Анализ тональности публикаций

Анализ результатов участия в мероприятиях

Данный алгоритм позволяет получать информацию о результатах выступлений на различных олимпиадах и конкурсах. Он состоит из двух модулей:

- парсинга содержимого страниц сайтов организаторов мероприятий;

- обработки результатов и протоколов, представленных в формате Excel.

Схема соответствующего алгоритма представлена на рис. 12.

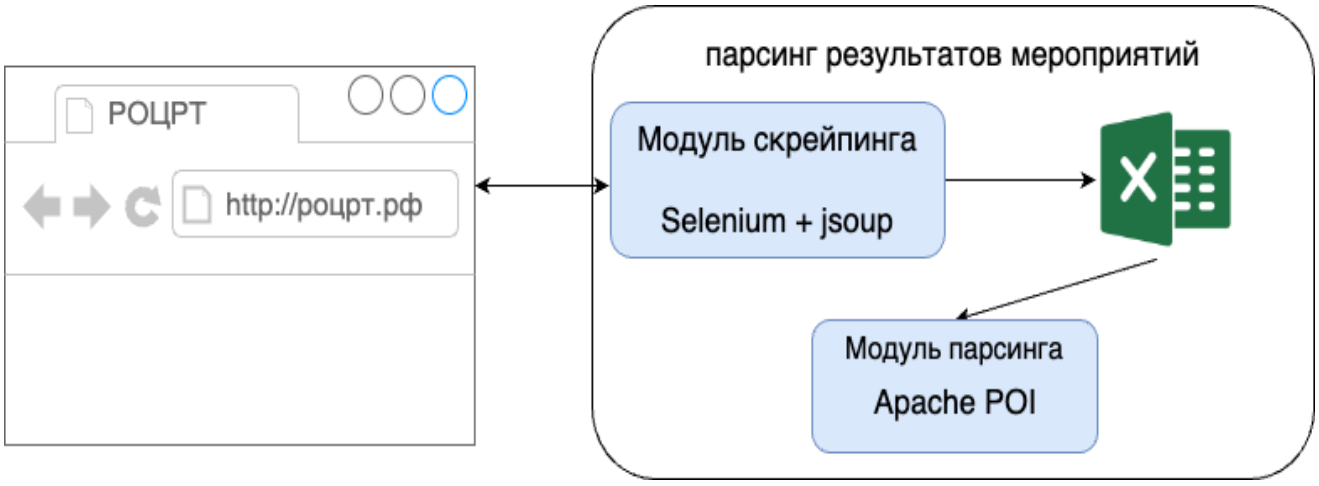

Рис. 12. Алгоритм парсинга результатов

Модуль парсинга страниц реализует нахождение URL необходимых файлов с результатами. Для прохождения по содержимому сайтов организаторов используются Selenium, а также библиотека jSoup [22, 23]. Selenium дает возможность выполнять JavaScript код для получения доступа к определённому полю, в то время как более легковесная библиотека Jsoup дает возможность получения объектов HTML-страниц путем обращения по селектору. Selenium – это веб-драйвер, который использует для работы браузер, поэтому для его запуска необходимо больше ресурсов. Поэтому в системе для получения данных из объектов разметки в первую очередь используется jsoup, и только в случае необходимости выполняются JavaScript скрипты. В случае увеличения количества рассматриваемых сайтов организаторов олимпиад и других мероприятий можно использовать специальные хранилища для правил обработки, в данной же системе для хранения правил используется только память сервера.

Для каждого из организаторов мероприятий в базе данных хранится набор правил для обработки страниц их сайтов. Правила включают в себя селекторы и JavaScript код, который необходимо запустить для получения ссылок на файлы с результатами. После применения правил модуль парсинга предоставляет список ссылок на файлы с результатами.

После того, как ссылки получены, модуль обработки результатов скачивает

файлы с результатами, используя HTTP-клиент. Дальше Excel файлы трансформируются в формат, который может быть использован для хранения и дальнейшего использования. Пример содержимого файла с результатами представлен на рис. 2. После этого содержимое файла фильтруется и считывается в JSON-строки, которые затем записываются в базу данных. Чтение Excel-таблиц в Json происходит с помощью библиотеки Apache POI [24].

Собранная информация позволяет группировать пользователей по заинтересованности в определенных тематиках, а также по успешности выступлений. Согласно работе [25] плотность взаимодействия может трактоваться как сплоченность группы. В системе реализована функциональность, позволяющая, используя информацию о взаимодействии группы пользователей, визуализировать степень социальной сплоченности группы пользователей.

ЗАКЛЮЧЕНИЕ

Рассмотрены источники информации, необходимые для построения цифрового портрета учащегося. Был выделен перечень используемых характеристик. Разработаны методы получения, обработки, и анализа рассматриваемых данных.

Созданы инструменты, оценивающие результаты выступлений учащегося в различных мероприятиях и анализирующие данные из профилей социальных сетей для построения цифрового портрета. Реализованы алгоритмы, позволяющие при построении цифрового портрета оценивать тематику и тональность публикуемых учащимся текстов. Предложены варианты использования цифрового портрета в образовательных целях.

СПИСОК ЛИТЕРАТУРЫ

1. Постановление Правительства Российской Федерации от 07.12.2020 № 2040 «О проведении эксперимента по внедрению цифровой образовательной среды». URL: https://open.edu.gov.ru/files/faq/subjects.pdf (дата обращения: 28.10.2023).

2. Pepper.ninja [Электронный ресурс]. URL: https://pepper.ninja/ (дата обращения: 28.10.2023).

3. Segmento Target [Электронный ресурс].

URL: https://segmento-target.ru/ (дата обращения: 28.10.2023).

4. TargetHunter [Электронный ресурс]. URL: https://targethunter.ru (дата обращения: 28.10.2023).

5. Церебро Таргет [Электронный ресурс]. URL: https://церебро.рф (дата обращения: 28.10.2023).

6. Top 50 open-source web crawlers for data mining [Электронный ресурс]. URL: https://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining (дата обращения: 28.10.2023).

7. 8 Best Web Scraping Tools [Электронный ресурс]. URL: https://hevodata.com/learn/8-best-web-scraping-tools/ (дата обращения: 28.10.2023).

8. Обзор алгоритмов Data Mining [Электронный ресурс]. URL: https://www.intuit.ru/studies/courses/6/6/info (дата обращения: 28.10.2023).

9. Статистический портал «Statista» [Электронный ресурс]. URL: https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/ (дата обращения: 28.10.2023).

10. VK API [Электронный ресурс]. URL: https://vk.com/apiclub (дата обращения: 28.10.2023).

11. VK Java SDK [Электронный ресурс]. URL: https://vk.com/dev/Java_SDK (дата обращения: 28.10.2023).

12. ScribeJava. Simple OAuth library for Java [Электронный ресурс]. URL: https://github.com/scribejava/scribejava (дата обращения: 28.10.2023).

13. OAuth authorization framework [Электронный ресурс]. URL: https://oauth.net (дата обращения: 28.10.2023).

14. REST. Representational State Transfer [Электронный ресурс]. URL: https://restfulapi.net/ (дата обращения: 28.10.2023).

15. JSON. JavaScript Object Notation [Электронный ресурс]. URL: https://www.json.org/ (дата обращения: 28.10.2023).

16. *Черезов Д.С., Тюкачев Н.А.* Обзор основных методов классификации и кластеризации данных // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2009. №. 2. С. 25–29.

17. Scikit-Learn. Machine Learning in Python [Электронный ресурс].

URL: https://scikit-learn.org/stable (дата обращения: 28.10.2023).

18. Numpy. The fundamental package for scientific computing with Python [Электронный ресурс]. URL: https://numpy.org/ (дата обращения: 28.10.2023).

19. Keras. Python deep learning API [Электронный ресурс]. URL: https://keras.io/ (дата обращения: 28.10.2023).

20. Kaggle. the world's largest data science community [Электронный ресурс]. URL: https://keras.io/ (дата обращения: 28.10.2023).

21. Dostoevsky. Sentiment analysis library for Russian language [Электронный ресурс]. URL: https://github.com/bureaucratic-labs/dostoevsky (дата обращения: 28.10.2023).

22. Selenium. Automates browsers [Электронный ресурс]. URL: https://www.selenium.dev/ (дата обращения: 28.10.2023).

23. Jsoup. Java HTML Parser [Электронный ресурс]. URL: https://jsoup.org/ (дата обращения: 28.10.2023).

24. Apache POI. Java API for Microsoft Documents [Электронный ресурс]. URL: https://poi.apache.org/ (дата обращения: 28.10.2023).

25. *Печенкин В.В., Ярская-Смирнова Е.Р.* Сетевые подходы в анализе социальной сплоченности // Вестник Саратовского государственного технического университета. 2014. Т. 4. № 1 (77).

DEVELOPMENT OF METHODS AND SOFTWARE TOOLS FOR THE FORMATION OF A DIGITAL PORTRAIT OF STUDENTS

M. A. Solncev[1] [0009-0002-4106-3035], M. M. Abramskiy[2] [0000-0003-3063-8948]

[1, 2] *Institute of Information Technology and Intelligent Systems of Kazan Federal University*

[1]mrt.solncev@gmail.com, [2]ma@it.kfu.ru

Abstract

This paper considers the questions about the possibility of using data about the students presented in electronic form to build their digital portraits. A set of characteristics necessary for its construction is proposed, a data model is designated.

Implemented tools for collecting data about students from social networks and other Internet resources. Algorithms for constructing a digital portrait are proposed. The application of machine learning algorithms for these tasks is illustrated. Examples of the use of digital portraits in education are given.

Keywords: social networks, data retrieval, personal portrait of user, education

REFERENCES

1. Resolution of the Government of the Russian Federation dated 07.12.2020 No. 2040 "On conducting an experiment on the introduction of a digital educational environment". URL: https://open.edu.gov.ru/files/faq/subjects.pdf (date of access: 28.10.2023).

2. Project Pepper.ninja [Electronic resource]. URL: https://pepper.ninja/ (date of access: 28.10.2023).

3. Project Segmento Target [Electronic resource]. URL: https://segmento-target.ru/(date of access: 28.10.2023).

4. Project TargetHunter [Electronic resource]. URL: https://targethunter.ru (date of access: 28.10.2023).

5. Project Cerebro Target [Electronic resource]. URL: https://церебро.рф (date of access: 28.10.2023).

6. Top 50 open-source web crawlers for data mining [Electronic resource] // bigdata-madesimple.com.
URL: https://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining (date of access: 28.10.2023).

7. 8 Best Web Scraping Tools [Electronic resource] // hevodata.com URL: https://hevodata.com/learn/8-best-web-scraping-tools/(date of access: 28.10.2023).

8. Data Mining Algorithms Overview [Electronic resource] // intuit.ru: URL: https://www.intuit.ru/studies/courses/6/6/info (date of access: 28.10.2023).

9. Leading social media platforms in Russia in 3rd quarter 2022, by monthly penetration rate [Electronic resource] // statista.com.
URL: https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/ (date of access: 28.10.2023).

10. Project VK API [Electronic resource] // vk.com.
URL: https://vk.com/apiclub (date of access: 28.10.2023).

11. VK Java SDK Library [Electronic resource] // vk.com. URL: https://vk.com/dev/Java_SDK (date of access: 28.10.2023).

12. ScribeJava. Simple OAuth library for Java [Electronic resource] // github.com. URL: https://github.com/scribejava/scribejava (date of access: 28.10.2023).

13. OAuth authorization framework [Electronic resource] // oauth.net. URL: https://oauth.net (date of access: 28.10.2023).

14. REST. Representational State Transfer [Electronic resource] // restfulapi.net. URL: https://restfulapi.net/ (date of access: 28.10.2023).

15. JSON. JavaScript Object Notation [Electronic resource] // json.org. URL: https://www.json.org/ (date of access: 28.10.2023).

16. *Cherezov D.S., Tyukachev N.A* Overview of the main methods of data classification and clustering // Bulletin of the Voronezh State University. Series: System Analysis and Information Technologies. 2009. No. 2. P. 25–29.

17. Scikit-Learn. Machine Learning in Python [Electronic resource] // scikit-learn.org. URL: https://scikit-learn.org/stable (date of access: 28.10.2023).

18. Numpy. The fundamental package for scientific computing with Python [Electronic resource] // numpy.org. URL: https://numpy.org/ (date of access: 28.10.2023).

19. Keras. Python deep learning API [Electronic resource] // keras.io. URL: https://keras.io/ (date of access: 28.10.2023).

20. Kaggle. the world's largest data science community [Electronic resource] // kaggle.com. URL: https://keras.io/ (date of access: 28.10.2023).

21. Dostoevsky. Sentiment analysis library for Russian language [Electronic resource] // github.com. URL: https://github.com/bureaucratic-labs/dostoevsky (date of access: 28.10.2023).

22. Selenium. Automates browsers [Electronic resource] // selenium.dev. URL: https://www.selenium.dev/ (date of access: 28.10.2023).

23. Jsoup. Java HTML Parser [Electronic resource] // jsoup.org. URL: https://jsoup.org (date of access: 28.10.2023).

24. Apache POI. Java API for Microsoft Documents [Electronic resource] // apache.org. URL: https://poi.apache.org/ (date of access: 28.10.2023).

25. *Pechenkin V.V., Yarskaya-Smirnova E.R.* Network approaches in the analysis of social cohesion // Bulletin of the Saratov State Technical University. 2014. No. 1 P. 77.

СВЕДЕНИЯ ОБ АВТОРАХ

СОЛНЦЕВ Марат Альбертович – аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Marat Albertovich SOLNTSEV – Master, post graduate (Institute of Information Technology and Intelligent Systems, Kazan Federal University).

email: mrt.solncev@gmail.com

ORCID: 0009-0002-4106-3035

АБРАМСКИЙ Михаил Михайлович – директор Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, кандидат технических наук.

Mikhail Mikhailovich ABRAMSKIY – director of the Institute of Information Technology and Intelligent Systems, Kazan Federal University, PhD (Cand Sci. – Tech.)

email: mabramsk@kpfu.ru

ORCID: 0000-0003-3063-8948

Материал поступил в редакцию 30 октября 2023 года