Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2019, №2, Том 11 / 2019, No 2, Vol 11 https://esj.today/issue-2-2019.html

URL статьи: https://esj.today/PDF/102ITVN219.pdf

Ссылка для цитирования этой статьи:
Хлопотов М.В., Старцева Н.В., Макаренко А.А. Исследование кластеров кинолюбителей и их тематических сообществ в социальных сетях // Вестник Евразийской науки, 2019 №2, https://esj.today/PDF/102ITVN219.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

For citation:
Khlopotov M.V., Startseva N.V., Makarenko A.A. (2019). Analysis of movie lovers' preferences and their thematic communities in social networks. *The Eurasian Scientific Journal*, [online] 2(11). Available at: https://esj.today/PDF/102ITVN219.pdf (in Russian)

УДК 004

Хлопотов Максим Валерьевич
ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Санкт-Петербург, Россия
Доцент
Кандидат технических наук
E-mail: khlopotov@corp.ifmo.ru

Старцева Наталья Владимировна
ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Санкт-Петербург, Россия
Магистрант
E-mail: st.natalie.eilatan@gmail.com
SCOPUS: http://www.scopus.com/authid/detail.url?authorId=57195920234

Макаренко Анастасия Александровна
ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Санкт-Петербург, Россия
Магистрант
E-mail: AnastasiaMakarenko95@yandex.ru

Исследование кластеров кинолюбителей и их тематических сообществ в социальных сетях

Аннотация. В статье описано исследование предпочтений наиболее активных пользователей «Кинопоиска» и процесс выявления их взаимосвязей с тематическими сообществами таких пользователей в сети «ВКонтакте».

Для проведения исследования были собраны данные об оценках кинолюбителей на сайте «Кинопоиск». Были использованы лишь оценки для недавно вышедших картин для поддержания актуальности данных. Более десяти тысяч полученных профилей пользователей были поделены на группы с помощью различных алгоритмов кластеризации. Для определения оптимального числа кластеров был использован пакет NbClust, применяющий 30 разных методов для определения количества подгрупп данных. Большинство алгоритмов выбора оптимального числа кластеров выбрали число 3 как наилучший вариант.

Для измерения эффективности кластеризации была использована метрика силуэта. Данная метрика показала, что наиболее эффективным алгоритмом является алгоритм k-средних. Таким образом, итоговая модель кластеризации делила набор данных на 3 группы и была построена при помощи k-средних.

Для пользователей выборки были собраны сообщества с прикрепленной страницы «ВКонтакте», если таковая имелась. При анализе наиболее популярных сообществ среди кластеров были исключены сообщества кинолюбителей, поскольку они присутствовали в каждом кластере и были помехой при определении специфики подгруппы. Таким образом, анализ групп «ВКонтакте» и самых высоко оцененных фильмов помог выявить тенденции тематик каждого кластера. Среди пользователей были выделены три группы – группа «фанатов комиксов», группа «интеллектуалов» и группа «юмористов». Результаты исследования могут быть использованы при разработке маркетинговых стратегий и персонализации.

Ключевые слова: анализ социальных сетей; онлайн-сообщество; кластеризация; k-средних; силуэт; анализ данных; Кинопоиск; ВКонтакте

Введение

Социальные сети предоставляют практически безграничные возможности для выявления тенденций современного общества, выделения социальных подгрупп и анализа предпочтений пользователей. Каждая социальная сеть концентрируется на нескольких сегментах потребностей пользователей и предоставляет сервисы, основываясь на них. Собрав данные из двух различных социальных сетей и соединив их в единую модель, можно выявить взаимосвязи и получить более глобальную картину сведений о пользователях и сообществах, что может стать полезным инструментом для формирования маркетинговых стратегий и алгоритмов персонализации. Нами выбраны два сервиса, распространенных на территории Евразии: «ВКонтакте» и «Кинопоиск».

«ВКонтакте» – социальная сеть, имеющая огромную популярность на территории России и стран СНГ. «ВКонтакте» зарегистрировано более 543 000 000 пользователей. Заполненная страница пользователя располагает данными о его интересах, хобби и различных активностях. Уже проведены разнообразные исследования этого веб-сервиса, например, анализ коэффициентов вовлеченности пользователей и тематических профилей [1], а также социальных структур сообществ [2].

«Кинопоиск» – русскоязычный интернет-проект, посвященный кинематографу. Является самым посещаемым ресурсом про кинематограф в России и русскоязычных странах. Каждый фильм на «Кинопоиске» может быть оценен по 10-ти балльной шкале – от единицы (хуже некуда) до десятки (шедевр). Количество оценок, которое ставят пользователи ежемесячно, в среднем превышает два миллиона [3].

Можно смело предположить», что пользователи, часто оценивающие кино на портале, являются кинолюбителями. Однако предпочтения пользователей могут сильно варьироваться – кому-то нравятся фильмы одного жанра или совокупности жанров, кому-то – совсем других. Эти зависимости могут быть неявными, скрытыми – здесь могут иметь влияние не только жанры, но и множество других факторов, различных характеристик кинокартин. Также анализ этих предпочтений может быть использован в рекомендациях и оценках фильмов и может помогать директорам и маркетологам кинотеатров [4].

Пользователи могут делиться на подгруппы не только в зависимости от их предпочтений, выраженных оценками на портале «Кинопоиск», но также по сообществам интересов «ВКонтакте». Для интерпретации полученных результатов и определения потенциальных «мест скопления» представителей тех или иных кластеров мы сопоставляем данные двух социальных сетей.

Цель работы. Исследовать, насколько хорошо коррелируют подгруппы, выделенные на основании членства в тематических сообществах, с подгруппами по кинематографическим интересам.

1. Данные

Первым шагом, как и в любой задаче по анализу данных, стал сбор данных и их подготовка. Сбор проводился параллельно для обеих социальных сетей, после этого полученные данные сравнивались и подвергались предварительной очистке.

1.1 Данные из «Кинопоиска»

Для того, чтобы **кластеризовать** пользователей «Кинопоиска», необходимо было собрать информацию об их оценках различным кинокартинам. Оценки на «Кинопоиске» ранжируются от 0 до 10, что представляет обширные возможности для выделения схожих по вкусам групп пользователей.

Поскольку многие аккаунты могут быть заброшенными или неактуальными, было решено отталкиваться от обратного – собрать недавно вышедшие фильмы и сериалы, а затем выгрузить пользователей, их оценивших. Это бы обеспечило релевантность и относительную актуальность вкусовых предпочтений пользователей.

Все дальнейшие манипуляции для сбора, обработки данных и построения моделей были выполнены с помощью языка программирования R в среде разработки RStudio. В частности, с использованием библиотеки «RSelenium» [5].

К сожалению, у «Кинопоиска» нет открытого API для разработчиков, поэтому необходимо было разработать парсер для сбора информации со страниц «Кинопоиска» при помощи CSS и XPath селекторов, специфицированный под нужды данного исследования. Разработанный парсер помог извлечь следующую информацию:

1. информацию об оценках заданного фильма;
2. общую информацию о заданном пользователе (например, имя или ссылку на «ВКонтакте»);
3. информацию об оценках заданного пользователя.

При помощи построенного парсера, на этапе сбора и обработки данных были выполнены следующие шаги:

- Собраны фильмы, вышедшие за 2017 и 2018 гг., имеющие хотя бы одну тысячу оценок на портале. Порог был необходим для того, чтобы отсеять малоизвестные картины, которые могли бы «засорить» выборку и представлять собой наименее информативные признаки. Всего под эти критерии подошло 1479 кинолент.
- Для каждого фильма были собраны пользователи, поставившие ему оценку или «просмотр» (в случае «просмотра» оценка считалась нулем, что равнозначно отсутствию оценки). К сожалению, «Кинопоиск» предоставляет возможность просматривать оценки к фильмам лишь последней тысячи пользователей. Всего в наборе данных получилось более 250 тысяч уникальных пользователей.
- Лимит портала на «последнюю тысячу оценок» фильма означал, что не все оценки пользователей присутствуют в полученной матрице. Необходимо было

обратиться к аккаунту конкретного пользователя и найти «пропущенные» оценки.

- Чтобы сразу отсеять случайных или неактивных пользователей, были отфильтрованы все пользователи, у которых в первоначальной таблице было отмечено менее 20 кинолент. Таким образом, набор данных из 11538 пользователей и их оценок к 1479 фильмам и сериалам был дополнен пропущенными оценками и стал основой для дальнейшей обработки и кластеризации пользователей.

Также для каждого из пользователей финальной выборки была собрана привязанная страница «ВКонтакте». Всего пользователей, имеющих ссылку на такую страницу, оказалось 2232.

1.2 Данные из «ВКонтакте»

Для каждого пользователя из списка, полученного на первом этапе сбора данных, были собраны основные характеристики (такие как имя, фамилия, пол и статус активности) и сообщества «ВКонтакте», в которых пользователь состоит. Сбор данных также осуществлялся с помощью языка программирования R в среде разработки RStudio. Дополнительно был использован пакет, предназначенный для взаимодействия среды R с API ВКонтакте – «vkR» [6]. В отличие от API «Кинопоиска», API «ВКонтакте» поддается более удобному парсингу – для разработчиков существует целый раздел с документацией и доступными методами.

Полученный набор данных подвергся очистке, так как в изначальном виде он имел следующие несовершенства:

1. Некоторые идентификаторы повторялись в списке. Вероятно, пользователи давали ссылку не на свою страницу «ВКонтакте» или имели несколько аккаунтов. Дубликаты были удалены из списка.
2. Встречались идентификаторы, указывающие на удаленные или заблокированные страницы. Такие пользователи также были исключены из списка.

В итоге, из изначально имевшихся 2258 идентификаторов пользователей для дальнейшего анализа остались 2082.

Следующим шагом стало создание таблицы признаков. Членство в каждой из групп стало бинарным признаком, каждый из которых в последствии будет использован для деления на кластеры.

2 Моделирование и оценка

2.1 Проблемы кластеризации

Кластеризация – это метод машинного обучения без учителя для разделения набора данных на набор групп или кластеров. Одной из проблем является то, что методы кластеризации будут возвращать кластеры, даже если данные их не содержат [7]. Таким образом, необходимо:

1. оценить тенденцию кластеризации перед анализом;
2. проверить качество результата после кластеризации.

В литературе было предложено множество мер для оценки результатов кластеризации. Термин «проверка кластеризации» используется для обозначения процедуры оценки результатов алгоритма кластеризации.

Среди задач этапа моделирования мы выделяем следующие пункты:

1. Подобрать оптимальное количество кластеров для анализируемых данных.
2. Сравнить качество результатов кластеризации, полученных с помощью различных алгоритмов кластеризации.
3. Провести валидацию кластеров для алгоритма с наилучшими показателями качества.

На рассмотрение внесены следующие алгоритмы: k-means, k-medoids (PAM) и иерархическая кластеризация.

2.2 Выбор количества кластеров

Определение оптимального количества кластеров в наборе данных является фундаментальной задачей в кластеризации. Как пример, метод k-means требует от пользователя указания количества кластеров k, которые должны быть сгенерированы. В большинстве случаев оптимальное количество кластеров в некоторой степени субъективно и зависит от метода, используемого для измерения сходства, и параметров, используемых для разбиения.

Мы использовали функцию NbClust из пакета NbClust R [8]. Она предоставляет 30 показателей для определения оптимального количества кластеров и предлагает пользователям лучшую схему кластеризации, полученную при варьировании всех комбинаций количества кластеров, мер расстояния и методов кластеризации.

Применив функцию к обоим наборам данных, мы получили следующие результаты (таблица 1).

Таблица 1

Результаты работы алгоритма NbClust

Данные из Кинопоиска	Данные из ВКонтакте
* Among all indices: * 9 proposed 3 as the best number of clusters * 5 proposed 4 as the best number of clusters * 6 proposed 5 as the best number of clusters * 1 proposed 6 as the best number of clusters * 1 proposed 9 as the best number of clusters * 1 proposed 10 as the best number of clusters	* Among all indices: * 10 proposed 3 as the best number of clusters * 1 proposed 4 as the best number of clusters * 1 proposed 5 as the best number of clusters * 1 proposed 6 as the best number of clusters * 1 proposed 8 as the best number of clusters * 3 proposed 9 as the best number of clusters * 5 proposed 10 as the best number of clusters

Составлено автором

На разных наборах данных большинство алгоритмов показывает, что наиболее оптимальным количеством кластеров будет три кластера.

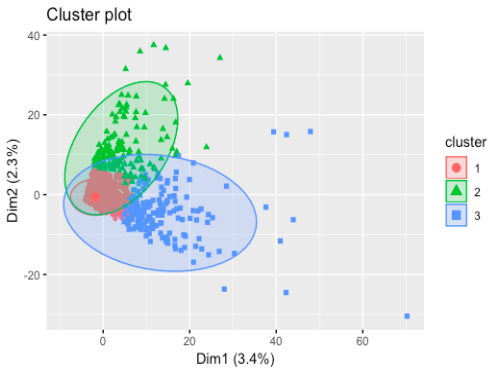
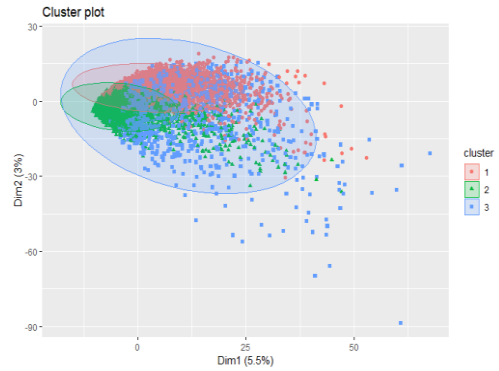
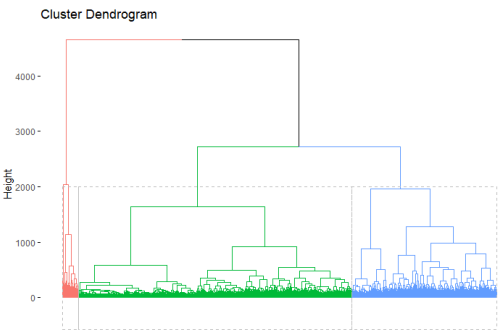
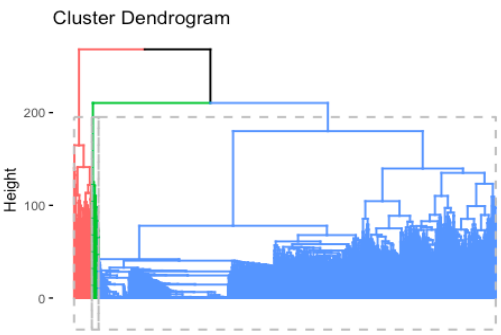
2.3 Оценка точности и выбор модели

Так как на предыдущем шаге мы выбрали 3 как оптимальное количество кластеров, это число теперь будет являться входным параметром для всех функций кластеризации.

Модель кластеризации была построена при помощи нескольких наиболее распространенных алгоритмов – алгоритм k-средних (k-means), алгоритм k-медоидов и реализации одного из методов иерархической кластеризации. Была построена сравнительная таблица (таблица 2) для упрощения визуального представления данных.

Таблица 2

Визуализация результатов кластеризации пользователей

Метод	КиноПоиск	ВКонтакте
k-means		
k-medoids		
Hierarchical clustering		

Составлено автором

По таблице можно оценить, насколько плотно расположены объекты разных кластеров друг к другу и насколько велики или малы полученные кластеры. Чтобы оценить качество построенных моделей, обратимся к известным мерам валидации кластеров.

102ITVN219
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

2.4 Валидация кластеров

Одной из показательных мер для валидации кластеров является «силуэт». Силуэт – это метод интерпретации и проверки согласованности в кластерах данных [9]. В отличие от других метрик, силуэт не предполагает знания истинных классов объектов, и позволяет оценить качество кластеризации, используя только саму (неразмеченную) выборку и результат кластеризации.

Силуэтом выборки называется средняя величина силуэта объектов данной выборки. Таким образом, силуэт показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров. Данная величина лежит в диапазоне [-1,1]. Если значение силуэта стремится к -1, значит, данные кластеризуются плохо. Нулевое значение силуэта означает, что данные лежат на пересечении двух кластеров. Чем ближе значение силуэта к единице, тем качественнее и четче выделенные кластеры [10].

В таблице 3 представлены результаты валидации построенных моделей кластеризации с помощью метрики силуэта.

Таблица 3

Значение силуэтов для построенных моделей кластеризации

Метод	«Кинопоиск» – силуэты	«ВКонтакте» – силуэты
k-means	0.2313694	0.3436076
k-medoids	0.119297	0.004151373
hierarchical clustering	0.1449562	0.2052583

Составлено автором

Таким образом, следует брать за основу метод k-means, как метод, показавший самые высокие метрики качества на обоих наборах данных. Нельзя сказать, что выделенные кластеры сгруппированы очень плотно и компактно, однако данные кластеризуются, а закономерности выделяются.

3. Интерпретация результатов

Для того, чтобы интерпретировать полученные результаты и проанализировать полученные кластеры вручную, были построены визуализации для каждого определенного кластера – были визуализированы первые 15 наиболее популярных групп среди каждой группы пользователей, полученные в результате сбора информации о сообществах «ВКонтакте» пользователей «Кинопоиска», описанном в пункте 2.2.

Для большей репрезентативности результатов наиболее крупные и популярные сообщества, связанные с кино – «Кинопоиск», «Cinemaholics», «Киномания» и пр. – были исключены из списка, поскольку их можно было найти в топе каждого кластера.

Среди наиболее популярных сообществ первого кластера (рисунок 1) можно отметить множество групп, посвященных комиксам и смежным темам, в частности, DC и Marvel, и групп, нередко публикующих подобный контент (например, «CLIQUE»). Также закономерно много групп, связанных с кино.

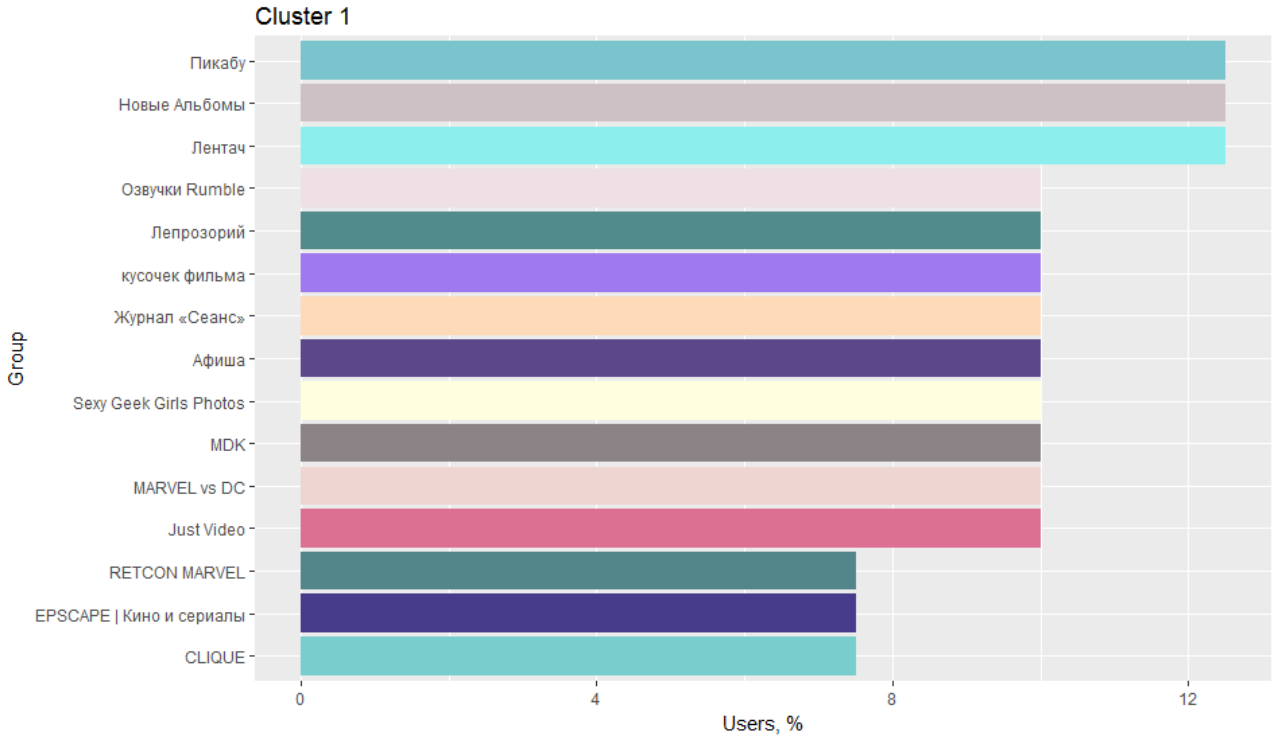


Рисунок 1. *Самые популярные группы пользователей первого кластера (разработано автором)*

В кластере номер 2 (рисунок 2) преобладают информационные и новостные сообщества – например, «Медуза», «TJ», «The Village», «Esquire» – отличающиеся содержательным, интересным и временами научным контентом. В пользу любви к интеллектуальным сообществам говорят также группы «Клуб National Geographic Россия» и «Science|Наука».

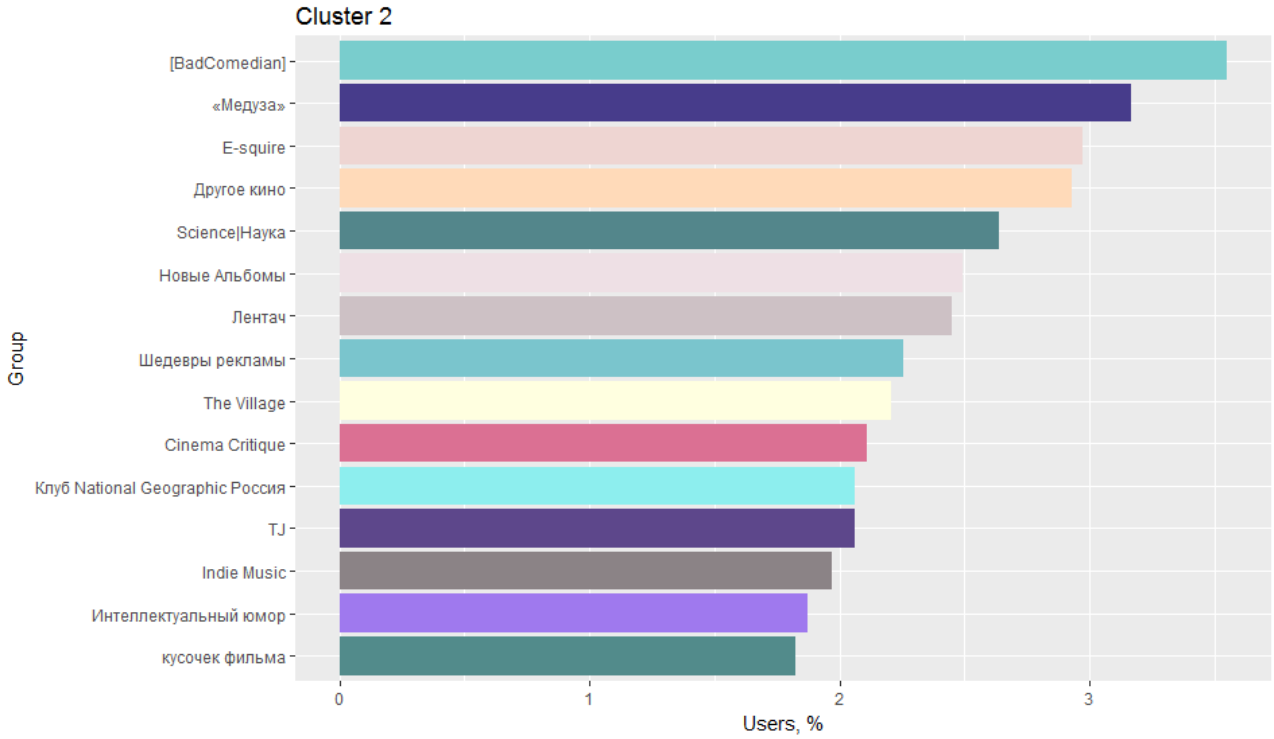


Рисунок 2. *Самые популярные сообщества пользователей второго кластера (разработано автором)*

102ITVN219
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

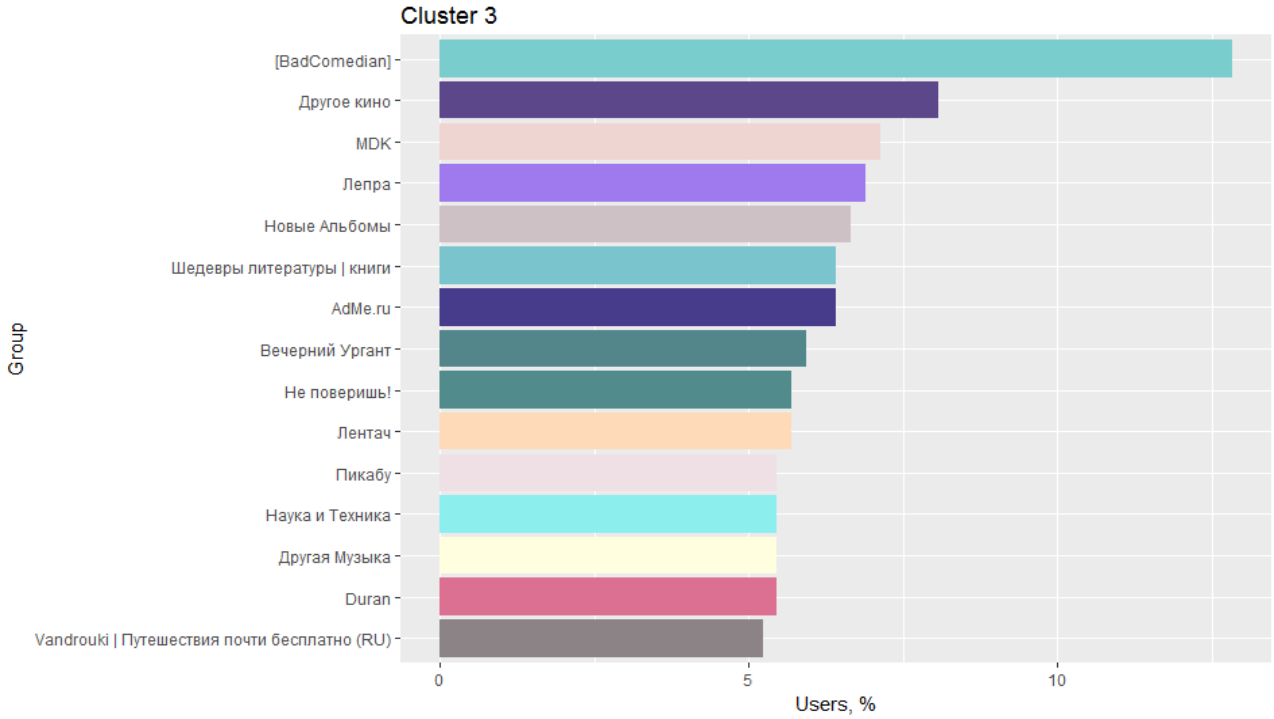


Рисунок 3. *Самые популярные сообщества пользователей третьего кластера (разработано автором)*

В третьем кластере (рисунок 3) в глаза бросается наличие групп, публикующих юмористический контент: так сообщества «MDK», «Лепра», «Пикабу» и «Duran» специализируются на шутках и «мемах», а «Вечерний Ургант» – популярнейшее в России вечернее развлекательное шоу.

Для упрощения интерпретации были выделены топ-5 наиболее высоко оцененных внутри каждого кластера фильмов (таблица 4).

Таблица 4

Топ-5 высоко оцененных фильмов у найденных кластеров

Кластер 1	Кластер 2	Кластер 3
1. Мстители: Война бесконечности	1. Человек	1. Песочник
2. Стражи Галактики. Часть 2	2. Песочник	2. Тайна Коко
3. Веном	3. Гранд тур	3. Нарко
4. Сорвиголова	4. Твин Пикс	4. Удивительная миссис Мейзел
5. Человек-паук: через Вселенные	5. Ванпанчмен	5. Головоломка

Составлено автором

Топ «любимых» фильмов лишь подтвердил сформированную теорию о «любителях комиксов», «интеллектуалах» и «юмористах».

Все фильмы первого кластера являются экранизациями комиксов и/или картинами о супергероях. Во втором можно отметить наличие документального фильма «Человек» и британскую автомобильную телепрограмму, созданную бывшими ведущими Top Gear, «Гранд тур». В третьем кластере преобладают легкие комедии, даже мультфильмы.

Выводы

Таким образом, результаты кластеризации показали, что кинолюбители могут относительно хорошо делиться на 3 кластера. Полученные кластеры можно интерпретировать как группу любителей комиксов и поклонников Marvel/DC, группу «интеллигентов» – людей, следящих за новостями, научными событиями и трендами, и группу «юмористов». Не исключено, что деление по интересам могло бы стать более специфичным, если при запуске алгоритмов кластеризации указать другое количество кластеров. Тем не менее, на наших данных наиболее качественным и репрезентативным оказалось деление именно на 3 кластера, однако при любых изменениях данных необходимо будет заново применить к ним алгоритм расчёта оптимального количества кластеров, чтобы при любых условиях иметь наиболее качественное деление.

ЛИТЕРАТУРА

1. Телевной А.Д., Хлопотов М.В. Исследование тематических профилей и способов расчета вовлеченности аудитории в сообществах социальной сети «ВКонтакте» [Электронный ресурс] // Вестник Евразийской науки, 2018 №2, URL: https://esj.today/PDF/43ITVN218.pdf (дата обращения: 10.12.2018).
2. Рыков Ю.Г. Структура социальных связей в виртуальных сообществах: сравнительный анализ онлайн-групп социальной сети «ВКонтакте» // дис. … канд. соц. наук Нац. иссл. ун-т Высшая школа экономики, Москва, 2016.
3. Сухагузов М. Цифры и факты: что сейчас происходит с сайтом «Кинопоиск» [Электронный ресурс] // 2017. URL: https://daily.afisha.ru/cinema/4450-cifry-i-fakty-chto-seychas-proishodit-s-saytom-kinopoisk (дата обращения: 05.11.2018).
4. Wang Y., Ru Y., Chai J. Time series clustering based on sparse subspace clustering algorithm and its application to daily box-office data analysis // Neural Computing and Applications, September 2018. DOI: 10.1007/s00521-018-3731-7.
5. Harrison J. RSelenium: R Bindings for 'Selenium WebDriver'. R package version 1.7.4 [Электронный ресурс] // 2018. URL: https://CRAN.R-project.org/package=RSelenium (дата обращения: 15.10.2018).
6. Sorokin D. vkR: Access to VK API via R. R package, version 0.1., 2016 [Электронный ресурс] // URL: https://github.com/Dementiy/vkR (дата обращения: 10.01.2019).
7. Kaufman L., Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley. 1990.
8. Charrad M., Ghazzali N., Boiteau V., Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1–36, 2014.
9. Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics, 20, November 1987, pp. 53–65.
10. Сивоголовко Е.В. Методы оценки качества чёткой кластеризации. Индекс оценки силуэта. (Silhouette index) С. 20 // Компьютерные инструменты в образовании, №4 2011 г.

Khlopotov Maxim Valerievich
Saint-Petersburg national research university of information technologies, mechanics and optics, Saint Petersburg, Russia
E-mail: khlopotov@corp.ifmo.ru

Startseva Natalia Vladimirovna
Saint-Petersburg national research university of information technologies, mechanics and optics, Saint Petersburg, Russia
E-mail: st.natalie.eilatan@gmail.com

Makarenko Anastasiia Aleksandrovna
Saint-Petersburg national research university of information technologies, mechanics and optics, Saint Petersburg, Russia
E-mail: AnastasiaMakarenko95@yandex.ru

Analysis of movie lovers' preferences and their thematic communities in social networks

Abstract. The article describes the study of the preferences of the most active Kinopoisk users and their correlation with these users' communities on the social network VKontakte.

To conduct the research, we collected the data on ratings of movie-lovers from the website Kinopoisk. Over 10000 user profiles were divided into groups using various clustering algorithms. To define the optimal number of clusters, the package «NbClust» was used. This package applies 30 algorithms to the dataset in order to identify the best number of sub-groups in the data. Most of the algorithms from the package proposed 3 as the best number of possible clusters.

The silhouette coefficient was used to measure clustering efficiency. The metric identified k-means as the best clustering algorithm. Thus, the final clustering model divided the dataset into 3 groups and was constructed using k-means algorithm.

The data about the linked VKontakte profiles of the users was collected in order to receive information about their communities. Analysis of the most popular communities among representatives of each cluster and the most highly rated films helped to identify trends in the topics of each cluster. Among the users, three groups were singled out – a group of «fans of comics», a group of «intellectuals», and a group of «humorists». The results of the study can be useful for marketing strategies and in personalization purposes.

Keywords: analysis of social networks; online community; silhouette; clustering; data science; Kinopoisk; VKontakte; k-means