Решение задачи сбора открытых данных пользователей в социальной сети ВКонтакте с целью дальнейшего исследования возможности использования цифровых следов для решения проблемы профориентации

Э. В. Обухова, email: obukhova.elina@yandex.ru
А. В. Швырева, email: shvyreva@cs.vsu.ru

Воронежский государственный университет

Аннотация. *Цифровой след – это данные, которые пользователи оставляют при использовании сети Интернет. Разделяют «активные» и «пассивные» цифровые следы. К «активным» относится информация, которой пользователь намеренно делится о себе, например, личная информация, публикации, сообщения, комментарии. Пассивный цифровой след собирается без ведома пользователя, и может использоваться, например, для подбора контекстной рекламы. В данной работе рассматривается задача сбора и подготовки набора данных, на основе активных цифровых следов пользователей в социальных сетях для дальнейшего его использования с целью решения проблемы профориентации. Проводится формирование набора критериев для выбора подходящей социальной сети, а также определяется перечень необходимой собираемой информации.*

Ключевые слова: *цифровые следы, социальная сеть, набор данных, профориентация.*

Введение

Невозможно представить современное цифровое пространство без социальных сетей, где пользователи открыто размещают личную информацию, общаются, заявляют о своих интересах и высказывают собственное мнение. Вся опубликованная информация составляет цифровой след человека, который позволяет составить цифровой портрет человека, рассказать о его характере и предрасположенностях. В результате возникает вопрос о возможности использования цифровых следов из социальных сетей для предсказания профессиональных склонностей человека.

Недостатком существующих профориентационных мероприятий является использование наборов стереотипных вопросов с нечеткими альтернативами, где испытуемые могут скомпрометировать результат, что обеспечивает низкий индекс доверия к полученной оценке. В то же

время использование цифровых следов в сети Интернет для анализа позволит обеспечить большую надежность и объективность полученных данных.

В данной работе рассматривается решение задачи подготовки набора данных [1] для проведения дальнейшего исследования, в том числе формирование набора требований к необходимой пользовательской информации в социальных сетях, изучение популярных социальных сетей.

1. Выбор социальной сети

Для получения единообразной, полной информации были сформированы критерии, которым должна отвечать социальная сеть, используемая для сбора данных:

1. Наличие открытого API для отправки запросов.
2. Широкая аудитория.
3. Наличие сообществ образовательных организаций в социальной сети.
4. Наличие возможности указать информацию о полученном образовании в профиле пользователя.

При анализе существующих социальных сетей было определено, что только социальная сеть ВКонтакте отвечает всем представленным требованиям: пользователь имеет возможность подробно указать информацию об образовании (вуз, факультет, направление, год начала и окончания обучения, возможность добавления нескольких образований); а также во ВКонтакте располагается множество тематических сообществ, включая сообщества образовательных учреждений.

Использование информации из профилей ВКонтакте не нарушает законов о передаче и использовании персональной информации, поскольку открытый API, предоставляемый ВКонтакте для разработчиков, не позволяет получить доступ к информации из закрытых профилей, а информация, публикуемая в открытых профилях, доступна для просмотра всем и не является приватной.

2. Определение требований к собираемым данным

ВКонтакте предоставляет пользователю возможность не только очень подробно описать свою личность, но и публиковать посты, фотографии, слушать музыку, вступать в тематические сообщества. Все перечисленные действия могут рассматриваться в качестве цифровых следов, однако для нашей постановки задачи было решено остановиться только на перечне тематических сообществ, которые наиболее точно отражают интересы и увлечения пользователя.

Таким образом, сформируем набор данных из профиля респондента, необходимых для проведения исследования:

- числовой идентификатор вуза, в котором учился респондент;
- название вуза;
- числовой идентификатор факультета;
- название факультета;
- список из пар: числовой идентификатор группы + название группы.

Персональный идентификатор пользователя фиксироваться не будет, что позволит сохранить анонимность используемых данных.

3. Обзор инструментов взаимодействия с ВКонтакте

Методы VK API разнообразны и делятся на десятки категорий [2-4], но для данной задачи представляют интерес методы классов Groups и Users: Groups.getMembers() и Users.getSubscriptions().

Метод Groups.getMembers() позволяет извлечь список членов группы, а также основную информацию о пользователях. По-умолчанию извлекается:

- цифровой идентификатор, ФИО;
- основная информация об образовании: идентификатор вуза и его название + идентификатор школы и ее название;
- дата последнего появления в сети.

Также существует возможность указать дополнительные поля, чтобы расширить список возвращаемой информации о члене группы. В нашем случае для образовательных целей указывается параметр education, который возвращает полную информацию об образовании:

- идентификатор вуза и его название;
- идентификатор кафедры и ее название;
- год поступления и год выпуска;
- идентификатор школы и ее название;
- года поступления и выпуска.

Метод Users.getSubscriptions() возвращает список всех подписок пользователя, включая людей и сообщества вперемешку, так как ВКонтакте позволяет подписываться не только на сообщества, но и на других пользователей.

4. Первичный анализ данных

В результате сбора и первичной обработки данных [5-6], включающей удаление дубликатов, некорректных и неполных записей данных было получено 8536 профилей пользователей, гистограмма распределения которых по факультетам представлена на рисунке 1. Полученное распределение неравномерно и содержит шумовые

одиночные значения, которые также являются выбросами и не подходят для собираемого набора данных.

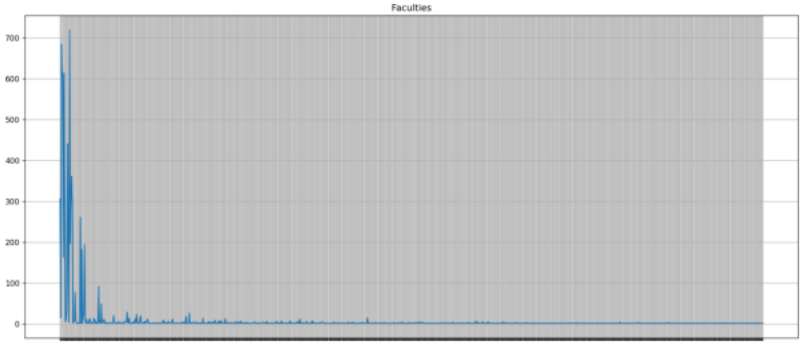

Рис. 1. Распределение респондентов по факультетам

Зная, что всего в ВГУ 18 факультетов, не составит труда отсечь лишнее. В результате получаем 6067 профилей:

- медико-биологический – 33 профиля,
- военного образования – 9 профилей,
- географии, геоэкологии и туризма – 182 профиля,
- геологический – 194 профиля,
- журналистики – 361 профилей,
- исторический – 306 профилей,
- компьютерных наук – 441 профиль,
- математический – 208 профилей,
- международных отношений – 261 профиль,
- прикладной математики, информатики и механики – 601 профиль,
- романо-германской филологии – 618 профилей,
- фармацевтический – 197 профилей,
- физический – 719 профилей,
- филологический – 301 профилей,
- философии и психологии – 163 профиля,
- химический – 204 профиля,
- экономический – 614 профилей,
- юридический – 685 профилей.

Фрагмент полученного набора данных представлен на рисунке 2. Первая строка в нем является заголовками столбцов, где нулевой столбец содержит в себе порядковый номер факультета, а столбцы 1-947 содержат в себе уникальные идентификаторы групп.

0	1	2	3	...	943	944	945	946	947
2	160905377	67363111	93077522	...	0	0	0	0	0
9	10889156	182875281	41768412	...	0	0	0	0	0
8	40836944	23213239	96245789	...	0	0	0	0	0
9	150718140	39009769	86179596	...	0	0	0	0	0
1	65178314	15755094	121258913	...	0	0	0	0	0
1	54361599	206961598	79268570	...	0	0	0	0	0
0	190339907	187218622	179169425	...	0	0	0	0	0
6	38629367	66420111	170182113	...	0	0	0	0	0
2	34215577	139923997	93330757	...	0	0	0	0	0
1	72918254	60841972	125387396	...	0	0	0	0	0

Рис. 2. Фрагмент полученного набора данных

5. Семантический анализ названий сообществ

Руководствуясь идеей, что для каждого факультета характерны уникальные термины, был составлен словарь слов для каждого факультета, определяемый общими словами среди десяти наиболее часто встречаемыми слов в названиях сообществ среди подписок каждого представителя факультета.

В результате проведенного эксперимента было определено:

– наиболее часто встречающимися словами являются разделители, не несущие смысловой нагрузки (например: -, |, ♥, ●);

– на втором месте по частоте использования находятся предлоги, союзы и междометия;

– третьими по частоте являются слова, указывающие на географическое положение (Воронеж, Липецк, Старый Оскол, Российская и т.д.);

– часто встречаются аббревиатуры, прямо указывающие на вуз и факультет (ВГУ, ФКН, Истфак, ПММ);

– использование специальных терминов (история, наука, психология, английский, немецкий, проект);

– единичные слова, не несущие информации о профессиональных склонностях респондентов.

Полученный анализ семантического состава групп пользователей ВКонтакте, вступивших в группу ВГУ, свидетельствует о наличии взаимосвязи между профессиональными склонностями личности и названиями тематических сообществ в его профиле.

Заключение

В результате работы было разработано программное средство, позволяющее собрать публично доступную информацию о студентах Воронежского государственного университета. Также была проведена первичная обработка данных, а также семантический анализ названий тематических сообществ.

Первичный анализ полученных данных свидетельствует о возможности использования цифровых следов для составления не только психологического, но и профессионального портрета человека, что потенциально может помочь многим молодым людям определиться с выбором программы обучения.

Список литературы

1. О важности датасета и о том, как сделать его лучше [Электронный ресурс] : блог компании «Технологика». – Режим доступа: https://www.technologika.ru/blog/how-to-create-great-dataset
2. Библиотека Requests: HTTP for Humans [Электронный ресурс] : блог сообщества MoscowPython. – Режим доступа: https://python.ru/post/97/
3. VK library [Электронный ресурс] : документация сервиса ВКонтакте. – Режим доступа: https://vk.readthedocs.io/en/latest/index.html
4. Python Requests Tutorial [Электронный ресурс] : сайт GeeksForGeeks. – Режим доступа: https://www.geeksforgeeks.org/python-requests-tutorial/
5. Как подготовить датасет к Machine Learning с PySpark и построить систему потоковой аналитики больших данных на Apache Kafka и ELK: пример прогнозирования CTR [Электронный ресурс] : блог компании «BD School». – Режим доступа: https://www.bigdataschool.ru/blog/ctr-prediction-with-kafka-spark-elk-case.html
6. Как собрать датасет за неделю: опыт студентов магистратуры «Наука о данных» [Электронный ресурс] : блог компании «SkillFactory». – Режим доступа: https://habr.com/ru/company/skillfactory/blog/534682/