Федеральное государственное автономное
образовательное учреждение
высшего профессионального образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Базовая кафедра «Интеллектуальные системы управления»

УТВЕРЖДАЮ

Заведующий кафедрой

_____

Ю.Ю.Якунин

«___»_____2020г.

МАГИСТЕРСКАЯ ДИСЕРТАЦИЯ

«Кластеризация данных для сайта-агрегатора»

09.04.04 – «Программная инженерия»
09.04.04.02 – «Технологии индустриального производства программного обеспечения интеллектуальных систем управления»

Руководитель	_____	Доцент, канд. техн. наук	А.А. Даничев
	подпись, дата	должность, ученая степень	инициалы, фамилия
Выпускник	_____		Д.Н. Галин
	подпись, дата		инициалы, фамилия
Рецензент	_____	Канд. физ. - мат. наук	А.Л. Двинский
	подпись, дата	должность, ученая степень	инициалы, фамилия

Красноярск 2020

РЕФЕРАТ

Магистерская диссертация по теме «Кластеризация данных для сайта-агрегатора» содержит 59 страниц текстового документа, 15 рисунков, 9 формул, 1 таблицу, 11 использованных источников.

МЕТОДЫ КЛАССИФИКАЦИИ, МЕТОДЫ КЛАСТЕРИЗАЦИИ ДАННЫХ, RFM-АНАЛИЗ, МАРКЕТИНГОВЫЕ ТЕХНОЛОГИИ, БИЗНЕС ТЕХНОЛОГИИ, PYTHON, EXCEL, МОДУЛЬНОЕ ПРОГРАММИРОВАНИЕ.

Объект – Desktop-приложение, позволяющее анализировать данные о покупках пользователей путем их кластеризации и тем самым оценивать результативность проводимых акций.

Предмет – методы классификации и кластеризации клиентской базы данных.

Цель:

Разработать Desktop-приложение кластеризации данных для сайта-агрегатора.

Задачи:

- Выполнить анализ требований к разрабатываемому программному продукту;
- Изучить теоретическую составляющую методов классификации и кластеризации для дальнейшего использования в маркетинговых технологиях;
- Разработать архитектуру программы;
- Программно реализовать выбранные и разработанные математические решения;
- Проанализировать полученные результаты работы.

В результате проделанной работы были проведен анализ технического задания и системный анализ функционирования сайта-агрегатора. Так же принято решение о реализации набора инструментов для аналитика, придерживаясь парадигмы модульного программирования.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Наука о данных, в частности наука о больших данных давно заинтересовала аналитиков из разных сфер. Значительный интерес к работе с внушительным объемом данных проявляют в сфере торговли, что неудивительно. Обычным явлением в наши дни является сохранение данных покупателей и их покупок для дальнейшей работы с ними. Согласно анализу интернет-рынка, проведенному в ходе выполнения исследования, предложение провести кластеризацию данных является достаточно актуальным и пользуется спросом, однако все эти предложения платные и принцип их работы показан лишь в общих чертах; программный код, используемый для обработки данных, скрыт.

Актуальностью выбранной темы является необходимость внедрения современных технологий и алгоритмов для повышения эффективности продаж товаров и услуг. Маркетологам необходим инструментарий как для анализа имеющейся базы покупателей, так и для оценки результатов своей деятельности.

Объект – Desktop-приложение, позволяющее анализировать данные о покупках пользователей путем их кластеризации и тем самым оценивать результативность проводимых акций.

Предмет – методы классификации и кластеризации клиентской базы данных.

Цель:

Разработать Desktop-приложение кластеризации данных для сайта-агрегатора.

Задачи:

- Выполнить анализ требований к разрабатываемому программному продукту;

- Изучить теоретическую составляющую методов классификации и кластеризации для дальнейшего использования в маркетинговых технологиях;
- Разработать архитектуру программы;
- Программно реализовать выбранные и разработанные математические решения;
- Проанализировать полученные результаты работы.

Для создания программы, решающей задачу кластеризации пользователей сайта-агрегатора необходимо выбрать алгоритм и удобный для его реализации язык разработки. Так же необходимо учесть, что работа будет производиться с большим объемом данных, поэтому стоит озаботиться быстродействием системы.

В данном конкретном случае ведется разработка программного обеспечения для сайта, собирающего у себя информацию с множества других сайтов, своего рода торговой площадки. Основываясь на этом, принято решение произвести RFM-анализ и совместить его с кластеризацией с использованием иерархического алгоритма.

Таким образом, целью данной работы является разработка методики, объединяющей два подхода в сегментации клиентской базы данных – RFM-анализ и кластеризация покупателей по категории приобретаемых товаров и разработка модульного Desktop-приложения для сайта-агрегатора, реализующего разработанную методику.

Результаты теоретических исследований апробированы на международной научной конференции «Проспект свободный 2020». Тезисы докладов находятся в печати.

1 Анализ технического задания

Необходимо разработать Desktop-приложение для сайта-агрегатора, реализующее следующий функционал:

- RFM-анализ;
- Кластеризация клиентов по предпочитаемому товару;
- Предобработка данных для анализа.

Данное приложение разрабатывается для сайта nazya.com

Программа, разработанная в результате выполнения данной работы, должна иметь модульную структуру, позволяющую без глубоких знаний программирования изменять ядро алгоритма анализа, в зависимости от желаний маркетолога. Кроме того частью системы модулей, которые обеспечивают работу сайта nazya.com. таким образом, модель системы удовлетворяет условиям модульного программирования, а сама программа является модулем.

Рисунок 1 – модульная структура системы

1.1 Анализ технологий

Настольное приложение – программа, обрабатываемая на стороне клиента и запускаемая в виде обыкновенного исполняемого файла на устройстве пользователя. В качестве такого устройства чаще всего выступает компьютер. Рассмотрим средства для реализации таких приложений, а также выявим вспомогательные программы.

1.1.1 Python

Python – это универсальный современный язык программирования высокого уровня, к преимуществам которого относят высокую производительность программных решений и структурированный, хорошо читаемый код. Ядро имеет очень удобную структуру, а широкий перечень встроенных библиотек позволяет применять внушительный набор полезных функций и возможностей. Язык программирования может использоваться как для написания прикладных приложений, так и для разработки WEB-сервисов.

В научной среде Python активно применяют для проведения различных расчетов при подключении к пакетов NumPy, SciPy и MatPlotLib. Такой набор позволяет заменять даже специализированные коммерческие пакеты уровня Matlab.

Python – достаточно молодой язык. Вследствие этого, в нем можно найти многое от других языков, то, что создателями Python считается лучшим в этих языках. В частности, от ABC взята идеология отступов для группировки операторов и высокоуровневые структуры данных, от C, C++ - некоторые синтаксические конструкции, которые показались создателю языка наименее противоречивыми, и другие особенности, присущие другим языкам, как

например возможность как придерживаться парадигм ООП, так и некоторые черты функционального программирования.

Наиболее часто Python сравнивают с Perl и Ruby. Эти языки также являются интерпретируемыми и обладают примерно одинаковой скоростью выполнения программ. Как и Perl, Python может успешно применяться для написания скриптов (сценариев).

Как и Ruby, Python является хорошо продуманной системой для ООП. При этом реализация ООП в Python отличается от многих других объектно-ориентированных языков. В частности:

- В отличие от Ruby, Python не придерживается идеологии «всё — объект», и поддерживает встроенные примитивные типы, не входящие в иерархию классов. Такое решение упрощает и делает более технически эффективным межъязыковое взаимодействие, хотя может быть сочтено неудобным фанатами объектного подхода;
- В отличие от некоторых ООЯП (Java, Object Pascal, Ruby, …) в Python нет реального общего базового класса, от которого все объекты наследуют общие методы. Хотя формально новый класс в Python наследует (прямо или косвенно) тип object, это является только синтаксическим приёмом, так как методы, которые являются общими для всех объектов — id, type, isinstance, issubclass, str, repr, getattr, … не наследуются от object, а реализованы в виде глобальных функций. Такое решение приводит к тому, что изменение поведения этих методов производится не перегрузкой, а определением специальных методов класса.

В среде коммерческих приложений скорость выполнения программ на Python часто сравнивают с Java-приложениями.

Простота языка позволяет слабо разбирающемуся в программировании человеку легко добавить, заменить или удалить модуль программы.

1.1.2 LibreOffice

LibreOffice — мощный офисный пакет, полностью совместимый с 32/64-битными системами. Переведён более чем на 30 языков мира. Поддерживает большинство популярных операционных систем, включая GNU/Linux, Microsoft Windows и Mac OS X.

LibreOffice бесплатен и имеет открытый исходный код, следовательно, его можно бесплатно скачивать и использовать. **LibreOffice** бесплатен как для частного, так и для образовательного или коммерческого использования.

Пакет программ содержит текстовый и табличный процессор, средство записи и просмотра презентаций, редактор векторной графики, систему управления базами данных и редактор формул. Основным форматом файлов, используемым в приложении, является открытый международный формат OpenDocument, но также возможна работа с другими популярными форматами, включая Office Open XML, DOC, XLS, PPT и CDR.

Офисный пакет распространяется по публичной лицензии MPL 2.0, что позволяет его свободно устанавливать и использовать в бюджетных и коммерческих организациях, а также на домашних компьютерах и в учебных заведениях.

LibreOffice Calc — табличный процессор и визуальный редактор HTML, входящий в состав офисного пакета LibreOffice. Является ответвлением табличного процессора OpenOffice.org Calc. LibreOffice Calc распространяется по свободной лицензии Mozilla Public License v2.0.

Возможности табличного процессора:

Возможность чтения/записи OpenDocument (ODF), Excel (XLS), CSV, и нескольких других форматов.

Поддерживает 1 миллион рядов/строк в электронной таблице, что делает Calc вполне подходящим для сложных научных или финансовых документов.

Количество столбцов не превышает 1024, что намного ниже чем в Excel (16384 столбцов).

1.1.3 Microsoft Office Excel

Microsoft Excel - это пакет программ, разработанный фирмой Microsoft для работы в среде Windows. Этот пакет позволяет автоматизировать учрежденческую, производственную или научную деятельность, связанную со сбором, хранением, переработкой, передачей и использованием информации. Многие документы (планы работ, ведомости, справки, отчеты, счета, финансовые документы и т.п.) для наглядности представляют текстовые и числовые данные в табличной форме. Для автоматизации работы с ними служат специальные программы (точнее пакеты программ), называемые процессорами электронных таблиц (ЭТ) по названию используемого в них метода представления и обработки данных, одной из которых и является пакет программ Excel.

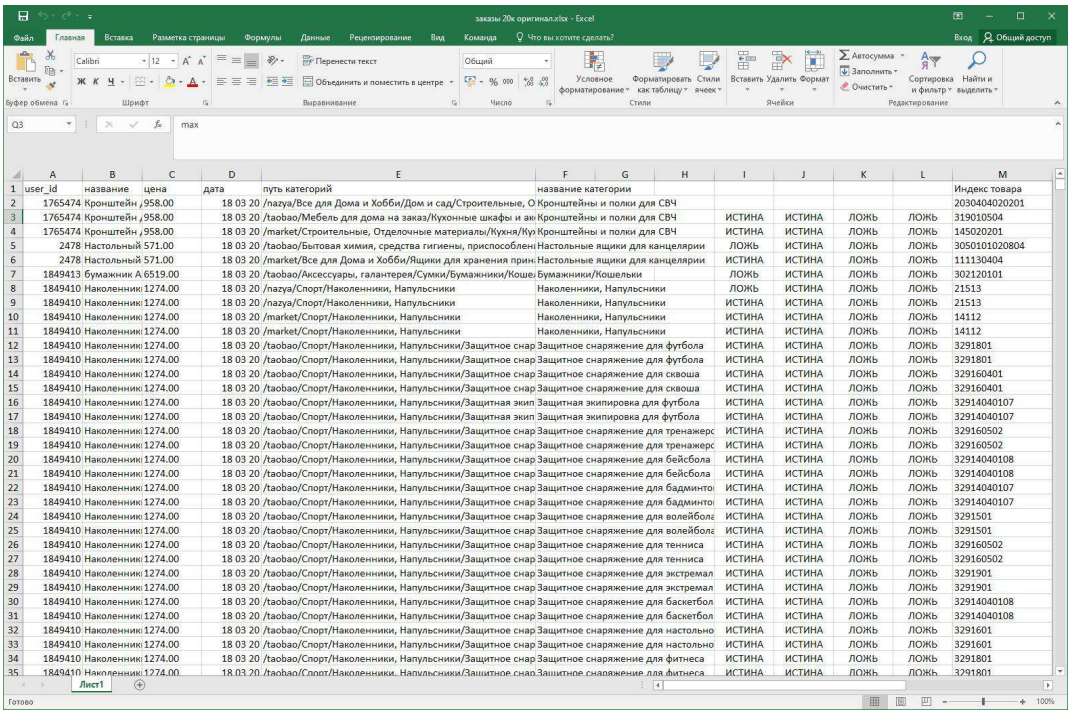


Рисунок 2 – пример данных в программе

Благодаря своему длительному присутствию на рынке и, вероятно, своей удобности, в настоящее время наиболее популярной программой среди пользователей является MS Excel. Данная программа предназначена для создания электронных таблиц любого уровня сложности, проведения математических расчетов, решения задач из области статистики и финансового анализа, построения графиков и диаграмм.

1.2 Выводы по первому разделу

В данной главе было проанализировано техническое задание и произведен обзор на средства его реализации.

Основное требование к программе – возможность модернизации ядра алгоритма пользователем, не обладающим глубокими знаниями программирования, например, в частности замена ядра алгоритма RFM – анализа на другой, которому аналитик, возможно, доверяет больше, чем тому, что используется сейчас. Это достигается модульным строением программы и использованием в качестве языка разработки Python, поскольку данный язык обладает большим количеством как встроенных, так и постоянно добавляющихся программистами по всему миру библиотек и прост даже для пользователя, слабо разбирающегося в программировании.

Так же, для подготовки данных, выгружаемых из базы, было принято решение использовать MS Office Excel, поскольку маркетологу компании он более привычен и удобен, чем возможный его бесплатный аналог LibreOffice Calc.

2 Системный анализ функционирования сайта-агрегатора

2.1 Основные понятия системного анализа

Термин системный анализ или системный подход, несмотря на многолетнее использование человеком, так и не получил общепринятого стандартного толкования. Причина заключается в динамизме всех процессов, происходящих практически в любой области человеческой деятельности, и, кроме того, в фундаментальной возможности использования системного подхода практически в любой проблеме, решаемой человеком [1].

Во всем мире системный подход как к управлению технологиями, так и к управлению организациями быстро развивался не как абстрактная теория, а как реальный инструмент управления, который делает работу человека более продуктивной. Системный подход использовался при разработке и реализации крупных военных проектов, программ полетов человека на Луну - во всех случаях, когда необходимо было планировать и организовывать деятельность сотен компаний с различными формами собственности и спецификой работы. Она использовалась для проектирования и управления отдельными крупными и малыми организациями, а теперь стала основой для международного языка лидеров, дав им возможность понять друг друга.

Организационное развитие как подход полностью основано на принципах систематики. Методы исследования операций основаны на идее системного подхода к анализу сложных задач и синтезу инструментов их решения. Математические модели были средством выявления связей между элементами сил и средств, используемых в операциях. [2]

В наши дни, системный анализ – это прикладная наука, цель которой - выяснение причин реальных сложностей, возникших перед «обладателем проблемы» (конкретной организацией, учреждением, предприятием,

коллективом) и на выработка вариантов их устранения. Существует большое количество определений термина «системный анализ», обобщая их, можно дать следующее толкование. Системный анализ – междисциплинарное научное направление. Предмет системного анализа – это концепции и причины постановки и разрешения практических проблем на основе системной идеологии. В качестве инструмента системного анализа используется широкий спектр математических методов: линейное и нелинейное программирования, теория принятия решения, имитационное моделирование и т.п.

Объект - это любая реальная сущность, которую можно отличить по некоторым признакам, изолированным в бесконечно разнообразном мире. Другой фундаментальной концепцией системного анализа является субъект. Субъект - это физическое или юридическое лицо, в состав которого входит лицо как неотъемлемая часть (группа, организация, население, нация, государство, объединение людей и т.д.). В чем разница между объектом и субъектом? Во-первых, субъект способен не только взаимодействовать со средой, характерной для всех объектов, но и оценивать свои взаимодействия со средой (это свойство присуще только субъекту). Во-вторых, субъект имеет возможность нацеливаться, способность к целеполаганию.

Объектом системного анализа являются прикладные задачи различного иерархического уровня (от государственного до личного), связанные с созданием новых и совершенствованием (модернизацией) существующих организационных, технических, технологических, концептуальных, информационных, экономических и иных систем. Отсюда следует следующее понятие системного анализа - сложная система.

«Система есть средство достижения цели» и «система есть совокупность взаимосвязанных элементов, обособленная от среды и взаимодействующая с ней как целое». Ф.П. Тарасенко раскрывает понятие «система» для целей системного анализа через перечисление свойств системы.[4] Эти свойства должны отвечать следующим требованиям:

1. Они должны быть присущи всем системам (естественным и искусственным, реальным и идеальным);

2. Знание каждого из них потребуется на какой-то стадии процесса решения проблемы.

Количество свойств равно 12, их можно объединить в 3 группы: статические, динамические, и синтетические свойства.

Тем не менее, данное определение не дает ясности, что является системой, а что нет. Для этого понятие системы дополняют классификациями и уточнениями. В литературе встречаются самые разные классификации:

- По характеру поведения (детерминированного, вероятностного, игрового);
- По типу назначения (открытое и закрытое);
- Сложность структуры и поведения (простые и сложные);
- По типу научного направления, используемого для их моделирования (математического, физического, химического и др.);
- По степени организации (хорошо организованная, плохо организованная и самоорганизующаяся).[5]

Вероятностные или стохастические системы - системы, поведение которых описывается законами теории вероятностей. Данных о текущем состоянии системы и взаимосвязях элементов недостаточно, чтобы с уверенностью предсказать будущее поведение системы. Для такой системы существует множество возможных переходов из одного состояния в другое, то есть серия преобразований состояния системы, и каждый сценарий имеет свою вероятность. Примером стохастической системы является деятельность туристического агентства. Количество проданных билетов будет зависеть от количества заявок, поведения объектов и т.д.

Следующий признак классификации – по типу целеустремленности. Открытые системы взаимодействуют с окружающей средой, происходит обмен энергией, массой, информацией. В закрытой системе изучаются

внутрисистемные обратные связи, они изолированы от внешнего мира. Разница между открытыми и замкнутыми системами определяется с точностью до принятой чувствительности модели.

Сложные системы реагируют по-разному на внешнее воздействие согласному своему внутреннему состоянию. Причем наблюдатель не может предугадать поведение системы. В двух на разных случаях система может дать одинаковый результат на одно и то же воздействие, а может – совершенно разный. Это объясняется тем, сложная система формирует свои законы и правила. Она характеризуется большим количеством внутренних связей.

Простая система имеет небольшое количество возможных состояний, их поведение легко описывается в рамках той или иной математической модели. Она вполне предсказуемо реагирует на внешнее воздействие.

Итак, в дисциплине системный анализ можно выделить три главных направления:

- Построение модели исследуемого объекта;
- Постановка задачи исследования;
- Решение поставленной задачи.

Кратко рассмотрим каждый этап. Построение модели – описание процесса на языке математики. При этом подробно описывают те процессы, которые интересуют исследователя. Описание выполняют согласно тем требованиям, которые предъявляются к исследователю, а качество модели определяется соответствием получаемых с помощью модели результатов ходу наблюдаемого процесса или явления. От качества модели зависит результат всего системного анализа.

На этапе постановки задачи исследования формулируется цель анализа. Цель – самостоятельный объект исследования, она должна быть формализована. Задача системного анализа состоит в проведении необходимого анализа неопределенностей, ограничений и формулирований, в конечном счете, некой оптимизационной задачи.

На этапе решения поставленной задачи используется в полной степени математические методы. Следует отметить, что задачи системного анализа могут иметь ряд особенностей, которые приводят к необходимости применения наряду с формальными процедурами эвристических подходов.

2.2 Понятие лояльности потребителя

Лояльность (loyalty) – качественная маркетинговая характеристика отношения потребителей к товарам и услугам по признаку их привязанности к определенной марке.

Потребители бывают:

Рисунок 3 – Три группы лояльности клиентов

Лояльные покупатели, это те, которые доверяют и постоянно пользуются услугами компании. Развивающиеся клиенты пользуются так же услугами других компаний. Нелояльные, или отток, это те клиенты, которые уже перестали или в скором времени совсем перестанут совершать покупки в целевом коммерческом предприятии.

В современном мире слово «loyal» с французского и английского переводится как «верный». Лояльность имеет как минимум 2 значения:

- Верность действующим законам, постановлениям органов власти;
- Корректное, благожелательное отношение к кому-либо или чему-либо.

Лояльность в маркетинге — это построение долгосрочных отношений, в которых клиент благосклонно относится к товару, бренду или магазину и становится постоянным клиентом.

Маркетинг лояльности еще называют маркетингом взаимоотношений. Понятие «лояльность» объединяет три ключевых элемента:

- Доверие;
- Приверженность ценностям;
- Долговременные отношения.

В современном мире, программы лояльности преследуют ряд задач, необходимых для выживания в конкурентной борьбе. Именно с помощью разработки и увеличения лояльности клиентов выполняются многие аспекты развития коммерческого предприятия. Задачи лояльности:

- Повышение узнаваемости бренда;
- Повышение лояльности посетителей к вашей продукции;
- Повышение продаж продвигаемого вами продукта на период проведения акции;
- Повышение частоты совершения покупок посетителем торговой сети.

При правильно выстроенной программе повышения лояльности клиентов, компания получает ряд выгод и преимуществ по сравнению с коммерческими предприятиями, не имеющими рабочих программ лояльности. Выгоды, получаемые от внедрения программ повышения лояльности:

- Прибыль;
- Уменьшение себестоимости отношений с клиентом;
- Стабильный денежный поток.

Существует 4 вида ресурсов, которыми обладает каждый потребитель:

- Временной ресурс;
- Денежный ресурс;
- Когнитивный ресурс (Познание);
- Аффективный ресурс (Эмоциональное отношение).

Главное желание покупателя — купить необходимый ему продукт *с* наименьшими потерями ценных для него ресурсов: быстрее, дешевле, проще и без эмоционального волнения.

Качественное удовлетворение главного желания потребителя и порождает лояльность клиента к торговому предприятию.

2.3 Задача увеличения лояльности

Лояльность клиента к компании или товару имеет две стороны – внешнюю и внутреннюю. Внутренняя сторона – сторона клиента. Лояльность клиента повышается при каждой совершенной покупке, при условии, что клиент ощущает экономию важных для него ресурсов – времени, денег и умственных усилий. Внешняя сторона – сторона продавца, то есть коммерческого предприятия, которое предоставляет услугу или товар. Каждая совершенная клиентом покупка увеличивает товарооборот создаваемый клиентом, посредством увеличения стоимости его среднего чека или частоты совершения покупок. Другими словами, для коммерческого предприятия лояльность клиента выражается в его покупательских характеристиках – частоте совершения покупок и создаваемом товарообороте.

Таким образом, задача увеличения лояльности клиентов достигается двумя путями – увеличением частоты совершения покупок клиентом, или увеличением среднего чека. Другими словами, показателем лояльности клиента является его деятельность в сети как активного покупателя, чем больше и чаще он покупает, тем более он лоялен к торговому предприятию.

Пусть $z_i \in Z$ – запись в базе данных (соответствующая дисконтной карте), где Z – множество записей, а $|Z| = n$

Для каждого $z_i \in Z$:

$$\begin{aligned} &F(z_i) \to \max, \\ &M(z_i) \to \max, \end{aligned} \qquad (1)$$

Где:

$F(z_i)$ – частота совершения покупок по карте

$M(z_i)$ – товарооборот, создаваемый клиентом, использующим карту

2.4 Программа увеличения лояльности клиентов

Программа лояльности представляет собой комплекс маркетинговых мероприятий, основной целью которых является привлечение и удержание клиентов для увеличения перепродаж, продажи им дополнительных товаров и услуг, а также повышение их интереса к деятельности компании.

Программа лояльности - это форма маркетинга, направленная на создание долгосрочных отношений с клиентами, чтобы сделать их постоянными покупателями. Лояльность позволяет понять потребности клиента и разработать нужные ему услуги. Программа лояльности направлена на повышение удовлетворенности клиентов вашей компанией. [7]

Задача программ лояльности - сформировать стабильную потребительскую базу. Согласно закону Парето (закон 80:20), основанному на статистических исследованиях, 20% покупателей обеспечивают 80% прибыли. Именно для удержания этих 20% потребителей следует рассчитать программы лояльности, так как расходы торговой компании на выигрыш новых клиентов в 6-11 раз превышают расходы на укрепление существующей клиентской базы. И

лучший способ сохранить клиента - предложить ему любую выгоду при покупке продукта или услуги у вашей компании.

При правильном планировании программы лояльности могут стать хорошим инструментом для увеличения клиентской базы. Например, более 60 % всех домохозяйств Канады принимают участие в программе Air Miles Canada.

В настоящее время существует множество видов программ лояльности: дисконтная карта, сберегательная система, система бонусов и подарков, реферальные программы и так далее. При входе в программу, как правило, клиент заполняет анкету, в которой указаны контакты получателя, что дает организации возможность после проведения предварительного анализа данных уведомить клиента о новых и потенциально интересных продуктах и услугах. Используя данные о покупках клиентов, можно создать профиль поведения покупателей, который позволяет компаниям лучше знать своих клиентов, их интересы и предпочтения, привычки и способы совершения покупок, в то время как им удобнее совершать покупки. Главный принцип - правильное использование информации о клиентах для повышения их интереса к розничной сети, тем самым увеличивая перепродажу и общий товарооборот сети.

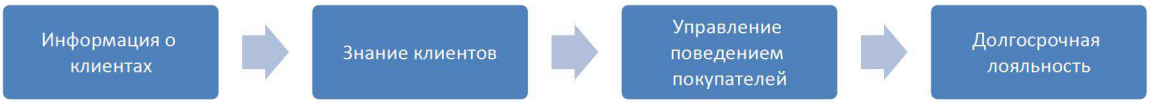


Рисунок 4 – Принцип работы программы лояльности

Преимущества программы лояльности:

- Расширение знаний клиентов за счет анализа данных;
- Анализ корзины;
- Создание профилей клиентов;
- Объединение клиентов в группы;
- Формирование выгодных торговых предложений;
- Привлечение и удержание клиентов;
- Увеличение объема продаж;
- Укрепление долгосрочных торговых отношений с клиентом.

Одна из главных задач программ лояльности - анализ целевой аудитории, то есть выявление групп людей, которые с наибольшей вероятностью купят товар или воспользуются услугой. В большинстве случаев основным способом анализа целевой аудитории является разделение (кластеризация) данных - разделение клиентов на группы с похожими свойствами, выявление потребностей группы и формирование предложения, ориентированного на целевой сегмент.

Одно из общеизвестных правил мотивационного менеджмента гласит: «эффективно управлять можно только тем, что можно измерить».

Зарубежные специалисты утверждают, что абсолютно лояльных покупателей нет. Настоящую целевую аудиторию формируют те клиенты, которые ходят постоянно в течение длительного времени.

Лояльность клиента может быть определена как положительное отношение покупателя к товару, бренду, магазину, сервису и т. д., которое, хотя и является следствием значимых для покупателя факторов, в большей степени лежит в эмоциональной сфере.

Рассмотрим подробнее модель программы лояльности. Принцип взаимодействия отделов компании показан на рисунке 5. Задача отдела информационных технологий – предоставить данные для анализа отделу

системного анализа. Отделу системного анализа, в свою очередь, необходимо из полученных данных получить знания о клиентах в удобном для маркетологов виде и передать эту информацию в отдел маркетинга. Маркетологи сети, исходя из ситуации, формируют конкретное маркетинговое предложение покупателям, и осуществляют взаимодействие с клиентом по реализации предложения.

В рамках данной работы отдел системного анализа представляет выполняющий исследование человек. Отдел системного анализа ответственен за составление технического задания ИТ-отделу по формированию первичных данных для анализа, а так же за предоставление конечных знаний о клиентах в удобном для маркетологов виде, для того, чтобы сотрудники отдела маркетинга, проанализировав эти данные, могли разработать конкретное маркетинговое предложение по реализации некоторой торговой акции на продукты сети. Это маркетинговое предложение может быть обсуждено с сотрудниками других отделов, с целью получения их экспертной оценки о выгоде, получаемой с рассматриваемой акции. В конечном итоге, после утверждения акции, отдел маркетинга составляет техническое задание по реализации акции и затем передает в ИТ-отдел, либо, если это возможно, выполняет связь с клиентом собственными силами.

Обобщенная схема взаимодействия отделов по программе лояльности представлена на рисунке 5. Следует отметить, что все части системы, участвующие в развитии программы лояльности являются равноправными и равнозначными, при нарушении работы хотя бы одного элемента системы, развитие клиентской базы будет нарушено.

Рисунок 5 – Взаимодействие отделов компании

Следует отметить, что кроме формирования и реализации маркетингового предложения необходимо так же отслеживать отклик клиентов, и как следствие – динамику изменения клиентской базы. Далее, при формировании новых маркетинговых предложений, учитывать так же и прошлый опыт по динамике и развитию клиентской базы.

В ходе проведения работы были рассмотрены главные принципы работы программы лояльности. Основная часть работы посвящена именно получению знаний о покупателях, в рамках работы отдела системного анализа.

При создании и развитии программ лояльности возникает проблема анализа целевой аудитории, т.е. определения группы людей, которые чаще всего покупают товары или пользуются услугой. В большинстве случаев основным способом анализа целевой аудитории является разделение (кластеризация) данных - разделение клиентов на группы с похожими свойствами, выявление потребностей группы и формирование предложения, ориентированного на целевой сегмент.

В данной работе были рассмотрены основные методы сегментации данных, их достоинства и недостатки относительно развития программы лояльности.

В ходе проведенной работы были изучены основные методы сегментации баз данных. Для каждого из этих методов были разобраны их подвиды. Данная научно-исследовательская работа проводилась на базе сайта-агрегатора Nazya.com в 2019-2020 году.

2.5 Выводы по второму разделу

В данной главе были проанализированы принципы функционирования программы лояльности, что позволит продолжить изучение поставленной проблемы, а так же сконцентрироваться на конкретном аспекте ее развития – эффективной сегментации и анализе данных для приобретения знаний о клиентах.

3 Кластеризация и обработка данных

3.1 Кластеризация данных

Одна из главных задач при разработке программы лояльности - анализ целевой аудитории, то есть выявление групп людей, которые с наибольшей вероятностью купят предлагаемый товар или воспользуются предлагаемым сервисом. В большинстве случаев основным способом анализа целевой аудитории является разделение (кластеризация) данных - разделение клиентов на группы с похожими свойствами, выявление потребностей группы и формирование предложения, ориентированного на целевой сегмент.

Кластерный анализ (англ. data clustering) - задача разделения заданного выбора объектов (ситуаций) на подмножества, называемые кластерами, так что каждый кластер состоит из одних и тех же объектов, а объекты различных кластеров значительно отличаются. Задача кластеризации относится к статистической обработке, а также к широкому классу образовательных задач без учителя. Кластерный анализ - многомерная статистическая процедура, которая собирает данные, содержащие информацию выборки объектов, а затем упорядочивает объекты по относительно однородным группам (кластерам).

Кластер - это группа элементов, характеризующихся общим свойством, основной целью кластерного анализа является поиск групп подобных объектов в выборке. Диапазон применения кластерного анализа очень широк. Анализ используется в археологии, медицине, психологии, химии, биологии, госуправлении, филологии, антропологии, маркетинге, социологии и других дисциплинах. "Темы исследования варьируются от анализа морфологии мумифицированных грызунов в Новой Гвинее до изучения результатов голосования сенаторов США, от анализа поведенческих функций замороженных тараканов при их оттепели до изучения географического

распределения некоторых видов депривации в Саскачеване". Однако универсальность применения привела к появлению большого числа несовместимых терминов, методов и подходов, которые затрудняют однозначное использование и последовательное толкование кластерного анализа.

Кластерный анализ выполняет следующие основные задачи:

- Разработка типологии или классификации;
- Изучение полезных концептуальных схем группировки объектов
- Формирование гипотезы на основе исследования данных;
- Проверка гипотез или исследований для определения того, действительно ли типы (группы), изолированные так или иначе, присутствуют в доступных данных.

Независимо от предмета исследования, использование кластерного анализа включает следующие шаги:

- Выбор выбора для кластеризации;
- Определение набора переменных, по которым будут оцениваться объекты выборки
- Расчет значений измерения подобия между объектами;
- Использование метода кластерного анализа для создания групп подобных объектов;
- Проверка результатов кластерного решения;
- Кластерный анализ имеет следующие требования к данным:
- Не следует сравнивать показатели;
- Показатели должны быть безразмерными;
- Распределение показателей должно быть близко к норме;
- Показатели должны отвечать требованию "устойчивости", а значит, влияния на их значения случайных факторов нет;
- Образец должен быть однородным, свободным от "выбросов".

Если кластерному анализу предшествует факторный анализ, то образец не нуждается в "ремонте" - заявленные требования автоматически выполняются самой процедурой моделирования факторов (есть и другое преимущество - z-стандартизация без негативных последствий для образца; если он выполняется непосредственно для кластерного анализа, это может привести к снижению ясности разделения групп). В противном случае шаблон должен быть исправлен.

Формальная постановка задачи кластеризации:

Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации.

Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки исходных объектов y_i изначально не заданы, и даже может быть неизвестно само множество Y.

Решение проблемы кластеризации принципиально неоднозначно, и для этого есть несколько причин:

- Нет однозначно наилучшего критерия качества для кластеризации. Известен ряд эвристических критериев, а также ряд алгоритмов, не имеющих

чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию "по построению". Все они могут дать различные результаты;

- Количество кластеров, как правило, заранее неизвестно и устанавливается по какому-то субъективному критерию;
- Результат кластеризации существенно зависит от метрики, выбор которой, как правило, тоже субъективен и определяется экспертом.

Существует несколько видов сегментации клиентских баз данных в связи с анализом клиентских баз розничных сетей:

- В соответствии с профилем клиента (социально-демографическая сегментация);
- По поведенческим данным клиентов (частота и объем закупок, анализ RFM);
- Стоимость клиента (анализ ABC);
- По предпочтениям в ассортименте;
- По динамике развития и деятельности;
- Сложная сегментация.

К сожалению, невозможно рассматривать какие либо манипуляции над базами данных, а тем более, такие как кластеризация базы данных, не столкнувшись с проблемой избыточности информации или с проблемой ограниченности вычислительных мощностей ЭВМ. Прежде чем приступить к разработке алгоритма кластеризации клиентской базы данных для торговой розничной сети, рассмотрим подробнее проблему избыточности данных.

3.2 Возможности Data-mining в розничной торговле

Методология Data Mining переводится как "добыча" или "раскопка данных". Нередко рядом с «Data Mining» встречаются слова "обнаружение знаний в базах данных" (knowledge discovery in databases) и "интеллектуальный анализ данных". Появление всех этих терминов связано с новым витком в разработке инструментов и методов обработки данных.

Сфера охвата "Data Mining" не ограничена – он есть везде, где есть какие-либо данные. Но прежде всего методы майнинга данных сегодня, мягко говоря, заинтриговали коммерческие предприятия, развертывающие проекты на основе хранения данных. Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Например, ежегодная экономия в размере 700 000 долл. США. через внедрение "Data Mining" в вагонную сеть в Великобритании.

Data Mining представляет большую ценность для менеджеров и аналитиков в их повседневной деятельности. Бизнесмены поняли, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе. Кратко опишите некоторые возможные бизнес-приложения для майнинга.

До начала 1990-х годов казалось, что особой необходимости переосмыслять ситуацию в этой области нет. Все шло в рамках направления под названием прикладная статистика. Теоретики проводили конференции и семинары, писали впечатляющие статьи и монографии, изобиловали аналитическими находками.

Однако пратики всегда знали, что попытки применить теоретические упражнения для решения реальных проблем в большинстве случаев безрезультатны. Но заботам практиков пока не уделяется большого внимания - они решали в основном свои частные проблемы обработки небольших локальных баз данных.

Из-за совершенствования технологий записи и хранения данных огромные потоки информационных руд в различных районах ударили по людям. Деятельность любого предприятия (коммерческого, промышленного, медицинского, научного и т.д.) теперь сопровождается регистрацией и регистрацией всех деталей его деятельности.

Специфика современных требований к переработке данных такова:

- Данные имеют неограниченный объем;
- Данные неоднородны (количественные, качественные, текстовые);
- Результаты должны быть конкретными и четкими;
- Средства обработки необработанных данных должны быть простыми в использовании.

Традиционная математическая статистика, долгое время претендовавшая на роль главного инструмента анализа данных, откровенно спасовала перед лицом проблем. Основная причина - понятие усреднения выборки, приводящее к операциям над фиктивными значениями (такими как средняя температура пациентов в больнице, средняя высота дома на улице, состоящая из дворцов и лачуг и т. д.). Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез (разработка данных на основе проверки) и для "грубого" анализа данных, который составляет основу обработки аналитических данных в режиме онлайн (OLAP).

Современная технология Data Mining основана на концепции шаблонов (шаблонов), которые отражают фрагменты многомерных отношений в данных. Эти паттерны являются паттернами, характерными для подзамеров данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов осуществляется методами, не ограничивающимися априорными допущениями о структуре выборки и типе распределений значений анализируемых показателей.

В целом технологию Data Mining достаточно точно определяет Григорий Пиатецкий-Шапиро — один из основателей этого направления:

Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Сегодня розничные продавцы собирают подробную информацию о каждой отдельной покупке с помощью кредитных карт под торговой маркой магазина и компьютеризированных систем управления. Вот типичные проблемы, которые можно решить с помощью интеллектуального анализа данных в розничной торговле:

- Анализ корзины покупок (анализ схожести) предназначен для идентификации товаров, которые покупатели хотят приобрести вместе. Знание корзины необходимо для улучшения рекламы, разработки стратегии создания запасов товаров и способов их размещения в торговых залах;
- Исследование шаблонов времени помогает маркетологам принимать решения по запасам. Он отвечает на такие вопросы, как "Если покупатель приобрел видеокамеру сегодня, завтра он, скорее всего, купит новые батареи и пленку?";

Создание прогнозных моделей позволяет маркетологам узнать характер потребностей различных категорий клиентов с определенным поведением, например, покупка товаров известными дизайнерами или посещение продаж. Эти знания необходимы для разработки точно ориентированных и экономически эффективных мероприятий по продвижению продукции. При разработке мер по продвижению и распределению товаров следует привлекать высококлассных маркетологов-специалистов.

3.3 RFM-анализ

Анализ RFM - это инструмент, позволяющий сегментировать потребителей по уровню лояльности на основе их прошлых действий,

прогнозировать их поведение. RFM - аббревиатура от слов Recency - новизна, Frequency - частота и Monetary, что означает затраты или инвестиции.

Если рассмотреть эти концепции подробнее, то Recency означает вероятность возвращения клиента, исходя из того, сколько времени прошло с момента его последней деятельности - чем меньше времени, тем больше вероятность того, что потребитель вернется снова.

Параметр Frequency - это количество действий, предпринятых клиентом за определенный период времени. Считается, что чем больше заказов делает тот или иной потребитель, тем больше вероятность, что в следующем периоде он снова сделает заказ.

Денежный параметр, Monetary, характеризуется количеством денег, потраченных клиентом в течение выбранного периода времени. Опять же, чем больше средств потратил потребитель, тем больше вероятность их повторного расходования. Стоит отметить, что этот пункт часто может отсутствовать в анализе, поскольку имеет прочную связь с Frequency. Денежные средства также могут отсутствовать в случаях, когда полученная от клиента выгода не может быть учтена в деньгах.

В некоторых ситуациях вместо Monetary может использоваться понятие Duration, которое описывает общую продолжительность работы с конкретным клиентом, длительность его подписки.

Первый шаг – исследование параметра Recency. Для начала здесь нужно определиться с тем, что будет являться критерием активности клиента. Это может быть покупка, посещение магазина, или даже переход по ссылке в интернете на сайт компании. Все зависит от того, какие цели преследует фирма, проводя анализ, и от специфики деятельности.

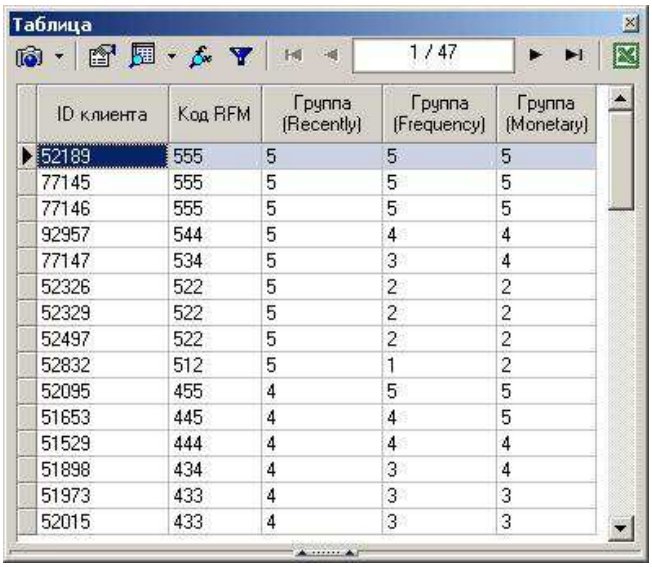
Далее необходимо распределить клиентов на пять групп на основе того, как давно они совершали те или иные действия.

Это распределение удобно осуществлять, определяя несколько временных циклов (в нашем случае - 5). Например, действия, совершенные за

прошедший месяц, от 1 до 2 месяцев назад, от 2 до 3 месяцев назад, от 3 месяцев до шести месяцев и от шести месяцев до года.

Группа, в которой собираются потребители, недавно совершившие действия, обычно обозначается номером 5. Номер 1 - это группа потребителей, которые не были активны дольше всех.

Следующим шагом в анализе является работа с Frequency. Принцип распределения клиентов по 5 группам здесь такой же, как в Recency. Только вместо распределения временных сегментов необходимо определить, сколько операций будет критерием присоединения клиента к определенной группе. Например, 20 или более покупок - 5 группа, 0-2 покупки - 1. В то же время следует понимать, что установление завышенных или заниженных пороговых значений ввода в группы может существенно повлиять на точность результатов. На рисунке 6 показана идеальная ситуация сегментации RFM.



ID клиента	Код RFM	Группа (Recently)	Группа (Frequency)	Группа (Monetary)
52189	555	5	5	5
77145	555	5	5	5
77146	555	5	5	5
92957	544	5	4	4
77147	534	5	3	4
52326	522	5	2	2
52329	522	5	2	2
52497	522	5	2	2
52832	512	5	1	2
52095	455	4	5	5
51653	445	4	4	5
51529	444	4	4	4
51898	434	4	3	4
51973	433	4	3	3
52015	433	4	3	3

Рисунок 6 – Идеальная ситуация RFM сегментации

Затем, следуя процедуре анализа RFM, необходимо определить денежный параметр. Здесь все так же, как и в предыдущих абзацах. Условием вхождения клиента в ту или иную группу является преодоление установленной планки в виде суммы потраченных средств.

В результате после анализа можно сформировать 125 групп с 111-й по 555-ю (числа являются комбинациями номеров групп для каждого индикатора). Но это не означает, что к потребителям каждой группы следует относиться индивидуально. Здесь будет наиболее полезно выделить различные тенденции в поведении клиентов, выявить наиболее важных, ключевых клиентов компании и тех, кто близок к тому, чтобы стать приверженцем компании.

Группа 555, например, являются самыми лояльными потребителями, в которых компания может быть уверена. Но при этом не нужно думать, что их можно "забыть", так как они никуда не уйдут. Таким клиентам нужно показать, что они действительно важны для компании, что компания благодарна им (создавая особые условия, программы лояльности).

Группа 111 - самые бесперспективные потребители. Однако нужно понимать, что хотя бы раз, но они обратились к услугам компании. Попытка привлечь их снова или понять причину низкой активности - возможные задачи для маркетологов компании.

Клиенты с Recency 5 находятся в состоянии оценки компании и могут появиться в качестве покупателей вновь. Их можно безопасно привлекать, осуществляя почтовые или онлайн-рассылки, поощрять совершать покупки.

Группе потребителей, которые совершают частые покупки с небольшими суммами, могут быть предложены сопутствующие товары и услуги для того, что они покупают.

Таким образом, с помощью RFM анализа можно добиться разбиения основной базы на рабочие сегменты. На рисунке 7 представлена сегментация базы данных на рабочие сегменты: лояльные, развивающиеся, отток.

	Частота покупок						
Размер средних трат за нед	Ежедневно	Дважды в неделю	Еженедельно	Нестабильно по неделям	В начале и в конце	Редко но с большим объемом	Редко
Большая корзина	Ежедневные средние покупки		Недельные основные	Разовые средние		Несколько больших корзин	Покидающие: Не совершал покупок за последние 6 недель
Средняя корзина	Лояльные			Развивающиеся			
Маленькая корзина	Малые средние покупки		Еженедельные средние		Несколько средних покупок		Нелояльные - Отток

Рисунок 7 – Разбиение клиентской базы на сегменты

Сегментация с помощью RFM-анализа позволяет легче отслеживать динамику клиентских сегментов. На рисунке 2.3.3 представлен пример реализации отслеживания динамики рабочих RFM-сегментов.

		Стало в новом периоде					
		1. Новички	2. Пот. отток	3. Лояльные клиенты	4. Новички VIP или кассиры	5. VIP клиенты	6. Отток
Было в предыдущем периоде	1. Новички	10%	25%	20%	10%	5%	30%
	2. Потенциальный отток			20%			60%
	3. Лояльные клиенты		15%	80%			5%
	4. Новички VIP или кассиры		10%	20%	50%	20%	
	5. VIP клиенты		20%			80%	
	6. Отток		20%				70%

Рисунок 8 – Динамика рабочих сегментов

Сегментация клиентской базы является одной из самых важных задач, с которой может столкнуться сеть магазинов розничной торговли.

В результате исследования, было выявлено, что ключевыми факторами лояльности клиента являются – частота покупок, длительность нахождения в базе, и количество потраченных клиентом денег. Сегментация по данным факторам является важнейшей частью для реализации цели создания лучшей программы лояльности. Так же, данный анализ позволяет легче отслеживать динамику лояльных и нелояльных сегментов покупателей.

Алгоритм RFM-анализа:

1. Классификация по параметру Recency:

- для каждого клиента определить дату последней покупки;
- для каждого клиента рассчитать давность покупки (Recency) как разность между текущей датой и датой последней покупки;
- разбить полученные данные на 5 групп (квантилей). Каждый клиент при этом получит идентификатор от 1 до 5 в зависимости от его активности. Тем, кто недавно осуществлял покупку, будет присвоен код R=5. Те, кто дольше всех не покупал ничего, получат R=1.

1. Классификация по параметру Frequency:

- для каждого клиента определить количество покупок за определённый период;
- разбить полученные данные на 5 групп (квантилей). Клиентам, совершившим наибольшее число покупок, будет присвоен код F=5, наименее активные покупатели получат F=1.

2. Классификация по параметру Monetary:

- для каждого клиента определить сумму потраченных денег;
- разбить полученные данные на 5 групп (квантилей). Клиентам, потратившим наибольшие суммы, будет присвоен код M=5, клиентам, потратившим наименьшие суммы – M=1.

3. Совместить полученные результаты, каждый клиент при этом получит код RFM, состоящий из трёх цифр.

3.4 Адаптация алгоритма

По результатам первых двух этапов – проведения RFM-анализа и кластеризации RFM-групп были получены неудовлетворительные результаты. Было принято решение о пересмотре алгоритма. В частности, возникла необходимость пересмотреть временные рамки отбора клиентских карт для

анализа, а так же сформировать более наглядные диапазоны для параметров Monetary и Frequency (товарооборот и частота совершения покупок).

Параметр R (Recency – актуальность клиента) был исключен из рассмотрения при проведении RFM-анализа. Было принято решение, что все клиенты, посетившие сайт за период, по которому формируется отчет – являются целевой аудиторией, и программа лояльности будет действовать на них только в рамках одного квартала. Исключение параметра актуальности клиента, помимо дополнительных вычислительных затрат позволило проиллюстрировать результаты RFM-анализа в обычной таблице.

Далее, для формирования новых диапазонов параметров Monetary и Frequency, был проведен частотный анализ данных. Для этого, для каждого аккаунта были подсчитаны показатели среднего чека и среднего товарооборота. В таблице 1 представлены показатели среднего чека и средней частоты покупок по нескольким аккаунтам.

К сожалению, средняя частота покупок, совершенных клиентами, имеет сильный перекос в сторону «купил один-два раза».

Таблица 1 – Средние показатели аккаунтов

ID аккаунта	Ср. чек 1 покупки период, (руб.)	Ср. частота покупок за период
1765474,1,2,1	2874	1
1849413,1,3,1	6519	1
1849410,1,5,1	34398	1
…	…	…
1840687,1,3,1	4902	1

Все аккаунты были сгруппированы в интервалы по средней частоте и среднему чеку соответственно. В таблице 1 представлена группировка клиентов по средней частоте покупок, совершенных клиентом за неделю.

Параметр R (Recency – актуальность клиента) был исключен из рассмотрения при проведении RFM-анализа. Было принято решение, что все клиенты, посетившие сайт за период, по которому формируется отчет – являются целевой аудиторией, и программа лояльности будет действовать на них только в рамках одного периода. Исключение параметра актуальности клиента высвобождает дополнительные вычислительные мощности, необходимые для проведения сегментации базы.

Алгоритм RFM:

1. Для каждого $z_i \in Z$:

- Определить $f_i = F(z)$ – частота совершения покупок по карте.
- Определить $m_i = M(z)$ – товарооборот создаваемый клиентом использующим карту.

2. Составить множество B вида

$$B = \{\{f_1; m_1\} \ldots \{f_n; m_n\}\}, \quad (2)$$

3. Разбить множество B на подмножества

$$S_{fm} \in B, \quad (3)$$

Так что:

$$z_i \in S_{fm}, \quad (4)$$

Если:

$$f_{\min f} \le f_i \le f_{\max f}, \quad (5)$$

$$m_{\min m} \le m_i \le m_{\max m}, \quad (8)$$

Где $[f_{\min}; f_{\max}]$ и $[m_{\min}; m_{\max}]$ – границы интервалов, задаваемые экспертом.

3.5 Кластеризация FM-сегментов

Для того чтобы объединить два подхода в сегментации клиентской базы данных, было принято решение о применении стандартного метода кластеризации. Предварительный этап – свертка товарных позиций в номенклатурные группы.

Особенность выбора метода кластеризации заключается в том, что априори не известно количество кластеров, на которое следует разбить выборку

FM-сегментов. Для проведения кластеризации был выбран метод агломеративной иерархической кластеризации.

Алгоритмы планарной кластеризации выполняют только разбиение исходного набора объектов на кластеры, каждый кластер которых представлен неструктурированным набором объектов. Алгоритмы иерархической кластеризации не страдают недостатком малой информативности - из-за дополнительных временных затрат на выходе таких алгоритмов весь набор объектов представлен в виде иерархии вложенных друг в друга кластеров, что может дать исследователю данных дополнительную полезную информацию о структуре данных. Также большинство алгоритмов иерархической кластеризации детерминированы, что в некоторых случаях имеет значение.

Алгоритмы иерархической кластеризации можно разделить на две большие группы:

1) Восходящие алгоритмы иерархической кластеризации;

2) Нисходящие алгоритмы кластеризации.

Первая группа алгоритмов действует по принципу первоначального объединения отдельных объектов в близкие пары объектов, затем пары объединяются с другими по тому же принципу и т. д., пока не появится только один кластер, содержащий все точки исходного набора объектов. Алгоритмы убывающей иерархической кластеризации работают по обратному принципу - сначала выбирается самый большой кластер, содержащий все объекты выборки, который затем постепенно делится на множество потомков кластера.

Объекты выборки представлены как кластерная иерархия, также называемая дендрограммой (рис. 9).

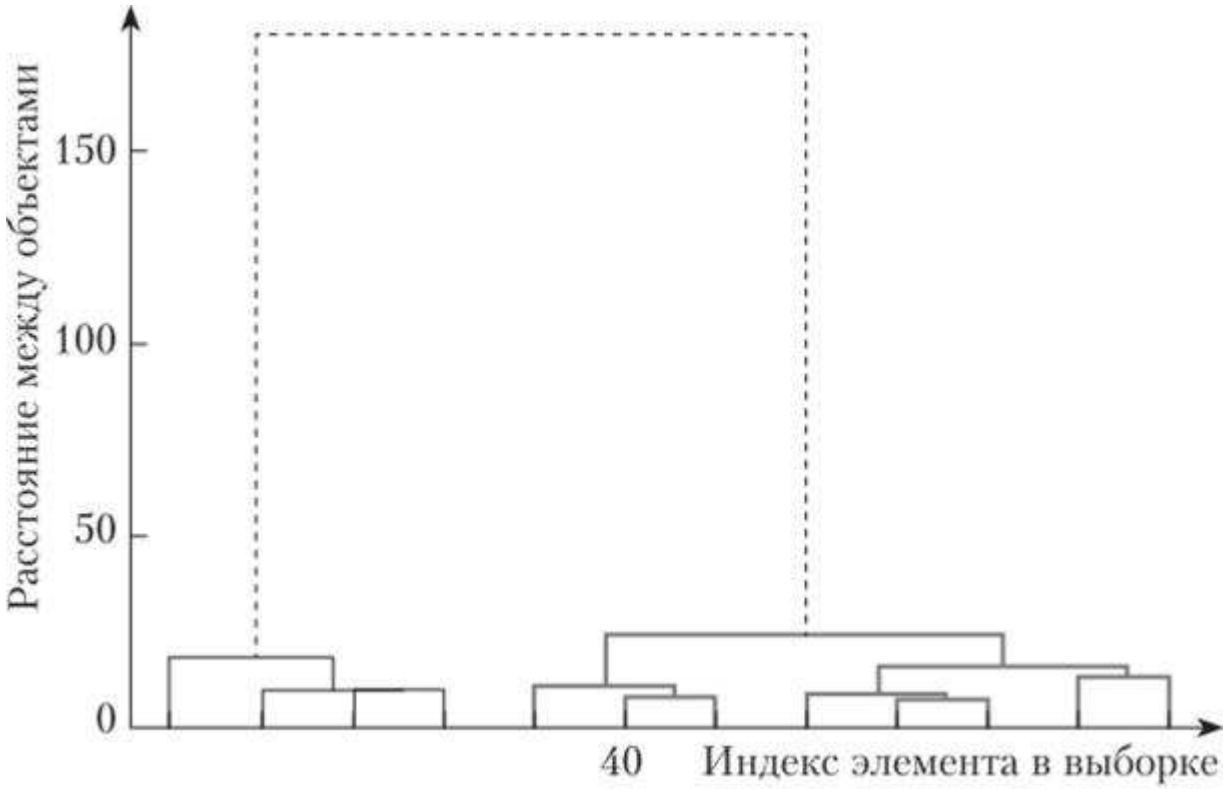


Рисунок 9 – пример дендрограммы

Как видно из рисунка 9, любое количество кластеров может быть построено для входной выборки - только объекты, поскольку кластеры содержатся на самом низком уровне. затем они присоединяются к ближайшим кластерам по одному, и если указать правило, из которого может быть сделан раздел дендрограммы, то есть нижние ветви могут быть "отрезаны", то такой алгоритм кластеризации сам сможет определить количество кластеров для минимизации целевой функции.

Дендрограмма обычно - дерево, построенное из матрицы измерений близости. Дендрограмма позволяет изобразить взаимные отношения между объектами из данного множества. Для создания дендрограммы требуется матрица подобия (или разности), определяющая уровень подобия между парами кластеров. Чаще используются агломерационные методы.

Для построения матрицы сходства (различия) необходимо задать меру расстояния между двумя кластерами. В данной работе был использован метод

Уорда. В отличие от других методов кластерного анализа, для оценки расстояний между кластерами здесь используются методы дисперсионного анализа. В качестве расстояния между кластерами берётся прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения (формула 9) $\Delta = \sum_i (x_i - \bar{x})^2 - \sum_{x_i \in A} (x_i - \bar{a})^2 - \sum_{x_i \in B} (x_i - \bar{b})^2$ На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению дисперсии. Этот метод применяется для задач с близко расположенными кластерами.

3.6 Выводы по третьему разделу

В данной главе были рассмотрены и проанализированы основные методы и способы сегментации клиентской базы данных в коммерческих предприятиях. Более подробно была рассмотрена методология RFM-анализа, позволяющая анализировать клиентскую базу данных по покупательскому поведению. Методика RFM-анализа идеально подошла к разрабатываемой на предприятии программе по увеличению лояльности клиентов. В данной главе представлен результат адаптации зарубежного опыта RFM-анализа в реалиях коммерческого предприятия на российском рынке.

Большая часть этой главы содержит информацию о том, как эффективно объединить два подхода в сегментации баз данных, а именно по покупательскому поведению, и по покупательской корзине. Эффективное совмещение данных подходов гарантирует огромный простор для дальнейшего развития целевой аудитории. Основной результат работы, представленный в третьей главе – решение отбросить параметр R и передать для дальнейшей иерархической кластеризации только индекс покупаемого пользователем товара и параметров F и M.

4 Программная реализация

Для реализации программы принято решение использовать Python версии 3, поскольку эта версия является актуальной на момент написания программы и поэтому эта версия имеет наилучшую поддержку и обладает наиболее активно развивающейся и обновляющейся базой библиотек. В качестве вспомогательного инструмента решено использовать Microsoft Office Excel.

Требования к системе: опытным путем установлено, что узким местом программы является оперативная память компьютера, на котором она исполняется. Минимальный объем оперативной памяти, необходимый, чтобы программа обработала данные и не зависла – 2Гб DDR3 с частотой 1333МГц.

Структура программы – модульный набор инструментов, а именно: основная функция, из которой предлагается провести подготовку данных, RFM-анализ и кластеризация по категориям с использованием данных RFM-анализа. Наглядно структура программы представлена на рисунке 10.

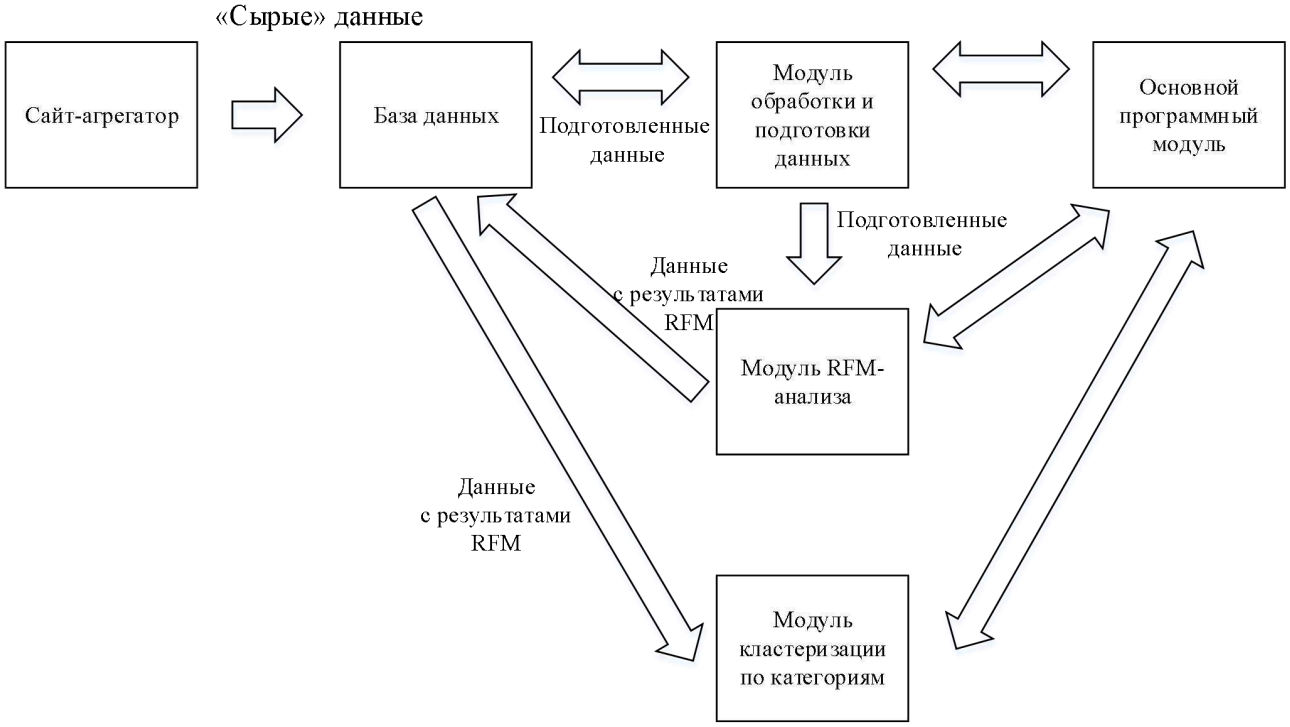


Рисунок 10 – структурная схема программы

Общий алгоритм работы программы: сайт-агрегатор хранит данные в свой базе. Из этой базы делается выгрузка «сырых» данных, которые, вызвав соответствующую команду из основной программы, модуль обработки и подготовки данных формирует в необходимый для дальнейшей работы вид и сохраняет рядом с БД. Эти подготовленные данные уже можно, вызвав соответствующую команду, подать либо в модуль RFM-анализа, либо напрямую в модуль кластеризации, если имеется такая цель. Модуль RFM-анализа в свою очередь проводит анализ данных и модифицирует результатами своей работы подготовленные данные, которые уже будет правильно запросить соответствующей командой из главного модуля подать в модуль кластеризации по категориям.

4.1 Модуль обработки и подготовки данных

Большинство алгоритмов кластеризации и анализа, находящихся в интернете в общем доступе, для своей работы требуют, чтобы данные, обрабатываемые ими, соответствовали определенным условиям. В соответствии с этим, решено в полуавтоматическом режиме производить адаптацию данных выгрузки базы данных в соответствии с этими требованиями:

- Файл переводится в формат .csv;
- Файл переводится в кодировку UTF-8.
- Из начала файла удаляется маркер последовательности байтов (BOM, byte order marker).
- Поля в строках разделяются запятыми: «,».
- Проверяется, что в качестве символа десятичной точки в числах должен использоваться символ `.` (точка), если это не так – исправляется.
- В полях, содержащих двойные кавычки `"`, все символы двойных кавычек задваиваются;

- Поля, содержащие запятую `,`, двойные кавычки `"` или переводы строки, заключаются в двойные кавычки `"`.

На рисунке 11 можно увидеть пример предоставленных данных, обработанный в соответствии с требованиями RFM-алгоритма и для наглядности открытый в программе «Блокнот»

```
order_date,user_id,order_value
2020-03-18,1765474,958.00
2020-03-18,1765474,958.00
2020-03-18,1765474,958.00
2020-03-18,2478,571.00
2020-03-18,2478,571.00
2020-03-18,1849413,6519.00
2020-03-18,1849410,1274.00
2020-03-18,1849410,1274.00
```

Рисунок 11 – пример предобработанных данных

Такой обработки достаточно для того, чтобы провести RFM-анализ, однако для проведения иерархической кластеризации необходимо перевести данные о категориях товаров из текстового в числовой формат.

Для этого принято решение произвести индексацию товаров по категориям и последующую нормализацию индексов. Для этого необходимо:

- Отделить категории друг от друга;
- Упорядочить по какому-либо признаку, в данном случае – по алфавиту;
- Пронумеровать базовые категории
- Если вложенная категория принадлежит к той же базовой категории, что и предыдущий, то его номер увеличивается, иначе нумерация начинается сначала;
- Если вложенная категория закончилась, она становится базовой, а вложенная в нее – новой вложенной;

- Повторять 2 предыдущих пункта, пока не останется вложенных категорий

В результате индексации мы получаем для каждой записи в выгрузке базы данных цифровую информацию о категории товара (рисунок 12), с которой может работать большее количество алгоритмов кластеризации, чем с текстовой.

101
101
102010101
102010102
102010102
102010103
102010103
102010201
102010202
102010202
102010203
102010301
102010301
102010301
102010301
102010301
102010301
1020201
1020201

Рисунок 12 – пример индексов товаров

user_id	название	цена	дата	путь категорий	название категории	Индекс товара
1765474	Кронштейн ,	958.00	18 03 20	/nazya/Все для Дома и Хобби/Дом и сад/Строительные, О	Кронштейны и полки для СВЧ	2030404020201
1765474	Кронштейн ,	958.00	18 03 20	/taobao/Мебель для дома на заказ/Кухонные шкафы и ак	Кронштейны и полки для СВЧ	319010504
1765474	Кронштейн ,	958.00	18 03 20	/market/Строительные, Отделочные материалы/Кухня/Ку	Кронштейны и полки для СВЧ	145020201
2478	Настольный	571.00	18 03 20	/taobao/Бытовая химия, средства гигиены, приспособлен	Настольные ящики для канцелярии	3050101020804
2478	Настольный	571.00	18 03 20	/market/Все для Дома и Хобби/Ящики для хранения прин	Настольные ящики для канцелярии	111130404
1849413	бумажник А	6519.00	18 03 20	/taobao/Аксессуары, галантерея/Сумки/Бумажники/Коше	Бумажники/Кошельки	302120101
1849410	Наколенник	1274.00	18 03 20	/nazya/Спорт/Наколенники, Напульсники	Наколенники, Напульсники	21513

Рисунок 13 – сводная таблица с данными о пользователе, категории и ее индексе

4.2 Модуль RFM-анализа

Модуль RFM-анализа не подразумевает модернизацию. Ниже представлен алгоритм проведение RFM-анализа.

Для анализа и интерпретации RFM групп была использована выгрузка из БД за предоставленный владельцем сайта интересующий его период.

Описание реализации алгоритма RFM-анализа:

1. Построчно читаем исходный файл с информацией о заказах:

    1. Пропускаем строки, в которых не удалось определить дату заказа.

    2. Для каждого идентификатора пользователя (`user_id`) собираем словарь (dictionary), в котором ключами являются даты заказов, а значениями — суммы заказов данного пользователя за данную дату. В дальнейшем считаем все заказы, сделанные в один и тот де день, одним заказом на общую сумму за день.

    3. Если из заказов, сделанных пользователем за день, нет ни одного заказа с ненулевой и непустой суммой, то сумма заказов за этот день считается равной `None`. Это делается для того, чтобы такие дни не учитывать при расчете средней суммы заказа (monetary), но учитывать при расчете частоты заказов (frequency) и давности последнего заказа (recency).

    4. Попутно запоминаем для каждого пользователя дату последнего заказа, а также дату последнего заказа в исходном файле без учета пользователя — эта последняя дата используется как _сегодняшняя_, т.е. как начальная точка отсчета по времени.

    5. Для каждого `user_id` также запоминаем номера сегментов пользовательских (не-RFM) измерений, если они присутствуют в исходном файле с заказами и соответствующим образом перечислены в конфигурационном файле.

2. На временном периоде от сегодняшней даты до `look_back_period` дней назад рассчитываем номера сегментов по измерениям «Recency» (давность последнего заказа), «Frequency» (частота заказов) и «Monetary» (средняя сумма заказа):

1. Сначала для каждого пользователя расчитываем абсолютные значения метрик по измерениям:

- Количество заказов за период;
- Количество дней, прошедших со дня последнего заказа до сегодняшней даты;
- Среднюю сумму заказа без учета нулевых заказов и заказов, для которых в исходном файле не указана сумма.

2. Для пользователей, у которых нет заказов за период, но есть более старые заказы, устанавливаем все метрики в значение `'stale'` — «протухшие». Информацию о пользователях, у которых нет заказов за период, но есть более _новые_ заказы, удаляем из массива пользователей. На первом проходе этого не происходит, поскольку последняя дата периода совпадает с датой последнего заказа в исходном файле, но на втором проходе — при расчете ценности сегментов (см. п. 4 ниже) — это позволяет рассчитать среднюю сумму заказов без учета пользователей, у которых все заказы оказались будущими по отношению к рассматриваемому периоду.

3. Сегментируем пользователей по каждому измерению:

1. Пользователей со значением соответствующей метрики, равным `'stale'`, относим к нулевому сегменту.

2. Пользователей со значением соответствующей метрики, равным `None`, относим к первому сегменту.

3. Остальных пользователей упорядочиваем в порядке возрастания соответствующей метрики.

4. Считаем минимальное количество пользователей в первом сегменте: делим общее количество упорядоченных пользователей на `segments_count` для данного измерения.

5. Проходим массив упорядоченных пользователей от начала и помечаем пользователей, как относящихся к первому сегменту до тех пор, пока не будут удовлетворедны два условия:

- Количество пользователей будет не меньше минимального;
- Значение метрики очередного пользователя будет отличаться от значения метрики предыдущего пользователя.

6. Запоминаем значение метрики, при котором «закончились» пользователи, относящиеся к первому сегменту.

7. Для следующих сегментов выполняем пункты 2.3.4–2.3.4 на оставшихся пользователях.

3. Сохраняем в файл рассчитанные значения номеров сегментов для каждого `user_id`. Помимо расчитанных сегментов сохраняем также номера сегментов пользовательских (не-RFM) измерений, если они присутствовали в исходном файле с заказами и были соответствующим образом перечислены в конфигурационном файле.

4. Выполняем п. 2, приняв за точку отсчета дату, отстоящую на `prediction_period` дней назад от сегодняшней (см. п. 1.iv), т.е. проводим сегментацию пользователей по RFM-измерениям на более старом временном периоде от `prediction_period` дней назад до `prediction_period + look_back_period` дней назад. При этом используем границы сегментов по каждому измерению, рассчитанные на предыдущем проходе, т.е. на периоде от сегодняшней даты (см. п. 1.iv) до `look_back_period` дней назад.

5. Подсчитываем среднюю сумму заказов для всех пользователей на периоде от сегодняшней даты до `prediction_period` дней назад.

6. Подсчитываем среднюю сумму заказов для пользователей каждого сегмента по каждому измерению отдельно на периоде от сегодняшней даты до `prediction_period` дней назад. Расчет выполняется как для RFM-измерений, так и для пользовательских измерений.

7. Для каждого сегмента каждого измерения подсчитываем относительную ценность сегмента — отношение средней суммы заказов пользователей в сегменте к средней сумме заказов для всех пользователей.

8. Перебираем все возможные комбинации сегментов по всем измерениям и определяем ценность микросегментов, перемножая относительные ценности сегментов, определенные в п. 7. Под микросегментами подразумеваем пересечения сегментов разных измерений.

9. Сохраняем в файл информацию об относительно ценности каждого микросегмента. Пример файла со значениями относительной ценности каждого сегмента:

`recency,frequency,monetary,bid ratio`

`1,1,1,0.9`

`2,1,1,1.2`

`3,1,1,1.1`

Наибольшую ценность для дальнейшей работы представляют данные, хранящиеся в файле _mapping, поскольку это данные о пользователях и их принадлежности к конкретному сегменту RFM.

4.3 Иерархическая кластеризация

С точки зрения данной работы данный элемент является «черным ящиком», поскольку программа специально так структурирована, что можно взять любой модуль алгоритма иерархической кластеризации. В данном случае отдано предпочтение модулю, в котором в качестве инструмента измерения расстояний между кластерами используется метод Уорда как

зарекомендовавший себя наилучшим образом. На рисунке 14 можно увидеть результат кластеризации с использованием данных о FM-принадлежности пользователя и приобретаемому им товару

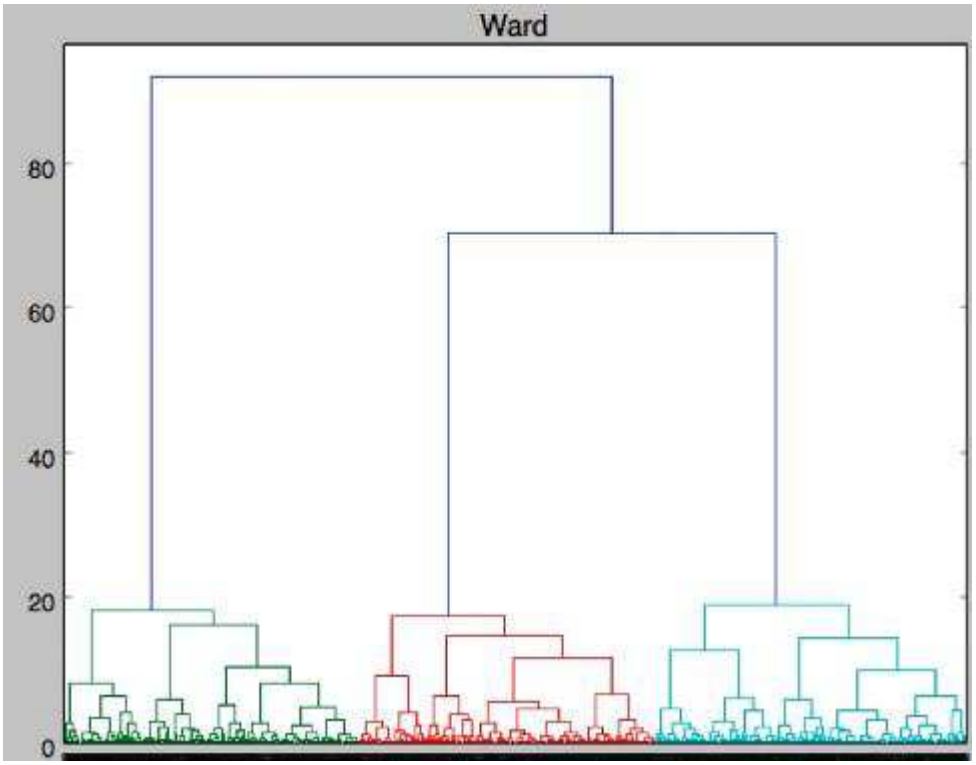


Рисунок 14 – кластеризация пользователей

На рисунке 15 можно увидеть файловую структуру программы, придерживаясь которой достигается взаимозаменяемость модулей

Рисунок 15 – файловая структура программы

4.4 Выводы по четвертому разделу

В данной главе были реализованы в программе приемы методы и технологии, описанные в третьей главе, а именно, был применен RFM-анализ, адаптация FM-сегментов, свертка по номенклатурным группам, кластеризация FM-сегментов.

Разработанный программный продукт позволяет производить подготовку выгруженных данных и приведение их к виду, необходимому для дальнейшей работы модулей, в частности модуля RFM-анализа и модуля иерархической кластеризации. Кроме того, программа позволяет выполнять RFM-анализ и производить кластеризацию подготовленных ею же данных; кроме того, модуль кластеризации обладает средством визуализации результата своей работы.

В данной главе представлен практический результат адаптации зарубежного опыта RFM-анализа в реалиях коммерческого предприятия на российском рынке. Проблема заключалась в том, что зарубежный опыт не полностью отражает особенности покупательского поведения, и профиль зарубежного покупателя не совпадает с профилем отечественного потребителя.

Основной результат, представленный в данной главе – программа, формирующая после выполнения всех ее шагов дендрограмму, представляющая собой объединение результата FM-сегментации базы данных сайта Nazya.com и кластеризации пользователей по товарным предпочтениям. Данная дендрограмма представляет огромный интерес для маркетологов, так как она отображает множество новых знаний о клиентах.

ЗАКЛЮЧЕНИЕ

В ходе исследований в рамках магистерской диссертации был разработан модульный набор инструментов сегментации клиентской базы данных, основанный на стандартных методах кластеризации и объединяющий два подхода в сегментации клиентских баз данных – по покупательскому поведению и по товарным предпочтениям, а так же реализован набор вспомогательных инструментов для его реализации. Данный набор инструментов позволяет производить обработку данных, характерных для сайтов-агрегаторов. Программа позволяет использовать различные программные модули в зависимости от предпочтений маркетолога. Разработанная программа является гибкой, и может быть подстроена под реалии практически любого предприятия, поскольку позволяет заменять алгоритмы на любые другие, которые со значительной долей вероятности будут работать без каких-либо тонких настоек программистом.

Программа повышения лояльности оказывает огромное влияние на развитие сетей розничной торговли. Правильно построенная программа может в корне изменить ситуацию в конкурентной борьбе. Чем лучше будет проведена сегментация базы данных, тем более высокий отклик будут проявлять клиенты при получении предложения о покупке товара.

Стоит отметить особую роль программы увеличения лояльности в отслеживании динамики сегментов в базе данных. Дальнейшее развитие программы напрямую связано с наблюдением за развитием этих сегментов. Программа повышения лояльности не является чем-то постоянным, она должна постоянно совершенствоваться, ориентируясь на отклик покупателей, мировые тенденции сегментации баз данных, развитие технологий.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – Москва: Финансы и статистика, 1989. – 232 с.

2. Гиг Дж., Ван. Прикладная общая теория систем. / Дж. ван Гиг. – Москва: Мир, 1981. – 733 с.

3. Журавлев, Ю. И. Распознавание. Математические методы. Программная система. Практические применения. / Ю. И. Журавлев, В. В. Рязанов, О. В. Сенько. – Москва: Фазис, 2006. – 159 с.

4. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний. / Н. Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

5. Киселев, М. Средства добычи знаний в бизнесе и финансах. / М. Киселев, Е. Соломатин // Открытые системы. – 1997. – № 4. – С. 41–44.

6. Кречетов, Н. Продукты для интеллектуального анализа данных. / Н. Кречетов // Рынок программных средств. – 1997. – № 14–15. – С. 32–39.

7. Мандель, И. Д. Кластерный анализ. / И. Д. Мандель. – Москва: Финансы и статистика, 1988. – 176 с.

8. Олдендерфер, М. С. Кластерный анализ / Факторный, дискриминантный и кластерный анализ: пер. с англ. / М. С. Олдендерфер, Р. К. Блэшфилд; под. ред. И. С. Енюкова. – Москва: Финансы и статистика, 1989. – 215 с.

9. Паклин, Н. Б. Бизнес-аналитика: от данных к знаниям. / Н. Б. Паклин, В. И. Орешков. – Санкт-Петербург: Питер, 2009. – 706 с.

10. Хайдуков, Д. С. Применение кластерного анализа в государственном управлении. Философия математики: актуальные проблемы. / Д. С. Хайдуков. – Москва: МАКС Пресс, 2009. – 287 с.

11. Шуметов, В. Г. Кластерный анализ: подход с применением ЭВМ. В. Г. Шуметов, Л. В. Шуметова. – Орел: ОрелГТУ, 2000. – 118 с.

12. LibreOffice Draw Википедия [Электронный ресурс] : - Режим доступа: https://ru-wiki.ru/wiki/LibreOffice_Draw

Федеральное государственное автономное
образовательное учреждение
высшего профессионального образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Базовая кафедра «Интеллектуальные системы управления»

УТВЕРЖДАЮ
Заведующий кафедрой
_____
Ю.Ю.Якунин
«25» июня 2020г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

«Кластеризация данных для сайта-агрегатора»

09.04.04 – «Программная инженерия»
09.04.04.02 – «Технологии индустриального производства программного обеспечения интеллектуальных систем управления»

Руководитель	25.06.20	Доцент, канд. техн. наук	А.А. Даничев
	подпись, дата	должность, ученая степень	инициалы, фамилия
Выпускник	_____		Д.Н. Галин
	подпись, дата		инициалы, фамилия
Рецензент	_____	Канд. физ. - мат. наук	А.Л. Двинский
	подпись, дата	должность, ученая степень	инициалы, фамилия

Красноярск 2020