УДК 004.67
DOI: 10.46548/21vek-2020-0950-0027

ВОЗМОЖНОСТЬ АВТОМАТИЗАЦИИ ПРОЦЕССА ПОИСКА ЦЕЛЕВОЙ АУДИТОРИИ В «ВКОНТАКТЕ» С ИСПОЛЬЗОВАНИЕМ ДАННЫХ ОБ ОЦЕНКАХ ПОЛЬЗОВАТЕЛЕЙ ДРУГОЙ СОЦИАЛЬНОЙ СЕТИ

Коростелев Александр Владимирович, аспирант кафедры «Информационные технологии и системы»
Мартышкин Алексей Иванович, кандидат технических наук,
доцент, доцент кафедры «Вычислительные машины и системы»
Пензенский государственный технологический университет
(440039, Россия, Пенза, проезд Байдукова/ул. Гагарина, д. 1а/11,
e-mails: bezpaniki@hotmail.com, alexey314@yandex.ru)

Аннотация. Статья посвящена исследованию вопросов, связанных с возможностью автоматизации процесса поиска целевой аудитории в социальной сети «Вконтакте» с использованием данных об оценках пользователей другой социальной сети. Установлено, что одним из наиболее популярных и эффективных видов маркетинга является таргетированный маркетинг, позволяющий оптимизировать рекламные кампании и нацелить их на потенциальных клиентов. Применяя технологии интеллектуального анализа данных на оценках пользователей, можно установить различные группы кинолюбителей, схожих по интересам. Имея информацию о сообществах этих групп в социальных сетях, представляется возможным установить оптимальные ресурсы для проведения рекламных кампаний и найти большое количество потенциальных клиентов. Подробно описаны данные с портала «Кинопоиск», необходимые для дальнейшего анализа и кластеризации, а также методы их сбора. Построены парсеры, специализированные под нужды исследования, при помощи которых получена информация об оценках пользователей, проставленных недавно вышедшим фильмам. Рассмотрены основные методы кластеризации, а также способы валидации результатов кластерного анализа. В качестве способа сокращения признакового пространства предложен метод главных компонент. Алгоритмы k-means, pam, CLARA и иерархическая кластеризация применены к набору данных с оценками пользователей «Кинопоиска». Согласно определенным метрикам качества, оптимальной кластеризацией оказалась иерархическая кластеризация на 3 кластера датасета, преобразованного с помощью метода главных компонент. Модель классификации была построена при помощи алгоритмов бэггинга, бустинга, дерева решений и метода опорных векторов. Наилучший результат показал классификатор, основанный на дереве решений – его средняя точность составила более 80%.

Ключевые слова: автоматизация поиска, алгоритм, валидация, кластеризация, классификация, модель, нормализация, профиль пользователя, социальная сеть, статистические данные, Big Data.

THE ABILITY TO AUTOMATE THE PROCESS OF SEARCHING FOR A TARGET AUDIENCE IN VKONTAKTE USING DATA ABOUT THE RATINGS OF USERS OF ANOTHER SOCIAL NETWORK

Korostelev Alexander Vladimirovich, postgraduate of sub-department «Information technologies and systems»
Martyshkin Alexey Ivanovich, candidate of technical sciences,
docent, associate Professor of sub-department «Computers and systems»
Penza state technological University
(440039, Russia, Penza, Baydukov Proyezd / Gagarin Street, 1a/11,
e-mails: bezpaniki@hotmail.com, alexey314@yandex.ru)

Abstract. The article is devoted to the study of issues related to the possibility of automating the process of searching for a target audience in the social network «Vkontakte» using data on the ratings of users of another social network. It is established that one of the most popular and effective types of marketing is targeted marketing, which allows you to optimize advertising campaigns and target them to potential customers. By using data mining technologies based on user ratings, you can identify different groups of movie fans with similar interests. With information about the communities of these groups in social networks, it is possible to establish optimal resources for advertising campaigns and find a large number of potential customers. Data from the «Kinopoisk» portal that is necessary for further analysis and clusterization, as well as methods for collecting them, are described in detail. Built parsers that are specialized for the needs of research, which provides information about user ratings assigned to recently released movies. The main clustering methods and methods for validating the results of cluster analysis are considered. The principal component method is proposed as a method for reducing the feature space. The K-means, pam, and CLARA algorithms and hierarchical clustering are applied to a dataset with «Kinopoisk» user ratings. According to certain quality metrics, the optimal clustering was hierarchical clustering into 3 clusters of a dataset converted using the principal component method. The classification model was constructed using bagging, boosting, decision tree, and support vector machine algorithms. The best result was shown by the classifier based on the decision tree – its average accuracy was more than 80%.

Keywords: search automation, algorithm, validation, clustering, classification, model, normalization, user profile, social network, statistics, Big Data

Введение. Сегодня персонализация занимает важное место в сфере веб-маркетинга. Компании хотят оптимизировать свою рекламу и сделать ее максимально гибкой, подстраиваясь под вкусы и предпочтения разных категорий людей, которые, в свою очередь, хотят получать персонализированные предложения и видеть рекламу того, что им интересно. Статистика показывает, что большинство потребителей (более 71%) предпочитают персонализированные рекламные объявления [1]. В свою очередь доступность интеллектуальных технологий, больших данных, методов и инструментов для их сбора открывают возможности для персонализации и таргетинга. Область киноиндустрии в России стремительно развивается и требует поддержки современных способов маркетинга для продолжения и оптимизации этого роста. Поэтому поиск целевой аудитории в этой сфере – актуальная и полезная задача, требующая максимальной сегментации потенциальных клиентов и наиболее персонализированных предложений для достижения высокой эффективности.

Цель проводимого исследования – автоматизация процесса поиска аудитории в сфере киноискусства в социальной сети (СС) «*ВКонтакте*» с использованием данных об оценках пользователей веб-ресурса «*Кинопоиск*». Предмет исследования – процесс автоматизации поиска целевой аудитории в СС «*ВКонтакте*» в сфере киноискусства, а объект – данные об оценках пользователей «*Кинопоиска*» и их сообществах в СС «*ВКонтакте*». Для достижения поставленной цели необходимо решить следующие частные задачи:

1. Собрать и обработать данные об оценках пользователей сети «*Кинопоиск*»;

2. Представить модель кластеризации пользователей «Кинопоиска» и составить основанные на их предпочтениях тематические профили;

3. Разработать алгоритм поиска тематических сообществ по выделенным кластерам;

4. Представить модель классификации кино по выделенным кластерам.

Материалы исследования. Сегодня Интернет представляет собой площадку для реализации возможностей *digital*-маркетинга. Присутствие компании стало практически обязательным условием для успешного ведения бизнеса. Эволюция *Big Data* и передовые аналитические технологии позволяют маркетологам разбираться в своей целевой аудитории как никогда прежде и выстраивать рекламные кампании, принимая во внимание особенности и интересы потенциальных или существующих клиентов. Согласно исследованиям [2], маркетологи все больше погружаются в так называемую «умную рекламу», основывая свои стратегии на данных аккаунтов, однако искусственный интеллект помогает перейти на совершенно новый уровень анализа. В последние время в России все большую популярность набирают онлайн-кинотеатры – стриминговые сервисы, где по подписки клиенты имеют возможность смотреть свежие фильмы и сериалы в хорошем качестве (рис. 1).

Рисунок 1 – Динамика российского рынка онлайн-кинотеатров, в млрд. рублей

На Западе такие площадки уже давно поглотили рынок [3]. Огромную роль в популяризации сервисов на Западе стало умение грамотно пользоваться интеллектуальными технологиями и *Big Data* при разработке маркетинговых кампаний. Ключевую роль в становлении *Netflix* стриминговым сервисом номер один сыграло использование *Big Data* и искусственного интеллекта для понимания своей аудитории и их интересов, в том числе и в маркетинговых стратегиях [4]. Для того, чтобы построить гибкий инструмент таргетирования, который смог бы учитывать разнообразие вкусов потенциальных клиентов-кинолюбителей, необходимо иметь большой набор данных, на которым можно было бы построить такого рода алгоритм. Набор должен учитывать специфики разных групп клиентов, их отношения к кинокартинам того или иного рода. Крупнейшим и наиболее известным является веб-платформа «*Кинопоиск*» – русскоязычный интернет-проект, посвященный кинематографу. Сайт предоставляет информацию о кинофильмах, телесериалах, а также о личностях, связанных с кино- и телепроизводством: актерах, режиссерах, продюсерах, сценаристах, операторах, композиторах, художниках и монтажерах, а также имеет возможности СС. Здесь можно ставить оценки, находить друзей по интересам и смотреть фильмы онлайн на партнерских ресурсах. Итак, каждый фильм на «*Кинопоиске*» может быть оценен по десятибалльной шкале – от единицы (хуже некуда) до десятки (шедевр). Оценивать фильмы могут только зарегистрированные пользователи. Чтобы зарегистрироваться на «Кинопоиске», можно использовать авторизацию через другие СС – «*ВКонтакте*», «*Facebook*», «*Twitter*». СС – один из самых эффективных инструментов для продвижения релизов с четко сформированной аудиторией, так как с их помощью можно максимально таргетировать рекламные сообщения на целевые группы. СС могут стать важным инструментом продвижения в случае как узкожанрового продукта, так и блокбастера. Зная с большой долей вероятности, что та или иная картина понравится некоторой группе пользователей, основываясь на их предыдущих оценках и предпочтениях, можно оптимизировать рекламную кампанию гораздо эффективнее, нежели используя

некоторые абстрактные характеристики, которые используются маркетологами без привлечения интеллектуальных технологий и анализа – например, возраст, пол, регион и прочее [5].

«*Вконтакте*» как СС, обладающая наиболее молодой и активной аудиторией, служит первым выбором у дистрибьюторов в подавляющем большинстве случаев. «*ВКонтакте*» является самой популярной СС в России, ежемесячная аудитория которой превышает 100 миллионов пользователей. Неудивительно, что «*ВКонтакте*» является одним из самых популярных инструментов для проведения рекламных кампаний. «*ВКонтакте*» предлагает несколько видов рекламы: таргетированная реклама, реклама в сообществах, специальные проекты (стикеры, чат-боты, трансляции, брендированные подарки, интеграция бренда в игры и приложения) [6]. Как для таргетированной рекламы, так и для рекламы в сообществах можно использовать данные о тех пользователях «*Кинопоиска*» с привязанной станицей «*ВКонтакте*», которые оценили похожие фильмы на высокую оценку. Тематики сообщества «кино» недостаточно, чтобы определить, насколько может понравиться участникам этого сообщества фильм того или иного жанра или совокупности жанров, той или иной страны, того или иного режиссера. «*ВКонтакте*» существует тысячи групп подобного жанра – необходимо узнать, в которых из них процент целевой аудитории оптимальный. Кроме того, с помощью потенциального маркетингового инструмента, основанного на данных из «*Кинопоиска*», который является целью данной научно-исследовательской работы, можно будет использовать не только сообщества тематики «кино»: выяснив класс людей, которым нравятся фильмы определенного типа, можно размещать рекламу в таких сообществах, где скопление людей этого «класса» максимально. В современном мире оформление страницы в СС, в частности, сообщества пользователя, очень хорошо характеризуют его интересы и даже тип мышления. Таким образом, необходимо находить некоторые неочевидные, скрытые зависимости между вкусами пользователя в кино и его группами «*ВКонтакте*». Собрав информацию о сообществах «*ВКонтакте*» пользователей сети «*Кинопоиск*», можно будет оценить распределение любителей тех или иных жанров, режиссеров, актеров в разных сообществах, используя оценки пользователей. Итак, для создания качественного инструмента по нахождению целевой аудиторию, необходимо собрать данные об оценках пользователей на «*Кинопоиске*», изучить корреляцию между жанрами, кластеризировать пользователей по предпочтениям, изучить страницы «*ВКонтакте*» пользователей определенных классов и найти оптимальные сообщества для рекламы.

Для сегментации аудитории с целью таргетирования рекламы существует некоторые сервисы, например, «*Pepper*» [2], «*Церебро*» [7], «*Сегменто таргет*» [8], «*Oktarget.ru*» [1]. Основные функции, преимущества и недостатки перечисленных сервисов сведены в таблице 1.

Таблица 1 – Сравнение аналогов

Программа	Критерий		
	Поддержка «ВКонтакте»	Учитывает специфику компании	Интеллектуализация поиска
Pepper	Да	Да	Нет
Церебро	Да	Нет	Нет
Сегменто таргет	Да	Нет	Нет
Oktarget.ru	Нет	Нет	Нет

Проанализировав таблицу 1, резюмируем, что данные инструменты неплохо взаимодействуют с СС, но все они используют набор примерно одинаковых статистических показателей, не применяя интеллектуальный анализ данных и не извлекая потенциальных знаний о пользователях и их настоящих предпочтениях, который смог бы максимально поднять уровень таргетизации для рекламных кампаний. Описываемый в статье инструмент позволяет обосновать сегментацию пользователей, учитывая не только такие статистические параметры, но и используя реально проставленные пользователем оценки, реальные данные о предпочтениях клиента, который он сам публично предоставил.

Методы и результаты исследования. Чтобы получить инструмент для эффективного таргетированного маркетинга в сфере киноискусства, необходимо собрать достаточно большое количество данных о пользователях и их предпочтениях в кино. Для этой цели было принято использовать портал «*Кинопоиск*». Имея информацию о предпочтениях тех или иных групп пользователей, мы имеем возможность проанализировать их профили в СС «*ВКонтакте*», определив, в каких сообществах «обитают» разные группы пользователей, можно найти похожие сообщества с предполагаемой целевой аудиторией, таким образом, определив искомые группы с целевой аудиторией.

Сервис «*Кинопоиск*» хранит огромное количество данных о фильмах и пользователях, представляющих собой потенциал для анализа, и формируют перспективную *Big Data*, которую можно использовать в целях кастомизации, персонализации и маркетинга. Информация о пользователях также представляет собой огромный потенциал: здесь можно найти их оценки по каждому фильму, количество оцененных картин, друзей, любимые фильмы, средние оценки по жанрам, годам и многое другое. Для нашей конкретной цели было решено собирать данные именно об оценках пользователей и их близости между собой, не беря в расчет характеристики самих фильмов на первом этапе. Далее, имея сегментацию пользователей и относящиеся к каждому сегменту фильмы, можно будет выявить закономерности между характеристиками этих кинокартин, чтобы в дальнейшем иметь возможность классифицировать новый вышедший фильм. На «*Кинопоиске*» заре-

гистрировано более 7 миллионов пользователей, многие профили которых были удалены или не используются. Некоторые профили заброшены и имеют лишь по одной или две оценки за все время существования. Поэтому параметр «дата последней записи», существующий в характеристиках профиля, тоже не может однозначно ответить, что человек действительно пользуется «*Кинопоиском*». На ресурсе имеется большое количество чатов и списков фильмов, в том числе и список по годам. Списки по годам выпуска картин содержат все фильмы и сериалы, выпущенные в этот год, чьи профили существуют на сайте [7]. Пользуясь таким списком, можно собрать все оцененные фильмы за определенный год. Именно такой подход к сбору данных был выбран в рамках статьи: собрав данные о всех картинах и оценках пользователей за последние 6 лет (с 2014 по 2020 года), можно избежать совсем неактуальных данных – пользователей, не пользовавшихся сайтом многие годы. Естественно, будет необходима фильтрация, но количество избыточных данных будет меньше, чем если бы собирать данные о всех пользователях сайта вообще.

У «*Кинопоиска*» отсутствует в свободном доступе *API*, поэтому необходимо было построить парсер для сбора данных веб-страниц «*Кинопоиска*», используя *css* и *XPath* селекторы [9]. Все манипуляции для сбора данных с веб-страниц были выполнены с использованием пакета *RSelenium* на языке программирования *R* в среде разработки *RStudio* [10]. Данная библиотека представляет собой расширение для автоматизации работы веб-браузера и, в рамках текущей задачи, позволяет имитировать действия пользователя для навигации по сайту и наиболее гибкого сбора данных. В целом данный пакет позволяет «доставать» данные с веб-страниц, когда обычные методы бессильны – например, когда веб-приложения используют *JavaScript* и *Ajax* и контент их веб-страниц генерируется динамически и простой HTML-код страниц не содержит всей нужной информации [11]. Примером может служить так называемый «*lazy loading*», когда новые данные подгружаются лишь в тот момент, когда пользователь пролистал вниз страницы, при этом адрес страницы остается тем же. Следующая диаграмма визуализирует количество собранных фильмов за каждый год (рис. 2). Количество оцененных фильмов за каждый год закономерно уменьшается, поскольку чем новее фильм, тем меньшее количество людей успело его посмотреть. Общее количество собранных картин составило 5871, что на момент собора являлось общим количеством оцененных фильмов за этот период в целом. Визуализация была выполнена на языке программирования *R* с использованием специального кастомного набора цветов «*wesanderson*» [8].

Далее из набора данных были удалены фильмы, общее количество оценок которых было менее 1000. Это было необходимо, чтобы уменьшить размерность финального датасета для кластеризации и исключить «редкие» фильмы, которые не были удостоены внимания широкого круга публики. Таким образом, в наборе данных осталось 4193 фильма. Парсер собирает оценку и идентификатор пользователя для каждого идентификатора фильма, создавая таким образом итоговый набор данных, на которым будет осуществляться кластеризация. Однако на «*Кинопоиске*» существует небольшое ограничение – можно посмотреть лишь последнюю тысячу оценок к определенному фильму. Это ограничение означало два важных момента: не все пользователи, оценившие фильмы последних лет попадут в финальный набор данных; не все оценки пользователей к выбранным фильмам могут попасть в финальный набор данных. При наличии 5871 фильма и, соответственно, более миллиона уникальных пользователей, такой объем работы был бы колоссальным. Было решено использовать информацию лишь и фильмах 2019 и 2020 годов – это 1477 фильмов. После сбора оценок пользователей данным фильмам и отбора уникальных пользователей в датасете оказалось 225 тысяч пользователей. Чтобы сократить время сбора данных – процесс парсинга с помощью автоматизированного веб-драйвера занимает немало времени и напрямую зависит от скорости интернета и качества оборудования – были отфильтрованы те пользователи, количество оценок в датасете которых составило менее 10. Также это добавило элемент рандомизации в данные и помогло отсеять людей с маленьким количеством оценок и, соответственно, потенциально загрязняющих данные.

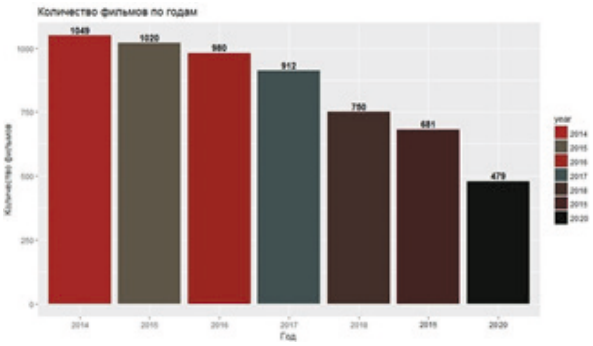


Рисунок 2 – Визуализация собранных данных

После всех описанных манипуляций с датасетом количество пользователей составило 11538 человек, а их оценок к фильмам 1477. Большие объемы информации не смогли бы быть обработаны на имеющемся компьютерном оборудовании и программном обеспечении. Также для дальнейшей интерпретации полученных результатов кластеризации был построен парсер для сбора дополнительной информации о фильмах. Согласно собранной статистике, количество пользователей с привязанной страницей «*ВКонтакте*» составляет примерно 15-20% от общего количество пользователей. Таким образом, для дальнейшего анализа сформирован набор данных, состоящий из 1477 кинокартин и оценок к ним 11538 пользователей «*Кинопоиска*». Для каждого пользователя также были найдены привязанные страницы «*ВКонтакте*», если

такие имелись, а для каждого фильма – некоторые описательные характеристики, призванные помочь интерпретировать полученные кластеры и классифицировать новые кинокартины в один из определённый кластеров.

Далее опишем процесс кластеризации пользователей. Кластеризация – метод машинного обучения без учителя и распространенный метод статистического анализа данных, используемый во многих областях. Его главной задачей является разделение множества объектов на группы (называемые «кластерами») таким образом, чтобы объекты внутри каждой группы были похожи между собой и отличались от объектов других групп. По сути, это отбор объектов на основе их сходства и различия друг с другом [12]. В отличие от классификации, для методов кластеризации не нужны знания о классах объектов – она работает с неразмеченными данными. Для кластеризации необходим лишь набор данных и их признаковое описание – некоторое количество характеристик, присущих каждому из объектов выборки. Признаки могут быть как числовыми, так и нечисловыми. Рассмотрим наиболее распространенные алгоритмы кластеризации:

1) Алгоритм *k*-средних – один из наиболее простых и распространенных алгоритмов кластеризации. Этот алгоритм разбивает множество объектов на заранее определенное число *k* кластеров, при этом каждый объект относится к тому кластеру, к центроиду которого он ближе всего [13]. В качестве меры близости чаще всего используется Евклидово расстояние

$$p(x,y) = \|x-y\| = \sqrt{\sum_{p=1}^{n}\left(x_p - y_p\right)^2} \quad (1)$$

где $x,y \in R^n$.

Метод *k*-средних разделяет *m* объектов на *k* подгрупп (кластеров) ($k \le m$) так, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров:

$$\min\left[\sum_{i=1}^{k}\sum_{x^{(j)}\in Si}\left\|x^{(j)}-\mu_i\right\|^2\right] \quad (2)$$

где $x^{(j)} \in R^n$, $\mu_i \in R^n$, μ_i– центроид для кластера S_i.

Основными минусами данного алгоритма является необходимость заранее знать оптимальное число кластеров и отсутствие гарантии нахождения оптимального разбиения, поскольку результаты сильно коррелируют с начальным выбором центроидов.

Агломеративная иерархическая кластеризация – различные вариации этого алгоритма отличаются правилами вычисления расстояния между кластерами. Например: алгоритм средней связи – на каждом шаге объединяет два ближайших кластера, оценивая арифметическое расстояние между всеми парами объектов; алгоритм одиночной связи («ближайшего соседа») – расстояние между кластерами рассчитывается как минимальное из расстояний между парами объектов из двух разных кластеров; алгоритм полной связи («дальнего соседа») – вычисляет расстояние между наиболее удаленными объектами.

2) Алгоритм *k*-медоидов – также называемый *PAM* (*Partitioning Around Medoids* [14]), почти идентичен методу *k*-средних, однако вместо вычисления центроидов осуществляется поиск *k* наиболее представительных объектов выборки, а разброс внутри кластеров может измеряться манхэттенским расстоянием (расстоянием городских кварталов), а не евклидовым:

$$d_{1(p,q)} = \|p-q\|_1 = \sum_{i=1}^{n}|p_i - q_i| \quad (3)$$

где p_i и q_i – *i*-я координата первого и второго объекта соответственно.

3) Алгоритм *CLARA (clustering large applications)* – является расширением методов *k-medoids* для работы с данными, содержащими большое количество объектов (более нескольких тысяч наблюдений), чтобы уменьшить время вычислений и проблему с объемом оперативной памяти. Это достигается с помощью метода выборки. Вместо того, чтобы находить медоиды для всего набора данных, *CLARA* берет небольшую выборку из датасета и применяет алгоритм *PAM* для генерации оптимального набора медоидов. Качество полученных медоидов измеряется по среднему различию между каждым объектом во всем датасете *D* и медиоидом его кластера, определяемым как следующая функция стоимости:

$$\mathrm{Cos}\,t(M,D) = \frac{\sum_{i=1}^{n} dissimiliarity(O_i, rep(M,O_i))}{n} \quad (4)$$

где *M* – набор выбранных медоидов; (O_i, O_j) – различие между объектами O_i и O_j; *rep* (*M*, O_i) – медоид в *M*, ближайший к O_i.

После построения модели кластеризации необходимо оценить, насколько качественно она была построена и насколько в действительности значимы ее результаты. Валидация результатов кластеризации важна не только для избегания нахождения закономерностей в случайных данных, но также для сравнения несколько алгоритмов кластеризации и выбо-ра наилучшей из моделей. Как правило, метрики валидации кластеризации разделяются на 3 группы [15]:

- *внутренняя валидация* – использует только информацию о внутренней структуре кластеров для оценки качества кластеризации. Может быть использован как для выбора оптимального числа кластеров, так и для выбора самого эффективного алгоритма;
- *внешняя валидация* – результаты кластеризации сравниваются с некоторыми внешними известными результатами – заранее известными классами объектом, т.е. размеченными данными. Поскольку «истинный» класс объекта известен заранее, этот подход в основном используется для выбора оптимального алгоритма кластеризации для определенного набора данных;
- *относительная валидация* – оценивает структуру кластеризации, варьируя различные значе-

ния параметров для одного и того же алгоритма (например, изменяя количество кластеров *k*). Обычно используется для определения оптимального количества кластеров.

Однако, прежде чем оценивать качество кластеризации, можно проверить, имеют ли данные тенденцию к группированию. Как уже было написано ранее, одной из главных проблем кластерного анализа является то, что алгоритмы кластеризации будут делить данные на группы вне зависимости от того, существуют ли в данных закономерности или нет. Таким образом, алгоритмы смогут кластеризовать даже рандомно сгенерированный набор данных. Поэтому первым делом перед началом кластеризации рекомендуется вычислить общую предрасположенность данных к объединению в группы. Для этого используется статистика Хопкинса [16]. Для подсчета этой метрики создается N сгенерированных случайным образом набор данных на основе распределения с таким же стандартным отклонением, что и исходный набор данных. Для каждого наблюдения i рассчитывается его расстояние до k ближайших соседей: w_i между объектами оригинальной выборки и q_i между случайными наблюдениями и их ближайшими реальными соседями.

$$H_{ind} = \frac{\sum_n w_i}{\sum_n q_i + \sum_n w_i} \quad (5)$$

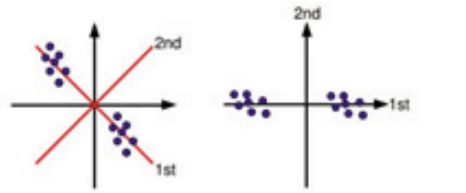
Если статистика Хопкинса превышает 0,5, это означает, что рандомно сгенерированный набор данных подобен оригинальному, а объекты группируются случайно. Если данная метрика меньше 0,25, можно с 90% вероятностью утверждать, что в данных есть закономерности и тенденции к группированию.

Одним из распространенных методов внутренней валидации является силуэт [17]. Анализ силуэтов кластеров помогает измерить, насколько хорошо сгруппированы данные, и оценивает расстояние между кластерами. Для каждого наблюдения i ширина силуэта может быть подсчитана следующим способом: пусть, a – среднее расстояние от данного объекта до других объектов кластера, а b – среднее расстояние от данного объекта до объектов другого ближайшего кластера, тогда силуэт объекта может быть подсчитан по следующему выражению

$$s = \frac{b-a}{\max(a,b)} \quad (6)$$

Чем больше значение силуэта, тем лучше кластеризуется объект: наблюдения со значением силуэта, близким к 1, сгруппированы очень хорошо и почти однозначно принадлежат к определенному кластеру. Небольшой силуэт, близкий к 0, означает, что объект лежит на границе двух кластеров. Наблюдения с отрицательным силуэт, вероятно, помещены в неправильный кластер. Качество кластеризации обычно оценивается средним значением силуэта по всей выборке. В случае большого количества признаков сокращение признакового пространства может повлиять на качество разбиения данные на группы при помощи удаления каких-либо неинформативных или засоряющих данные признаков [18]. В отличие от классификации, многие традиционные методы отбора признаков (*feature selection*) не могут быть применены к неразмеченному набору данных. Самым распространенным способом сокращения пространства является метод главных компонент.

Метод главных компонент (*principal component analysis,* PCA) позволяет уменьшить размерность больших наборов данных путем преобразования боль-шого набора переменных в меньший, потеряв при этом наименьшее количество информации [19]. Суть метода заключается в преобразовании некоторого набора возможно скоррелированных переменных в набор линейно нескоррелированных переменных главных компонент. В основе методы лежит переход к новой систем координат y_1, y_2,…,y_n от исходной системы координат x_1,x_2,…,x_n многомерного пространства признаков, которая является системой ортонормированных комбинаций. Линейные комбинации выбираются таким образом, чтобы первая главная компонента $y_1(x)$ имела бы максимальную дисперсию, т.е. объясняла вариабильность данных насколько это возможно. Число главных компонент всегда меньше или равно количеству начальных переменных. На рисунке 3 представлен пример применения метода главных компонент к данным с более чем тремя различными признаками. Первая и вторая главные компоненты описывают переменные на 96%, третья – на 3%, остальные же переменные являются не столь информативными и описывают датасет вкупе лишь на 1%.



Principal Component	1st	2nd	3rd	Other
Proportion of Variance	62%	34%	3%	rest

Рисунок 3 – Распределение информативности между главными компонентами

Также важным этапом *PCA* является нормализация. Метод является чувствительным к диапазону данных – так при наличии различия в диапазонах исходных переменных, переменные с большим диапазоном будут доминировать над переменными с небольшим диапазоном, что приведет с смещенному, неверному результату. Поэтому важно сперва преобразовать данные в сопоставимые масштабы.

Собранные ранее данные в контексте кластеризации представляют собой следующую структуру: объектами являются пользователи «*Кинопоиска*», оценившие фильмы. Признаками – просмотренные всеми пользователями фильмы, оцененные баллами от 0 до 10. Таким образом, для каждого из 11538 пользовате-

лей определяется класс на основе его оценок 1477 фильмам. Для определения оптимального количества кластеров применялся пакет для *R* под названием *NbClust* [20]. Данный пакет предоставляет реализацию 30 различных индексов для оценки качества результатов кластеризации. Любая комбинация индексов проверки и методов кластеризации может быть запрошена в одном вызове функции. Это позволяет одновременно оценивать несколько схем кластеризации, изменяя количество кластеров, чтобы помочь определить наиболее оптимальное число для загруженного датасета. В статье пакет *NbClust* использован только для предварительного определения тенденций кластеризации, а более детальный анализ результатов с более обширным кругом алгоритмов и методов кластеризации был проведен с помощью других пакетов и описан в исследовании далее.

Данные были кластеризованы при помощи *factoextra* с использованием реализации следующих алгоритмов: к-средних (*k-means*); *k*-медоидов (*pam*); агломеративная иерархическая кластеризация (*hierarchical clustering*); *CLARA*. В качестве меры качества была использована метрика силуэта. Валидация кластеризации необходима для сравнения различных моделей и выбора оптимальной из них. Еще одним фактором, влияющим на выбор итоговой модели кластеризации, станет количество объектов в кластерах, поскольку некоторые алгоритмы склонны определять большую группу объектов в один кластер и «откалывать» единичные экземпляры, создавая несколько других. Для этого введем ограничение: поскольку, как было отмечено ранее, прикрепленную ссылку «*ВКонтакте*» имеют лишь 15-20% всех пользователей «*Кинопоиска*». Однако кластеризацию алгоритмом *CLARA* нельзя назвать хорошей, несмотря на высокое значение силуэта – многие кластеры, определенные в данных моделях, имеют критически малое количество объектов, им принадлежащих – менее установленного порога в 500 объектов. Наиболее высокий силуэт показала модель *PAM* с четырьмя кластерами. В результате работы функции набор данных был разбит на главные компоненты, упорядоченные в порядке их важности для датасета. В исследуемом наборе данных первые 838 компонент описывают набор на 90%. Было решено кластеризовать данные с использованием этих компонент и сравнить результаты с кластеризацией на оригинальном наборе. Индекс Хопкинса для данного датасета оказался значительно выше – 0,156613, а *NbClust* показал те же рекомендации, что и для оригинального набора данных. Некоторые методы кластеризации показали лучшие результаты на обработанном наборе данных – например, точность иерархической кластеризации значительно повысилась. Согласно двум определенным ранее критериям – силуэту и размеру кластеров – наилучшим разделением можно считать иерархическую кластеризацию на 3 кластера. Таким образом, точность кластеризации меняется лишь незначительно в зависимости от моделей, а в рамках одного алгоритма точность между кластеризацией на 3, 4 и 5 кластеров также остается практически неизменной. Кластеризация на 10 групп дает значительно худшие результаты. Выделенные группы объектов также близки по размеру – в каждом разделении имеется одна группа, значительно больше других. Основываясь на метрике силуэта и размерах кластеров, модель иерархической кластеризации, разделяющая обработанный при помощи метода главных компонент набор данных на 3 кластера, была выбрана для дальнейшего анализа.

Для того чтобы собрать данные о сообществах пользователей, применялся пакет *vkR* [21], предназначенный для удобного взаимодействия с *API* «*ВКонтакте*» для языка программирования *R*. Пакет предоставляет множество функций для выгрузки данных «*ВКонтакте*», в том числе функцию для сбора сообществ пользователей. Самые распространенные сообщества для каждого кластера были визуализированы для упрощения интерпретации. Именно эти сообщества могут быть предложены алгоритмом рекламодателям для поиска потенциальной целевой аудитории той или иной картины.

На рисунке 4 представлены уникальные для первого кластера сообщества.

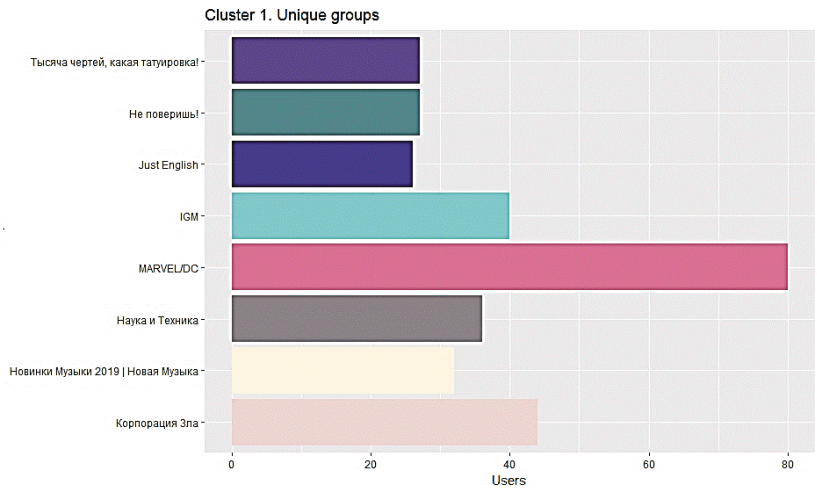


Рисунок 4 – Уникальные сообщества первого кластера

Выявили, что у каждого кластера существуют некоторые зависимости с группами «*ВКонтакте*», которые характеризуют как аудиторию сообщества, так и аудиторию кинокартин, оцененных представителями кластеров. Например, в первом кластере преобладают юмористические сообщества, сообщества по комиксам, а также группы, посвященные «новинкам». Рассмотрим несколько известных методов классификации, которые были рассмотрены и применены в настоящем исследовании. Метод опорных векторов – алгоритм обучения по прецедентам, используемый для бинарной классификации. Этот метод также называют классификатором с максимальным зазором. Суть метода заключается в том, что каждый объект обучающей выборки представляет собой точку в *n*-мерном пространстве, каждая из которых принадлежит одному из двух классов. Среди всех точек обучающей выборки алгоритм находит точки, лежащие на границе двух классов, и строит между ними разделяющую гиперплоскость размерности *n-1*. Эти точки называют опорными векторами [22]. Ансамблевые классификаторы – метод, в основе которого лежит идея обучения нескольких слабых классификаторов на одной и тоже выборке и объединение их предсказаний для новых тестируемых объектов с целью достижения более высокой точности. В задаче классификации алгоритм считается слабым, если его ошибка меньше 50%, но больше 0%. Наиболее популярными ансамблевыми методами являются *bagging* и *boosting* [23]. Деревья решений – метод осуществляющий процесс деления исходных данных на группы до тех пор, пока не будут получены однородные множества. Совокупность таких правил позволяет предсказывать наиболее вероятный класс объекта, основываясь на характеристиках этого объекта. Метод деревьев решений считается одним из самых эффективных и применим для решения задач классификации, возникающих в самых разных областях.

Чтобы определить, какие именно картины и премьеры рекламировать в соответствующих кластерам сообществах, необходимо выделить характеристики фильмов, которые предпочитают пользователи тех или иных кластеров. Самые популярные фильмы и жанры рассчитывались по количеству просмотров среди ядра кластера, а наиболее высоко оцененные – по среднему рейтингу среди этих же пользователей. Можно отметить, что наиболее популярные кинокартины в первом кластере являются экранизациями комиксов или фильмами про супергероев в целом. Среди предпочтений во втором кластере можно найти кино, высоко оцененное критиками, номинированное на престижные премии, а также «нетипичные» картины. В третьем преобладают популярные фильмы. Поскольку один и тот же фильм может приглянуться представителям разных классов, принято решение провести бинарную классификацию для каждого класса отдельно – таким образом, кино может принадлежать или не принадлежать каждому из классов. Одна и та же картина может принадлежать сразу трем классам. Чтобы построить модели классификации, необходимы размеченные данные. Чтобы определить принадлежность фильма для каждого кластера были отобраны фильмы, наиболее распространенные и наиболее высоко оцененные у ядра этих кластеров. Таким образом, в кластерах оказались фильмы, которые однозначно понравились доле пользователей кластера, и фильмы, которые не обязательно имеют высокую оценку среди представителей кластера, однако они все же посмотрели эти картины. Результаты представлены в таблице 2.

Таким образом, классификатором, показавшим наиболее высокое значение точности для большинства классов, оказался классификатор дерева решений. Средняя точность по трем классам составила более 80%, что можно назвать успешным результатом в рамках поставленной задачи. Наиболее важными признаками, согласно построенной модели, оказались режиссер, жанр и количество голосов. Наименее важным – страна производства.

Таблица 2 – Результаты классификации

Метод классификации	Класс		
	Cluster1	Cluster2	Cluster3
Bagging	0,7398	0,645	0,6829
Boosting	0,6585	0,4688	0,8401
Дерево решений	0,7696	0,8022	0,8347
SVM	0,6856	0,7398	0,5718

Заключение. В ходе исследования была рассмотрена актуальность таргетированного маркетинга, а также.

Было установлено, что большинство одной из наиболее популярных и удобных площадок для персонализированных рекламных кампаний является СС «*ВКонтакте*». Однако среди уже существующих инструментов для настройки такой рекламы и поиска пользователей, наиболее заинтересованных в предлагаемом продукте, отсутствуют инструменты, которые учитывали бы специфику области компании.

Для построения инструмента поиска целевой аудитории в сфере киноиндустрии был построен парсер данных с самого популярного российского портала о кино – «*Кинопоиска*».

Для определения схожих по интересам пользователей, данные были кластеризованы различными алгоритмами. Чтобы определить оптимальное количество кластеров, использован пакет *NbClust*, который показал, что набор данных в перспективе может хорошо делиться на 3, 4, 5 и 10 кластеров. Для сокращения признакового пространства был применен метод главных компонент – 1477 возможных признаков были преобразованы в 838 компонент, описывающих набор данных на 90%.

Модели кластеризации были построены как для оригинального, так и для преобразованного набора данных. Были использованы такие методы кластеризации, как *k*-средних, *k*-медоидов, иерархическая кластеризация и *CLARA*. Наилучшие результаты,

согласно определенным метрикам качества, показала модель, основанная на иерархической кластеризации, делящая набор на 3 разных кластера. Построена модель бинарной классификации при помощи алгоритмов *bagging*, *boosting*, дерева решений и метода опорных векторов. Модель, основанная на дереве решений, показала наилучший результат – более 80% в среднем по всем кластерам, что можно считать достаточно хорошим результатом.

Таким образом, совокупность построенных моделей и разработанных алгоритмов сбора и анализа данных позволяет осуществить автоматизированный поиск целевой аудитории в сфере киноискусства.

СПИСОК ЛИТЕРАТУРЫ:

1. Kirkpatrick D. Study: 71% of consumers prefer personalized ads [Электронный ресурс]. 2016. Режим доступа: https://www.marketingdive.com/news/study-71-of-consumers-prefer- personalized-ads/418831/ свободный. Яз. англ. (дата обращения: 14.05.2020).

2. Tjepkema L. What Is Artificial Intelligence Marketing & Why Is It So Powerful? [Электронный ресурс]. 2016. Режим доступа: https://www.emarsys.com/en/resources/blog/artificial-intelligence-marketing-solutions/ свободный. Яз. англ. (дата обращения: 14.05.2020).

3. Burks R. Netflix Is More Popular Than Broadcast, Cable & More In TV Viewing [Электронный ресурс]. 2018. Режим доступа: https://screenrant.com/netflix-popular-broadcast-cable-tv-viewing/ свободный. Яз. англ. (дата обращения: 14.05.2020).

4. Netflix's Use Of Big Data: Lessons For Brand Marketers [Электронный ресурс]. 2017. Режим доступа: https://adexchanger.com/data-driven-thinking/netflixs-use-big-data-lessons-brand-marketers/ свободный. Яз. англ. (дата обращения: 14.05.2020).

5. Harrison J. RSelenium: R Bindings for 'Selenium WebDriver'. R package version 1.7.7. 2020 [Электронный ресурс]. 2020. Режим доступа: https://CRAN.R-project.org/package=RSelenium свободный. Яз. англ. (дата обращения: 14.05.2020).

6. Развивайте свой бизнес ВКонтакте [Электронный ресурс]. 2020. Режим доступа: https://vk.com/biz свободный. Яз. русский. (дата обращения: 14.05.2020).

7. Все фильмы, 2018 год [Электронный ресурс]. 2020. Режим доступа: https://www.kinopoisk.ru/ lists/m_act[-year]/2018/m_act[all]/ok/ свободный. Яз. русский. (дата обращения: 14.05.2020).

8. Ram K., Wickham H. «wesanderson: A Wes Anderson Palette Generator. R package version 0.3.6. n'» [Электронный ресурс]. 2018. Режим доступа: https://cran.r- project.org/web/packages/wesanderson/wesanderson.pdf свободный. Яз. англ. (дата обращения: 14.05.2020).

9. Therneau T., Atkinson B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. [Электронный ресурс]. 2019. Режим доступа: https://cran.r-project.org/web/packages/rpart/index.html свободный. Яз. англ. (дата обращения: 14.05.2020).

10. Логинов А., Позина М., Манахова А., Дугинова А. Социальные сети: как заманить зрителя в кино [Электронный ресурс]. 2016. Режим доступа: http://kinometro.ru/analytics/show/name/social_media_distribution_9213 свободный. Яз. русский. (дата обращения: 14.05.2020).

11. Wu J. Advances in K-means Clustering A Data Mining Thinking. Springer, New York, 2012. 180 p.

12. Jain A., Murty M., Flynn P. Data clustering: A review // ACM Computing Surveys. – 1999. – Vol. 31, no. 3. – PP. 264-323.

13. Sahin K. Web Scraping: Handling AJAX website [Электронный ресурс]. 2018. // URL: https://ksah.in/web-scraping-handling-ajax-website/ свободный. Яз. англ. (дата обращения: 14.05.2020).

14. Kaufman L., Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley. 1990.

15. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия, алгоритмы Data Mining с использованием R [Электронный ресурс]. 2017. Режим доступа: https://github.com/ranalytics/data-mining свободный. Яз. русский. (дата обращения: 14.05.2020).

16. Banerjee, A. Validating clusters using the Hopkins statistic. IEEE International Conference on Fuzzy Systems: 149–153. 2004. doi: 10.1109/FUZZY.2004.1375706.

17. 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.

18. Alelyani S., Tang J., Liu H. Feature Selection for Clustering: A Review. Data Clustering: Algorithms and Applications 29, pp. 110-121, 2013.

19. Jolliffe I.T. Principal Component Analysis and Factor Analysis. In: Principal Component Analysis. Springer Series in Statistics. Springer, 1986, New York, NY.

20. Niknafs A., Charrad M., Ghazzali N., Boiteau V. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set // Journal of statistical software, October 2014. DOI: 10.18637/jss.v061.i06.

21. Sorokin D. vkR: Access to VK API via R. R package, version 0.1. [Электронный ресурс]. 2016. Режим доступа: https://github.com/Dementiy/vkR свободный. Яз. англ. (дата обращения: 14.05.2020).

22. Gunn S. R. Support Vector Machines for Classification and Regression. Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997.

23. В I полугодии 2018 года выручка российских онлайн-видеосервисов выросла на 32% [Электронный ресурс].2018. Режим доступа: https://telesputnik.ru/materials/video-v-internete/news/v-i-polugodii-2018-goda-vyruchka-rossiyskikh-onlayn-videoservisov-vyrosla-na-32/ свободный. Яз. русский. (дата обращения: 14.05.2020).

Статья публикуется при поддержке гранта РФФИ «Конкурс на лучшие проекты фундаментальных научных исследований» (Грант № 19-07-00516 А).

Статья поступила в редакцию 06.05.2020
Статья принята к публикации 10.06.2020