М.А. Пухачева[1], В.Т. Елиманова[1], И.Р. Мансурова[1], М.С. Воробьева[12]

[1] Тюменский государственный университет, г. Тюмень

[2] Научно-технический университет «Сириус», г. Сочи

УДК 004.912

ИССЛЕДОВАНИЕ ПОДХОДОВ ДЛЯ ОПРЕДЕЛЕНИЯ ИНТЕРЕСОВ ГРУПП СТУДЕНТОВ В СОЦИАЛЬНОЙ СЕТИ «ВКОНТАКТЕ»

Аннотация. В статье рассматривается исследование подходов для определения интересов на основе данных студентов ИМиКН в социальный сети «Вконтакте». Проанализированы результаты определения спрогнозированных интересов группы студентов с использованием кластерного анализа методом k-средних, кластеризации графов и словарного подхода, проведена экспертная оценка и сделаны выводы.

Ключевые слова: сбор данных, извлечение данных, анализ данных, классификация, кластеризация, графы, Python, социальная сеть.

Социальные сети направлены на создание сообществ со схожими интересами, взглядами, увлечениями. На сегодняшний день одними из самых активных пользователей таких интернет-площадок являются студенты, почти каждый из них зарегистрирован в одной из популярных социальных сетей для нахождения своих единомышленников, интересных сообществ, подходящих мероприятий. Поэтому для университета вопрос об определении интересов групп студентов становится важным, так как это позволяет отследить активность и увлеченность студента в разных областях, чтобы предложить соответствующие конкурсы и форумы, курсы по выбору и элективные дисциплины и улучшить поиск целевой аудитории для формирования студенческих сообществ.

Перед авторами работы была поставлена задача провести исследование методов для определения интересов на данных по студентам Института математики и компьютерных наук в социальной сети «Вконтакте», в которой

определено 28 тематик для сообществ и групп. В исследовании было решено выделить 17 из них: автомобили, спорт, игры, хореография, IT, кинематограф, образование и наука, дизайн и иллюстрация, литература, туризм и активный отдых, культура и искусство, животные, музыка, фотография, СМИ и блоггинг, мода, экономика и финансы.

Были использованы текстовые данные о подписках и записей студентов ИМиКН, полученные с помощью интерфейса VK API. Для работы определены классы тематик T= $[t_0, t_2, \ldots, t_{16}]$ и выделен класс студентов S = $[St_1, S_2, \ldots, S_{210}]$. Для каждого студента S_j определено количество групп в подписках в каждой тематике $[k_{t_0}, k_{t_1}, k_{t_2}, \ldots, k_{t_{16}}]$.

В качестве подходов для определения интересов пользователей были рассмотрены следующие методы: кластеризация графов, метод k-средних и словарный подход.

На первом этапе исследования был выполнен *кластерный анализ методом k-средних*, основной задачей которого является разбиение множества объектов определенной структуры на подмножества по некоторым комбинированным признакам, при этом объекты одного множества имеют примерно одинаковые характеристики [1].

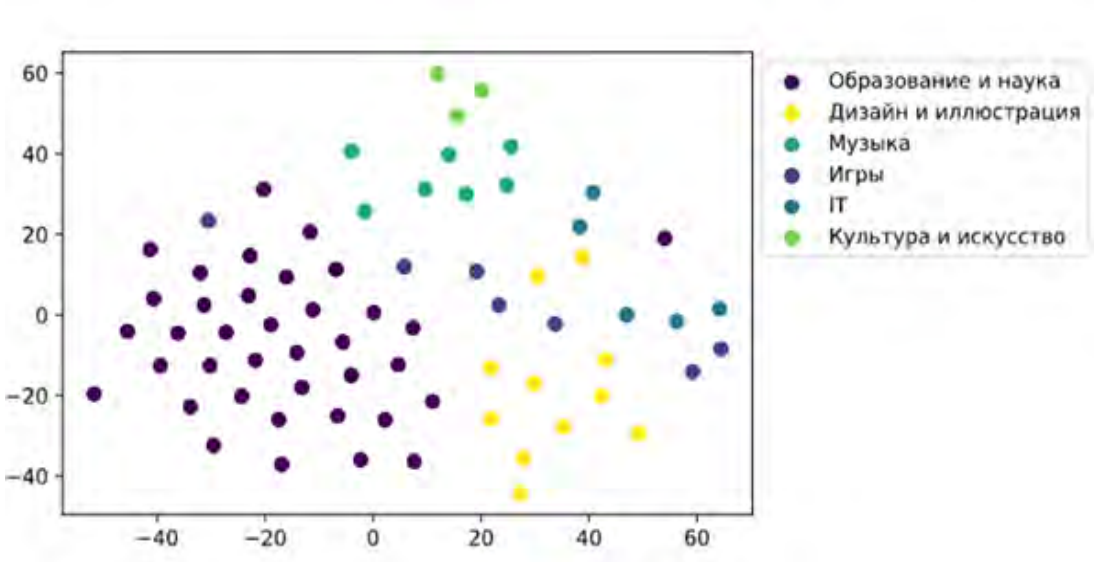

Рис. 1. Визуализация результата работы метода k-средних

Результат работы метода k-средних показывает соотношение студентов, относящихся к тематике в каждом кластере (см. рис. 1). Самым многочисленным оказался кластер с темой «Образование и наука», что объяснимо, так как исследуемыми пользователями были студенты.

На втором этапе исследования рассматривался *алгоритм кластеризации графов*, который можно представить в виде графа, вершинами которого будут являться студенты, объединенные общими интересами.

По входным данным был построен полный, неориентированный, взвешенный граф G, веса в котором между двумя вершинами рассчитывались как евклидово расстояние в 17-мерном пространстве [2].

Использовав алгоритм Р. Прима, было построено минимальное остовное дерево в графе G и отсечено (m-1) ребер с максимальным весом, где m – это количество желаемых кластеров, полученное эмпирическим анализом [2].

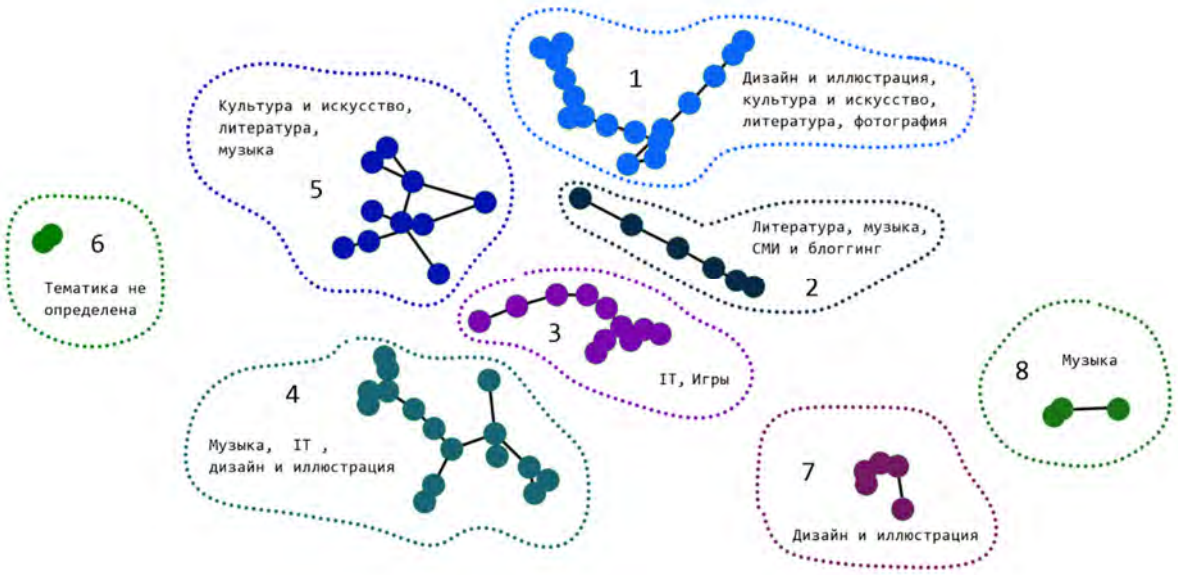

Рис. 3. Визуализация результата работы кластеризации графа

В результате работы *алгоритма кластеризации графов* обнаружилось, что шестой кластер оказался с неопределенной тематикой (см. рис. 3). Исследовав страницы студентов, вошедших в кластер, выяснилось, что у них меньше пяти сообществ в подписках, что затруднило поиск похожих пользователей (см. табл. 1).

Таблица 1. Описание кластеров

№	Тематика кластера	Количество пользователей
1	Дизайн и иллюстрация, Культура и искусство, Литература, Фотография	49
2	Литература, Музыка, СМИ и блоггинг	24
3	IT, Игры	29
4	Музыка, IT, Дизайн и иллюстрация	41
5	Культура и искусство, Литература, Музыка	28
7	Дизайн и иллюстрация	23
6	Тематика не определена	3
8	Музыка	13

На третьем этапе исследования для определения интересов студентов с помощью *словарного подхода* была произведена предобработка и группировка данных. По входным данным пользователя был получен «мешок слов» [4]. Для работы алгоритма был спроектирован словарь объемом 7645 слов со следующей структурой:

$$\{«W_1»: t_0,$$
$$«W_2»: t_1,$$
$$\dots\ \dots\ \dots\ \dots\ ,$$
$$«W_{7645}»: t_{16}\}.$$

Количественное распределение слов в словаре по тематикам приведено в таблице 2.

Таблица 2. Словарь по тематикам

№	Тема	Кол-во	№	Тема	Кол-во
0	IT	602	9	музыка	559
1	автомобили	503	10	образование и наука	577
2	дизайн и иллюстрация	497	11	сми и блоггинг	305
3	животные	344	12	спорт	572

№	Тема	Кол-во	№	Тема	Кол-во
4	игры	367	13	туризм и активный отдых	349
5	кинематограф	446	14	фотография	287
6	культура и искусство	470	15	хореография	362
7	литература	350	16	экономика и финансы	535
8	мода	520			

В результаты работы метода для каждого студента S_j построен список $[k_{t_0}, k_{t_1}, k_{t_2}, \dots, k_{t_{16}}]$, из которого выбираются 3 максимальных элемента такие, что $max_1 \geq max_2 \geq max_3$, и подсчитывается среднее арифметическое всех элементов V. Так как студенты могут иметь несколько интересов, был спроектирован алгоритм индивидуального отбора тематик, в котором C – количество интересов для каждого студента S_j (рис. 4).

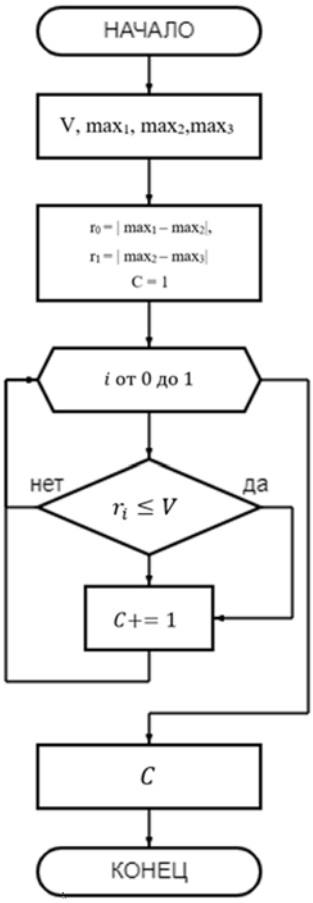

Рис. 4. Блок-схема алгоритма выбора количества интересов

Выходные данные представляют собой самые интересные тематики для каждого студента.

Результаты работы словарного подхода в разрезе тематик показали, что большинство студентов ИМиКН интересуется образованием и наукой, в то время как в данной группе хореографией никто не увлекается (рис. 5).

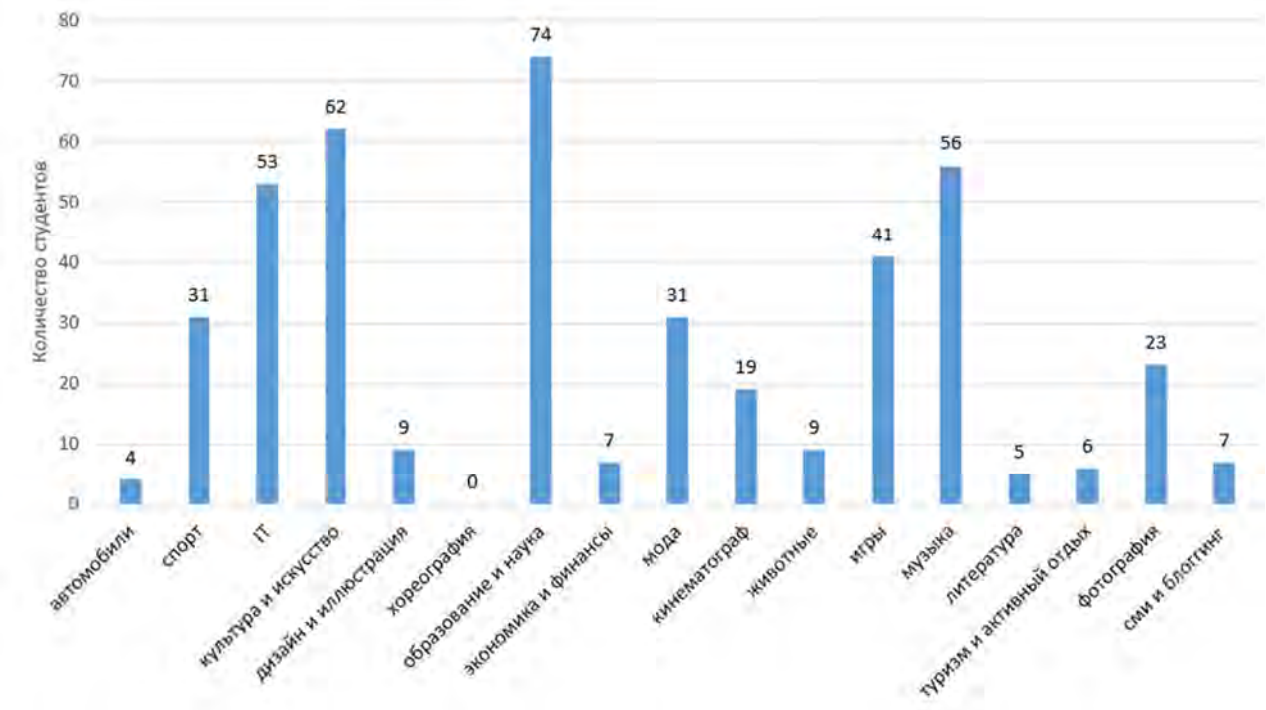

Рис. 5. Результаты работы словарного подхода

После проведения всех этапов исследования было сделано экспертное заключение. Сравнительный анализ экспертной оценки показал, что на текущих данных кластерный анализ методом k-средних определяет интерес верно в 67% случаев, кластеризация графов в 44%, а словарный подход в 61% (см. рис. 6).

Также стоит отметить, что кластерный анализ методом k-средних определяет только один интерес, поэтому в экспертной оценке результатом может быть только однозначный ответ (см. рис. 6).

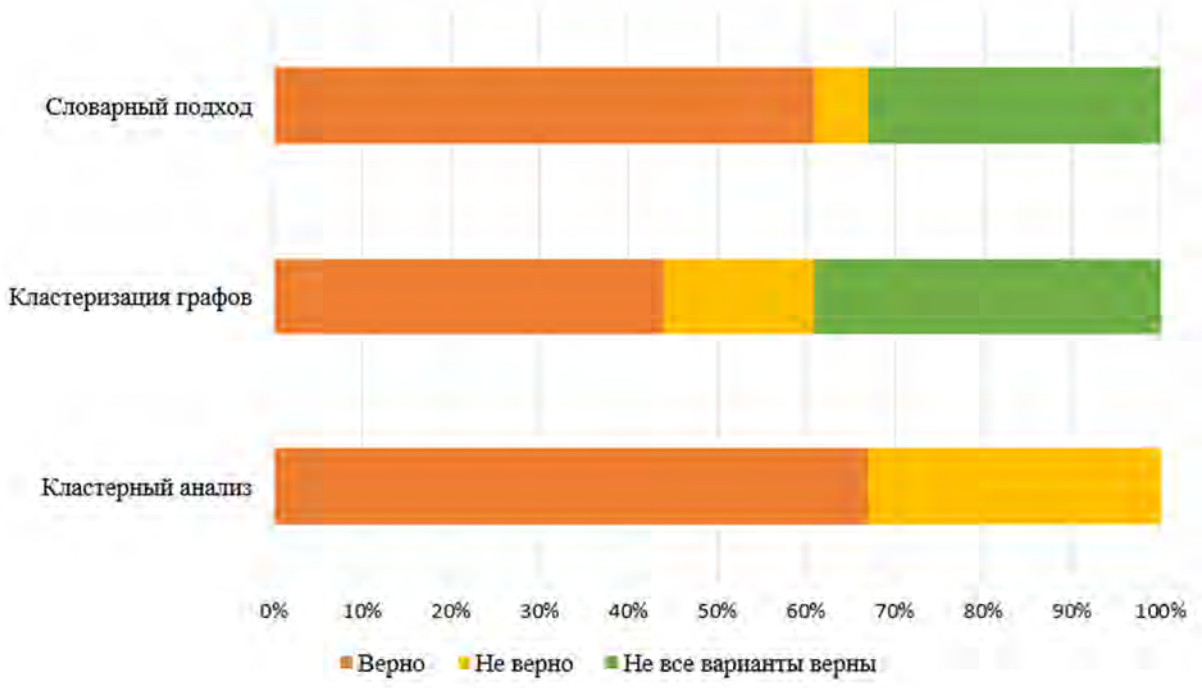

Рис. 6. Анализ экспертной оценки

Таким образом, реализовав и проанализировав методы для определения студентов ИМиКН на группы по интересам в социальной сети ВКонтакте, можно сделать вывод:

- кластерный анализ лучше справился с поставленной задачей, но, в отличие от остальных алгоритмов, интерес может быть только один;
- кластеризация графов больше подходит для выделения групп с похожими интересами из-за структуры модели алгоритма;
- словарный метод подходит для выявления нескольких интересов, но требует объемного словаря и дальнейших исследований с определением наилучшей метрики для выбора количества тематик.

Благодарности

Статья подготовлена в рамках разработки образовательного кейса для НТУ Сириус при финансовой поддержке РФФИ в рамках научного проекта № 19-37-51028.

СПИСОК ЛИТЕРАТУРЫ

1. Воробьева, М.С., Дубаков, А.А. Разработка приложения для определения тематики текста с использованием алгоритмов кластеризации [Электронный ресурс] / М.С. Воробьева, А.А. Дубаков. – Электрон. журн. – 2019. – No17. – С. 85-95. – Режим доступа: https://www.elibrary.ru/item.asp?id=41380880 (дата обращения: 29.04.2021).

2. Соломатин, Д.И., Гаршин, Т.С. Построение и анализ социальных графов пользователей сети “Вконтакте”. Сборник студенческих научных работ факультета компьютерных наук ВГУ / под ред. Д.Н. Борисова. Воронежский государственный университет. – Вып. 11. – Воронеж: Издательский дом ВГУ, 2017. – 389 с. С. 53-58. – Режим доступа: http://www.cs.vsu.ru/stwork/download.html (дата обращения: 29.04.2021).

3. Благов, А.В. Анализ социальных сетей: учебное пособие / А.В. Благов, И. А. Рыцарев. – Самара: Издательство Самарского университета, 2020. – 104 с.: ил. – Режим доступа: http://repo.ssau.ru/handle/Uchebnye-izdaniya/Analiz-socialnyh-setei-ucheb-posobie-Tekst-elektronnyi-88042 (дата обращения: 20.04.2021).

4. Воробьева, М.С., Якубов, Р.М., Бильдин, С.М., Дроздецкий, М.Д. Разработка веб-приложения для формирования цифрового профиля студента. [Электронный ресурс] / Математическое и информационное моделирование: материалы Всероссийской конференции молодых ученых, г. Тюмень, 2020 г. – Тюмень: Изд-во ТюмГУ, 2020. – Вып.18 – 735 с. С. 299-306. – Режим доступа: https://library.utmn.ru/dl/PPS/Bidulja_2020_18.pdf/info (дата обращения: 29.04.2021).

5. Еременко, В.Т., Сазонов, М.А., Шекшуев, С.В. Исследование подходов к созданию информационных моделей для сбора и обработки данных социальных сетей. [Электронный ресурс] / В.Т. Еременко, М.А. Сазонов, С.В. Шекшуев. – Электрон. журн. – 2019. – Т. 17. – С. 118-129. – Режим доступа: https://www.elibrary.ru/item.asp?id=39195692 (дата обращения: 20.04.2021).

6. Коршунов, А., Белобородов, И., Бузун Н., Аванесов, В., Пастухов, Р., Чихрадзе, К., Козлов, И., Гомзин, А., Андрианов, И., Сысоев, А., Ипатов, С., Филоненко, И., Чуприна, К., Турдаков, Д., Кузнецов, С. Анализ социальных сетей: методы и приложения. [Электронный ресурс] / Труды ИСП РАН. 2014. №1. – Режим доступа: https://cyberleninka.ru/article/n/analiz-sotsialnyh-setey-metody-i-prilozheniya (дата обращения: 29.04.2021).

7. Донцов, Д.Ю. Применение методов кластеризации для анализа использования интернет-ресурсов. – Красноярск : СФУ, 2017. – С. 13-21. – Режим доступа: http://elib.sfu-kras.ru/handle/2311/67621 (дата обращения: 29.04.2021).

8. Jinpeng, W., Wayne, X.Z., Yulan, H. Infer User Interests via Link Structure Regularization // ACM Transactions on Intelligent Systems and Technology (TIST) Volume 5 Issue 2, April 2014 Article No. 23. – Режим доступа: https://scholar.google.co.jp/citations?user=JNhNacoAAAAJ&hl=zh-CN (дата обращения: 29.04.2021).