Федеральное государственное автономное образовательное
учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра «Информационные системы»

УТВЕРЖДАЮ
Зав. кафедрой ИС
_____ С.А. Виденин
«___» _____ 2016 г.

БАКАЛАВРСКАЯ РАБОТА

09.03.01 Информатика и вычислительная техника

Применение методов интеллектуального анализа данных в сфере малого предпринимательства

Руководитель _____ П.А. Светашков
подпись, дата — инициалы, фамилия

Выпускник _____ В.А. Черноусов
подпись, дата — инициалы, фамилия

Нормоконтролер _____ Л.С. Троценко
подпись, дата — инициалы, фамилия

Красноярск 2016

РЕФЕРАТ

Выпускная квалификационная работа по теме «Применение методов интеллектуального анализа данных в сфере малого предпринимательства» содержит 63 страницы текстового документа, 1 приложение, 11 использованных источников, 18 листов графического материала.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, DATA MINING, АССОЦИАТИВНЫЕ ПРАВИЛА, БИЗНЕС-АНАЛИТИКА, ПРОГРАММЫЙ ПРОДУКТ, ПО, АВТОМАТИЗАЦИЯ, БАЗА ДАННЫХ, СУБД, SQLITE

Объектом исследования является сфера малого предпринимательства.

Предметом исследования является база данных фирмы.

Целью данной работы является обеспечение информационной поддержки представителей малого предпринимательства путём предоставления им упрощённого доступа к технологии интеллектуального анализа данных.

Задачи:

- разработка ИС, содержащей в себе базу данных транзакций фирмы и предоставляющей пользователю инструментарий для работы с этой базой;

- разработка ИС, содержащей в себе средства интеллектуального анализа данных.

Актуальность работы

Сфера предпринимательства сейчас решительнее, чем когда-либо, настроена на повышение конкурентоспособности своих активов с помощью современных технологий, а интеллектуальный анализ данных представляет большую ценность для руководителей и аналитиков в их повседневной деятельности.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

За последние пару десятилетий мы стали свидетелями повсеместного распространения новейших информационных технологий и их внедрения в различные области социальной и экономической сфер. Всё это вкупе с наличием свободного доступа к сети Интернет привело к огромному росту объёма информации, производимой человеком.

Только за пару дней человечество сегодня производит столько информации, сколько было создано за всё время существования Земли вплоть до нынешнего тысячелетия. По утверждениям исследовательских компаний, сейчас объём данных на планете удваивается каждые пару лет. И такая тенденция сохранится как минимум до 2020 года.

Из-за огромного количества существующей информации, лишь малая её часть будет увидена человеческим глазом, и, как следствие, значительные объёмы практичной и полезной информации просто пройдут мимо. Это приводит нас к пониманию важности проблем, сопряжённых с анализом накопленных данных для извлечения новых знаний.

Наша единственная возможность понять и найти что-то полезное в этом огромном океане – это широкое применение методов интеллектуального анализа данных.

Интеллектуальный анализ данных широко используется во многих областях с большим объёмом данных. В науке – биологии, медицине, астрономии и так далее. В бизнесе – в первую очередь в торговле, телекоммуникации, промышленном производстве и других. В рамках данной работы я ограничусь сферой малого предпринимательства, или, по-другому, малого бизнеса.

Сейчас идёт много разговоров о поддержке и содействии малому бизнесу в России, и я со своей стороны предпринял попытку внести в это дело посильный вклад соответственно своим возможностям и сложившейся в отрасли ситуации.

В настоящее время технология интеллектуального анализа данных представлена целым рядом коммерческих и свободно распространяемых продуктов. Изучая предметную область, я пришёл к выводу, что все они имеют помимо прочего одни общий недостаток – они требуют наличия специальных знаний и навыков для работы с ними. Крупные фирмы и предприятия решают эту проблему нанимая квалифицированных специалистов или обучая своих. В то время как представители малого предпринимательства, зачастую, не располагают необходимыми для этого ресурсами.

По итогу всего вышесказанного передо мной была поставлена следующая цель: обеспечение информационной поддержки представителей малого предпринимательства путём предоставления им упрощённого доступа к технологии интеллектуального анализа данных.

Для достижения поставленной цели были выдвинуты следующие задачи:

- создание свободно распространяемого программного продукта, не требующего специальных навыков для работы и максимально понятного;

- предоставление пользователю инструментария для работы с базой данных;

- применение методов интеллектуального анализа данных к этой базе данных, с целью анализа накопленной информации для извлечения новых и потенциально полезных знаний.

1 Обзор современного состояния сферы интеллектуального анализа данных

1.1 Обзор научной литературы

При написании данной работы были использованы научная и учебно-методическая литература, статьи на электронных ресурсах сети Интернет.

Основным источником послужило научное пособие «Бизнес-аналитика: от данных к знаниям» Николая Борисовича Палкина. В этом учебном пособии автор перечислил основные технологии, используемые сегодня в сфере бизнес-аналитики. Привёл наглядные примеры способов реализации данных технологий, воплощения их в жизнь и внедрения на производстве. Сама сфера бизнес аналитики находится на пересечении таких разных областей науки как информационные технологии, экономика и статистика. Их объединение позволило создать мощные инструменты и информационно-аналитические системы для современного бизнеса, о которых и идёт речь в учебном пособии. Книга знакомит читателя с областью бизнес-аналитики постепенно, всё дальше погружаясь в основные проблемы и объясняя пути их решения. Книга будет полезна в качестве учебного пособия как студентам информационно-технических специальностей, так и студентам экономических направлений и программ обучения. Также она будет полезна уже действительным бизнес-аналитикам, специалистам по анализу данных и профессионалам, связанным с этой областью для изучения вопросов внедрения систем бизнес-аналитики. Также книга пригодится всем тем, кто стремится больше узнать о проблемах интеллектуального анализа данных и бизнес-аналитики.

Отличным источником послужила монография «Методология информационной аналитики» от Алексея Курлова и Вячеслава Петрова. Аналитическая работа освещена в этой книге со всех возможных сторон. Поднимается широкий круг проблем, связанных с аналитической работой в целом, её организацией, методологией и информационно-технологическим

оснащением современных аналитиков. Часть книги авторы знакомят читателя со способами и методами эффективной аналитической деятельности (как в повседневной жизни, так и применимо к профессиональной деятельности), методам мыслительной работы в целом. Другую часть книги авторы посвятили освещению технологической стороны современной аналитики, наглядно разобрали вопросы, связанные с проектированием и разработкой автоматизированных информационных систем аналитической работы, их функционированием и эффективным использованием. Авторы раскрывают суть понятия интеллектуального анализа данных. Показывают, как в аналитической деятельности переплелись и взаимосвязаны такие различные области науки, как логика, математика, философия, сфера информационных технологий и даже психология и др. Разобраны различия русской и зарубежных аналитических школ, их история и формирование. Авторы раскрывают всю значимость аналитики для всех сфер нынешней человеческой деятельности, таких как социальная сфера, экономическая сфера, политическая сфера. Показаны возможности использования аналитики в этих сферах для повышения эффективности принятия управленческих решений, прогнозирования на основе внутренних и внешних данных. Книга будет интересна и полезна в основном специалистам, работающим в области разработки информационно-аналитических систем, систем поддержки принятия решений. Может использоваться в качестве научного пособия для аспирантов и студентов информационно-технических направлений и специальностей.

Ещё одним полезным пособием послужила книга Сергея Нестерова «Базы данных. Интеллектуальный анализ данных». В этом учебном пособии подробно излагаются способы интеллектуального анализа данных с помощью систем управления базами данных. Рассматриваются суть и значение интеллектуального анализа данных, приводится последовательное описание основных алгоритмов и их применение на практике. Рассмотрены способы создания и проектирования баз данных и последующего их анализа как с помощью уже существующих программных решений, так и с использованием

основных алгоритмов аналитики данных. Подробно освещены проблемы извлечения информации из баз данных и других источников. Автор учебного пособия подробно и понятно иллюстрирует каждое решение наглядными практическими примерами, разъясняющими суть применяемых методов и технологий. Книга написана простым и понятным языком и даже у начинающих изучать области баз данных и интеллектуального анализа не возникнет проблем с пониманием изложенного материала. Книга будет полезна для аспирантов и студентов информационно-технических направлений и специальностей.

Если говорить о состоянии литературы по интеллектуальному анализу данных вообще, я могу сказать, что при работе я ощутил нехватку современной актуальной литературы на русском языке.

Большая часть литературы по Data Mining на русском языке была написана в нулевых годах, и в основном это были оригинальные научные работы, а не перевод зарубежных книг, но, к сожалению, они уже потеряли свою актуальность. Если говорить о современной, не старше пяти лет литературе, что находится в свободном доступе и может быть бесплатно загружена или приобретена в магазинах и онлайн-сервисах – на русском языке она представлена в основном переводами зарубежных и научно-популярных книг, дающих в основном общий обзор предметной области, и не может быть использована в качестве серьёзного учебного пособия.

На сегодняшний день научная литература по Data Mining представлена в основном на английском языке. Это связано с тем, что английский язык де-факто стал языком современной информационной науки, т.к. в сфере информационных технологий большинство разработок, статей, научной литературы и технической документации ведутся на английском языке.

Что касается возможности обучения методам интеллектуального анализа данных на русском языке – с развитием сети Интернет появилось множество онлайн-курсов от крупных отечественных и зарубежных компаний, ведущих подготовку специалистов как своих потенциальных работников. К ним относятся «Школа анализа данных Яндекса», «Школа анализа данных Билайн»,

«Технопарк Mail.ru Group». Крупные вузы нашей страны также проводят подготовку бакалавров, магистров и специалистов по программам интеллектуального анализа больших данных. В частности, Московский государственный университет, Высшая школа экономики, Санкт-Петербургский государственный университет информационных технологий, механики и оптики, Московский физико-технический институт и Санкт-Петербургский государственный университет на сегодняшний день имеют свои образовательные программы.

1.2 Обзор предметной области

Обзор темы

За последние десятилетия человечество накопило огромные объёмы информации в цифровой форме. Всевозможные данные органов государственного управления, промышленных предприятий, крупных и малых фирм. Очевидно, что среди этого моря информации содержится значительный скрытый потенциал знаний, обладая которым можно повысить эффективность своей деятельности, какой бы она ни была. Сегодня вся деятельность любого предприятия фиксируется, записывается и хранится в архивах и электронных хранилищах. Разумеется, что такие данные без необходимого анализа и переработки представляют собой просто бесполезную свалку. Это и приводит к тому, что задача извлечения скрытых знаний из накопленных данных является на сегодняшний день остро необходимой.

К процессу переработки сырых данных в полезные знания предъявляются особые требования:

- обрабатываемые данные могут быть сколь угодно большого объёма;
- данные по сути своей не однородны;
- обнаруженные знания должны быть ясны и недвусмысленны;
- средства анализа данных должны быть просты и понятны в использовании.

Долгое время именно математическая статистика была основным инструментом аналитиков. Однако сегодня, перед лицом новых проблем, она справляется уже не так убедительно. Основная причина этого в том, что в математической статистике принята концепция усреднения по выборке, что заставляет аналитиков оперировать несуществующими значениями типа средней зарплаты дворника и олигарха. Это не означает, что математическая статистика слабый инструмент, но каждый инструмент должен использоваться для своей работы. Не нужно забивать гвозди микроскопом. Математическая статистика полезна для анализа уже высказанных гипотез и для разведочного анализа, который является основой технологии OLAP (аналитическая обработка в реальном времени).

Сегодня для решения актуальных проблем появляются другие методы и технологии анализа данных, получившие название Data Mining, или, что является наиболее общепринятым переводом на русский язык, интеллектуальный анализ данных. Эти методы направлены на те задачи, в которых сегодня больше всего нуждаются – анализ больших объёмов данных и поиск скрытых знаний внутри этих объёмов. Уровни знаний и применимые к ним аналитические инструменты приведены в таблице 1.

Таблица 1 – Уровни применения аналитических инструментов

Уровень знаний, извлекаемый из данных	Аналитические инструменты
Поверхностный	Язык простых запросов
Неглубокий	Оперативная аналитическая обработка (OLAP)
Скрытый	Интеллектуальный анализ данных (Data Mining)

Главная особенность интеллектуального анализа данных – нетривиальность обнаруживаемых знаний. Это значит, что добытые знания должны содержать в себе неочевидные на первый взгляд связи, составляющие так называемые скрытые знания. Сейчас большинству людей стало очевидно,

что глубоко под поверхностью накопленных данных находятся скрытые знания, которые могут быть извлечены при должной раскопке.

Исходя из приведённой информации, при использовании методов и технологий интеллектуального анализа данных появляется действительная возможность открывать неочевидные закономерности между накопленными данными и использовать полученные знания в системах принятия решений, помогая в решении управленческих задач где бы то ни было: на предприятиях, в бизнесе, муниципальных учреждениях.

Сегодня технологии и методы Data Mining используются практически везде, где собрана какая-либо информация. Data Mining может применятся везде, где есть данные.

Наибольшее применение и распространение интеллектуальный анализ данных получил в области бизнес-аналитики. Причина в том, что грамотный анализ данных фирмы позволяет не только повысить эффективность производства, но и значительно повысить его прибыльность, что делает системы аналитики быстро окупаемыми.

Широкое применение технологии интеллектуального анализа данных получили в области розничной торговли и маркетинга. Основные приёмы Data Mining, вроде кластерного анализа и поиска ассоциативных правил, позволяют здорово помочь при решении основных ежедневных проблем, вроде закупки и расходования товаров, их размещения в торговом зале. Они также позволяют повысить лояльность клиентов и увеличить доход от них, анализируя их активность.

Рынок услуг и технологий интеллектуального анализа данных в России

По сравнению с мировым рынком услуг и технологий Data Mining российский рынок выглядит ничтожно малым. Это выражается в основном в объёмах оборота мирового и нашего рынков. По исследованиям агентства Wikibon, оборот мирового рынка интеллектуального анализа данных за 2015 год составил $33,3 млрд. Для сравнения, компания IDC оценила российский рынок

всего в $340 млн. Но зато отечественный рынок имеет огромный потенциал для роста, чем и пользуется – по данным той же IDC его рост за 2015 составил 40% по сравнению с предыдущим годом.

Традиционный бизнес обратил свой взгляд на возможности анализа своих данных. В технологиях Data Mining в первую очередь заинтересованы представители рынков с высоким уровнем конкуренции, т.к. они остро нуждаются в новых способах повышения своей конкурентоспособности. В 2015 году агентство CNews Analytics провело опрос среди 108 компаний. Чуть меньше половины из них уже начали использование интеллектуальной аналитики. Основными пользователями этих технологий являются банки (пятая часть опрошенных) и телеком-операторы (одна десятая). Также технологии Data Mining получили широкое распространение в сфере онлайн-рекламы и розничной торговли.

По информации из открытых источников, системы интеллектуального анализа данных уже используются в таких компаниях как «М.Видео», Ozon, «Роснефть», X5 Retail Group, «ВТБ24», «Сбербанк», «Транснефть», «ОТП Банке», «Всероссийском банке развития регионов» и «Уральском банке реконструкции и развития», «Райффайзенбанке», сети гипермаркетов «Лента», банке «Уралсиб», «Альфа-Банке», «Азбуке вкуса» и др.

Сравнительно слабо технологии интеллектуального анализа данных используются в государственном секторе. Что удивительно, ведь именно здесь они могли бы значительно увеличить эффективность и продуктивность госслужб. Однако отдельные структуры всё же используют продукты Data Mining в своей работе, среди них ФНС, ФГУ аналитический центр при Правительстве РФ, Пенсионный фонд Российской федерации, ФСБ, ФОМС, СК и Служба внешней разведки. К сожалению, в сфере медицины внедрений систем аналитики нет и пока не ожидается, несмотря на высочайшую необходимость и огромный потенциал.

Проведём небольшой обзор, дающий общее представление о рынке обработки данных в России. Для удобства разделим участников рынка на категории, хоть и весьма условные:

- поставщики (SAP, Oracle, IBM, EMC, Microsoft и др.);
- разработчики алгоритмов (датамайнеры) (Yandex Data Factory, «Алгомост», Glowbyte Consulting, CleverData и др.);
- интеграторы («Форс», «Крок» и др.);
- потребители (телеком, банки, ритейл и др.);
- отдельные разработчики.

Поставщики занимаются продажей профессиональных СУБД, информационно-аналитических систем и аппаратных комплексов – самостоятельно или через своих агентов. Для того чтобы внятно разобраться в этих продуктах и использовать их с максимальной для себя выгодой, фирмы должны обладать собственными отделами экспертов и аналитиков. Это под силу только крупным компаниям, компании поменьше предпочитают пользоваться услугами интеграторов и ИТ-консультантов, которые подбирают программное и аппаратное обеспечение индивидуально под нужды заказчика.

Крупным поставщиком на рынке является немецкая компания SAP. На мировом рынке систем аналитики она успешно процветает с 2007 года. В России продукты этой компании приобретают как государственные структуры (ФНС и ПФ РФ), так и частные предприятия («Открытие», «Сибирская генерирующая компания»).

Как и немецкая SAP, американская Oracle предоставляет широкий ассортимент технологий для анализа и обработки данных. Компания достаточно известна во всём мире. Её клиентами являются, например, ФНС и «Альфа-Банк».

Продуктами и услугами мирового гиганта IBM пользуются «Вымпелком» и ПФ РФ. По данным агентства Wikibon, IBM с недавних пор является лидером по заработку на технологиях работы с данными.

Компания Microsoft предлагает технологии интеллектуального анализа данных для любого масштаба бизнеса, вследствие чего получила широкое

распространение как среди крупных игроков на бизнес-арене, так и среди небольших.

Компания SAS, одна из первых начавшая работу в сфере систем бизнес-аналитики, сейчас работает с такими клиентами как «Теле2», «Российские Железные Дороги», «Сбербанк». Помимо продажи своих систем, SAS также предоставляет клиентам помощь в обучении работе со своими продуктами.

Google вышла на рынок анализа данных для бизнеса совсем недавно – только в 2012 году. Она заняла относительно свободную нишу работы со средними и мелкими предприятиями, вроде «Связной», «Эльдорадо», «М.Видео». Информации о сотрудничестве Google с российскими государственными структурами в сфере аналитики обнаружено не было.

Датамайнеры (или разработчики алгоритмов), сами занимаются обнаружением знаний в данных клиента. Некоторые их сервисы позволяют пользователям просто загрузить исходные данные в облако и на выходе получить полезные знания. Главным преимуществом такого подхода является то, что нет необходимости приобретать инфраструктуру и нанимать дорогой персонал для работы с ней. Крупные компании в основном пользуются своей инфраструктурой и специалистами, в то время как услуги датамайнеров пользуются спросом в основном среди представителей малого предпринимательства.

Крупнейшим игроком среди датамайнеров на отечественном рынке является «Яндекс». Значительная часть продуктов самого «Яндекса» работает на технологиях Data Mining – такие как поиск, машинный перевод, почта, контекстная реклама, сервис пробок. Предоставлять свои технологии для сторонних заказчиков «Яндекс» начали практически в одно время с Google – в 2012 году. За прошедшие несколько лет Яндекс успели обзавестись крупными клиентами как в нефтедобывающей отрасли («Роснефть»), так и среди других отраслей. Не так давно «Яндекс» открыли международный отдел для работы с представителями крупного бизнеса. В основном они сотрудничают с банками, фирмами розничной торговли, промышленными предприятиями. Также

«Яндекс» успел поработать со «Сбербанком» и AstraZeneca. Для аналитики «Яндекс» использует собственные решения и разработки.

Среди игроков поменьше можно выделить такие отечественные фирмы, как «Алгомост», IBS, «Прогноз», AT Consulting, CleverData, EasyData, Double Data, DataMining Labs, MLClass, BaseGroup Labs, Global Innovation Labs и «Айкумен ИБС». Большинство компаний достаточно молоды, но все они уже прочно закрепились на российском рынке и постепенно пробиваются за рубеж.

Интеграторы и ИТ-консультанты помогают клиентам внедрить у себя систему аналитики. Они являются связующим звеном между бизнесом и технологиями. Их услугами пользуются те, кого не устраивают готовые решения поставщиков и услуги датамайнеров. Интеграторы занимаются в основном тем, что комбинируют системы и их части от различных поставщиков в единое целое, что позволяет удовлетворить все потребности заказчика относительно нужного ему функционала.

На отечественном рынке наиболее успешными интеграторами являются «Форс» и «Крок». Обе они относительно молоды, и начали свою деятельность в сфере аналитики больших данных в 2013 году.

Компания «Форс» сотрудничает с банками, сетями розничной торговли, телеком-операторами, госсектором. Кроме развёртывания решений сторонних поставщиков, они также занимаются собственными разработками. «Форс» в основном работают с продуктами компании Oracle.

«Крок» уже успел поработать с продуктами таких гигантов как с EMC, HP, Oracle и Microsoft, Intel. Услугами «Крок» пользовались банки, телеком-операторы и сектор здравоохранения.

Нужно также пару слов сказать о готовых сервисах на основе больших данных. На технологиях Data Mining работают такие привычные нам сервисы как антиспам-фильтр, контекстная реклама, антифрод. Данные для своей работы эти системы берут из свободных источников – социальных сетей, масс-медиа, интернет форумов. Это позволяет исключить расходы на инфраструктуру. Крупнейшие Российские ИТ-компании анализируют эти огромные объёмы

данных самостоятельно. Это позволяет им развивать свои сервисы, повышать эффективность рекламы и оптимизировать контент.

Mail.Ru Group применяли технологии Data Mining ещё тогда, когда даже сам термин не вошёл в обиход. Система веб-аналитики «Рейтинг Mail.Ru» была одним из их первых подобных сервисов. Сегодня практически все их сервисы в той или иной степени занимаются интеллектуальным анализом больших объёмов данных: поисковая машина, почта, таргетинговая реклама и др. Используя технологии Data Mining они избавляются от спама в своей почте, таргетируют рекламу, повышают эффективность своей поисковой машины, собирают и обрабатывают информацию об активности пользователей. У них также есть свои собственные разработки, например – NoSQL СУБД Tarantool.

«Рамблер» всегда использовал аналитику для улучшения своей поисковой машины, но в последние годы особенно активно стал применять интеллектуальный анализ. «Рамблер» использует методы Data Mining для тех же целей, что и Mail.Ru: оптимизация поиска, рекламы, фильтрация спама. Также компания недавно приобрела фирму, производящую продукты для интеллектуальной аналитики текстов (Text Mining), которыми пользуются крупные отечественные компании вроде Центробанка, Федеральной службы безопасности, «Газпрома» и др.

Бизнес-аналитика в сфере розничной торговли

К сожалению, среди представителей малого бизнеса в сфере ритейла системы бизнес-аналитики ещё недостаточно широко распространены, и их возможности ещё не в полной мере раскрыты пользователями. Хотя они и помогают увеличить эффективность ведения бизнеса, но почему-то недостаточно охотно внедряются на малых предприятиях..

В прошлом году компанией Qlik был проведён опрос, основными задачами которого было выяснить причины использования аналитических систем в розничной торговле и узнать, насколько хорошо их пользователи осведомлены о возможностях и эффективности таких систем.

Опрос проводился среди представителей управляющего звена розничной торговли, работающих с самыми разными товарами. Респонденты были ранжированы по размерам их бизнеса, исходя из числа торговых точек компании. Из всего числа опрошенных большинство составляют представители малого предпринимательства (29% респондентов) с небольшим числом торговых точек. Вторыми по численности среди опрошенных оказались сверхбольшие сети, имеющие более 500 торговых точек. На их долю пришлось 18%. На остальные категории пришлось примерно по 10% на каждую.

Проведём небольшой анализ полученных данных и выясним состояние сферы интеллектуального анализа данных в области ритейла.

Среди респондентов значительная часть уже применяет в своей работе системы анализа данных – 57,1% опрошенных. Среди тех компаний, кто ещё не пользуется бизнес аналитикой, значительная часть собирается внедрить такие системы у себя в ближайшем будущем. А полностью отказываются работать в ногу со временем всего 2,4% опрошенных. В основном это представители малого предпринимательства. Эти результаты можно увидеть на рисунке 1.

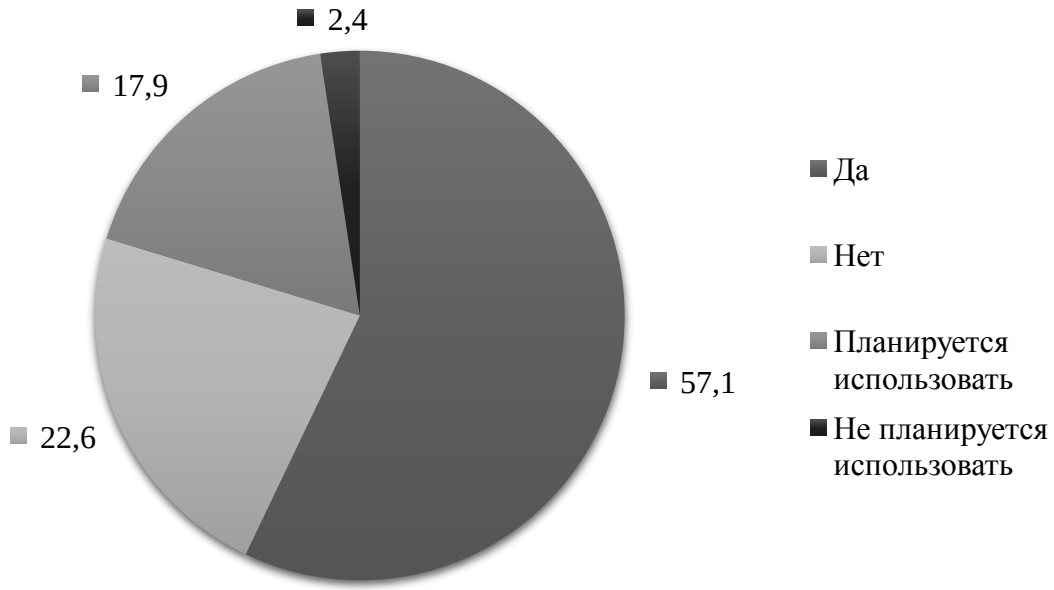

Рисунок 1 – Наличие системы бизнес-аналитики

Что касается основных поводов, побудивших компании использовать у себя системы анализа больших данных, здесь они во многом совпадают независимо от размеров бизнеса участников опроса. Респондентам предлагалось выбрать несколько главных на их взгляд причин. Приведу три наиболее популярные. Самой важной причиной, побудившей к использованию бизнес-аналитики, респонденты назвали желание повысить скорость принимаемых управленческих решений и повысить их эффективность. Эту необходимость отметили 88,6% участников опроса. Чуть менее важным, но всё же значимым, оказалось желание управляющих повысить производительность работы своих сотрудников, избавив их от рутинных действий по работе с отчётами (72,7%). Среди причин, которые повлияли на решение использовать на производстве системы бизнес-аналитики также была названа потребность объединения информации из разных источников – на её долю пришлось 65,9%. Распределение топ-3 причин наглядно отображено на рисунке 2.

Рисунок 2 – Причины внедрения бизнес-аналитики

Такие, казалось бы, очевидные причины, как повышение прибыли, увеличение эффективности работы с клиентами, к удивлению, оказались не

столь важны для участников опроса. Эти причины указывал куда меньший процент респондентов.

Такая тенденция может говорить о том, что среди тех, кто уже использует у себя системы бизнес-аналитики, многие не до конца понимают весь спектр их возможностей, рассматривая их в основном как инструмент для упрощения работы с отчётами и повышения эффективности взаимодействия сотрудников, не более.

Косвенно это подтверждают и результаты опроса по достигнутым результатам работы таких систем среди пользователей. Большинство опрошенных (70,7%) упомянули о сокращении времени, которое используется ими для сбора необходимой для принятия управленческих решений информации. Половина опрошенных отметили увеличение эффективности таких решений. А повышения доходности бизнеса путём снижения расходов, оптимизации логистики и оптимального ценообразования при помощи систем бизнес-аналитики удалось добиться не многим. Лишь около 10% участников опроса отметили у себя такие достижения. Причины этого уже объяснялись выше, просто для этих целей аналитикой пользуется лишь малый процент опрошенных, но, как видим, своих целей они достигают.

По сферам применения в ритейле систем бизнес-аналитики в первую тройку вошли анализ продаж (93,2% использующих компаний), финансовый анализ и на третьем месте почти поровну маркетинг и закупки. Реже всего возможности интеллектуального анализа данных используют в области безопасности и мерчандайзинга. Данные ранжированы по категориям. Также среди тех, кто ещё не использует у себя системы бизнес-аналитики, но планирует их внедрение, спросили, в какой сфере они планируют их использование. Результаты во многом совпали с ответами тех, кто уже использует системы анализа данных в своей работе. На рисунке 3 отображено соотношение респондентов, уже использующих систему бизнес-аналитики и только планирующих использование по сферам применения. Так же указаны проценты от общего числа респондентов.

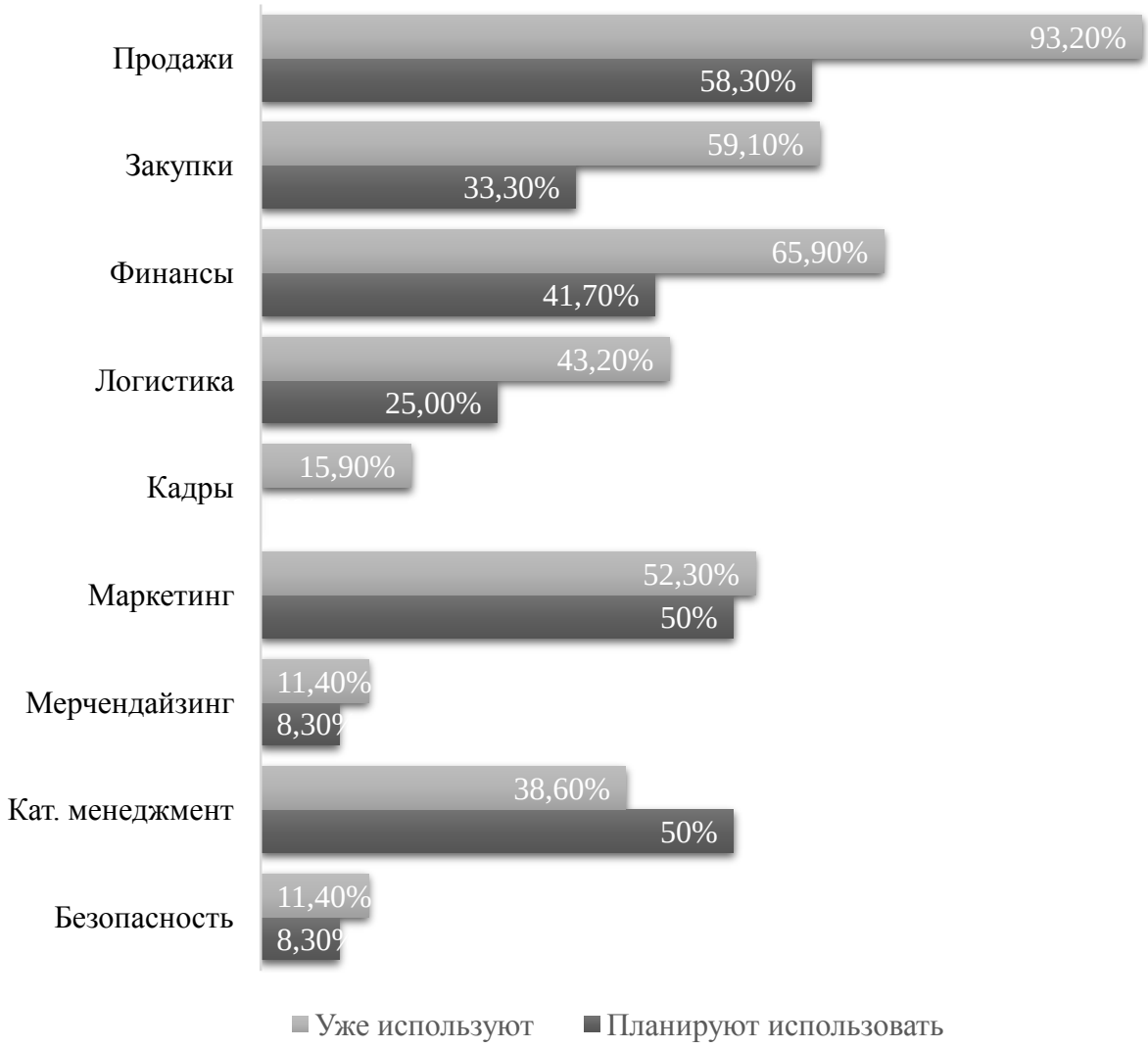

Рисунок 3 – Области использования бизнес-аналитики

Что касается основных факторов выбора системы бизнес-аналитики, большинство отдают своё предпочтение системам, позволяющим повысить скорость и упростить создание отчётов и работу с ними (65,9%). В необходимости анализа больших объёмов данных высказались 59,1% опрошенных. Немаловажным фактором также оказалось наличие в системе бизнес-аналитики наглядной и простой для восприятия визуализации данных. Замыкает пятёрку основных факторов при выборе системы аналитики возможность быстрой окупаемости системы, пятая часть респондентов сочла этот параметр значимым для себя. В последние годы также отмечается потребность в приложениях, способных работать на мобильных платформах –

11,4% указывают такую необходимость. Это связано с вовлечением в процесс аналитики сотрудников, работающих удалённо или вынужденных к частым разъездам. Наглядно эти данные отображены на рисунке 4.

Рисунок 4 – Основные факторы выбора системы бизнес-аналитики

Также участников опроса спросили, в каких функциях и возможностях они испытывают недостаток при использовании современных систем бизнес-аналитики. Острее всего респонденты ощутили нехватку предиктивного инструментария, позволяющего прогнозировать бизнес-процессы на основе накопленных данных. Почти половина опрошенных испытывают недостаток в средствах интеллектуального анализа данных. В топ-5 также вошли нехватка средств визуализации данных (напомню, что наличие в системе средств качественной визуализации является одним из главных факторов при выборе системы по результатам этого же опроса) и приложений, работающих на мобильных платформах. Обоснованность необходимости мобильных приложений отмечалась в предыдущем абзаце. Результаты этого опроса могут свидетельствовать о том, что многое пользователи систем бизнес-аналитики уже начали постигать весь потенциал возможностей своих систем. Каких

возможностей, по мнению респондентов, не хватает в системах бизнес-аналитики наглядно продемонстрировано на рисунке 5.

Рисунок 5 – Каких возможностей не хватает системам бизнес-аналитики

Но, несмотря на то, что респонденты высказались о нехватке этих средств, это, с другой стороны, не говорит однозначно о том, что современные системы ими не обладают. Как отмечалось ранее, даже используя у себя на производстве системы бизнес-аналитики, многие не всегда имеют полное представление о их потенциале и функционале, и не используют все возможности системы просто потому, что не знают о них или не умеют в должной мере ими пользоваться.

Организаторы опроса также выяснили, кому больше всего доверяют при выборе системы те, кто ещё никогда прежде ими не пользовался. В основном, выбирая для своего бизнеса систему аналитики, люди обращаются к сети Интернет (75%). Почти половина (41%) доверяют мнению своих коллег, уже внедривших у себя подобные решения. А вот интеграторы и ИТ-консультанты находятся по уровню доверия только на третьем месте. Такой низкий процент доверия к ИТ-консультантам и интеграторам отчасти объясняет слабую

осведомлённость представителей малого бизнеса о возможностях и потенциале систем интеллектуального анализа данных.

Среди факторов, которые останавливают многих от внедрения у себя систем бизнес-аналитики, основными являются высокая стоимость таких систем, трудности их внедрения отсутствие структуры в своих хранимых данных. Главные факторы, удерживающие потенциальных клиентов от внедрения у себя систем бизнес-аналитики проиллюстрированы на рисунке 6.

Рисунок 6 – Сдерживающие факторы для внедрения бизнес-аналитики

Исходя из результатов проведённого компанией Qlik опроса можно сделать вывод, что несмотря на достаточно широкое распространение систем анализа данных в сфере розничной торговли, многие используют их не в полной мере эффективно, пользуясь лишь частью предоставляемых им возможностей. Такая ситуация, вероятнее всего, является следствием нехватки осведомлённости о возможностях, предоставляемых бизнес-аналитикой, что косвенно подтверждается разницей в результатах опроса среди тех, кто уже знаком с такими системами и тех, кто знает о них лишь понаслышке. Однако несмотря на это большинству использующих системы бизнес-аналитики удалось

добиться поставленных ими задач, что в немалой степени является также заслугой аналитиков и ИТ-консультантов, повышающих уровень осведомлённости о возможностях таких систем. При это у систем бизнес-аналитики остаётся огромный потенциал, т.к. пока что они используются далеко не во всех направлениях и сферах бизнеса, хотя их возможности гораздо шире.

Возможности применения интеллектуального анализа данных в сфере ритейла

Ритейл является одной из самых динамично развивающихся сфер отечественной экономики. Постоянное развитие требует поиска новых способов увеличения своей конкурентоспособности на общем рынке и быстрой реакции на изменяющиеся условия.

Интеллектуальный анализ данных является одним из таких способов, позволяя фирмам использовать всю собранную ими информацию по своим транзакциям, клиентам, отчётам и др. для собственного роста и повышения продуктивности.

Раньше аналитики ещё могли обрабатывать показатели фирмы самостоятельно, но сейчас это стало нетривиальной задачей с огромным ростом накопленной фирмами информации и скоростью её пополнения.

Следствием этого стало активное внедрение информационно-аналитических систем в области розничной торговли. Необходимость оперативного получения новых данных, позволяющих увеличить эффективность и ускорить процесс принятия управленческих решений объясняет растущую популярность систем бизнес-аналитики. Постоянное непрерывное изменение условий, в которых приходится работать бизнесменам, вынуждает их искать способы наиболее быстрого реагирования на окружающую обстановку.

Одним из нагляднейших примеров использования интеллектуального анализа данных в области розничной торговли служит анализ покупательской корзины, анализирующий зависимости между товарами с точки зрения спроса на них.

Очевидно, что для увеличения прибыли от продажи, товары должны быть размещены в торговом зале не просто так, а правильно и обдуманно. Системы интеллектуального анализа данных позволяют решить эту проблему путём поиска сопутствующих товаров – где один объект или набор объектов является основным приобретаемым покупателем, а другой – сопутствующий ему, который с большой долей вероятности будет приобретён, если приобретён и первый. Такие группы сопутствующих товаров определяются при помощи анализа покупательской корзины, путём анализа прошлых чеков и транзакций фирмы.

Информация, полученная в результате анализа покупательской корзины, может использоваться:

- маркетологами, для увеличения продаж товара путём его продвижения через таргетированную рекламу;

- мерчендайзерами, для эффективного и грамотного размещения товаров на торговой площади фирмы;

- при формировании цен, снижая цену на основной товар и повышая на сопутствующий и наоборот.

К сожалению, в нашей стране представители малого бизнеса не замечают ценности интеллектуального анализа данных, в отличии от зарубежных стран. Это видно хотя бы по объёму средств, выделяемых на информационно-техническую инфраструктуру. Для многих бизнес-аналитика остаётся чем-то чужим и непонятным. А среди тех, кто всё же внедрил у себя технологии информационного анализа, далеко не все в полной мере раскрыли и используют их возможности. Зачастую, аналитикам и ИТ-специалистам приходится самим доказывать значимость бизнес-аналитики для повышения производственной эффективности.

К счастью, в последние годы ситуация начала меняться в лучшую сторону, и всё больше управленцев замечают преимущества использования систем бизнес-аналитики.

Сейчас анализ данных, накопленных фирмой, в условно можно разделить на две категории:

- анализ внутренних данных, накопленных фирмой за годы своей работы. Сюда относят все отчёты, реквизиты, чеки, транзакции, анкеты клиентов, складской учёт и др.;

- анализ внешних данных. Сюда относят информацию о конкурентах и сфере деятельности в целом, анализ целевой аудитории.

Разумеется, это лишь основные направления, выделенные весьма условно. Каждое из них можно рассматривать более детально и углублённо.

Анализ внутренних данных является сегодня основным источником информации для розничной торговли, на его долю приходится около 90% всей аналитики, решаемой средствами интеллектуального анализа данных в сфере ритейла.

Анализ потребительской корзины и активности покупателей являются на сегодня главной задачей аналитических систем розничной торговли. Они останутся такими и в ближайшем будущем, только лишь укрепив свои позиции, став незаменимой составляющей бизнес-аналитики в розничной торговле.

1.3 Методы интеллектуального анализа данных

Главное преимущество методов Data Mining заключается в том, что они сочетают в себе как математический инструментарий (начиная от статистического анализа и заканчивая новейшими эвристическими методами), так и последние постижения в области ИТ. В методах интеллектуального анализа данных переплелись формализованные и неформальные способы и техники аналитики, различные способы анализа данных.

Методы Data Mining в своём родами являются логичным продолжением и развитием методов математической статистики, в связи с чем большое количество её методов нашли здесь применение.

Основными методами интеллектуального анализа данных являются в первую очередь методы, основанные на переборе. Обычный перебор всех вариантов занимает $O(2^N)$ операций (где N – общее количество объектов), а это значит, что с ростом числа объектов перебора вычислительная сложность растёт экспоненциально, что при значительном количестве объектов делает решение такой задачи практически невыполнимым.

Для того чтобы уменьшить количество переборов и снизить вычислительную сложность таких алгоритмов, методы интеллектуального анализа данных используют различные эвристические подходы.

Одним из преимуществ методов интеллектуального анализа данных является их лёгкость как для восприятия, так и для программной реализации. Из минусов методов Data Mining можно отметить отсутствие формализованного математического аппарата и строго оформленной теории на основании которых строятся методы, и как из этого следует – трудности в их развитии.

Перечислим основные типы знаний, которые добываются при помощи методов интеллектуального анализа:

- ассоциативные правила;
- деревья решений;
- кластеры;
- математические функции.

Методы поиска этих знаний используют наработки различных сфер науки, включая теорию вероятностей, математический анализ, статистический анализ, теорию множеств, нейронные сети, генетическое и эволюционное программирование.

Многие технологии, используемые в интеллектуальном анализе данных уже давно известны и широко применяемы в математике. Однако в сфере Data Mining они находят новое применение, открывающее ранее неизвестные возможности, что в первую очередь связано с ростом уровня технических и программных средств за последние десятилетия. Многие методы

интеллектуального анализа данных работают и в рамках теории искусственного интеллекта.

Основные задачи, решаемые методами интеллектуального анализа данных, это регрессионный анализ, задача классификации, задача кластеризации и поиск ассоциативных правил в базах данных.

Эти задачи принято делить на описательные и предиктивные.

Описательные задачи направлены на усовершенствование осмысления информации. Получаемые в результате работы над этими задачами данные должны быть просты и понятны для человека. К этому виду относятся поиск ассоциативных правил в базах данных и задача кластеризации.

Задачи классификации и регрессионный анализ относят к предиктивным задачам. При их решении используются методы, направленные на анализ текущей информации с целью прогноза информации в будущем.

2 Проектирование автоматизированной информационной системы

2.1 Постановка задач и анализ существующих решений

Требования, предъявляемые к системе

Прежде чем приступать к анализу существующих решений и постановке задач, перечислим требования, предъявленные к проектируемой информационной системе.

Ранее мы уже установили, какие требования предъявляются к системам аналитики в ритейле (см. стр. 16 – Бизнес-аналитика в сфере розничной торговли), чего от них ожидают пользователи и в каких функциях испытывают недостаток.

На основании этих данных, выделим основные критерии, которым должна удовлетворять проектируемая ИС.

Среди основных сдерживающих факторов для внедрения системы бизнес-аналитики респондентами были названы такие как: высокая стоимость, длительный период внедрения, неструктурированность данных компании. Проектируемая ИС должна быть максимально избавлена от этих недостатков.

Во-первых, она должна распространятся свободно. Это удовлетворит почти половину (41,7%) респондентов, сомневающихся в выборе системы бизнес-аналитики по причине высокой стоимости большинства существующих аналогов. Более того, т.к. 20% опрошенных были заинтересованы в быстрой окупаемости системы, свободное распространение проектируемой ИС так же будет преимуществом, исключая этап окупаемости системы.

Во-вторых, необходимо максимально упростить этап внедрения, недовольство длительностью которого высказали четверть респондентов. Внедрение ИС, разработанной самостоятельно, нередко приводит к изменению уже сложившихся на предприятии процессов работы. Приходиться менять их в соответствии со стандартами и принципами работы информационной системы. Несмотря на то, что внедрение информационной системы помогает решить

многие управленческие задачи, основной проблемой остаётся человеческий фактор. Кроме того, при внедрении информационной системы на производство необходимо обязательно обучить персонал работе с этой системой, но, нередко случается, что персонал не очень то желает переучиваться. Изменение старых привычек и вырабатывание новых – долгий и трудный процесс. На основании всего вышесказанного, для проектируемой ИС необходимо максимально упростить взаимодействие с конечным пользователем, сделать его интуитивно понятным, а назначения функций и элементов интерфейса программы очевидными и однозначными. Это позволит значительно сократить этап подготовки и обучения сотрудников работе с ИС, и как следствие – сократить время внедрения информационной системы.

В-третьих, информационная система должна располагать собственными инструментами работы с данными компании, хранить и обрабатывать их в чётко структурированном виде. К этому можно предъявить следующие основные требования:

- данные должны быть определённым образом структурированы, это повысит эффективность и скорость поиска;

- данные должны храниться в одном месте;

- хранимая информация должна быть полезной в данный момент, или иметь перспективы применения в будущем;

- должна быть реализована возможность фильтрации данных;

- новая информация должна записываться максимально просто и быстро.

Среди всё тех же участников опроса почти половина (44,2% респондентов) обеспокоены нехваткой инструментария интеллектуального анализа данных в современных системах бизнес-аналитики. Проектируемая ИС будет содержать в себе возможности для интеллектуального анализа данных посредством поиска ассоциативных правил в базе транзакций компании, т.к. это позволит в перспективе повысить рентабельность бизнеса. Данное решение является особенно актуальным и востребованным, учитывая, что 93,2% уже внедривших у себя системы бизнес-аналитики используют их именно для анализа продаж, а

среди только планирующих внедрение больше половины (58,3%) собираются использовать их с той же целью.

Итак, обозначим основные функции, которыми должна обладать проектируемая ИС:

- система должна содержать в себе базу данных транзакций фирмы и предоставлять пользователю инструментарий для работы с этой базой;

- система должна содержать средства интеллектуального анализа данных, а именно поиска ассоциативных правил в базе данных.

Для проектирования и разработки ИС необходимо конкретизировать эти функции и грамотно поставить задачи, что и будет сделано далее.

Проведённый ранее анализ области позволил сформулировать основные требования к ИС на основе актуальных потребностей пользователей, их опыта и пожеланий. Для дальнейшей работы необходимо провести анализ уже существующих на рынке продуктов, выявить их плюсы и минусы, что позволит объединив достоинства имеющихся аналогов и исправив их недостатки создать конкурентоспособную информационную систему.

Обзор существующих решений

Рынок программного обеспечения средств интеллектуального анализа данных представлен огромным множеством вариантов. С каждым годом их число только растёт, но, к сожалению, не всегда растёт качество.

Сегодня наиболее популярными на рынке являются такие продукты как Oracle Data Mining, STATISTICA Data Miner, Microsoft SQL Server Analysis Services и другие. Основные поставщики готовых продуктов были перечислены ранее при анализе рынка.

Рассмотрим возможности наиболее популярных на сегодняшний день продуктов.

SAS Enterprise Miner. Программа американской компании включает в себя широкий инструментарий для интеллектуального анализа данных, включая такие распространённые на сегодняшний день методы интеллектуального анализа данных, как:

- деревья решений;
- нейронные сети;
- регрессионный анализ;
- кластеризацию;
- поиск ассоциативных правил;
- секвенсиальный анализ.

STATISTICA Data Miner. Продукт компании СтатСофт, о которой я упоминал при обзоре научной литературы. За годы своего существования платформа STATISTICA обросла множеством всевозможных модулей, реализующих самые различные методы и технологии аналитики и статистики. Очень трудно не потеряться в таком разнообразии. Программный продукт включает в себя работу с нейронными сетями, построение и анализ всевозможных моделей, регрессионный анализ, деревья решений, методы кластеризации, классификации, ассоциации, широкий набор средств визуализации. Основные компоненты программы обеспечивают работу со следующими задачами:

- классификация;
- моделирование;
- предиктивный анализ;
- нейронные сети;
- поиск ассоциативных правил;
- построение всевозможных деревьев решений.

Oracle Data Mining. Oracle является крупнейшим игроком на рынке услуг и технологий интеллектуального анализа данных. В Oracle Data Mining реализованы такие технологии, как:

- классификация;
- кластеризация;
- поиск ассоциаций;
- выделение признаков.

KXEN Analytic Framework. KXEN Analytic Framework является по сути своей платформой для различных модулей, позволяющих проводить описательный и предиктивный анализ. Продукт использует регрессионные алгоритмы, позволяет выявить естественные кластеры в наборе данных, позволяет производить бинарную классификацию, позволяет проводить предиктивный анализ.

Microsoft SQL Server Analysis Services. Продукт компании Microsoft, как и ранее перечисленные продукты, включает в себя основные технологии для интеллектуального анализа баз данных на своей же платформе.

Программные продукты SPSS. Они позволяют строить всевозможные деревья решений, проводить кластерный анализ, находить ассоциативные правила в базах данных и проводить предиктивный анализ. Содержит большинство технологий, предоставляемых его прямыми конкурентами.

Сравнительный анализ методов Data Mining, используемых в наиболее распространённых продуктах извлечения знаний можно обобщить в таблице 2.

Таблица 2 – Методы Data Mining в программных продуктах

	SAS Enter-prise Miner	STATI-STICA Data Miner	Oracle Data Mining	KXEN Analytic Frame-work	Micro-soft SQL Server Analysis Services	ПП SPSS	Встре-чае-мость
Адаптивная Байесовская сеть			+				1
Анализ временных рядов		+		+	+	+	4
Граничные методы		+	+	+			3
Деревья решений	+	+			+	+	4

Продолжение таблицы 2

	SAS Enter-prise Miner	STATI-STICA Data Miner	Oracle Data Mining	KXEN Analytic Frame-work	Micro-soft SQL Server Analysis Services	ПП SPSS	Встре-чае-мость
Иерархическая кластеризация	+	+	+	+			4
Линейная регрессия	+	+	+	+	+	+	6
Логистическая регрессия	+	+	+	+	+	+	6
Наивный Байесовский алгоритм			+		+	+	3
Неиерархическая кластеризация	+	+	+	+	+		5
Нейронные сети	+	+			+		3
Поиск ассоциативных правил	+	+	+		+		4

Как видно из таблицы 2, выбранное нами для реализации в ИС решение реализовать поиск ассоциативных правил является одним из наиболее удачных. С одной стороны, данная технология встречается достаточно часто чтобы считаться востребованной, но не так часто реализуется программно.

Постановка задач

В своей работе я задался целью спроектировать и реализовать информационную систему, обеспечивающую информационную поддержку представителей малого предпринимательства путём предоставления им

возможностей доступа к методам и технологиям интеллектуального анализа данных.

Проведя анализ предметной области, изучив требования конечных пользователей и состояние рынка программных продуктов схожего назначения, можно поставить более конкретные задачи.

Таким образом передо мной была поставлена задача спроектировать и реализовать информационную систему, решающую определённые задачи и выполняющую конкретно обозначенные функции. А именно:

- система должна обладать возможностью работы с базой транзакций фирмы. Должны быть реализованы возможности создавать БД, добавлять в неё новые транзакции, редактировать и удалять существующие, проводить выборку из БД по заданным условиям;

- система должна содержать средства интеллектуального анализа данных, а именно поиска ассоциативных правил в базе данных. Должны быть реализованы связанные с этим подготовка и преобразования данных;

- пользователь должен иметь возможность получать результаты работы программы как в виде отчёта, так и с помощью наглядной и понятной визуализации.

Проектируемая ИС должна быть выполнена максимально просто и понятно для пользователя, реализуя достижение поставленных задач не усложняя себя лишними функциями. Интерфейс ИС должен быть интуитивно понятен и дружелюбен.

2.2 Проектирование информационной системы

Обоснование и выбор способа решения обозначенной проблемы

С учётом поставленных цели и задач, необходимо спроектировать систему, предоставляющую пользователю возможности ведения базы транзакций компании и получения из неё полезных скрытых данных через использование методов интеллектуального анализа данных.

При рассмотрении требований, предъявляемых к системе (стр. 29), мною были перечислены основные требования как к самой системе, так и к структуре информации, которую предстоит обрабатывать проектируемой ИС. Это ведёт к необходимости использования в проектируемой ИС максимально простой встроенной БД, отвечающей поставленным требованиям. Применение встраиваемой СУБД даёт проектируемой ИС следующие преимущества:

- система управления базой данных связана с прикладным приложением и работает на той же рабочей станции, не требуя специального администрирования;

- отсутствует программа-сервер. Это значительно упрощает работу с БД, т.к. к ней всё равно не требуется доступ с нескольких рабочих станций;

- высокая скорость работы и небольшое потребление ресурсов системы. Благодаря специализированному API количество операций чтения-записи минимально.

При постановке задач (стр. 34) было приведено высокоуровневое описание основных функций системы, обоснован их выбор.

Выбор метода интеллектуального анализа данных, применяемого в системе, был обоснован при обзоре существующих решений (стр. 31). Для этой работы был выбран метод поиска ассоциативных правил в базе транзакций фирмы.

Высокоуровневое описание алгоритма

Основные функции проектируемой ИС: ведение базы данных, содержащей сведения о транзакциях фирмы; поиск ассоциативных правил в базе данных.

Работа с базой данных, содержащей сведения о транзакциях фирмы, будет осуществляться посредством SQL запросов программы к своей встроенной базе данных, содержащей транзакции фирмы. При выборе пользователем определённого действия, направленного на работу с базой данных (будь то добавление новой записи или любое другое действие) программа формирует запрос на языке SQL к встраиваемой СУБД, направленный на осуществление

выбранного пользователем действия. Применение встраиваемой СУБД было обосновано ранее.

Среди действий, направленных на работу пользователя с базой, ИС должна содержать следующие:

- создание новой БД;
- открытие уже существующей БД;
- добавление новой транзакции в БД;
- редактирование существующей транзакции в БД;
- удаление из БД существующей транзакции;
- отображение всего содержимого БД в виде наглядной таблицы;
- фильтрация отображаемого содержимого по номеру транзакции;
- фильтрация отображаемого содержимого по наименованию товаров;
- фильтрация отображаемого содержимого по дате транзакции.

Перечисленных действий достаточно для приведения информации фирмы в структурированный вид и обеспечения пользователю комфортной работы.

При создании новой базы данных пользователь указывает её имя. Программа посылает SQL запрос встраиваемой СУБД, которая создаёт БД с двумя таблицами в ней: первая таблица хранит все транзакции фирмы, о назначении второй будет сказано позднее при рассмотрении алгоритма поиска ассоциативных правил.

При открытии БД программа проверяет наличие в ней необходимых таблиц, и, в случае отсутствия, создаёт их.

При добавлении новой транзакции пользователь вводит в окно программы следующие данные: список товаров и дату совершения транзакции. По умолчанию устанавливается текущая дата, но пользователь в праве изменить её по своему усмотрению. После того, как пользователь подтвердил ввод данных. Программа посылает запрос к БД, содержащий введённые пользователем данные. В таблицу транзакций заносится строка, содержащая уникальный номер транзакции, список товаров и дату совершения транзакции. Уникальный номер

присваивается транзакции автоматически без участия пользователя и не может быть им изменён.

При редактировании транзакции пользователь вначале указывает уникальный номер транзакции, которую требуется отредактировать. Затем в окне программы пользователь вносит изменения в требуемые поля. После подтверждения программа посылает запрос к БД и запись обновляется. Номер транзакции при этом не изменяется.

Удаление транзакции происходит по уникальному номеру транзакции. Пользователь указывает номер, после чего программа посылает к БД запрос, удаляя требуемую транзакцию.

Содержимое БД отображается в основном окне программы. Каждое действия пользователя, приводящее к изменению содержимого базы, сразу же отображается в окне программы.

При фильтрации содержимого базы в основном окне программы отображаются только записи, удовлетворяющие выбранным условиям. При фильтрации по номеру, пользователь указывает интервалы. Транзакции, номера которых попадают в указанный интервал, будут отображены в основном окне программы. Фильтрация по дате происходит схожим образом, только вместо интервала номеров пользователь указывает интервал дат. При фильтрации по товару пользователь пишет в поле программы наименование требуемого товара. Все транзакции, содержащие этот товар, будут отображены в основном окне программы.

Алгоритм поиска ассоциативных правил будет подробно рассмотрен в одном из следующих пунктов. Что касается описания способа его работы в проектируемой ИС – в ней реализуется инструментарий, позволяющий пользователю выполнить поиск ассоциативных правил и получить результат в наглядном и понятном виде.

Алгоритм из-за своих особенностей должен работать с особым форматом данных, а именно – БД транзакций должна быть преобразована в файл, содержащий только список транзакций, без указания номеров транзакций и даты

их совершения. Наименования товаров должны быть заменены целочисленными кодами. Использование целочисленных данных вместо строковых позволяет колоссально увеличить скорость работы алгоритма и сократить количество требуемых вычислительных ресурсов. Для преобразования базы транзакций в целочисленный формат, каждому товару, содержащемуся в базе, должен быть присвоен уникальный целочисленный номер. Это происходит автоматически при помощи второй таблицы, содержащей список товаров и их уникальные номера. При добавлении новой транзакции в базу, программа проверяет, имеются ли введённые товары в таблице товаров. Если программа находит новый товар, ранее не встречавшийся в базе, она добавляет его в таблицу товаров и присваивает ему уникальный номер. Таблица товаров изменяется только программой, и пользователь не может получить к ней доступ. Программа содержит функцию конвертирования стандартной базы данных в формат, требуемый для работы алгоритма.

Пользователь должен указать программе пути к двум файлам: входному, содержащему список транзакций в целочисленном конвертированном формате; выходному, куда будут записаны найденные правила.

Пользователь обязательно должен указать программе два основных параметра: минимальную поддержку и минимальную достоверность. В случае, если пользователь не укажет нужные значения, используются значения по умолчанию.

Помимо обязательных основных, пользователь может также указать необязательные дополнительные параметры: минимальный лифт и максимальную поддержку.

После выбора всех нужных параметров, пользователь должен запустить процесс поиска ассоциативных правил путём выбора этого действия в интерфейсе программы. Процесс поиска ассоциативных правил происходит автоматически и не требует вмешательства пользователя. По завершении процедуры поиска, пользователю отображается список найденных правил.

Построение моделей

Для лучшего понимания сути работы проектируемой системы, рассмотрим несколько моделей в разных нотациях.

Для начала рассмотрим, как происходит передача данных внутри информационной системы. Для этого воспользуемся моделью представления данных в DFD, изображённой на рисунке 7. Модель представления данных в DFD очень хорошо подходит для высокоуровневого моделирования системы.

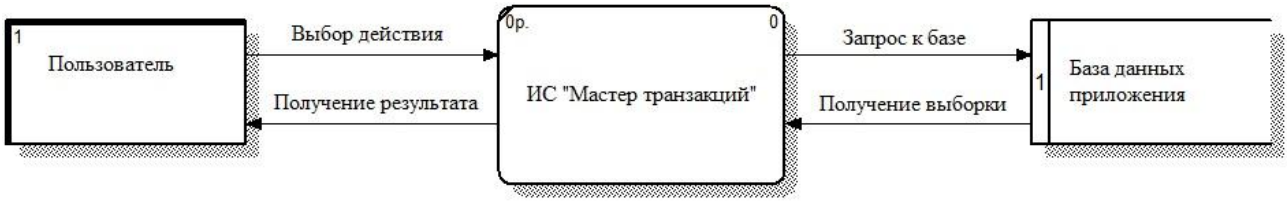

Рисунок 7 – Модель передачи данных ИС в нотации DFD

ИС получает от пользователя команды, а затем обрабатывает их. Взаимодействие пользователя с информационной системой происходит через графический интерфейс информационной системы. При необходимости информационная система обращается к своей встроенной базе данных, содержащей сведения о транзакциях фирмы, посылая ей запросы на языке SQL и получая в ответ требуемую выборку. Запросы информационной системы к своей базе данных программа формирует на основании выбранного пользователем действия (например, редактирование транзакции в базе данных). Полученная от базы данных выборка обрабатывается информационной системой. Обработав данные, полученные от пользователя и базы данных, информационная система возвращает пользователю результат своей работы в соответствии с выбранным пользователем действием.

Для отображения структуры и функций системы, обратимся к нотации IDF0 и построим функциональную модель проектируемой информационной системы. Данная модель изображена на рисунке 8.

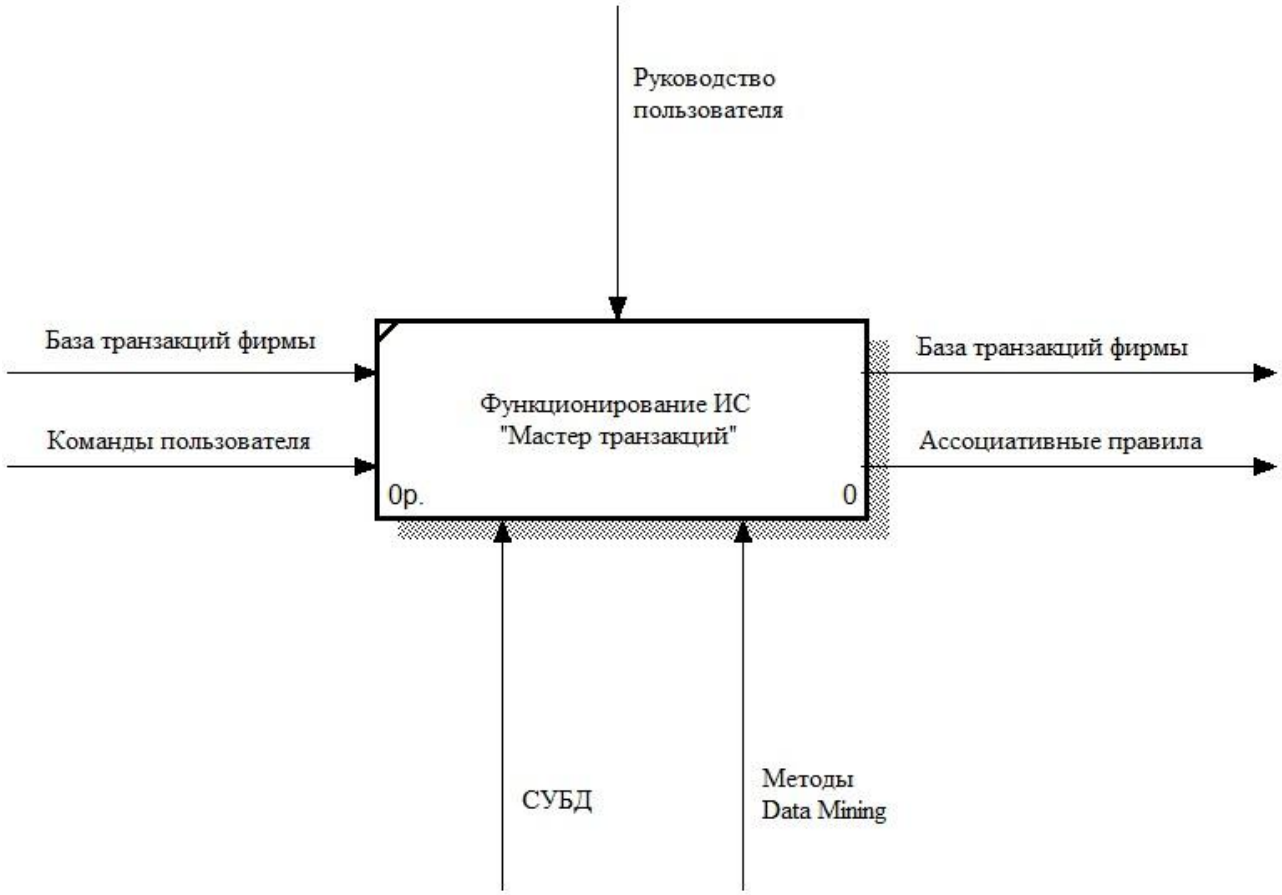

Рисунок 8 – Функциональная модель ИС в нотации IDF0

Как видно из контекстной диаграммы, изображённой на рисунке 8, для своего функционирования информационная система получает на вход базу транзакций фирмы и команды пользователя. Входные данные в последствии преобразуются системой в то, что мы получаем на выходе системы. База транзакций фирмы используется в качестве входя при выполнении задачи поиска ассоциативных правил. Содержимое самой базы при этом не изменяется. Оно может быть изменено только в соответствии с командами пользователя информационной системы. На выходе системы мы получаем преобразованную в соответствии с командами пользователя базу транзакций фирмы и ассоциативные правила. Стрелки, подключённые к нижней стороне блока, представляют механизмы – средства, используемые для выполнения функций программы. Руководство пользователя в данном случае является управлением, т.е. условием, при выполнении которого выход информационной системы будет правильным.

Поиск ассоциативных правил

Для решения задачи поиска ассоциативных правил в базах данных, как и для решения любой другой задачи, необходимо обработать исходные данные и получить результат. Результаты поиска представляются в форме ассоциативных правил. При их поиске принято выделять два основных этапа:

- поиск всех часто встречающихся наборов объектов;

- поиск ассоциативных правил по найденным часто встречающимся наборам объектов.

Опишем задачу поиска всех часто встречающихся наборов объектов в общем виде. Объекты, которые составляют исследуемые наборы (itemsets), обозначим как множество:

$$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}, \quad (1)$$

где i_j – объекты, входящие в наборы;

n – общее количество объектов.

В области ритейла, например, такими объектами являются товары, продаваемые фирмой.

Наборы объектов, состоящих из множества I, хранящиеся в базе данных и в последствии анализируемые программой, называются транзакциями. Транзакцию можно представить как подмножество множества I:

$$T = \{i_j \mid i_j \in I\}. \quad (2)$$

Транзакции в ритейле представляют собой наборы товаров, приобретаемые покупателем единовременно. Транзакции сохраняются в базе данных фирмы.

Набор транзакций, который хранится в базе данных и который доступен нам для анализа мы обозначим множеством:

$$D = \{T_1, T_2, \ldots, T_r, \ldots, T_m\}, \quad (3)$$

где m – количество доступных для анализа транзакций.

Множество транзакций, в которые входит объект i_j, обозначим следующим образом:

$$D_{ij} = \{T_r \mid i_j \in T_r;\ j = 1..n;\ r = 1..m\} \subseteq D. \quad (4)$$

Некоторый произвольный набор объектов (itemset) обозначим следующим образом:

$$F = \{i_j \mid i_j \in I;\ j = 1..n\}. \quad (5)$$

Набор, состоящий из k объектов, называется k-элементным набором.

Множество транзакций, в которые входит набор F, обозначим следующим образом:

$$D_F = \{T_r \mid F \subseteq T_r;\ r = 1..m\} \subseteq D. \quad (6)$$

Отношение количества транзакций, в которое входит набор F, к общему количеству транзакций называется поддержкой (support) набора F и обозначается Supp(F):

$$\mathrm{Supp}(F) = \frac{|D_F|}{|D|}. \quad (7)$$

При поиске ассоциативных правил в базе данных аналитик указывает значение минимальной поддержки для наборов объектов Supp_{min}. Если значение поддержки набора больше, чем минимальное значение поддержки Supp_{min}

указанное аналитиком – то такой набор принято называть часто встречающимся набором.

Решая задачу поиска ассоциативных правил нужно прежде всего найти такое множество часто встречающихся наборов.

Ассоциативные привала обыкновенно имеют вид «если (условие), то (результат)». Здесь «условие» – набор любых объектов из множества всех объектов I, с которым непосредственно связан «результат», также состоящий из объектов множества I. Ассоциативное правило можно представить как импликацию над множеством I. Главным преимуществом ассоциативных правил является их наглядность и простота как для восприятия человеком, так и для программной реализации и интерпретации. Однако, не все обнаруженные в процессе поиска правила несут в себе пользу. Принято выделять три типа ассоциативных правил:

- полезные правила. Полезные правила содержат в себе информацию, которая при их применении принесёт очевидную пользу. Эти правила раньше не были известны аналитикам и были открыты в процессе интеллектуального анализа данных, но эти правила логически объяснимы;

- тривиальные правила. Тривиальные правила содержат в себе информацию, которая является истинной, но уже известной и очевидной аналитикам. Эти правила легко объяснить с логической точки зрения, но их использование не принесёт ощутимой пользы, т.к. большинство этих правил уже используются в силу их известности. Иногда такие правила могут применять для проверки правильности уже принятых решений;

- непонятные правила. Непонятные правила содержат информацию, которая не может быть объяснена аналитиками. Такие правила могут быть получены в результате анализа глубоко скрытых знаний или в результате анализа аномальных и нестандартных значений. Такие правила нельзя использовать в чистом виде, т.к. из-за невозможности объяснить их с логической точки зрения их применение также приведёт к неизвестным и непредсказуемым результатам.

Для генерации самих ассоциативных правил необходимо предварительно найти все часто встречающиеся наборы. Вследствие большого количества таких наборов при большом множестве объектов, количество сгенерированных ассоциативных правил также может быть слишком велико для понимания человеком. К тому же, как было замечено ранее, не все ассоциативные правила полезны. Для сокращения числа правил и отсеивания бесполезных вводятся дополнительные величины.

Поддержка (support) – показывает, какой процент транзакций поддерживает данное правило. Так как правило строится на основании набора, то, значит, правило имеет поддержку, равную поддержке набора.

$$\mathrm{Supp}_{X \Rightarrow Y} = \mathrm{Supp}_{F} = \frac{|D_{F=X \cup Y}|}{|D|}, \tag{8}$$

где X – набор объектов, $X \in I$;

Y – набор объектов, $Y \in I$.

Очевидно, что правила, построенные на основании одного и того же набора, имеют одинаковую поддержку.

Достоверность (confidence) – показывает вероятность того, что из наличия в транзакции набора X следует наличие в ней набора Y. Достоверностью правила является отношение числа транзакций, содержащих наборы X и Y, к числу транзакций, содержащих набор X:

$$\mathrm{Conf}_{X \Rightarrow Y} = \frac{|D_{F=X \cup Y}|}{|D_X|} = \frac{\mathrm{Supp}_{X \cup Y}}{\mathrm{Supp}_X} \tag{9}$$

Очевидно, что чем больше достоверность, тем правило лучше, причём у правил, построенных на основании одного и того же набора, достоверность будет разная.

Достоверность позволяет отсеять часть слабых правил, но она не показывает, насколько то или иное правило является полезным. Для определения полезности правила была введена такая мера, как лифт правила (оригинальное название – интерес, также встречается термин «улучшение»).

Лифт правила показывает, случайно ли правило или же оно имеет под собой логичное обоснование. Лифт правила является отношением числа транзакций, содержащих наборы X и Y, к произведению количества транзакций, содержащих набор X, и количества транзакций, содержащих набор Y:

$$\mathrm{Lift}_{X \Rightarrow Y} = \frac{|D_{F=X \cup Y}|}{|D_X| \cdot |D_Y|} = \frac{\mathrm{Supp}_{X \cup Y}}{\mathrm{Supp}_X \cdot \mathrm{Supp}_Y} \quad (10)$$

Если значение лифта правила больше одного, то это значит, что правило более вероятно имеет под собой логичное обоснование, чем получено случайно. Если лифт меньше единицы – с большой вероятностью можно говорить о том, что правило является всего-навсего случайным совпадением. Лифт показывает, насколько ассоциированы между собой объекты правила.

Зачастую количество получаемых в результате работы какой-либо программы правил заранее ограничивается аналитиком для упрощения анализа и восприятия. Обычно задаются минимальные и максимальные значения поддержки и достоверности. Если установить слишком большие значения ограничения, то будут найдены тривиальные и хорошо известные правила. Если установить слишком низкие значения – будет сгенерировано огромное количество правил, большинство из которых будут непонятны и необоснованны, и найти в этом море полезные правила будет трудновыполнимой задачей. Следовательно, аналитику необходимо подобрать оптимальные значения ограничений, которые приведут к генерации полезных и необходимых ассоциативных правил.

Алгоритм Apriori – самый распространённый алгоритм для поиска ассоциативных правил в базах данных. Алгоритм был впервые описан в 1994 году и с тех пор вышло множество его модификаций.

Первый этап поиска ассоциативных правил – поиск всех часто встречающихся наборов – нетривиальная задача, т.к. при сравнительно больших объёмах данных она требует немалых вычислительных ресурсов и времени. Самый простой способ её решения – простой перебор всех возможных вариантов – при большом объёме данных становится трудновыполнимой задачей. т.к. с ростом количества объектов число всех переборов растёт экспоненциально. Алгоритм Apriori использует эвристический подход для решения этой задачи. Одно из свойств поддержки набора заключается в том, что что поддержка какого-либо набора не может быть больше поддержки любого из его подмножеств. Например, поддержка 3-элементного набора {a, b, c} будет всегда меньше или равна поддержке 2-элементных наборов {a, b}, {a, c}, {b, c}. Дело в том, что любая транзакция, содержащая {a, b, c}, также должна содержать {a, b}, {a, c}, {b, c}, причём обратное не верно.

Это свойство называется свойством анти-монотонности и значительно сужает пространство поиска, делая задачу нахождения всех часто встречающихся наборов относительно легко выполнимой. Если бы не свойство анти-монотонности – поиск ассоциативных правил был бы практически не осуществимой задачей.

Свойство анти-монотонности можно перефразировать несколько иначе: с ростом размерности набора его поддержка либо уменьшается, либо остаётся прежней. Расти она не может. Резюмируя всё вышесказанное, можно смело утверждать, что любой набор будет часто встречающимся тогда и только тогда, когда все его подмножества также будут часто встречающимися.

На начальном этапе работы алгоритма Apriori находятся все 1-элементные наборы и подсчитывается их поддержка. На этом этапе происходит обращение к базе данных; в дальнейшем оно не потребуется. Наборы, поддержка которых меньше минимально допустимой – отсекаются.

Следующие шаги состоят из двух этапов:

- генерация потенциально часто встречающихся наборов (такие наборы называются кандидатами);

- подсчёт поддержки кандидатов.

Функция генерации потенциально часто встречающихся кандидатов выглядит следующим образом. Каждый кандидат формируется с помощью расширения исходного набора добавлением к нему одного элемента другого набора. Затем, на основании вышеописанного свойства анти-монотонности происходит удаление избыточных правил. Набор удаляется из списка кандидатов, если хотя бы одно из его подмножеств не соответствует требованиям минимальной поддержки и не является часто встречающимся.

После того, как все возможные кандидаты были сгенерированы, подсчитывается их поддержка. Кандидаты, поддержка которых оказывается меньше допустимого значения – отсекаются, остальные же используются в дальнейшей работе алгоритма.

Когда этап поиска всех часто встречающихся наборов завершён, переходят к этапу поиска ассоциативных правил.

Генерация правил не такая сложная задача, как поиск часто встречающихся наборов. Для вычисления достоверности правила необходимо знать поддержку набора и поддержку условия. Т.к. все подмножества – часто встречающиеся, их поддержка нам уже известна, что избавляет нас от необходимости каждый раз обращаться к базе данных.

Чтобы сгенерировать правило для множества, нужно найти все его непустые подмножества. Правила генерируются для каждого подмножества, и правила, удовлетворяющие условию минимальной достоверности, заданной аналитиком – сохраняются. Правила, имеющие достоверность ниже минимально допустимой – отбрасываются. Здесь следует отметить одну важную особенность: любое правило составленное из определённого набора должно содержать все элементы этого набора. Например, если набор состоит из элементов {a, b, c}, то правило a $\Rightarrow$ b не должно рассматриваться.

Рассмотрим работу алгоритма Apriori на примере, она изображена на рисунке 9. Минимальный уровень поддержки равен 3.

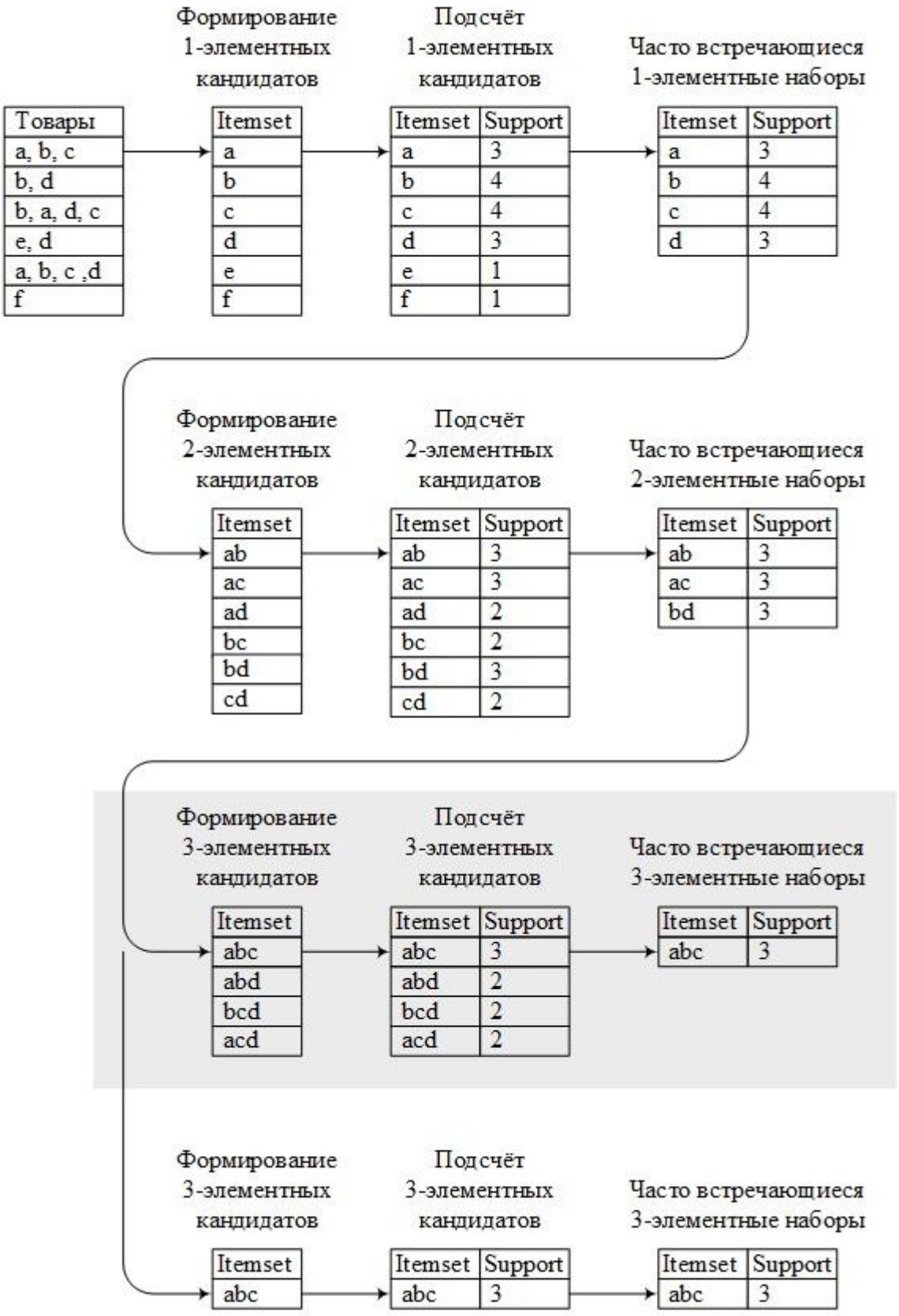

Рисунок 9 – Иллюстрация работы алгоритма Apriori

На начальном этапе работы алгоритма происходит генерация одноэлементных кандидатов. После этого вычисляется их поддержка. Если поддержка набора меньше минимально допустимой – набор исключается из работы. Все остальные наборы являются часто встречающимися одноэлементными наборами.

Затем мы повторяем те же действия: формируем кандидатов, в этот раз уже двухэлементные наборы, и вычисляем их поддержку. Наборы, не удовлетворяющие условию минимальной поддержки отсекаются.

Далее рассмотрим, как происходила бы работа алгоритма, не используй мы правило анти-монотонности. Взглянем на серую область. Здесь происходят действия, аналогичные предыдущим этапам: мы формируем кандидатов и вычисляем их поддержку. Неподходящие под условие наборы отсекаются и оставшийся набор является часто встречающимся. Т.к. больше не из чего формировать новых кандидатов, поиск всех часто встречающихся наборов завершён.

Теперь взглянем ниже серой области, как будет развиваться ситуация используй мы алгоритм Apriori и его правило анти-монотонности. Алгоритм отсекает, априори, те наборы, которые не могут стать часто встречающимися, т.к. содержат подмножества, ранее отсечённые нами как нечасто встречающиеся. При относительно большом множестве объектов это позволяет избежать огромного количества вычислений и значительно сэкономить ресурсы системы, сделав поиск всех часто встречающихся наборов выполнимой задачей.

3 Программная реализация информационной системы

3.1 Работа с базой данных

В предыдущих главах работы были поставлены задачи, которые должна решать информационная система; определены её функции, принципы построения и работы, выбраны и обоснованы методы программной реализации. В соответствие со всей проведённой работой была программно реализована информационная система, отвечающая всем ранее поставленным требованиям.

Информационная система «Мастер транзакций» была реализована на языке программирования C++ с использованием кроссплатформенного инструментария Qt. В качестве базы данных была выбрана SQLite – компактная встраиваемая реляционная база данных.

Программа предоставляет пользователю возможности работы с базой данных, а именно:

- создавать новые базы;
- добавлять, редактировать и удалять транзакции из базы;
- проводить выборку по базе с использованием фильтров.

Также программа имеет аналитический инструментарий, а именно:

- поиск ассоциативных правил по базе данных;
- визуализация динамики продаж товаров.

Рассмотрим работу программы подробнее.

На рисунке 10 изображено окно программы непосредственно после запуска. Основную область главного окна программы занимает объект класса QTabWidget, предоставляющий собой набор виджетов со вкладками, по сути являющийся контейнером. Содержимое первой вкладки, названной «Транзакции», позволяет пользователю работать с базой данных. Вторая вкладка, «Аналитика», как следует из названия, предоставляет пользователю доступ к аналитическому инструментарию информационной системы. Функции,

представленные в этих вкладках были перечислены выше, чуть позже я подробно опишу работу каждой из них.

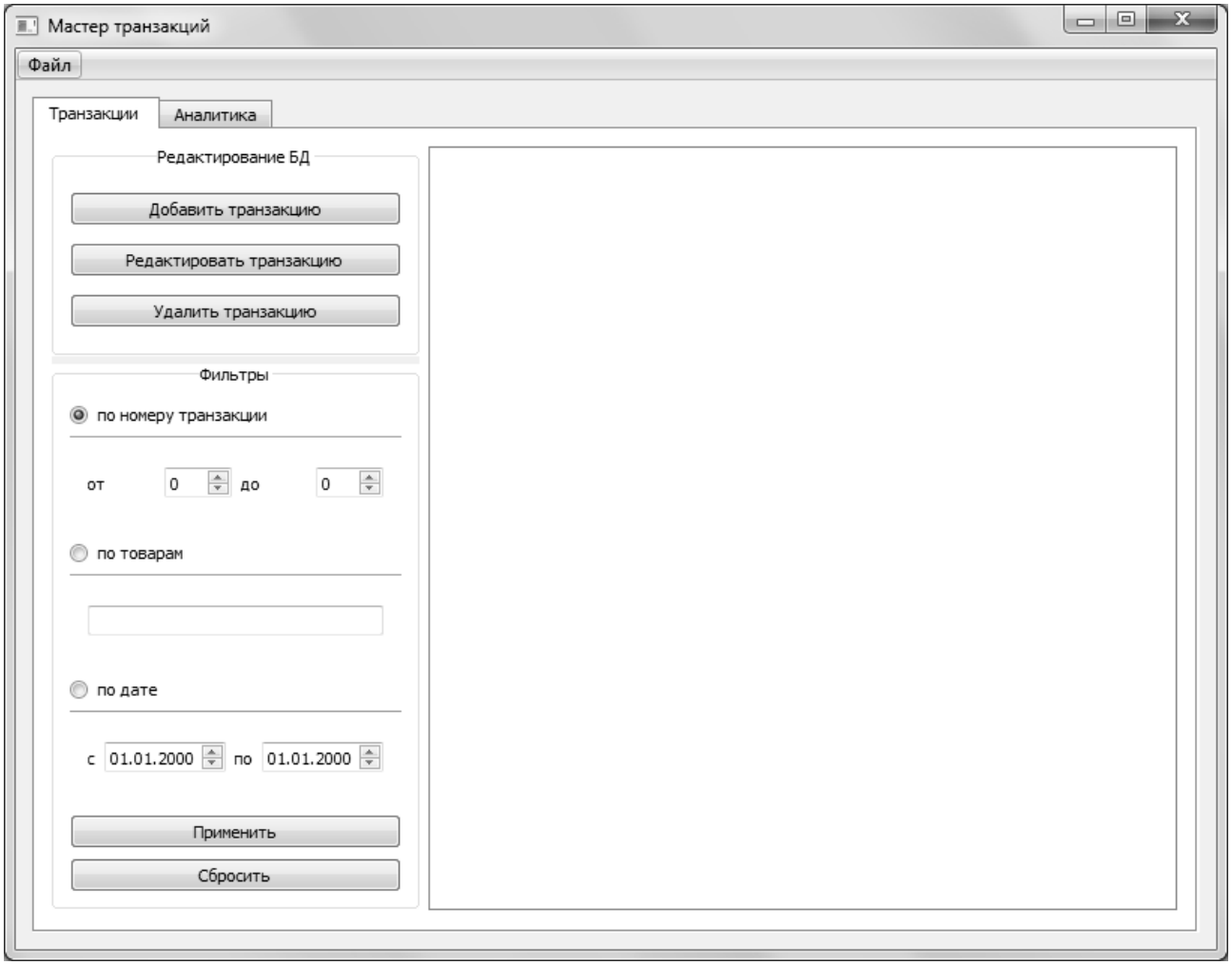

Рисунок 10 – Окно программы

Для того, чтобы начать работу, необходимо подключить базу данных. Пользователь может открыть существующую базу данных или же создать новую через пункт меню «Файл», содержащий два подпункта: «Создать базу транзакций» и «Открыть базу транзакций».

При выборе подпункта «Создать базу транзакций» пользователю открывается стандартное диалоговое окно выбора каталога, где будет создана база данных. Пользователю остаётся только ввести имя новой базы и нажать кнопку подтверждения.

При выборе подпункта «Открыть базу транзакций» пользователю открывается стандартное диалоговое окно выбора файла, где пользователю предлагается выбрать базу данных для использования в программе.

На рисунке 11 изображён процесс открытия уже существующей базы данных.

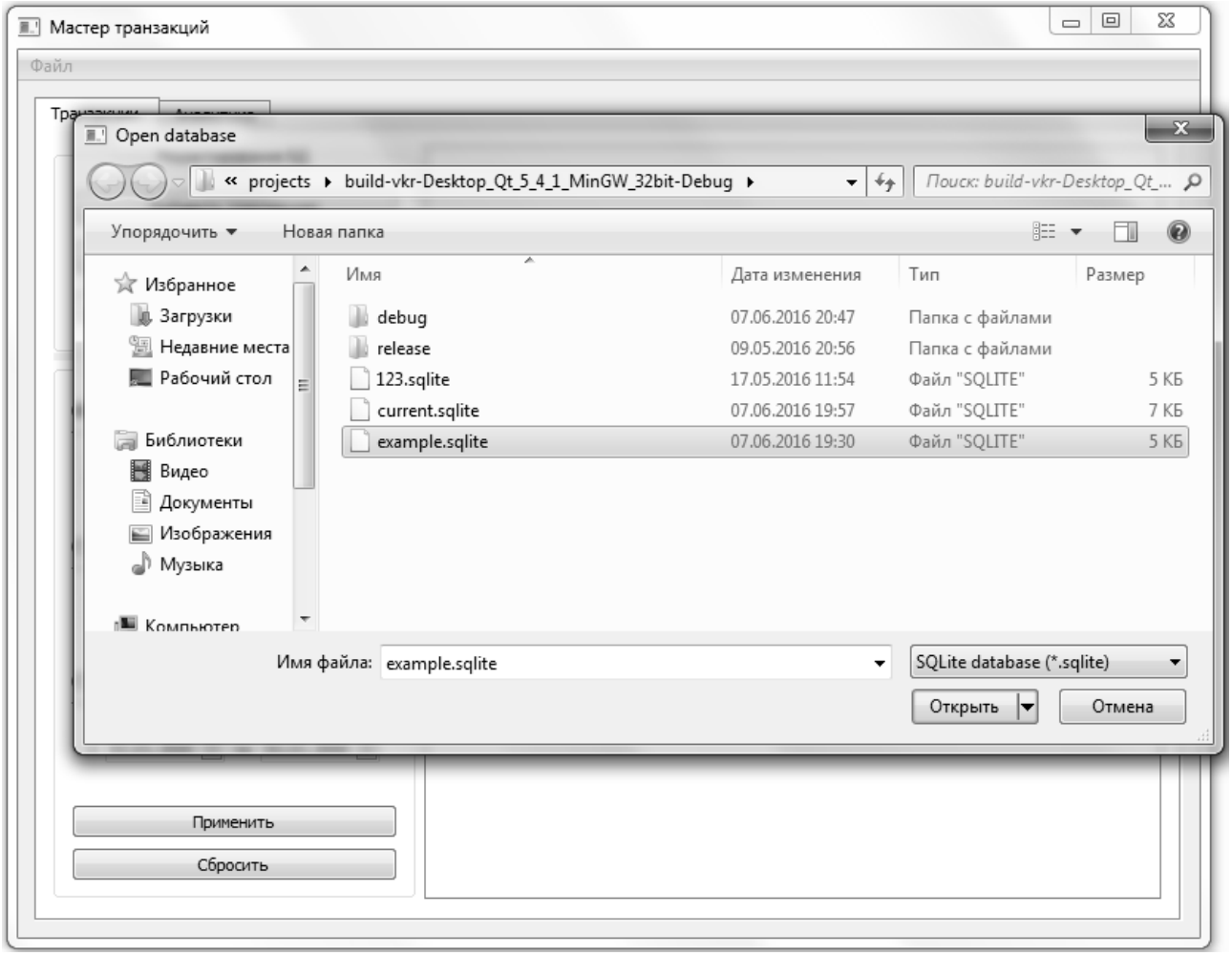

Рисунок 11 – Открытие базы транзакций

Программа работает с файлами с расширением «.sqlite», которое является стандартным расширением файлов для баз данных SQLite.

Соединение с базой данных осуществляется при помощи внутренних средств Qt. При открытии уже существующей базы данных требуется только открыть соединение с ней. При создании новой базы, эта база сначала создаётся

посредством выполнения sql запроса, затем с созданной базой устанавливается соединение.

Редактирование базы данных транзакций происходит при помощи элементов управления, объединённых в группу «Редактирование БД». Здесь располагаются три элемента управления, названные соответственно выполняемым им функциям:

- добавить транзакцию;
- редактировать транзакцию;
- удалить транзакцию.

Эти действия выполняются при помощи диалоговых окон, которые вызываются при нажатии пользователем одного из этих элементов управления.

Для того, чтобы добавить транзакцию в базу данных, пользователю необходимо нажать на кнопку «Добавить транзакцию». Откроется диалоговое окно, куда пользователю нужно будет ввести параметры добавляемой транзакции. Это диалоговое окно изображено на рисунке 12.

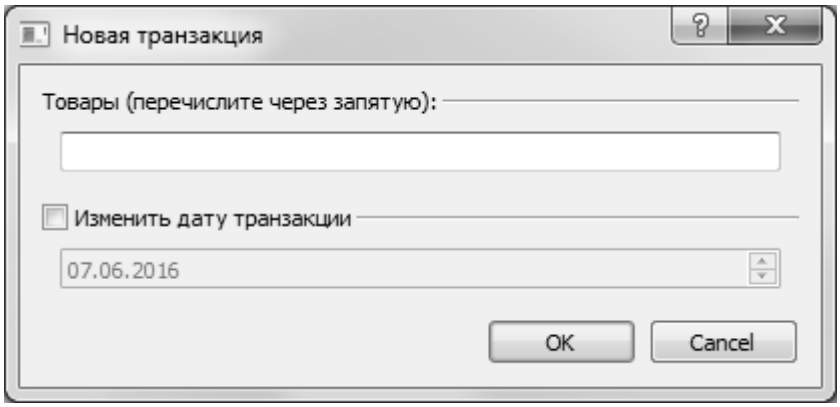

Рисунок 12 – Диалоговое окно добавления новой транзакции

В поле «Товары» пользователь вводит список товаров транзакции, перечисляя их через запятую. Программа воспринимает все символы, заключённые между запятыми, как наименование одного товара. Дата совершения транзакции по умолчанию устанавливается соответственно текущей дате, но при необходимости пользователь может её изменить, активировав поле редактирования даты выбором флажка «Изменить дату транзакции». При

нажатии кнопки «Cancel», введённая пользователем информация никуда не сохраняется. Транзакция добавляется в базу данных после нажатия пользователем кнопки «OK».

Помимо перечня товаров и даты совершения транзакции, все транзакции, хранящиеся в базе, так же имеют уникальный номер транзакции. Он присваивается транзакции автоматически и доступен пользователю только для просмотра, изменить его средствами программы пользователь не может.

Удаление транзакции происходит по её уникальному номеру. Пользователю открывается окно, где ему предлагается ввести номер транзакции, подлежащей удалению. Изображение этого окна приведено на рисунке 13.

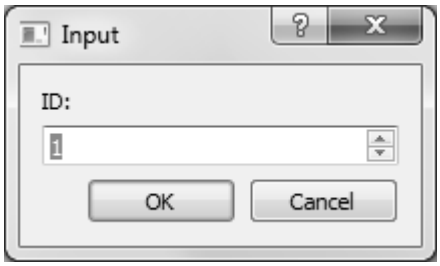

Рисунок 13 – Диалоговое окно удаления транзакции

При нажатии пользователем кнопки «OK» транзакция с таким номером удаляется из базы данных.

Для того чтобы внести изменения в уже существующую транзакцию, пользователю необходимо нажать кнопку «Редактировать транзакцию». Диалог с пользователем комбинирует решения, применяемые при удалении и добавлении транзакции. Редактирование происходит по номеру транзакции. Сначала пользователю отображается диалоговое окно, изображённое на рисунке 13, где ему необходимо ввести номер транзакции, которую требуется отредактировать. Затем пользователю отображается диалоговое окно, изображённое на рисунке 12, где он вводит новые требуемые данные.

Программой предусмотрены случаи, когда транзакции под требуемым номером не существует в базе данных, или же база данных пуста. В этом случае пользователю выдаётся сообщение об ошибке с описанием возникшей

проблемы, и программа автоматически возвращается к предыдущему стабильному состоянию.

При подключении базы данных, её содержимое незамедлительно отображается в правой части окна программы. Для того, чтобы уточнить выборку из базы, к ней можно применить имеющиеся фильтры. Пример этого изображён на рисунке 14: отображаются только те транзакции базы, которые среди перечня продуктов содержат хлеб.

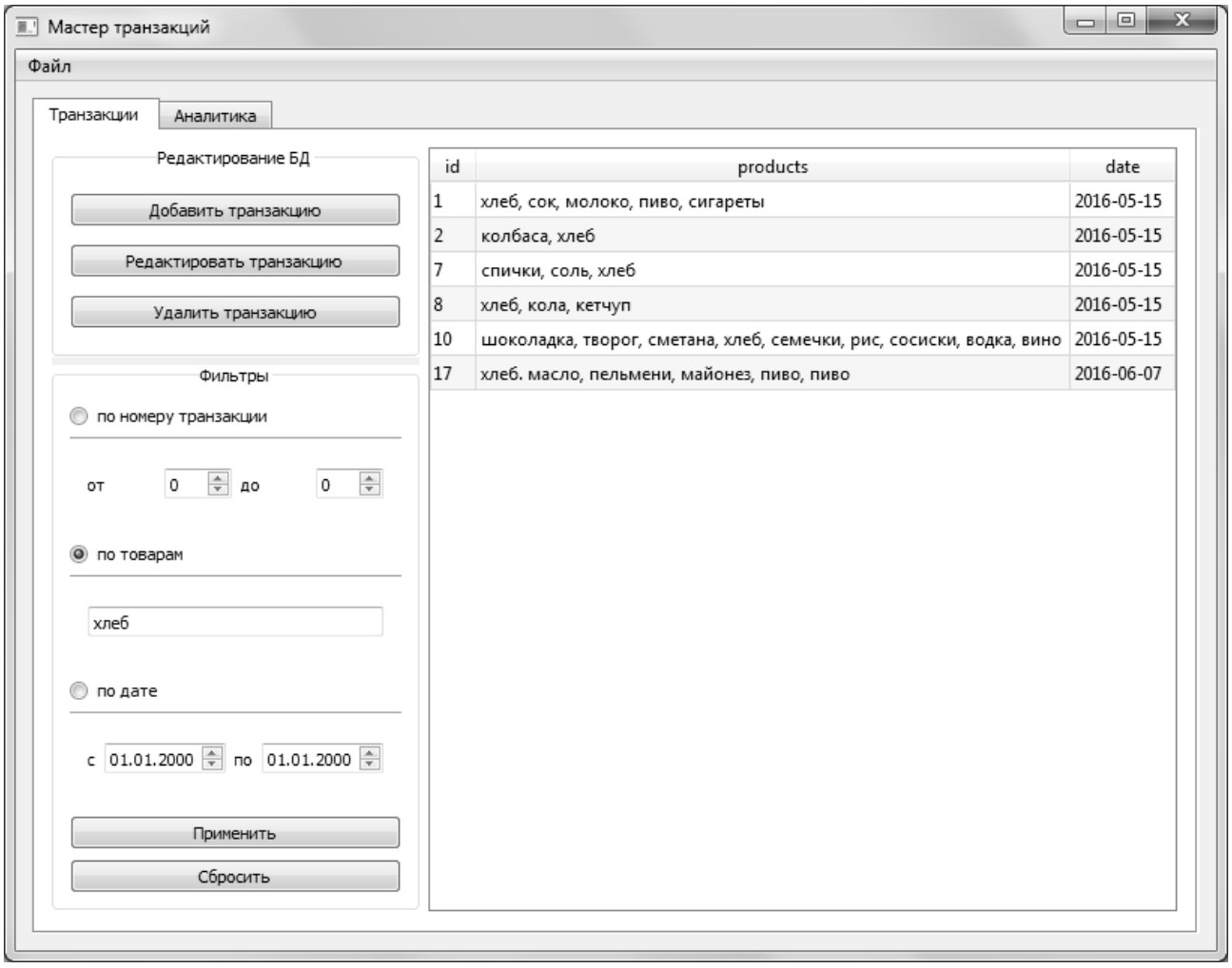

Рисунок 14 – Применение фильтров

Программой предусмотрена возможность фильтрации транзакций по трём параметрам: по номеру транзакции, по товарам и по дате.

При фильтрации по номеру транзакции пользователь указывает требуемый интервал. Все транзакции, номера которых попадают в указанный

интервал, отображаются в окне программы. Фильтрация по дате происходит схожим образом, но вместо интервала номеров пользователь указывает интервал дат. При фильтрации по товару отображаются все транзакции, содержащие указанный пользователем товар. При нажатии кнопки «Сбросить» происходит сброс всех фильтров и возврат к отображению полной базы транзакций.

3.2 Интеллектуальный анализ данных

Теперь рассмотрим вкладку «Аналитика», изображённую на рисунке 15:

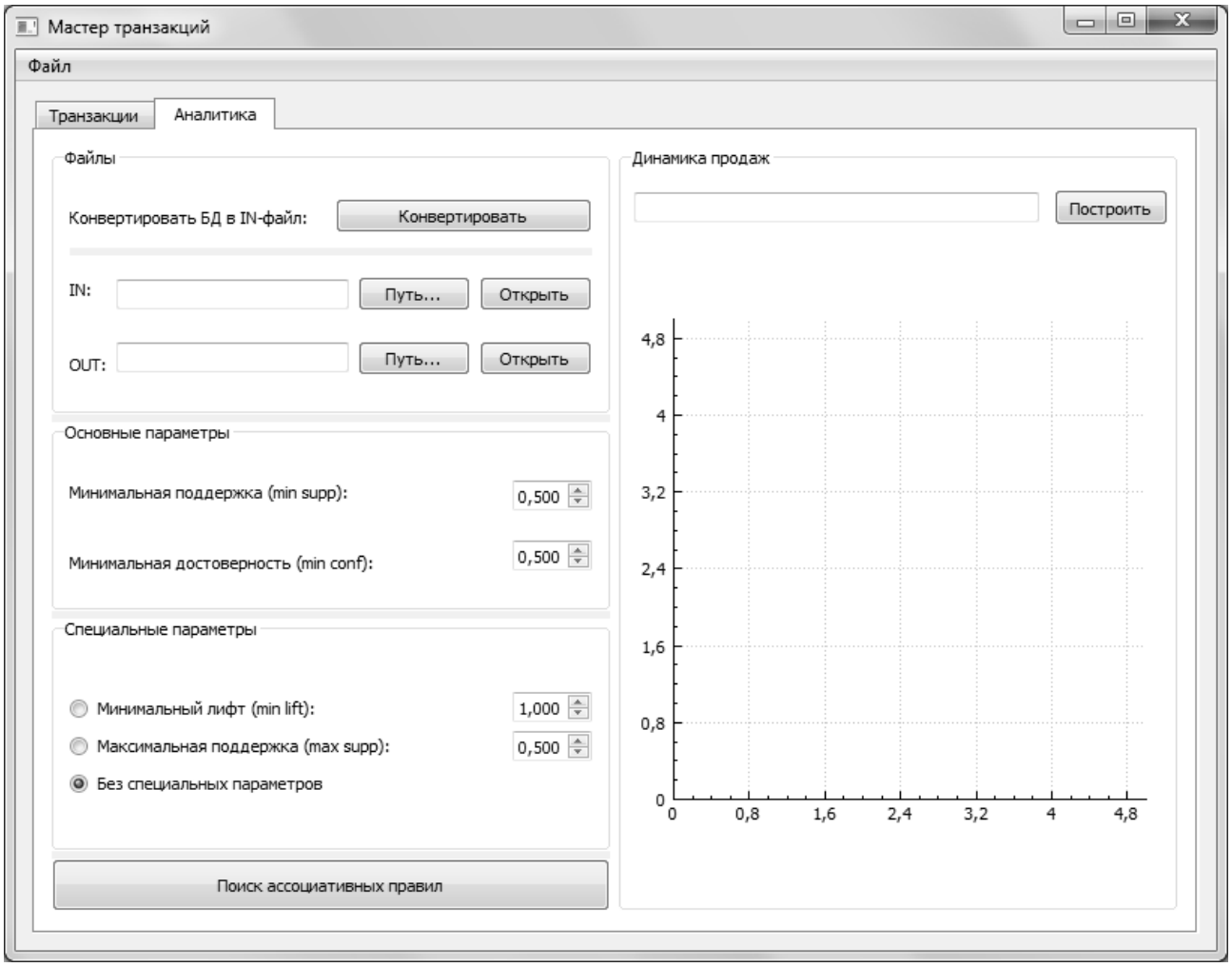

Рисунок 15 – Вкладка «Аналитика»

В левой части окна располагаются средства настройки поиска ассоциативных правил. А в правой – визуализация динамики продаж товаров.

Как упоминалось ранее на странице 36, при высокоуровневом описании алгоритма, для своей работы алгоритму поиска ассоциативных правил требуется особый формат данных. Кнопка «Конвертировать» производит преобразование стандартной базы данных во входной файл, требуемый алгоритму. В нём каждая строчка содержит только перечень товаров, без указания номера транзакции и даты её совершения, причём наименования товаров представлены их целочисленными кодами. Процесс формирования этих кодов был описан при высокоуровневом описании алгоритма (стр. 36). Пример входного файла, сформированного программой на основании стандартной базы данных, приведён ниже на рисунке 16.

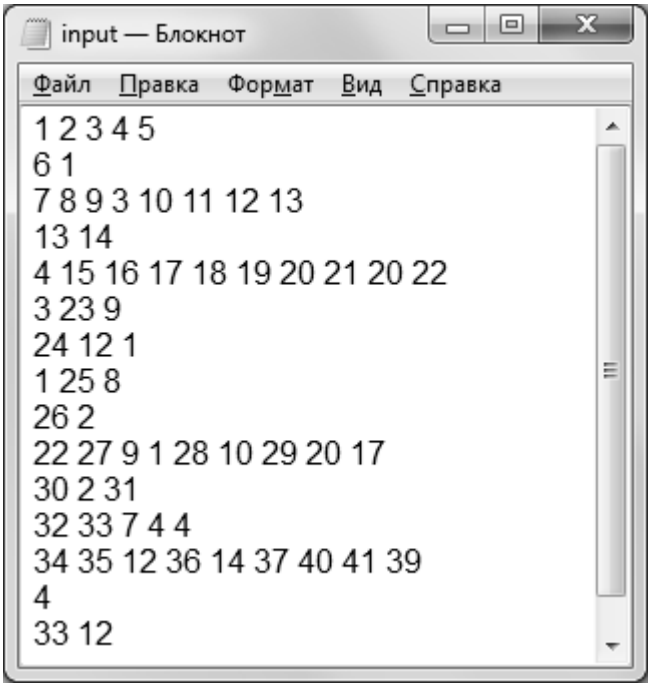

Рисунок 16 – Входной файл

В группе управляющих элементов «Файлы» пользователем указываются входной файл, содержащий конвертированную базу транзакций, и выходной файл, в который будут записанные найденные в результате работы алгоритма ассоциативные правила.

При нажатии на кнопку «Путь» пользователю открывается стандартное диалоговое окно выбора файла, где ему необходимо выбрать требуемый файл.

При нажатии на кнопку «Открыть» происходит открытие файла средствами операционной системы для просмотра его содержимого.

Если входной файл был сконвертирован программой, поле пути входного файла заполняется автоматически.

Если пользователь не указал выходной файл – он будет создан автоматически.

Все файлы создаются программой в том же каталоге, где находится исполняемый файл.

Группа элементов «Основные параметры» содержит элементы управления, позволяющие пользователю задать основные обязательные параметры работы алгоритма – минимальную поддержку и минимальную достоверность. О сути и назначении этих параметров говорилось ранее в этой работе. По умолчанию значения установлены на 0,5. Интервал значений, которые могут быть указаны – от 0,001 до 1.

Также программа позволяет указать несколько дополнительных параметров работы алгоритма. Это позволяет снизить количество ассоциативных правил, отбросив тривиальные и малополезные правила. В данной информационной системе существует возможность выбора минимального лифта или максимальной поддержки. О сути и назначении этих правил также говорилось ранее. Значением лифта по умолчанию является 1, значением максимальной поддержки – 0,5. Выбор требуемого дополнительного параметра работы алгоритма осуществляется путём выбора соответствующего переключателя в группе элементов «Специальные параметры».

Программа имеет возможность строить правила не только по подключённой базе данных. Если требуется анализ какой-либо другой базы данных, то нужен лишь её входной файл. Однако, в таком случае пользователь вряд ли получит полезную информацию, т.к. найденные ассоциативные правила не смогут быть преобразованы обратно к текстовому виду в соответствии с кодами товаров, т.к. без самой базы сделать это не получится, и найденные ассоциативные правила также будут в виде целочисленных кодов.

После того, как выбраны входной и выходной файл и указаны параметры работы алгоритма – можно приступать непосредственно к поиску ассоциативных правил. Для этого пользователю необходимо нажать кнопку «Поиск ассоциативных правил». Процедура происходит автоматически и не требует участия пользователя. По окончании процедуры поиска пользователю будет представлен текстовый файл, содержащий результаты работы программы. Вид этого файла изображён на рисунке 17.

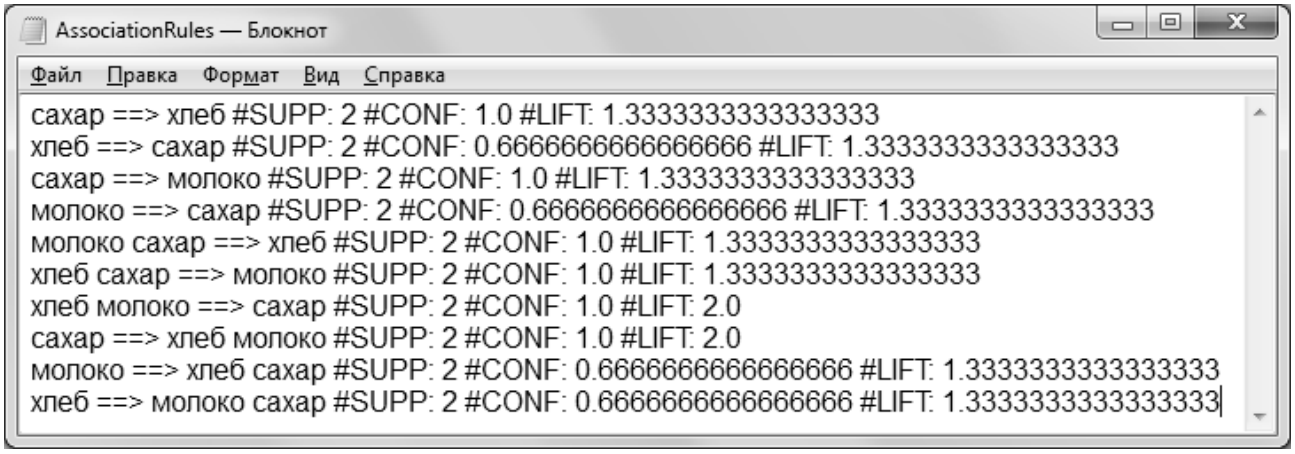

Рисунок 17 – Найденные ассоциативные правила

На рисунке хорошо видны сами правила в простой и понятной форме, а также их свойства (такие как поддержка, достоверность и лифт). Как уже отмечалось ранее, такие правила имеют вид «если (условие) – то (следствие)». Например, из результата работы программы и найденных правил следует, что клиенты, купившие хлеб и молоко с большой долей вероятности также приобретут и сахар. Это правило имеет очень хороший показатель лифта. Такие правила после их анализа могут использоваться для повышения рентабельности бизнеса. Способы практического применения ассоциативных правил ранее неоднократно упоминались и разъяснялись на протяжении всей работы.

Теперь рассмотрим визуализацию динамики продаж. Для того, чтобы посмотреть, какую динамику продаж имеет тот или иной товар, пользователь указывает интересующий его товар в поле группы управляющих элементов

«Динамика продаж» и нажимает кнопку «Построить». Для примера, динамика продаж хлеба изображена на рисунке 18.

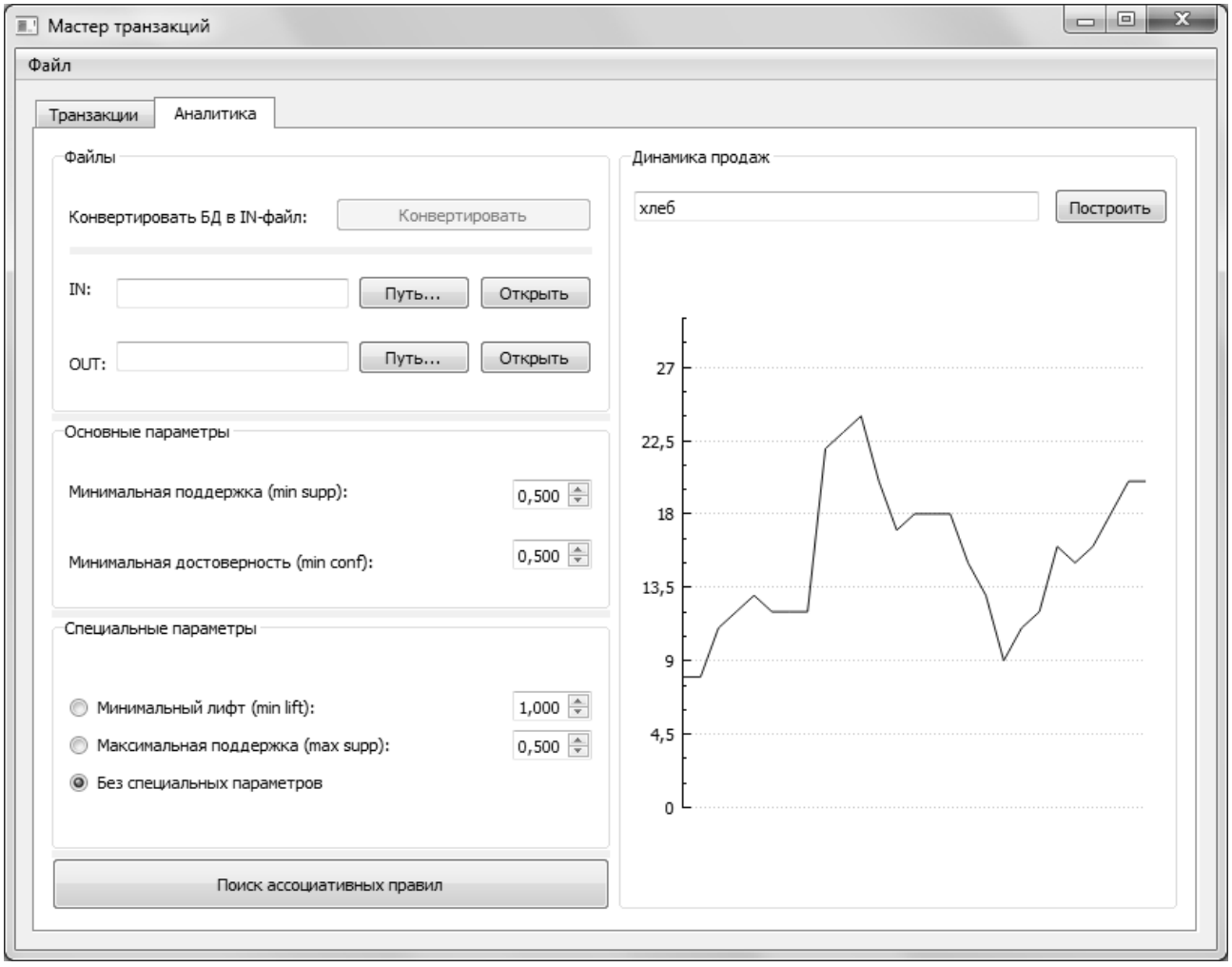

Рисунок 18 – Визуализация динамики продаж

Визуализация динамики продаж позволяет пользователю оценить количество проданного товара в хронологическом порядке.

Как видно из всего вышесказанного, программа весьма проста в использовании, интуитивно понятна и не требует наличия каких-либо специальных навыков или времени для обучения работе с ней. Но при этом она в полной мере справляется с возложенными на неё задачами.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были решены все поставленные задачи и достигнута поставленная цель, а именно:

- создан программный продукт, не требующий специальных навыков для работы и максимально понятный для пользователя;

- пользователю предоставлен инструментарий для работы с базой данных;

- к базе данных применены методы интеллектуального анализа данных с целью анализа накопленной информации для извлечения новых и потенциально полезных знаний.

Также было сделано несколько основных выводов относительно предметной области:

- повсеместное использование компьютеров привело к пониманию важности задач, связанных с анализом накопленной информации для извлечения новых знаний;

- сфера предпринимательства сейчас решительнее, чем когда-либо, настроена на повышение конкурентоспособности своих активов с помощью современных технологий, поэтому ритейл становится одной из передовых отраслей в области использования технологии интеллектуального анализа данных;

- Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Александров, Д. В. Инструментальные средства информационного менеджмента. CASE-технологии и распределённые информационные системы: учебное пособие / Д. В. Александров. – М.: Финансы и статистика, 2011. – 225 с.

2. Буреш, О. В. Интеллектуальные информационные системы управления социально-экономическими объектами: монография / О. В. Буреш, М. А. Жук. – М.: Красанд, 2012. – 192 с.

3. Верещагина, Е. А. Корпоративные информационные системы: учебно-методический комплекс / Е. А. Верещагина. – М.: Проспект, 2015. – 104 с.

4. Иванова, В. В. Основы бизнес-информатики: учебник / В. В. Иванова, Т. А. Лезина, А. А. Салтан – СПб.: СПбГУ, 2014. – 244 с.

5. Информационные системы и технологии в экономике и управлении: учебник / В. В. Трофимов [и др.]. – М.: Юрайт, 2015. – 542 с.

6. Курлов, А. Б. Методология информационной аналитики: монография / А. Б. Курлов, В. К. Петров. – М.: Проспект, 2014. – 384 с.

7. Нестеров, С. А. Базы данных: учебник / С. А. Нестеров. – М.: Юрайт, 2016. – 230 с.

8. Нестеров, С. А. Базы данных. Интеллектуальный анализ данных: учебное пособие / С. А. Нестеров. – СПб.: Издательство Политехнического университета, 2011. – 272 с.

9. Палкин, Н. Б. Бизнес-аналитика: от данных к знаниям: учебное пособие, 2-е изд., испр. / Н. Б. Палкин, В. И. Орешков. – СПб.: Питер, 2013. – 704 с.

10. Рафалович, В. И. Data Mining, или Интеллектуальный анализ данных для занятых: практический курс / В. И. Рафалович. – М.: SmartBook, 2014. – 96 с.

11. Самуйлов К. Е. Бизнес-процессы и информационные технологии в управлении современной инфокоммуникационной компанией: практическое руководство / К. Е. Самуйлов, А. В. Чукарин, Н. В Яркина. – М.: Альпина Паблишер, 2016. – 512 с.

ПРИЛОЖЕНИЕ А

Техническое задание

Полное наименование системы и её условное обозначение

Мастер транзакций (МТ).

Основание для разработки

Приказы о практике и дипломном проектировании.

Назначение системы

АИС МТ предназначена для автоматизации процессов ведения базы транзакций предприятий малого бизнеса и интеллектуального анализа базы транзакций.

Требования к программе

Программа должна выполнять функции простейшей СУБД, позволяя пользователю создавать и удалять базы транзакций, а также заниматься их наполнением и редактированием. Программа должна обеспечивать информационную поддержку представителей малого предпринимательства путём предоставления им упрощённого доступа к технологии интеллектуального анализа данных.

Требования к функциональным характеристикам

Программа должна позволять пользователю создавать и удалять базы транзакций. Добавлять транзакции в БД, редактировать и удалять существующие транзакции. Проводить выборку по базе с помощью фильтрации транзакций как отдельно по дате совершения, перечню товаров и уникальному идентификационному номеру транзакции, так и по нескольким фильтрам сразу. Эти функции должны осуществляться путём SQL запросов к компактной встраиваемой реляционной базе данных SQLite.

Программа должна обеспечивать информационную поддержку представителей малого предпринимательства путём предоставления им упрощённого доступа к технологии интеллектуального анализа данных. В

частности, должен быть реализован поиск ассоциативных правил по базе транзакций с использованием таких настраиваемых пользователем параметров, как: минимальная поддержка набора, максимальная поддержка набора, минимальная достоверность правила, лифт правила.

Должно быть реализовано представление результатов работы программы в виде понятной пользователю и информативной визуализации.

Условия эксплуатации

На объектах автоматизации должны отсутствовать такие воздействия, как: механический резонанс, синусоидальная вибрация, механические удары, атмосферное пониженное давление, плесневые грибы, рабочие растворы и агрессивные среды.

Электропитание на стационарных объектах эксплуатации осуществляется от электрической сети напряжением 380/220В, частотой 50 Гц с глухозаземлённой или изолированной нейтралью.

Рабочие станции должны размещаться в отапливаемых помещениях, в отдалении от отопительных приборов. Отапливаемые помещения должны быть оборудованы системами электроснабжения, связи, отопления, вентиляции и поддержки климатических условий:

- диапазон рабочих температур от +5°С до +35°С;
- относительная влажность до 80% при температуре +25°С;
- запылённость до 0,4 г/м3.

Требования к зданиям и помещениям, в которых располагаются рабочие станции с функционирующей на них АИС МТ, определяются следующими стандартами:

- РД 45.120-2000 «Нормы технологического проектирования. Городские и сельские телефонные сети. НТП 112-2000, (утверждены Минсвязи РФ 12.10.2000 г.);

- СН 512-78 «Строительные нормы. Инструкция по проектированию зданий и помещений для электронно-вычислительных машин» (утверждены Постановлением Госстроя СССР от 22 декабря 1978 г. №244), (в ред. Изменения

№1, утв. Постановлением Госстроя СССР от 27.02.1989 г. №33, Изменения №2, утв. Постановлением Госстроя РФ от 24.02.2000 г. №17);

- TIA-942 «Телекоммуникационная инфраструктура центров обработки данных» (редакция 7.0, февраль 2005 г.), утв. подкомитетом TLA TR 42.2, техническим комитетом TLA TR 42 и Американским национальным институтом стандартов (ANSI).

Требования к системным средствам

Информационная подсистема является Windows-приложением, работающим в среде Windows 7. В качестве БД используется компактная встраиваемая реляционная база данных SQLite.

Требования к средствам разработки

Информационная система разрабатывается в кроссплатформенной свободной интегрированной среде разработки Qt Creator.

Требования к техническим средствам

Для нормального функционирования АИС рекомендуется конфигурация компьютера, приведённая ниже:

ОС:	Windows 7/10/XP
Процессор:	Класса Pentium IV 1 ГГц или выше
Объем RAM:	Рекомендованный для установленной ОС
Видеоадаптер:	Разрешение экрана не менее 800*600, 256 цветов
Монитор:	Не менее 19”

Стадии и этапы разработки

Программное обеспечение разрабатывается в два этапа.

Этап 1. Июль 2015 года. Анализ объекта автоматизации. Проектирование структуры информационных ресурсов, входных и выходных форм интерфейса. Проектирование бизнес-процессов, алгоритма работы системы. Разработка программного обеспечения. Отладка системы на тестовых данных

Этап 2. Сентябрь 2015 г. – май 2016 года. Формирование информационных ресурсов. Разработка аналитических моделей с использованием технологий интеллектуального анализа данных. Подготовка программной документации.

Характеристика пользователей системы

Специальных требований к пользователю системы по подготовленности в области компьютерных технологий не предъявляется. Необходимо общее знакомство с правилами работы на компьютере, с функциональными возможностями системы и навыки работы в среде Windows. Интерфейс системы должен обеспечивать свободную работу пользователя и сопровождаться контекстной системой помощи.

Требования к персоналу

Квалификационные характеристики персонала должны соответствовать действующим на момент внедрения системы нормативным документам (общероссийский классификатор ОК 016-94, постановление Министерства труда РФ «Об утверждении разрядов оплаты труда и тарифно-квалификационных характеристик (требований) по общеотраслевым должностям служащих» от 06.06.1996г. №32).

Режим работы персонала должен соответствовать требованиям Трудового кодекса Российской Федерации, включая работу в условиях аварийных ситуаций, в том числе:

- требования к организации труда и режима отдыха персонала должны устанавливаться, исходя из требований к организации труда и режима отдыха при работе с персональными компьютерами;

- для обеспечения максимальной работоспособности и сохранения здоровья профессиональных пользователей на протяжении рабочей смены должны устанавливаться регламентированные перерывы: через 2 часа после начала рабочей смены и через 1,5 – 2,0 часа после обеденного перерыва продолжительностью 15 минут каждый или продолжительностью 10 минут через каждый час работы;

- продолжительность непрерывной работы персонала с разрабатываемой системой и персональными компьютерами без регламентированного перерыва не должна превышать 2 часа;

- деятельность персонала по эксплуатации системы должна регулироваться должностными инструкциями.

Требования к безопасности при эксплуатации технических средств

Организация рабочих мест пользователей, режим их работы должны соответствовать требованиям санитарных правил и норм (СанПиН) 2.2.2.542-96 «Гигиенические требования к видеодисплейным терминалам, персональным электронно-вычислительным машинам и организации работы», утверждённым постановлением Госкомсанэпиднадзора РФ от 14.07.96г. №14.

Требования по сохранности информации

Для защиты информации в случаях отказов технических средств, аппаратных и программных сбоев, некорректных действий пользователей система должна иметь инструментальные средства:

- для сохранения/восстановления рабочей конфигурации системы,

- для сохранения/восстановления отчётов и баз данных.

Требования к надёжности системы

Показатели надёжности системы определяются следующими требованиями:

- среднее время безотказной работы – не менее 10 часов,

- среднее время восстановления работоспособности системы при функциональных отказах – не более 5 минут.

Целевое назначение системы должно сохраняться на протяжении всего срока её эксплуатации. Срок эксплуатации системы определяется сроком устойчивой работы аппаратных средств вычислительных комплексов, своевременным проведением работ по замене (обновлении) аппаратных средств, по сопровождению программного обеспечения системы и его модернизации.

Требования по стандартизации и унификации

Классификация и кодирование информации в АИС МТ должны осуществляться на основе общероссийских классификаторов.

Технорабочая документация АИС МТ должна разрабатываться в соответствии с государственными стандартами.

Требования к эргономике и технической эстетике

Взаимодействие пользователей с прикладным программным обеспечением, входящим в состав системы должно осуществляться посредством визуального графического интерфейса (GUI). Навигационные элементы должны быть выполнены в удобной для пользователя форме. Средства редактирования информации должны удовлетворять принятым соглашениям в части использования функциональных клавиш, режимов работы, поиска, использования оконной системы. Ввод-вывод данных системы, приём управляющих команд и отображение результатов их исполнения должны выполняться в интерактивном режиме. Интерфейс должен соответствовать современным эргономическим требованиям и обеспечивать удобный доступ к основным функциям и операциям системы.

Интерфейс должен быть рассчитан на преимущественное использование манипулятора типа «мышь», то есть управление системой должно осуществляться с помощью набора экранных меню, кнопок, значков и т. п. элементов. Клавиатурный режим ввода должен используется главным образом при заполнении и/или редактировании текстовых и числовых полей экранных форм.

Система должна обеспечивать корректную обработку ситуаций, вызванных неверными действиями пользователей, неверным форматом или недопустимыми значениями входных данных. В указанных случаях АИС МТ должна выдавать пользователю соответствующие сообщения, после чего возвращаться в рабочее состояние, предшествовавшее неверной (недопустимой) команде или некорректному вводу данных.

Экранные формы должны проектироваться с учётом требований унификации.

Система должна соответствовать требованиям эргономики и профессиональной медицины при условии комплектования высококачественным оборудованием (ПЭВМ, монитор и прочее оборудование),

имеющим необходимые сертификаты соответствия и безопасности Госстандарта.

Порядок контроля и приёмки

Система сдаётся в соответствии с порядком, предусмотренным заданием на дипломное проектирование.

Для осуществления приёма создаётся комиссия из представителей Заказчиков и Исполнителя.

Комиссии предъявляются:

- документация на систему на машинном носителе, и два экземпляра на бумажном носителе;

- действующий программный комплекс на CD-диске;

- инструкция по установке и запуску системы в бумажном варианте;

- тексты программ на машинном носителе.

Программа испытаний системы составляется совместно Заказчиками и Исполнителем.

После подписания акта о приёмке назначается опытная эксплуатация сроком на 3 месяца, в ходе которой Разработчик сопровождает систему на консультативном уровне совместно со специалистами Заказчика.

При выявлении в период опытной эксплуатации недостатков и сбоев в работе программного обеспечения, возникших по вине Разработчика, последний обязан их устранить по первому письменному требованию Заказчика в разумные сроки.

В период опытной эксплуатации никакие доработки программного обеспечения, не связанные с недостатками, возникшими по вине Разработчика, не допускаются.

Сдача в промышленную эксплуатацию осуществляется после завершения опытной эксплуатации и устранения всех недоработок. Сдача в промышленную эксплуатацию оформляется актом о внедрении системы и актом о завершении работ по договору.